GENERALIZED METHOD OF MOMENTS

Advanced Texts in Econometrics

ALASTAIR R. HALL

ADVANCED TEXTS IN ECONOMETRICS

This page intentionally left blank

Generalized Method of Moments

ALASTAIR R. HALL

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Great Clarendon Street, Oxford OX2 6DP

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide in

Oxford New York

Auckland Bangkok Buenos Aires Cape Town Chennai
Dar es Salaam Delhi Hong Kong Istanbul Karachi Kolkata
Kuala Lumpur Madrid Melbourne Mexico City Mumbai Nairobi
São Paulo Shanghai Taipei Tokyo Toronto

Oxford is a registered trade mark of Oxford University Press
in the UK and in certain other countries

Published in the United States
by Oxford University Press Inc., New York

First published 2005

British Library Cataloguing in Publication Data
Data available

Library of Congress Cataloging in Publication Data
Data available

ISBN 0-19-877521-0 (hbk.)
ISBN 0-19-877520-2 (pbk.)

1 3 5 7 9 10 8 6 4 2

Typeset by Newgen Imaging Systems (P) Ltd., Chennai, India
Printed in Great Britain
on acid-free paper by
Biddles Ltd., King's Lynn, Norfolk

To Ada and Marten

This page intentionally left blank

Preface

Generalized Method of Moments (GMM) has become one of the main statistical tools for the analysis of economic and financial data. Accompanying this empirical interest, there is a growing literature in econometrics on GMM-based inference techniques. In fact, in many ways, GMM is becoming the common language of econometric dialogue because the framework subsumes many other statistical methods of interest, such as Least Squares, Maximum Likelihood and Instrumental Variables.

This book provides a comprehensive treatment of GMM estimation and inference in time series models. Building from the instrumental variables estimator in static linear models, the book presents the asymptotic statistical theory of GMM in nonlinear dynamic models. This framework covers classical results on estimation, such as consistency and asymptotic normality, and also inference techniques, such as the overidentifying restrictions test and tests of structural stability. The finite sample performance of these inference methods is also reviewed. Additionally, there is detailed discussion of recent developments on covariance matrix estimation, the impact of model misspecification, moment selection, the use of the bootstrap, and weak instrument asymptotics. There is also a brief exploration of the connections between GMM and other moment-based estimation methods such as Simulated Method of Moments, Indirect Inference and Empirical Likelihood.

The computer scientist Jan van de Snepscheut once admonished that "in theory, there is no difference between theory and practice. But, in practice, there is." Arguably a universal truth, this statement is certainly true about econometrics. Therefore, throughout the text, we focus not only on the theoretical arguments but also on issues that arise in implementing the statistical methods in practice. All the inference techniques are illustrated using empirical examples in macroeconomics and finance.

The text assumes a knowledge of econometrics, statistics and matrix algebra at the level of a course based on text such as William Greene's *Econometric Analysis*. All the main statistical results are discussed intuitively and proved formally. The presentation is designed to be accessible to a first- or second-year student in a graduate economics program at an American university.

This book developed out of lectures given at North Carolina State University. Parts of the material was also used as a basis for short courses at: the Division of Research and Statistics at the Board of Governors of the Federal Reserve

System in Washington D.C.; the Netherlands Graduate School of Economics; the Mansholt Graduate School of Social Sciences at Wageningen University in the Netherlands; the Department of Economics and Management at Wageningen University. Earlier drafts of the book were used by Eric Ghysels in a graduate econometrics course taught at Pennsylvania State University. I am very grateful to the participants in these courses for many useful comments and suggestions that have improved the book.

I made considerable progress in translating these lecture notes into the chapters of this book during my tenure of a research fellowship at the Department of Economics at the University of Birmingham. I am indebted to this department for both this support and also the colleagial atmosphere that made my visit both productive and pleasurable. I also worked on the book while a short-term visitor at the Department of Economics and Management at Wageningen University and gratefully acknowledge this support. The rest of the work was undertaken at the Department of Economics at North Carolina State University, and I happy to have this opportunity to record my gratitude to the department and university for their support over the years of both my own work and also econometrics more generally.

In the course of preparing the manuscript, a number of questions arose for which I had to turn to others for help. I would like to record my sincere gratitude to the following for generously sharing their time in order to provide me with the answers: John Aldrich, Anil Bera, Ron Gallant, Eric Ghysels, Atsushi Inoue, Essie Maasoumi, Louis Maccini, Angelo Melino, Benedikt Pötscher, Bob Rossana, Steve Satchell, Wally Thurman, Ken West, Ken Vetzal, and Tim Vogelsang. A number of people have read various drafts of this work and provided comments. This feedback was invaluable and I wish to thank particularly Ron Gallant, Eric Ghysels, Sanggohn Han, Atsushi Inoue, Kalidas Jana, Alan Ker, Kostas Kyriakoulis, Fernanda Peixe, Barbara Rossi, Amit Sen and Aris Spanos.

This book took far longer to complete than I ever imagined at the outset of the project. Over the years, I have accumulated a considerable debt of gratitude to: Lee Craig, who provided sagacious advice on various aspects of book authorship and literary style; Andrew Schuller, the editor, who provided continual encouragement; and Jason Pearce who patiently answered my questions about LaTeX. I have pleasure in thanking all three for their help.

However, my greatest debt is to my family. My wife Ada provided unfailing support throughout, and I dedicate this book to her and our son, Marten, as a token of my heartfelt gratitude.

Raleigh, NC

Contents

1

Introduction

1.1 Generalized Method of Moments in Econometrics

Generalized Method of Moments (GMM) was first introduced into the econometrics literature by Lars Hansen in 1982. Since then it has been widely applied to analyze economic and financial data. This interest has both stimulated and been facilitated by the development of numerous statistical inference techniques based on GMM estimators. These applications have been in very diverse areas spanning macroeconomics, finance, agricultural economics, environmental economics and labour economics. Depending on the context, GMM has been applied to time series, cross sectional, and panel data. In this book we focus on the use of GMM estimation with time series data and illustrate the various inference procedures using examples from macroeconomics and finance.[1] These areas are arguably the ones in which GMM has been most widely applied and, consequently, has had the biggest impact. Table 1.1 gives a list of various areas of economics to which GMM has been applied; inevitably this list is not exhaustive. Many of the studies have been published in top economic journals, which is one measure of the importance of the technique. Nearly all the studies have been published since the early 1990s and this testifies to the increasing impact of GMM on empirical analysis in economics.

It is natural to wonder why Hansen's 1982 paper had such an impact. After all, Maximum Likelihood estimation (MLE) has been around since the early part of twentieth century and it is the best available estimator within the Classical statistics paradigm. The optimality of MLE stems from its basis on the joint probability distribution of the data, which in this context becomes known as the likelihood function. However, in some circumstances, this dependence on the probability distribution can become a weakness. In the models in Table 1.1, two particular problems are present and these have motivated the use of GMM.

[1] For discussions of GMM with panel data, see Baltagi (2001) or Wooldridge (2002).

These are as follows.

1. *Sensitivity of statistical properties to the distributional assumption*
   The desirable statistical properties of MLE are only attained if the distribution is correctly specified. Unfortunately, economic theory rarely provides the complete specification of the probability distribution of the data. One solution is to choose a distribution arbitrarily. However, unless this guess coincides with the truth, the resulting estimator is no longer optimal and, worse still, its use may lead to biased inferences.

2. *Computational burden*
   For many of the models in Table 1.1, Maximum Likelihood estimation would be computationally very burdensome. Two types of problem tend to occur. In some cases, the economic model coincides with the joint probability distribution of the data but the implied likelihood function is extremely difficult to evaluate numerically with available computer technology. In other cases, the economic model only involves some aspects of the probability distribution and the completion of the specification introduces many additional parameters which must also be estimated. Often in these latter cases, the likelihood function must be maximized subject to a set of nonlinear constraints implied by the economic model, which further adds to the computational burden.

In contrast, the GMM framework provides a computationally convenient method of performing inference in these models without the need to specify the likelihood function.

The cornerstone of GMM estimation is a set of population moment conditions which are deduced from the assumptions of the econometric model. The exact nature of these conditions varies from application to application but, whatever they are, their validity is crucial for the properties of the resulting estimator. The potential of moment conditions for estimation has been recognized since the 1890s when a technique known as Method of Moments was first proposed. In fact, many estimation techniques familiar in econometrics are based either explicitly or implicitly on the information in population moment conditions. However, prior to Hansen's work, the statistical theory of these estimators tended to be restricted to the moment conditions of a particular functional form. One of the main contributions of Hansen's paper was to emphasize the common underlying structure of these previous analyses and to develop a statistical theory which can be applied to any set of moment conditions. Inevitably, GMM builds on these earlier analyses and so to help put GMM in perspective, it is useful to understand its statistical antecedents. Therefore, we start by briefly summarizing in Section 1.2 how the use of moment conditions has evolved in statistics and econometrics. This provides a first illustration of how moment conditions can be used as a basis for estimation. It also links GMM to a number of estimators familiar in econometrics. After this historical review, a set of contemporary examples from Table 1.1 are provided in Section 1.3. At this stage, the focus is on showing how the population moment conditions arise in

Table 1.1
Applications of GMM

Agriculture	Thijssen (1996), Chavas and Thomas (1999), Bourgeon and Le Roux (2001)
Business cycles	Singleton (1988), Christiano and Eichenbaum (1992), Burnside, Eichenbaum, and Rebelo (1993), Braun (1994), Boldrin, Christiano, and Fisher (2001)
Commodity markets	Deaton and Laroque (1992), Bjornson and Carter (1997), Considine and Heo (2000), Haile (2001)
Consumption	Miron (1986), English, Miron, and Wilcox (1989), Campbell and Mankiw (1990), Runkle (1991), Blundell, Pashardes, and Weber (1993), Blundell, Browning, and Meghir (1994) Attanasio and Browning (1995), Attanasio and Weber (1995), Ni (1995), Meghir and Weber (1996), Dynan (2000), Fuhrer (2000), Weber (2000)
Cost/Production frontiers/functions	Kopp and Mullahy (1990), Blundell and Bond (2000), Ahn, Good, and Sickles (2000)
Development	Jalan and Ravallion (1999), Hansen and Tarp (2001), Ogaki and Zhang (2001)
Economic growth	Caselli, Esquivel, and Lefort (1996)
Education/human capital	Angrist and Krueger (1992), Palacios-Huerta (2003)
Environmental economics	Smith and Pattanayak (2002)
Equity pricing	Hansen and Singleton (1982), Singleton (1985), Finn, Hoffman, and Schlagenhauf (1990), Ghysels and Hall (1990a,b) Ferson (1990), Bodurtha and Mark (1991), Epstein and Zin (1991), Ferson and Constantinides (1991), Harvey (1991), MacKinlay and Richardson (1991), Snow (1991), Bessembinder and Chan (1992), Ferson and Harvey (1992), Ilmanen (1992), Marshall (1992), Bansal, Hsieh, and Viswanathan (1993), Bansal and Viswanathan (1993), Cecchetti, Lam, and Mark (1993), Ferson, Foerster, and Keim (1993), Fisher (1994), Zhou (1994), Campbell (1996), Cochrane (1996), Hansen and Singleton (1996), He, Kan, Ng, and Zhang (1996), Ho, Perraudin, and Sørensen (1996), Hagiwara and Herce (1997), Hansen and Jaganathan (1997), Ghysels (1998), Garcia and Bonomo (2001), Timmerman (2001), Jiang and Knight (2002), Vissing-Jørgenson and Attanasio (2003)
Exchange rates	Hansen and Hodrick (1980), Mark (1985), Melino and Turnbull (1990), Modjtahedi (1991), Bekaert and Hodrick (1992), Cumby and Huizinga (1992), Backus, Gregory, and Telmer (1993), Imrohoroglus (1994), Dumas and Solnik (1995), Hartmann (1999), Bekaert and Hodrick (2001), Groen and Kleibergen (2003)

continued over

Table 1.1 (continued)
Applications of GMM

Health care	Windmeijer and Silva (1997), Schellhorn (2001), Silva and Windmeijer (2001)
Import demand	de la Croix and Urbain (1998)
Interest rates	Dunn and Singleton (1986), Diba and Oh (1991), Lee (1991), Chan, Karolyi, Longstaff, and Sanders (1992), Longstaff and Schwartz (1991), Cushing and Ackert (1994), Vetzal (1997), Green and Odegaard (1997)
Inventories	Miron and Zeldes (1988), Eichenbaum (1989), Kayshap and Wilcox (1993), Durlauf and Maccini (1995), Fuhrer, Moore, and Schuh (1995*a*), Bils and Kahn (2000)
Investment	Gordon (1992), Hubbard and Kayshap (1992), Whited (1992), Bond and Meghir (1994), Gilchrist and Himmelberg (1995), Oliner, Rudebusch, and Sichel (1996), Chirinko and Schaller (1996), Ogawa and Suzuki (1998), Chirinko and Schaller (2001)
Labour demand	Pindyck and Rotemberg (1983), Arellano and Bond (1991), Pfann and Palm (1993)
Labour market	Yashiv (2000), Yuan and Li (2000)
Labour supply	Mankiw, Rotemberg, and Summers (1985), Eichenbaum, Hansen, and Singleton (1988), Kahn and Lang (1991), Angrist (2001)
Macroeconomic forecasts	Keane and Runkle (1990), Bonham and Cohen (1995, 2001)
Microstructures in finance	Madhavan and Smidt (1993), Huang and Stoll (1997), Madhavan, Richardson, and Roomans (1997), Biasis, Hillion, and Spatt (1999), Grammig and Wellner (2002)
Money	Eckstein and Leiderman (1992), Dutkowsky (1993), Holman (1998), Clarida, Gali, and Gertler (2000)
Mutual fund performance	Chen and Knez (1996), Bekaert and Urias (1996)
Product demand	Berry, Levinsohn, and Pakes (1995)
Productivity	Bernstein (1994), Atkinson, Cornwell, and Honerkamp (2003)
R & D spending	Himmelberg and Petersen (1994)
Resources	Young (1991, 1992), Green and Mork (1991), Popp (2001)
Technological innovation	Blundell, Griffith, and Vanreenen (1995)
Trading volume of financial assets	Foster and Viswanathan (1993), Bessembinder, Chan, and Seguin (1996)
Transportation	Nevo (2003)

in these models. Later in the book, we return to these models to illustrate the various estimation and inference procedures discussed. The development of

these procedures requires certain statistical concepts and results. Section 1.4 provides a review of some background statistical theory which is needed for the introduction of the basic GMM framework in Chapters 2 and 3. More advanced statistical theory is developed as necessary in subsequent chapters. Section 1.5 concludes the chapter with an overview of the remainder of the book.

1.2 Population Moment Conditions and the Statistical Antecedents of GMM

The term *population moment* was originally used in statistics to denote the expectation of the polynomial powers of a random variable. So if v_t is a discrete random variable with probability mass function $P(v_t = v)$ defined on a sample space $\mathcal{V}$ then its r^{th} population moment is given by

$$E[v_t^r] = \sum_{\{v \in \mathcal{V}\}} v^r P(v_t = v) = \nu_r$$

where the summation is over all values in $\mathcal{V}$ and r is a positive integer. If v_t is a continuous random variable with probability density function $p(v)$ then its r^{th} moment is given by

$$E[v_t^r] = \int_{-\infty}^{\infty} v^r p(v) dv = \nu_r$$

From these definitions it is easily recognized that the population mean is just the first population moment and the population variance is $\nu_2 - \nu_1^2$. The term (population) moment has been in the statistical lexicon since at least the work of A. Quetelet who lived from 1796 to 1874 and was inspired by the concept of moments in physics, see Stuart and Ord (1987, p.53).[2]

Karl Pearson[3] (1893, 1894, 1895) was the first person to recognize the potential of population moments as a basis for estimation. In this series of articles, he introduced Method of Moments estimation. To understand his original motivation, it is necessary to consider briefly the state of statistical analysis in the late nineteenth century. During that century, a lot of natural phenomena were thought to be well summarized by a normal distribution. This belief can be attributed to at least two reasons. First, the actual evidence was limited, because only a few data sets had been collected. Secondly, the available diagnostic tests were very rudimentary and could only detect very dramatic departures from normality; see Stigler (1986, p.330) . However, as interest in statistics – and

[2] Adolphe Quetelet was a Belgian with far ranging interests. He wrote the libretto of an opera, a historical survey of romance and poetry as well as his scientific work in astronomy, sociology and statistics. Pearson (1895) described him as a man "who often foreshadowed statistical advances without providing the method by which they might be dealt with" (Pearson, 1895, p.381). For an interesting discussion of Quetelet's contributions see Stigler (1986).

[3] Karl Pearson (1857–1936) was an Englishman trained as a mathematician whose interests also included physics, German history, folklore and philosophy. Apart from Method of Moments, his numerous contributions to statistics included chi-squared goodness of fit tests, correlation and the Pearson family of distributions.

science – grew, more datasets were collected. With this growing body of empirical evidence, researchers became aware that many natural phenomena showed departures from normality and in particular exhibited skewness. This raised the challenge of finding theoretical probability distributions which could adequately capture this behaviour. Karl Pearson was in the forefront of this research and developed what has become known as the Pearson family of distributions, e.g. see Stuart and Ord (1987, pp.210–20). This family is characterized by a probability density function which is indexed by a vector of four parameters. Different values of the parameters can yield a wide variety of distributions, including the normal, beta and gamma.

The practical problem was to find the most appropriate member of this family for the data set in hand – or in other words, to estimate the parameter vector. The existing techniques for fitting normal distributions were not suited to these more general types of distribution. Instead, Pearson suggested calculating estimates based on moments. The idea is simple. Population moments implied by the family of distributions are functions of the unknown parameter vector. Pearson proposed estimating the parameter vector by the value implied by the corresponding sample moments. His approach is best understood by considering a simple example. For the purposes of our discussion we can abstract from the generality of the Pearson family and just focus attention on a particular member, the normal distribution. This distribution depends on just two parameters:[4] the population mean, μ_0, and the population variance, σ_0^2. These two parameters satisfy the population moment conditions

$$\begin{aligned} E[v_t] - \mu_0 &= 0 \\ E[v_t^2] - (\sigma_0^2 + \mu_0^2) &= 0 \end{aligned} \tag{1.1}$$

Pearson's method involves estimating (μ_0, σ_0^2) by the values $(\hat{\mu}_T, \hat{\sigma}_T^2)$ which satisfy the analogous sample moment conditions and we have indexed the estimators by the sample size T. Therefore $(\hat{\mu}_T, \hat{\sigma}_T^2)$ are the solutions to

$$\begin{aligned} T^{-1} \sum_{t=1}^{T} v_t - \hat{\mu}_T &= 0 \\ T^{-1} \sum_{t=1}^{T} v_t^2 - (\hat{\sigma}_T^2 + \hat{\mu}_T^2) &= 0 \end{aligned}$$

and so, with some rearrangement, it follows that

[4] The normal distribution is obtained from the generic form of the Pearson family by setting two of the four parameters to zero.

$$\begin{aligned} \hat{\mu}_T &= T^{-1}\sum_{t=1}^{T} v_t \\ \hat{\sigma}_T^2 &= T^{-1}\sum_{t=1}^{T}(v_t - \hat{\mu}_T)^2 \end{aligned} \tag{1.2}$$

Pearson called this approach the "Method of Moments" for obvious reasons. Pearson (1895) demonstrated the power of this technique with an analysis of the distributions of such diverse phenomena as barometric pressures, the sizes of the carapace of crabs, the heights of recruits to the U.S. army, the valuation of house prices and the number of divorces granted.

This approach is very intuitive but not without its weaknesses. For example, all the higher moments of the normal distribution depend on (μ_0, σ_0^2); e.g. see Stuart and Ord (1987, p.78). Therefore, this technique could have been applied equally well to the third and fourth moments, say, of the distribution. The problem is that the resulting estimators of (μ_0, σ_0^2) would be different from those given in (1.2). Which estimators should be used? This question is hard to address within the Method of Moments framework. In fact, it was this question which led R. A. Fisher[5] to analyze how information from a probability distribution can be channeled most effectively into parameter estimation. The result was the Maximum Likelihood principle; see Fisher (1912, 1922, 1925). In fact, MLE can also be interpreted as a special case of GMM based on a population moment condition whose derivation requires the specification of the probability distribution of the data. However, it is pedagogically most convenient to postpone further discussion of this interpretation until the complete GMM framework has been introduced in Chapter 3.[6] For our purposes here, it is more relevant to consider another weakness inherent in the Method of Moments framework. Suppose that it is desired to base estimation of (μ_0, σ_0^2) on the first three moments of v_t, that is (1.1) plus

$$E[v_t^3] - 3E[v_t^2]\mu_0 + 3E[v_t]\mu_0^2 - \mu_0^3 = 0 \tag{1.3}$$

In this case, the sample analogs to (1.1)–(1.3) form a system of three equations in two unknowns, and such a system typically has no solution. Therefore, the Method of Moments is infeasible. It is easily recognized that this problem is not specific to this example. Clearly, some modification is needed in order to

[5] Ronald Fisher (1890–1962) was an English scientist who made fundamental contributions to statistics, probability, genetics and the design of experiments. He is regarded by many as the founder of mathematical statistics. Apart from Maximum Likelihood, he developed the general framework of estimation theory including the concepts of consistency, information, sufficiency, efficiency, ancillarity and pivotal statistics. His other famous contributions include the analysis of variance method and the F-distribution.

[6] For completeness, we note that if it is assumed in our simple example that $\{v_t, t = 1, 2..., T\}$ are also independently distributed then $(\hat{\mu}_T, \hat{\sigma}_T^2)$ are the MLE's ; e.g. see Stuart and Ord (1987, p.287). However, this coincidence is the exception rather than the rule. In general, ML estimation does not involve matching these type of simple population moment conditions; see Section 3.6 for further discussion.

produce estimates of p parameters based on more than p population moment conditions. This brings us to the second important statistical antecedent of GMM, namely the method of Minimum Chi-Square.

In a series of articles in the late 1920s and the 1930s, Neyman and Pearson laid the foundations for the framework of "classical" hypothesis testing.[7] One side product of this research was the Minimum Chi-Square method of estimation. The method was originally proposed to facilitate inference about whether or not an observed sample was generated from a particular distribution, but the basic idea can be applied to estimation in a wide variety of problems including the estimation of (μ_0, σ_0^2) based on (1.1)–(1.3). However, it is instructive to introduce the method in the context of the specific example considered by Neyman and Pearson.

Neyman and Pearson (1928) considered the particular case in which a researcher wishes to model the probability that the outcome of an experiment lies in one of k mutually exclusive and exhaustive groups. If p_i is used to denote the probability the outcome lies in the i^{th} group then the null hypothesis of interest is that

$$p_i \;=\; h(i, \theta_0) \tag{1.4}$$

where $h(.)$ is some specified functional form indexed by an unknown parameter vector θ_0. The question was how to test this hypothesis. In 1928, the challenging feature of this problem was that the null hypothesis only specified the form of the probability function up to some unknown parameter vector. At that stage, the problem had only been solved if the null specified a particular value of θ_0 as well. In the latter case, Karl Pearson (1900) had shown that inference could be based on the goodness of fit statistic,

$$GF_T(\theta_0) \;=\; \sum_{i=1}^{k} \frac{[T_i \;- Th(i;\theta_0)]^2}{T_i} \tag{1.5}$$

where T_i is the frequency of outcomes in the i^{th} group in a sample of size T. Pearson (1900) showed that this statistic was approximately distributed χ^2_{k-1} under the null hypothesis.[8] Neyman and Pearson (1928) recognized that if θ_0 is unknown then the goodness of fit statistic can provide the basis for estimation of θ_0 as well as inference about the null hypothesis. Their idea was to estimate θ_0 by $\hat{\theta}_T$, the value of θ which minimizes the goodness of fit statistic.[9] In view of Pearson's (1900) aforementioned distributional result, Neyman and Pearson

[7] Jerzy Neyman (1894–1981) was born in Russia but came from a Polish family. Egon S. Pearson (1895–1980) was the son of Karl Pearson. Their collaboration began in the mid-1920s when Neyman held a post doctoral fellowship to study under Karl Pearson at University College of London where Egon Pearson was also on the faculty. Apart from their seminal work together, both made numerous other contributions to statistics including Neyman's work on the theory of survey sampling, estimation by confidence sets and best asymptotically normal estimators, and Pearson's work on quality control and operations research.

[8] Notice that the degree of freedom of the distribution is only $k-1$ and not k because once the frequencies in $k-1$ groups are known then the frequency in the k^{th} group is automatically determined by $T_k = T - \sum_{i=1}^{k-1} T_i$.

[9] This insight was not completely new even in 1928. Smith (1916) discussed the idea of

(1928) refered to $\hat{\theta}_T$ as a "Minimum Chi-Square estimator". Furthermore, they showed that under the null hypothesis in (1.4), $GF_T(\hat{\theta}_T)$ is approximately distributed χ^2 with $k-1-p$ degrees of freedom where p denotes the dimension of θ_0.

At first glance, it may not be readily apparent that there is any connection between the estimation problem considered by Neyman and Pearson (1928) and the problem of how to estimate (μ_0, σ_0^2) based on the first three moments of the normal distribution. However, both problems actually have the same underlying structure. To uncover this connection, it is necessary to view Neyman and Pearson's (1928) method from a slightly different perspective. To develop this new interpretation, it is necessary to rewrite the goodness of fit statistic and introduce a set of indicator variables. First, note the goodness of fit statistic can be written as

$$GF_T(\theta_0) \;=\; T\sum_{i=1}^{k} \frac{[\hat{p}_i \;- h(i;\theta_0)]^2}{\hat{p}_i} \tag{1.6}$$

where $\hat{p}_i = T_i/T$, the relative frequency in the sample of outcomes in the i^{th} group. Now consider the set of indicator variables $\{D_t(i); i = 1, 2, \ldots k;\ t = 1, 2, \ldots T\}$ which take the value one if the t^{th} outcome of the experiment lies in the i^{th} group and takes the value zero otherwise. Notice that if (1.4) is true then it follows that $P(D_t(i) = 1) = h(i;\theta_0)$, and hence that $E[D_t(i)] = h(i;\theta_0)$. So, using these indicator variables, it can be seen that (1.4) implies the following vector of k population moment conditions

$$E\begin{bmatrix} D_t(1) \;-\; h(1;\theta_0) \\ D_t(2) \;-\; h(2;\theta_0) \\ \qquad\quad . \\ \qquad\quad . \\ D_t(k) \;-\; h(k;\theta_0) \end{bmatrix} \;=\; 0 \tag{1.7}$$

Since $\sum_{i=1}^{k}\{D_t(i) - h(i;\theta_0)\} = 0$ by definition, only $k-1$ of the population moment conditions actually provide unique information about θ_0. However, we retain all k to elicit the connection with the goodness of fit statistic. If $k-1 \geq p$ – which we have assumed implicitly all along – then these population moment equations can be used to estimate θ_0. The sample analogs to (1.7) are given by

$$\begin{bmatrix} \hat{p}_1 \;-\; h(1;\theta) \\ \hat{p}_2 \;-\; h(2;\theta) \\ \qquad\quad . \\ \qquad\quad . \\ \hat{p}_k \;-\; h(k;\theta) \end{bmatrix} \;=\; 0 \tag{1.8}$$

choosing estimators to minimize the goodness of fit statistic. However, her focus was on trying to uncover a sense in which Method of Moments estimators could be considered optimal. In fact, she found that Method of Moments estimators gave a good approximation to the values which minimized the goodness of fit statistic in the examples considered in her paper. This finding may explain why this alternative method of estimation was not explored more fully until twelve years later. See Bera and Bilias (2002) for further discussion of the origins of Minimum Chi-Square.

The elements on the left hand side of (1.8) can be recognized as the same terms which appear inside the square in the numerator of the version of the goodness of fit statistic in (1.6). We are now in a position to establish the connection between Minimum Chi-Square estimation of θ_0 and estimation based on the population moment conditions in (1.7). First consider the case in which there are as many unique moment conditions as unknown parameters, that is $k-1=p$. By definition, the Method of Moments estimator, $\hat{\theta}_T$ say, satisfies $\hat{p}_i - h(i,\hat{\theta}_T) = 0$ for $i = 1, 2 \ldots p$.[10] This property implies that $GF_T(\hat{\theta}_T) = 0$, and since $GF_T(\theta) \geq 0$, it must follow that $\hat{\theta}_T$ also minimizes $GF_T(\theta)$. So if $k-1=p$ then the Minimum Chi-Square estimator is just the Method of Moments estimator based on (1.7). Now consider the case in which there are more unique moment conditions than parameters, that is $k-1>p$. In this case, the principle of Method of Moments estimation does not work, but Minimum Chi-Square is still valid. The key difference is that Method of Moments is defined as the solution to a set of moment conditions and this solution only exists if $k-1=p$, whereas Minimum Chi-Square is defined in terms of a minimization, which can be performed for any $k-1 \geq p$. This suggests that to estimate (μ_0, σ_0^2) from the first three moments of the normal distribution, it is necessary to formulate the estimation in terms of a minimization. To implement such a strategy, it is necessary to specify an appropriate minimand. Once again, Minimum Chi-Square provides the answer. It is easily verified that

$$GF_T(\theta) = T \begin{bmatrix} \hat{p}_1 - h(1;\theta) \\ \hat{p}_2 - h(2;\theta) \\ \vdots \\ \hat{p}_k - h(k;\theta) \end{bmatrix}' \begin{bmatrix} \hat{p}_1^{-1} & 0 & \ldots & 0 \\ 0 & \hat{p}_2^{-1} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \hat{p}_k^{-1} \end{bmatrix} \begin{bmatrix} \hat{p}_1 - h(1;\theta) \\ \hat{p}_2 - h(2;\theta) \\ \vdots \\ \hat{p}_k - h(k;\theta) \end{bmatrix} \tag{1.9}$$

and so $GF_T(\theta)$ can be interpreted as a quadratic form in the sample moment condition (1.8). Notice that the matrix in the centre of (1.9) is positive definite[11] by construction and so ensures that $GF_T(\theta) \geq 0$. This structure leads to the following intuitively appealing interpretation of the Minimum Chi-Square estimator: it is the value of θ which is *closest* to solving the sample moment conditions in the metric of $GF_T(\theta)$.

It takes only a little reflection to realize that the same approach can be applied to the estimation of any problem in which there are more moments than parameters to be estimated. To illustrate how, let us return to estimation of (μ_0, σ_0^2) based on (1.1)-(1.3). For this problem, the minimand takes the form

$$MC_T(\mu, \sigma^2) = \begin{bmatrix} m_v(1) - \mu \\ m_v(2) - (\sigma^2 + \mu^2) \\ m_v(3) - 3m_v(2)\mu + 3m_v(1)\mu^2 - \mu^3 \end{bmatrix}' M_T \times$$

[10] Note that we can obtain $\hat{\theta}_T$ by solving any $k-1$ of the sample moment conditions in (1.8), and that the estimator must satisfy the remaining sample moment condition because $\sum_{i=1}^{k}\{\hat{p}_i - h(i;\hat{\theta}_T)\} = 0$ by construction.

[11] The goodness of fit statistic is undefined unless $\hat{p}_i > 0$ for all i.

$$\begin{bmatrix} m_v(1) - \mu \\ m_v(2) - (\sigma^2 + \mu^2) \\ m_v(3) - 3m_v(2)\mu + 3m_v(1)\mu^2 - \mu^3 \end{bmatrix} \tag{1.10}$$

where M_T is a positive definite matrix which may depend on T, and $m_v(i) = T^{-1}\sum_{t=1}^{T} v_t^i$. Notice that this minimand embodies two modifications of (1.9) beyond the choice of sample moments. First, the scaling factor, T, has been omitted, because it has no impact on the minimization. Secondly, we have not specified an exact form for the matrix in the quadratic form; it can be any positive definite matrix. The Minimum Chi-Square estimators of (μ_0, σ_0^2) are the values of (μ, σ^2) which minimize $MC_T(\mu, \sigma^2)$.

This connection between Minimum Chi-Square and moment based estimation seems to have been made first during the late 1940s and the 1950s. It was certainly at this time that researchers began to realize the potential generality of the method, although their perspective was limited inevitably by the computational constraints of that time. Ferguson (1958) developed the statistical theory for the estimator in the case where the population moment condition takes the form $E[g(v_t)] - h(\theta) = 0$ and v_t is an i.i.d. process.[12] However, for some reason, his contribution appears not to have impacted on econometrics – perhaps because the functional form of the moment condition was not particularly appropriate for econometric applications of that time. However, with hindsight, it can be recognized that the statistical framework developed by Ferguson (1958) contains many of the elements which reappeared in the GMM literature twenty-five years later albeit in a far more general context.

The third important antecedent of GMM is the method of Instrumental Variables (IV) estimation. Unlike Method of Moments and Minimum Chi-Square, IV was specifically developed to exploit the information in moment conditions for the estimation of structural economic models. This method appears to have been first applied in an analysis of demand and supply of agricultural commodities in the 1920s. In both an U.S. Department of Agriculture Bulletin (Wright, 1925), and also in the appendix to his father's book, *The Tariff on Animal and Vegetable Oils* (Wright 1928), Sewall Wright showed how Method of Moments could be used to estimate the parameters of supply and demand equations.[13] He presented these estimators using a technique known as "Path Analysis", but it is most convenient to adopt an alternative approach which has become the standard derivation in econometric textbooks. To illustrate we consider the system of equations

[12] Ferguson (1958) also considers a number of variations on this estimation problem, some of which had been analyzed earlier by Barankin and Gurland (1951). Also see Neyman (1949).

[13] Sewall Wright (1889–1988) was an American who is best known for his work on population genetics. Following his position at the USDA, he became Professor of Zoology at the University of Chicago and is considered to be of the three founders of modern theoretical population genetics.

$$\begin{aligned} q_t^D &= \alpha_0 p_t + u_t^D \\ q_t^S &= \beta_{0,1}^{'} n_t + \beta_{0,2} p_t + u_t^S \\ q_t^D &= q_t^S = q_t \end{aligned} \tag{1.11}$$

where q_t^D, q_t^S represent demand and supply in year t, p_t is the price of the commodity in that year and n_t is a vector containing factors that affect supply. The market is assumed to clear and the total quantity produced is denoted q_t. For our purposes here, it suffices to consider the problem of how to estimate α_0 given a sample of T observations on q_t and p_t. An Ordinary Least Squares (OLS) regression of q_t on p_t runs into problems here because price and output are simultaneously determined and this causes OLS estimates to be biased, e.g. see Judge, Griffiths, Hill, Lutkepohl, and Lee (1985, p.570). Sewall Wright solved these problems as follows. Suppose there is an observable variable z_t^D which is related to price but whose covariance with u_t^D, $Cov[z_t^D, u_t^D]$, is zero. An example would be any of the factors that affect supply, such as an input price or yield per acre. Then by taking the covariance of z_t^D with both sides of the demand equation in (1.11) it follows that

$$Cov[z_t^D, q_t] - \alpha_0 Cov[z_t^D, p_t] = 0 \tag{1.12}$$

It is convenient to simplify this moment condition using other properties of the model. Typically, it is assumed that $E[u_t^D] = 0$ and so $E[q_t] = \alpha_0 E[p_t]$. Using this identity in (1.12), the moment condition can be rewritten as[14]

$$E[z_t^D q_t] - \alpha_0 E[z_t^D p_t] = 0 \tag{1.13}$$

Equation (1.13) provides a population moment condition involving the observable variables and the unknown parameter, α_0, which can be used as a basis for estimation. Pearson's Method of Moments principle leads to the estimation of the parameters by the values which solve the analogous sample moments, namely

$$\hat{\alpha}_T = \sum_{t=1}^{T} z_t^D q_t / \sum_{t=1}^{T} z_t^D p_t \tag{1.14}$$

This equation can be recognized as what is known today as an Instrumental Variables estimator with z_t^D being refered to as the "instrument". However this term was not coined until the 1940s when IV was rediscovered and came to stay in econometrics. In fact, Wright's work was largely ignored by economists until Goldberger (1970) returned it to its rightful place in the history of econometrics.

A similar Method of Moments reasoning was used in the 1940s. However, this time, IV was proposed as a solution for the problems caused by errors in variables. To illustrate, consider the case in which

$$y_t = \gamma_0 x_t^0 + u_{1,t} \tag{1.15}$$

[14] Recall that for any two random variables a and b, $Cov[a, b] = E[ab] - E[a]E[b]$.

but x_t^0 is only observed with error,

$$x_t = x_t^0 + u_{2,t}$$

Since the regressor is unobserved, equation (1.15) cannot be estimated directly. Instead inference is based on

$$y_t = \gamma_0 x_t + u_t \tag{1.16}$$

Ordinary Least Squares estimation of (1.16) is biased because x_t and $u_t = u_{1,t} - \gamma_0 u_{2,t}$ are correlated; *e.g.* see Judge, Griffiths, Hill, Lutkepohl, and Lee (1985, p.705–8). Reiersøl (1941) and Geary (1942, 1943) independently proposed solving this problem by introducing a variable z_t which is correlated with x_t but uncorrelated with u_t.[15] Using the same intuition as Wright, Reierosol and Geary deduced the moment condition

$$Cov[z_t, y_t] - \gamma_0 Cov[z_t, x_t] = 0$$

The Method of Moments estimation principle leads to the analogous formula to (1.14) for the IV estimator of γ_0.

Reiersøl (1945) introduced the term "instrumental variables" and Geary (1949) derived certain statistical properties of the estimator in the context of the errors in variables model. Durbin (1954) extended the method to simultaneous equation models, and Sargan(1958, 1959) provided the first complete theoretical analyses of the estimator.[16] Building from this basis, the IV framework has become so developed that, prior to the introduction of GMM, it was typically treated in econometrics as an estimation technique in its own right rather than being perceived as an example of the Method of Moments.[17] Within this literature on IV, Amemiya (1974) and Jorgenson and Laffont (1974) played an important role in extending the method to nonlinear models, and the statistical theory employed in these papers is an important precursor to the arguments used to analyze the properties of GMM.

The above discussion has illustrated some of the problems to which moment based estimation has been applied. Over the years, considerable attention has been focused on analyzing the properties of these estimators and various associated inference techniques. However, this theory has tended to place restrictions on the functional form of the population moment condition. One of

[15] See Morgan (1990, p.220–8) and Aldrich (1993) for more detailed discussions of the emergence of IV in the 1940s. Olav Reiersøl (1908–) is a Norwegian statistician who made a number of important contributions to econometrics, most notably through his work on IV and identification. He also contributed to other areas of statistics as well as genetics. Robert (Roy) Geary (1896–1983) was an Irishman who worked as a government statistician in Dublin for most of his career. Apart form his work in mathematical statistics, he is also known for being one of the pioneers in the field of national income accounting.

[16] See Arellano (2002) for an appraisal of the connection between Sargan's work and GMM.

[17] There are some exceptions. For instance, Burguette, Gallant, and Souza (1982) use the term "Method of Moments" to denote a class of estimators of the parameters of nonlinear static simultaneous equation model which includes IV estimators.

the main contributions of GMM is to provide a framework for the statistical analysis based on essentially any population moment condition. Accordingly, it is necessary to adopt a broad definition of a population moment condition.

Definition 1.1 Population Moment Condition
Let θ_0 be a vector of unknown parameters which are to be estimated, v_t be a vector of random variables and $f(.)$ a vector of functions then a population moment condition takes the form

$$E[f(v_t, \theta_0)] = 0 \tag{1.17}$$

for all t.

This definition encompasses the examples discussed above. For instance, the moment condition in (1.1) can be obtained from (1.17) by putting

$$f(v_t, \theta) = \begin{bmatrix} v_t - \mu_0 \\ v_t^2 - (\sigma_0^2 + \mu_0^2) \end{bmatrix}$$

where $\theta_0 = (\mu_0, \sigma_0^2)'$. Wright's example in (1.13) is obtained by putting

$$f(v_t, \theta) = z_t^D q_t - \alpha_0 z_t^D p_t$$

where $v_t = (z_t^D, q_t, p_t)'$ and $\theta_0 = \alpha_0$.

Just as in Minimum Chi-Square, GMM involves choosing parameter estimators to minimize a quadratic form in a weighting matrix, W_T, and the sample moment $T^{-1}\sum_{t=1}^{T} f(v_t, \theta)$.

Definition 1.2 Generalized Method of Moments Estimator
The Generalized Method of Moments estimator based on (1.17) is the value of θ which minimizes:

$$Q_T(\theta) \;=\; T^{-1}\sum_{t=1}^{T} f(v_t, \theta)' W_T T^{-1} \sum_{t=1}^{T} f(v_t, \theta) \tag{1.18}$$

where W_T is a positive semi-definite matrix which may depend on the data but converges in probability to a positive definite matrix of constants.

The restrictions on the weighting matrix are required to ensure that $Q_T(\theta)$ is a meaningful measure of distance. Notice that the positive semi-definiteness of W_T ensures both that $Q_T(\theta) \geq 0$ for any θ, and also that $Q_T(\hat{\theta}_T) = 0$ if $T^{-1}\sum_{t=1}^{T} f(v_t, \hat{\theta}_T) = 0$. However, positive *semi*-definiteness leaves open the possibility that $Q_T(\hat{\theta}_T)$ is zero at a value of $\hat{\theta}_T$ which does not satisfy the sample moment conditions. Since all our analysis is based on asymptotic theory, it is only necessary to rule out this eventuality in the limit as $T \to \infty$.

A comparison of (1.10) and (1.18) indicates that Minimum Chi-Square and GMM are essentially the same method. With hindsight, it might be argued that a new terminology was not really needed. However, Hansen (1982) refered

to the estimator in Definition 1.2 as "Generalized Method of Moments", and that is the name by which the method is known in econometrics.[18] We shall, therefore, follow this practice.

The next section presents five examples of moment conditions from models in Table 1.1. These models have been carefully selected because they provide convenient illustrations of many of the issues discussed in this book. Here, the focus is on showing how the population moment conditions arise and the potential problems encountered with maximum likelihood estimation in these models.

1.3 Five Examples of Moment Conditions in Economic Models

1.3.1 Consumption-Based Asset Pricing Model

The consumption-based asset pricing model is used by financial economists to explain how assets are priced and by macroeconomists to explain the evolution of consumption spending. To see how this can be done, it is necessary first to present the model formally and derive the population moment conditions which are the basis for GMM estimation. The ultimate aim of the model is to explain aggregate movements. This is done using a framework in which aggregate outcomes are assumed to be the result of the decisions made by a single "representative" agent. This representative agent approach is certainly open to criticism, e.g. see Kirman (1992), but nevertheless has received considerable attention in the literature. The general theoretical structure was first developed by Lucas (1978). However, Hansen and Singleton (1982) were first to highlight and exploit the potential of GMM in these types of models.

Consider the case where a representative agent makes decisions about consumption expenditures and investment to maximize his/her expected discounted utility

$$E[\sum_{i=0}^{\infty} \delta_0^i U(c_{t+i})|\Omega_t]$$

where c_t is consumption in period t, $U(.)$ is a strictly concave utility function, δ_0 is a constant discount factor and Ω_t is the information set available to the agent at time t. In any period the agent can choose to spend his/her income on either goods for consumption or investments in a collection of N assets with maturities $m_j, j = 1, 2, \ldots N$. Let $q_{j,t}$ be the quantity of asset j held at the end

[18] In fact, this terminology originates from a set of unpublished lecture notes produced by Christopher Sims for his graduate econometrics course at the University of Minnesota. Interestingly, Sims used the term to denote an estimator which is obtained by solving a linear combination of moment conditions rather than via the minimization in Definition 1.2. Hansen developed certain statistical results for Sim's estimator as part of his Ph.D. thesis submitted to the University of Minnesota in 1978. Hansen and Sims provide interesting background on the genesis of the method in interviews published in the October 2002 issue of the *Journal of Business and Economic Statistics*.

of period t, $p_{j,t}$ be the price of asset j at time t, $r_{j,t}$ be the period t payoff from a unit of the j^{th} asset purchased in period $t-m_j$, and w_t be real labour income in period t. All prices are denominated in terms of the consumption good.[19] The budget constraint is

$$c_t + \sum_{j=1}^{N} p_{j,t} q_{j,t} = \sum_{j=1}^{N} r_{j,t} q_{j,t-m_j} + w_t$$

for all t. The optimal path of consumption and investment satisfies

$$p_{j,t} U'(c_t) = \delta_0^{m_j} E[r_{j,t+m_j} U'(c_{t+m_j}) | \Omega_t] \tag{1.19}$$

for all t and $j = 1, 2, \ldots, N$, where $U'(c)$ denotes the marginal utility of consumption. This condition states that the utility lost by foregoing consumption in period t to purchase a unit of asset j, $p_{j,t}U'(c_t)$, must equal the value in period t of the expected utility gained from consuming the return on the investment in period $t+m_j$, $\delta_0^{m_j} E[r_{j,t+m_j} U'(c_{t+m_j})|\Omega_t]$. Equation (1.19) can be rewritten as

$$E[\delta_0^{m_j} (r_{j,t+m_j}/p_{j,t}) \{U'(c_{t+m_j})/U'(c_t)\} | \Omega_t] - 1 = 0 \tag{1.20}$$

for j=1,2,...N. Equation (1.20) is refered to as the Euler equation of the system, after the mathematician Leonhard Euler (1707–83) who derived an analogous equation to characterize the solution path in the calculus of variations problem. The Euler equation places a restriction on the co-movements of consumption and asset prices and so can be used by macroeconomists and financial economists to learn about these variables.

So far, the analysis has been in terms of a general utility function, but to make (1.20) operational it is necessary to specify a particular functional form. At this stage it is most convenient to follow Hansen and Singleton (1982) and define

$$U(c_t) = \frac{c_t^{\gamma_0} - 1}{\gamma_0} \tag{1.21}$$

The parameter γ_0 must be less than one for the utility function to be concave. This functional form is known as the constant relative risk aversion (CRRA) utility function because the relative risk aversion of the representative agent is $(1-\gamma_0)$ at any level of consumption. Differentiating (1.21) and making the appropriate substitutions into (1.20), the Euler equation becomes

$$E[\delta_0^{m_j} (r_{j,t+m_j}/p_{j,t}) (c_{t+m_j}/c_t)^{\gamma_0 - 1} | \Omega_t] - 1 = 0 \tag{1.22}$$

Clearly with this specification there are two parameters to be estimated, namely (γ_0, δ_0). Taking unconditional expectations of the Euler equation provides one population moment condition involving these parameters, but, in fact, (1.22) implies many more moment conditions. If we set

$$u_{j,t}(\gamma_0, \delta_0) \;=\; \delta_0^{m_j} (r_{j,t+m_j}/p_{j,t}) (c_{t+m_j}/c_t)^{\gamma_0 - 1} - 1$$

[19] In other words, p_{jt} is the price of the asset in dollars divided by the price of the consumption good in dollars.

then an iterated conditional expectations argument can be used in conjunction with the Euler condition in (1.22) to show that

$$E[u_{j,t}(\gamma_0, \delta_0)\, z_t] \;=\; E\,[\,E[u_{j,t}(\gamma_0, \delta_0)\,|\Omega_t]\, z_t] \;=\; 0 \tag{1.23}$$

for any vector $z_t \in \Omega_t$. In this context, z_t might include a constant, which amounts to taking the unconditional expectation of the Euler equation, and variables such as $r_{j,t}/p_{j,t-m_j}$, c_t/c_{t-m_j} or indeed any other macroeconomic variables contained in the representative agent's information set. The moment conditions in (1.23) provide the basis for GMM estimation of the parameters (γ_0, δ_0).

In contrast, Maximum Likelihood estimation would involve specifying the conditional distribution for $\{(r_{j,t+m_j}/p_{j,t}, c_{t+m_j}/c_t); j = 1, 2, \ldots N\}$ and maximizing the likelihood subject to the constraint in (1.22) for each t. The latter would involve numerical integration in most cases and is consequently computationally very burdensome.[20] Furthermore, due to the inherent nonlinearlity of the model, Hansen and Singleton (1982) show that MLE is unlikely to yield unbiased inferences unless the distribution its correctly specified.[21] The potential for this bias can be reduced by using a flexible functional form which is capable of approximating a wide class of probability density functions; e.g. see Gallant and Tauchen (1989). However this further adds to the computational burden.

1.3.2 Evaluation of Mutual Fund Performance

Mutual funds consist of a portfolio of financial assets administered by a fund manager.[22] The role of the manager is to vary the composition of this portfolio in response to any relevant economic or financial information to meet some specified criterion. An investor can purchase shares in the fund and thereby acquire an asset whose rate of return is that of the portfolio. The incentive for investing in the fund stems from the ability of the manager to acquire and efficiently process market information. However, in practice managers may misread their information or simply be the victims of unpredictable events. In this case the average investor may have received a better return by constructing his/her own portfolio based on a more restricted information set. Naturally there is considerable interest in identifying which funds have yielded superior returns compared to some suitably chosen benchmark. This topic received some attention in the 1970s, but interest has increased recently in response to the massive growth in assets managed by such funds in the U.S. In this section we describe a measure of fund performance proposed by Chen and Knez (1996). These authors actually propose a number of related measures but at this time it is sufficient to focus on the simplest because it illustrates how the moment condition arises.

[20] One exception is the model studied by Hansen and Singleton (1983). They estimate the CRRA model described above by Maximum Likelihood under the assumption that $(\{r_{j,t+m_j}/p_{j,t}\}, c_{t+m_j}/c_t)$ have a lognormal distribution.

[21] See Section 3.8.

[22] In practice funds may be administered by a team of managers, but for expositional convenience we refer to a single manager.

To begin with, it is useful to review two very fundamental results from finance. The "Law of One Price" states that any two investments with the same payoff in every state of the world must have the same price (e.g. see Ingersoll, 1987, p.59). The second fundamental result is deduced from this law. Chamberlain and Rothschild (1983) show that the Law of One Price implies a useful characterization of the relationship between the price and return of a financial asset. To flesh out this asset pricing equation, it is necessary to introduce some notation. Let X_t be a vector of $(N \times 1)$ payoffs on N traded assets with n^{th} element $x_{n,t}$ which is the time t return per time $t-1$ dollar invested in asset n. Notice that each payoff , $x_{n,t}$, can be interpreted as an asset with a price of \$1. Chamberlain and Rothschild (1983) show that the Law of One Price implies there exists a unique scalar random variable $d_t = X_t'\delta_0$ such that

$$E[X_t d_t] = 1_N \tag{1.24}$$

where 1_N is a $N \times 1$ vector of ones and δ_0 is an $N \times 1$ vector of constants. The variable d_t is known as the stochastic discount factor.[23] As we shall see, this asset pricing equation is central to Chen and Knez's method.

To evaluate the performance of a mutual fund it is necessary to have some benchmark. Since managers are essentially selling their ability to gather and process information, it is natural to compare the fund's return to that achievable by an investor with no such information. This "uninformed" investor is taken to be an individual who holds a constant composition portfolio and hence never buys or sells assets in response to new information. Let the weights of this portfolio be collected into an $N \times 1$ vector α whose n^{th} element is α_n. The return on such a passively held portfolio in period t is given by

$$R_t(\alpha) = \sum_{n=1}^{N} \alpha_n x_{n,t}, \quad \text{for} \quad \sum_{n=1}^{N} \alpha_n = 1 \tag{1.25}$$

Notice the weights have been normalized to sum to one and so $R_t(\alpha)$ can be interpreted as the payoff achievable from an initial investment of \$1. Also, the weight on $x_{n,t}$ can be positive indicating a long position in the asset or it can be negative indicating a short position.[24] In contrast to the uninformed investor, the fund manager has the option of updating the composition of the fund's portfolio and this is reflected by making the weights in his/her portfolio time dependent. Accordingly, the return on the fund is

$$r_t^m = \sum_{n=1}^{N} \theta_{n,t} x_{n,t}, \quad \text{for} \quad \sum_{n=1}^{N} \theta_{n,t} = 1 \tag{1.26}$$

[23] It is also known as the "pricing operator" or the "pricing kernel".

[24] An investor holds a long position in an asset if he/she owns units of the asset. An investor holds a short position in an asset if he/she has sold units of an asset that they did not own, say by borrowing it from a broker, and must return the borrowed units at some point in the future.

where the superscript m represents "mutual fund". Again the weights, $\{\theta_{n,t}\}$ this time, sum to one and so r_t^m represents the return on a \$1 investment. Clearly, the manager has the option to leave the weights unchanged over time. However, if he/she follows this strategy then the fund does not increase the opportunity set for investors. In this case, Chen and Knez argue the manager has provided no service and so should receive a performance measure of zero. Furthermore, they argue that the manager should receive the same evaluation if he/she changes the weights of the fund's portfolio but this only leads to a return which could have been earned by some passively held portfolio over the same period. A positive performance measure is only earned if the fund return exceeds that on *any* passively held portfolio over the same period.

It is clearly desirable to identify which funds have positive performance measures. It turns out to be most convenient to address this issue by reversing the question and seeking to identify funds with a zero performance measure. Chen and Knez (1996) show that the fund has a zero performance measure relative to the benchmark set of passively held portfolios in (1.25) if

$$\lambda(r_t^m, d_t) = E[r_t^m X_t' \delta_0] - 1 = 0 \tag{1.27}$$

To assess whether (1.27) is true, an estimate of δ_0 is needed. Chen and Knez solve this problem by combining (1.24) and (1.27) into the augmented population moment condition

$$E[Q_t X_t' \delta_0] - 1_{N+1} = 0 \tag{1.28}$$

where $Q_t = (X_t', r_t^m)'$. These equations provide a basis for the estimation of δ_0. At first glance this appears to impose the very hypothesis that we wish to test. However, (1.28) is a vector of $N+1$ moment conditions in N parameters and so the sample moments are not zero when evaluated at the estimated value of δ_0. As we shall see, this leaves scope for testing whether the data are actually consistent with (1.28) and hence the hypothesis that the fund has a performance measure of zero.

This problem could be approached using Maximum Likelihood estimation. It would involve specifying the conditional distribution of Q_t given the information available at time t-1 and assessing whether the estimated distribution satisfied the moment restriction in (1.27). However, this approach encounters both the types of problem described in Section 1.1. First, it is necessary to make a distributional assumption. A natural choice is normality but, unfortunately, this is not appropriate for stock return data; see Richardson and Smith (1993). To date there is no consensus on the appropriate choice; see Fama (1976, p.26) and Bollerslev, Engle, and Nelson (1994) for discussions of common features of the distribution of asset return data. Of course, unless the true distribution is used there is no guarantee that MLE yields more precise estimators than those obtained by GMM. Second, such estimation will involve significantly more parameters than the N involved in Chen and Knez's approach and so will be more computationally intensive.

1.3.3 Conditional Capital Asset Pricing Model

Harvey (1991) investigates whether the conditional Capital Asset Pricing Model (conditional CAPM hereafter) can explain the differences in the average returns across financial markets in industrialized countries. The original, or unconditional, CAPM is one of the main models in finance and has received a lot of academic and non-academic attention; e.g. see Malkiel (1987). Its importance stems from its provision of an explicit relationship between the expected rate of return on an asset and the sytematic risk of holding that asset. In this context risk is measured by the variance of the asset return and derives from two sources. There is *systematic risk* which derives from the inherent uncertainties in the macroeconomy and there is *unsystematic risk* which is specific to the stock in question.[25] Systematic risk is measured as the variance of the so-called "market portfolio". This portfolio consists of all the assets in the market and so represents the most diversified portfolio it is possible to hold. By holding a suitably large portfolio the investor can diversify away the unsystematic risk and so he/she is only compensated for bearing the systematic risk in holding an asset. Systematic risk is present in all risky assets but to different degrees depending on the nature of the asset. Another attractive feature of CAPM is that it provides a measure of the degree of the systematic risk present in an asset; this measure is known as the *investment beta*.

One weakness of the original CAPM is its implicit assumption that the level of systematic risk in an asset stays constant over time. Intuition suggests this risk should vary in response to changes in the macroeconomy and decisions made by the firm issuing the asset. This type of behaviour can be incorporated into the theory and yields the conditional CAPM. To introduce the model it is necessary to define first some notation. Let $R_{i,t}$ be the return in period t on investing \$1 in the asset in question in period t $-$ 1, $R_{m,t}$ be the corresponding return on investing \$1 in the market portfolio in period t $-$ 1 and $R_{f,t}$ be the return in period t from investing \$1 in the the risk free asset in period t $-$ 1.[26] The excess returns on the asset and the market portfolio are defined respectively as $r_{i,t} = R_{i,t} - R_{f,t}$ and $r_{m,t} = R_{m,t} - R_{f,t}$. The conditional CAPM implies

$$E[r_{i,t}|\Omega_{t-1}] = \beta_{i,t} E[r_{m,t}|\Omega_{t-1}] \tag{1.29}$$

where the conditional investment beta is

$$\beta_{i,t} = Cov[r_{i,t}, r_{m,t}|\Omega_{t-1}]/Var[r_{m,t}|\Omega_{t-1}] \tag{1.30}$$

and $E[.|\Omega_{t-1}]$, $Var[.|\Omega_{t-1}]$ and $Cov[.|\Omega_{t-1}]$ denote respectively the expectation, variance and covariance conditional on an information set Ω_{t-1}.[27]

We can now return to the specifics of Harvey's (1991) study. He examines whether the model in (1.29)–(1.30) can explain the variation in the returns

[25] Systematic and unsystematic risk are also refered to as market and idiosyncratic risk respectively.

[26] A risk-free asset is one whose return is known at the time of purchase.

[27] The original CAPM can be obtained from (1.29)–(1.30) by replacing the conditional expectations, variance and covariance by their unconditional counterparts.

across seventeen international stock markets. In this context $r_{i,t}$ becomes the excess return on holding the market portfolio for country i. The variable $r_{m,t}$ is the excess return from holding a "world market" portfolio that is weighted combination of the returns on a variety of world-wide investments; see Harvey (1991) for details. To make the model operational it is necessary to specify the conditional means of the excess returns. To this end, let z_{t-1} be the vector of relevant economic and financial variables contained in Ω_{t-1}. Harvey assumes that

$$\begin{aligned} E[r_{i,t}|\Omega_{t-1}] &= z'_{t-1}\delta_{i,0} \\ E[r_{m,t}|\Omega_{t-1}] &= z'_{t-1}\delta_{m,0} \end{aligned} \tag{1.31}$$

where $\delta_{m,0}$, $\{\delta_{i,0}\}$ are unknown vectors of constants. The parameters to be estimated are $\delta_{m,0}$ and $\{\delta_{i,0}; i = 1, 2, ...17\}$. The estimation is based on two types of moment conditions: those implied by the specification of the conditional means, (1.31), and those implied by the conditional CAPM, (1.29)–(1.30). To present the moment conditions it is convenient to define

$$\begin{aligned} u_{i,t} &= r_{i,t} - z'_{t-1}\delta_{i,0} \\ u_{m,t} &= r_{m,t} - z'_{t-1}\delta_{m,0} \end{aligned} \tag{1.32}$$

The first set of moments comes from using iterated conditional expectations and $E[u_{i,t}|\Omega_{t-1}] = 0$ to show that

$$E[u_{i,t}z_{t-1}] = E[E[u_{i,t}z_{t-1}|\Omega_{t-1}]] = E[E[u_{i,t}|\Omega_{t-1}]z_{t-1}]] = 0 \tag{1.33}$$

Using a similar argument for $u_{m,t}$ and substituting from (1.32) yields the moment conditions

$$\begin{aligned} E[(r_{i,t} - z'_{t-1}\delta_{i,0})z_{t-1}] &= 0 \\ E[(r_{m,t} - z'_{t-1}\delta_{m,0})z_{t-1}] &= 0 \end{aligned} \tag{1.34}$$

for i=1,2,...,17. The second set of moment conditions comes directly from the conditional CAPM structure. The substitution of (1.30) into (1.29) plus some rearrangement yields

$$Var[r_{m,t}|\Omega_{t-1}]E[r_{i,t}|\Omega_{t-1}] - Cov[r_{i,t}, r_{m,t}|\Omega_{t-1}]E[r_{m,t}|\Omega_{t-1}] = 0 \tag{1.35}$$

Employing a similar iterated conditional expectations argument as in (1.33) and substituting from (1.31), it can be deduced that

$$\begin{aligned} E[\{(r_{m,t} - z'_{t-1}\delta_{m,0})^2 z'_{t-1}\delta_{i,0} - (r_{m,t} - z'_{t-1}\delta_{m,0}) \times \\ (r_{i,t} - z'_{t-1}\delta_{i,0})z'_{t-1}\delta_{m,0}\}z_{t-1}] = 0 \end{aligned} \tag{1.36}$$

for i=1,2,...17, which constitute the second set of moment conditions used in estimation.

This model can be estimated by Maximum Likelihood but, again, this approach will encounter the problems mentioned in Section 1.1. The endogenous

variables are $r_t = (r_{1,t}, r_{2,t}, ..., r_{17,t}, r_{m,t})'$. To implement MLE the conditional distribution for r_t must be specified so that it satisfies both the conditional mean specification in (1.31) and the relationship between the conditional means, conditional variances and covariances in (1.35). Once again, the normal distribution is a natural first choice, but just as in the mutual fund example, these asset returns do not possess this distribution. Therefore, MLE under the assumption of normality is not necessarily more precise than GMM although it should lead to unbiased inferences provided the variances are correctly calculated.[28] MLE would be also slightly more computationally burdensome than GMM due to the imposition on the likelihood of the restrictions between first and second moments implied by the conditional CAPM.

1.3.4 Inventory Holdings by Firms

A firm can choose to use its output to meet current demand or hold it as inventory. There is a considerable literature in macroeconomics which seeks to explain the level of inventory holdings in the aggregate economy; e.g. see the survey by Blinder and Maccini (1991). These studies typically proceed by modelling the sales and inventories of a particular industry as if they are the outcome of decisions made by a single representative firm. One popular line of theory is based on the assumption that the representative firm uses inventories to smooth production levels. Although intuitively reasonable, the production smoothing model has had mixed success in explaining aggregate inventory behaviour; see Blinder and Maccini (1991). One response to this evidence has been to argue that firms smooth production costs and not levels. To test if either of these hypotheses can explain the data it is desirable to perform inference within a model which allows both types of behaviour. Eichenbaum (1989) presents such a model and uses it to analyze the inventory holdings in a number of two digit SIC industries in the U.S. This section outlines Eichenbaum's model.

The representative firm is assumed to face two types of costs: production costs and inventory holding costs. The production costs are assumed to be:

$$C_{Q,t} = \nu_t Q_t + (\alpha_0/2) Q_t^2 \tag{1.37}$$

where Q_t is the firm's output at time t and ν_t is a random variable capturing stochastic shocks to the marginal cost of production. Since ν_t is random, marginal costs are a random function and so there is an incentive for holding inventories to smooth production costs. However, if $\nu_t = 0$ then marginal cost is a deterministic function of output and so the only incentive for holding inventories is to smooth the level of production. The constant α_0 controls the slope of the marginal cost schedule: if α_0 is positive then the marginal costs are increasing with output and if α_0 is negative then the marginal costs are

[28] If the distribution is misspecified then in general the information matrix identity does not hold. This affects the formulae for the variances of the estimators; see White (1982) and Section 3.8.

decreasing with output. The inventory holding costs are assumed to be

$$C_{I,t} = (\delta_0/2)(I_t - \gamma_0 S_t)^2 + (\eta_0/2)I_t^2 \tag{1.38}$$

where I_t, S_t are the inventories and sales of the firm at time t respectively.[29] The constants γ_0, δ_0 and, η_0 are all nonnegative. The first term in (1.38) captures the cost to the firm of inventories deviating from the desired fraction of sales, $\gamma_0 S_t$. The second term in (1.38) captures the storage costs associated with holding inventories. The combination of the production and inventory costs yields the total cost function of the firm:

$$C_t = C_{Q,t} + C_{I,t} \tag{1.39}$$

By definition, sales, inventories and production are fundamentally related by: $Q_t = S_t + I_t - I_{t-1}$. Using this identity Q_t can be explicitly eliminated from the model. Therefore, the firm is assumed to choose I_{t+1} and S_{t+1} to maximize future discounted profits, denoted

$$E[\sum_{j=0}^{\infty} \beta_0^j (p_{t+j} S_{t+j} - C_{t+j}) | \Omega_t] \tag{1.40}$$

where p_t is the price in period t of the good produced by the firm, β_0 is the discount factor and Ω_t is the firm's information set at time t.

To characterize the optimal path for inventories and sales it is necessary to make some assumption about the random variable ν_t. Eichenbaum (1989) assumes that

$$\nu_t = \rho_0 \nu_{t-1} + \epsilon_t \tag{1.41}$$

where $E[\epsilon_t | \Omega_{t-1}] = 0$, $Var[\epsilon_t | \Omega_{t-1}] < \infty$ and $|\rho_0| < 1$. In this case the Euler equation implies the following condition:

$$E[h_{t+2}(\psi_0) - \rho_0 h_{t+1}(\psi_0) | \Omega_t] = 0 \tag{1.42}$$

where

$$h_{t+1}(\psi_0) = I_{t+1} - \{\lambda_0 + (\lambda_0 \beta_0)^{-1}\} I_t + \beta_0^{-1} I_{t-1} + S_{t+1} - \phi_0 \beta_0^{-1} S_t \tag{1.43}$$

and the parameters of the system are ρ_0 and the cost function parameters $\psi_0 = (\lambda_0, \beta_0, \phi_0)'$ where $\phi_0 = (1 - \delta_0 \gamma_0 / \alpha_0)$ and λ_0 is a root from the second order autoregressive polynomial governing the time series properties of the inventory series; see Eichenbaum (1989) for details. Using a similar iterated expectations argument as in (1.23), it can be shown that

$$E[\{h_{t+2}(\psi_0) - \rho_0 h_{t+1}(\psi_0)\} z_t] = 0 \tag{1.44}$$

[29] Eichenbaum includes a term $\eta_{1t} I_t$ where η_{1t} is a parameter which depends on t. However this parameter is argued to be eliminated by a data transformation prior to estimation. So for expositional simplicity this parameter has been set to zero.

for any vector $z_t \in \Omega_t$. For example, Eichenbaum estimates the parameters using the lagged values of inventories and sales, $\{S_{t-i}, I_{t-i}; i = 1, 2, \ldots k\}$, in z_t.

Maximum Likelihood would involve estimation of the bivariate vector autoregressive system for (S_t, I_t) subject to the nonlinear cross equation restrictions on the parameters implied by the model. This is likely to be more computationally burdensome with the exact degree depending on the choice of distribution. Unfortunately, economic theory provides no guidance on this choice. Once again, unless the chosen distribution is correct then the resulting MLE's are unlikely to have the anticipated optimal properties.

1.3.5 Stochastic Volatility Models of Exchange Rates

The preceding models have all been developed from economic theory. In some circumstances, it may be desired to capture the time series properties of an economic variable using a purely statistical model. An example of such a model would be the autoregressive integrated moving average (ARIMA) class developed by Box and Jenkins (1976). However, ARIMA models are not particularly appropriate for many financial assets because they do not allow the conditional variance to change over time. This has led to considerable interest in statistical models which can capture this type of behaviour. The most prominent of these models are the autoregressive conditional heteroscedasticity (ARCH) models introduced by Engle (1982), which have been applied very widely in finance, see the survey by Bollerslev, Chou, and Kroner (1992). More recently, a second class is receiving considerable attention and these are known as stochastic volatility models; see the survey by Ghysels, Harvey, and Renault (1996).

In this section we describe the stochastic volatility model used by Melino and Turnbull (1990) to analyze daily exchange rates. The model has its origins in a stochastic differential equation for the evolution of the exchange rate over time. However, we focus directly on the discrete time stochastic process which is used to approximate this underlying continuous time process. Let $y(\tau)$ denote the exchange rate at time τ and assume that the exchange rate is observed at times $\{\tau_1, \tau_2, \ldots \tau_T\}$. These observations are not at evenly spaced intervals because there are days on which no trading occurs, such as weekends and holidays. To accomodate these effects, it is useful to denote the distance between observations by $d_t = \tau_t - \tau_{t-1}$, and the minimum distance by $d = min_t(d_t)$. The discrete time approximation takes the form

$$\begin{aligned} y(\tau_t) \;=\; \alpha_0 d_t \;\; + \;\; & (1 + \beta_0 d_t) y(\tau_{t-1}) \\ & + x(\tau_{t-1}) y(\tau_{t-1})^{\gamma_0/2} d_t^{1/2} e(\tau_t) \end{aligned} \tag{1.45}$$

where the latent process $x(\tau_t)$ is generated by

$$ln[x(\tau_t)] \;=\; \delta_0 d \;+\; (1 + \eta_0 d) ln[x(\tau_t - d)] \;+\; \zeta_0 d^{1/2} u(\tau_t) \tag{1.46}$$

and

$$\begin{bmatrix} e(\tau_t) \\ u(\tau_t) \end{bmatrix} \;\sim\; IN\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix} \begin{bmatrix} 1 & \rho_0 \\ \rho_0 & 1 \end{bmatrix} \right) \tag{1.47}$$

Given that the model includes a distributional assumption, it is natural to use Maximum Likelihood. However, the evaluation of the conditional likelihood at time t involves a T-dimensional numerical integration which is computationally extremely burdensome – if not infeasible – on many currently available computer systems. However the normality assumption implies various population moment conditions which can form the basis of GMM estimation of the parameter vector $\theta_0 = (\alpha_0, \beta_0, \delta_0, \eta_0, \zeta_0, \rho_0)$.[30] For example, Melino and Turnbull (1990) show that the following population moment conditions hold:[31]

$$\begin{aligned}
E[w_t(\theta_0)] &= 0 \\
E[w_t^2(\theta_0)] - exp[2\mu_x + 2\sigma_x^2] &= 0 \\
E[w_t^3(\theta_0)] &= 0 \\
E[w_t^4(\theta_0)] - 3exp[4\mu_x + 8\sigma_x^2] &= 0 \\
E[|w_t(\theta_0)|] - (2/\pi)^{1/2} exp[\mu_x + 0.5\sigma_x^2] &= 0 \\
E[|w_t(\theta_0)|^3] - 2(2/\pi)^{1/2} exp[3\mu_x + 4.5\sigma_x^2] &= 0 \\
E[|w_t(\theta_0)|w_t(\theta_0)] &= 0 \\
E[w_t(\theta_0)w_{t-j}(\theta_0)] &= 0 \\
E[|w_t(\theta_0)w_{t-j}(\theta_0)|] - \ell_{1,j}(\theta_0) + \ell_{2,j}(\theta_0) &= 0 \\
E[|w_t(\theta_0)|w_{t-j}(\theta_0)] - m_j(\theta_0) &= 0 \\
E[w_t^2(\theta_0)w_{t-j}^2(\theta_0)] - n_j(\theta_0) &= 0
\end{aligned} \tag{1.48}$$

for $j = 1, 2, \ldots$ where

$$w_t(\theta_0) = \frac{y(\tau_t) - \alpha_0 d_t - (1 + \beta_0 d_t) y(\tau_{t-1})}{[d_t \{y(\tau_{t-1})\}^{\gamma_0}]^{1/2}} \tag{1.49}$$

and

$$\begin{aligned}
\ell_{1,j}(\theta_0) &= (2/\pi)^{1/2} exp[2\mu_x + \sigma_x^2(1 + (1+\eta_0 d)^j) - 0.5\rho_0^2\zeta_0^2 d(1+\eta_0 d)^{2(j-1)}] \\
\ell_{2,j}(\theta_0) &= (2/\pi)^{1/2} \rho_0 \zeta_0 d^{1/2} (1+\eta_0 d)^{j-1} (1 - 2\Phi(\rho_0\zeta_0 d^{1/2}(1+\eta_0 d)^{j-1}) \\
&\qquad \times exp[2\mu_x + \sigma_x^2(1 + (1+\eta_0 d)^j)] \\
m_j(\theta_0) &= (2/\pi)^{1/2} \rho_0 \zeta_0 d^{1/2} (1+\eta_0 d)^{j-1} exp[2\mu_x + \sigma_x^2(1 + (1+\eta_0 d)^j)] \\
n_j(\theta_0) &= \{4\rho_0^2\zeta_0^2 d(1+\eta_0 d)^{2(j-1)} + 1\} exp[4\mu_x + 4\sigma_x^2(1 + (1+\eta_0 d)^j)] \\
\mu_x &= -\delta_0/\eta_0 \\
\sigma_x^2 &= \zeta_0^2 d/[1 - (1+\eta_0 d)^2]
\end{aligned}$$

and $\Phi(.)$ denotes the cumulative distribution function of a standard normal random variable.

[30] In their estimations, Melino and Turnbull (1990) fix the value of γ_0 and so we omit this term from θ_0. See Section 9.4 for further discussion of this issue.

[31] These expressions are not actually presented in the published version of Melino and Turnbull's paper but are contained in an unpublished appendix by Ken Vetzal which was kindly sent to the author by Angelo Melino.

1.4 Review of Statistical Theory

To develop the theory of GMM estimators it is necessary to appeal to various statistical concepts and results. This section briefly reviews some basic ideas which are used throughout the text; other results are explained as they become needed. A more thorough review of these topics can be found in many econometric or statistical texts such as Davidson and MacKinnon (1993), Fuller (1976), Judge, Griffiths, Hill, Lutkepohl, and Lee (1985), and, for more rigorous treatments, Davidson (1994) and White (1984). All the results are based on asymptotic, or in other words, large sample theory. In the majority of our analysis, this type of analysis involves an examination of what happens to various statistics as the sample size, T, tends to infinity. Asymptotic is the adjective derived from "asymptote", the noun for the line which acts as a limit for a curve. According to the *American Heritage Dictionary*, asymptote comes from the Greek "asumptotos" in which "a" means not, "sun" means together and, "ptotos" means likely to fall. In spite of these unpromising origins, asymptotic analysis is used to approximate the behaviour of statistics in large, but finite, samples. An important secondary issue is the accuracy of this approximation and this is discussed in detail in Chapter 6.

Before reviewing this theory, it is useful to emphasize an item of notation. In the preceeding sections, it has been shown that statistical or economic models imply a set of population moment conditions involving the parameters and the data. It is important to realize that these moment conditions only hold at the *true* value of the parameters. A zero subscript is used to emphasize the true value of the parameter vector. This notation is neccessary to avoid ambiguity in the formal discussion of statistical estimation. As we have seen in Section 1.2, GMM estimation involves finding the value of the parameters which minimize $Q_T(\theta)$ given in (1.18). Formally, this will involve considering the behaviour of $Q_T(\theta)$ over a set of possible values for θ, known as the *parameter space* and denoted Θ. The notation θ is reserved to refer to an arbitrary element of Θ. As above, the notation $\hat{\theta}_T$ is used to denote the parameter estimator based on a sample of size T. Both θ_0 and $\hat{\theta}_T$ are individual elements of Θ.

The IV estimator in (1.14), $\hat{\alpha}_T$, can be used to illustrate several key features of asymptotic analysis of GMM estimators. It is of interest to analyze what happens to $\hat{\alpha}_T$ as $T \to \infty$ and for this we require the concept of *convergence in probability*. This analysis is facilitated by analyzing the limiting behaviour of the sums in the numerator and denominator separately using the *Weak Law of Large Numbers* and then taking the ratio of these limits to deduce the limiting behaviour of $\hat{\alpha}_T$. This last step can be justified using *Slutsky's Theorem*. In particular, it is of interest to examine whether the estimator converges in probability to the true population value of that coefficient; if so, then it is said to be *consistent*. For the purposes of constructing confidence intervals and hypothesis tests about α_0, it is necessary to find some transformation of $\hat{\alpha}_T$ which *converges in distribution* to a known probability distribution. For our purposes the appropriate transformation is $T^{1/2}(\hat{\alpha}_T - \alpha_0)$ and this statistic can be shown to converge to a normal distribution as $T \to \infty$ using the *Central Limit Theorem*.

In the remainder of this section these and certain other statistical concepts are defined more formally. It is most convenient to split the discussion into two parts. The first part deals with the properties of random sequences such as convergence in probability or distribution which can be discussed in abstract. The second part deals with results such as the Weak Law of Large Numbers and Central Limit Theorem for which it is neccessary to place restrictions on the nature of the random variables in the model.

1.4.1 Properties of Random Sequences

To fix ideas, consider the case where the sequence is deterministic and so not random. Let $\{h_T; T = 1, 2, \ldots\}$ be a sequence of real numbers. If this sequence has a limit, h, then this is denoted by

$$\lim_{T\to\infty} h_T = h$$

This implies that for every $\epsilon > 0$ there is a positive, finite integer T_ϵ such that

$$|h_T - h| < \epsilon \quad \text{for} \quad T > T_\epsilon \tag{1.50}$$

Note (1.50) does not imply $|h_T - h|$ becomes monotonically smaller as T increases. However, it does tell us that $|h_T - h|$ is smaller than ϵ for all $T > T_\epsilon$, and so conveys a sense in which h_T is becoming closer to h as T tends to infinity. Often, it is useful to characterize the behaviour of a sequence with respect to T regardless of whether it converges or not. This can be achieved using large and small orders of magnitude. The sequence is said to be of *large order of magnitude* c_T if there exists a real number m such that $|h_T|/c_T < m$ for all T. This is denoted by $h_T = O(c_T)$. The sequence is said to be of *small order of magnitude* c_T if the limit of h_T/c_T is zero as $T \to \infty$. This is denoted by $h_T = o(c_T)$.

In these definitions, the deterministic nature of the sequence is reflected in the way it can be stated with certainty that h_T satisfies the property in question. With sequences of random variables it is necessary to attach a probability to such events occuring. This leads us to the concept of convergence in probability. For notational convenience the results are also stated in terms of "h_T" but this is now a random variable.

Definition 1.3 Convergence in Probability
The sequence of random variables $\{h_T\}$ *converges in probability to the random variable* h *if for all* $\epsilon > 0$

$$\lim_{T\to\infty} P[|h_T - h| < \epsilon] = 1$$

In this case h *is known as the probability limit or plim of* h_T *and is denoted by* $plim\, h_T = h$ *or* $h_T \xrightarrow{p} h$.

The definition of convergence in probability implies that for each $\epsilon > 0$ there exists a finite T_ϵ such that the probability of $|h_T - h| < \epsilon$ is arbitrarily close to one for all $T > T_\epsilon$. So convergence in probability can be recognized as the natural extension of the concept of of convergence for deterministic sequences. The concepts of order of magnitude can be similarly extended to sequences of random variables.

Definition 1.4 Orders in Probability

1. *The sequence of random variables* $\{h_T\}$ *is said to be of large order in probability* c_T *if for every* $\epsilon > 0$ *there exists positive real numbers* m_ϵ *and* T_ϵ *such that* $P[|h_T|/c_T > m_\epsilon] \leq \epsilon$ *for all* $T \geq T_\epsilon$. *This is denoted by* $h_T = O_p(c_T)$.

2. *The sequence of random variables* $\{h_T\}$ *is said to be of small order in probability* c_T *if* $plim(h_T/c_T) = 0$. *This is denoted by* $h_T = o_p(c_T)$.

Both types of order in probability are very useful in asymptotic analysis because they can be linked to consistency and convergence in distribution as will be shown below. However, first it is necessary to extend the notion of convergence in probability to vectors and matrices. A vector (or matrix), h_T, is said to converge in probability to h if the i^{th} (or $(i,j)^{th}$) element of h_T converges in probability to the i^{th} (or $(i,j)^{th}$) element of h for all i (or (i,j)). The extension of orders in probability is a little more tricky because in general there is no guarantee that all elements of a random vector or matrix are of the same order. However, in the majority of our analysis this will be the case and so we use the notation $h_T = O_p(c_T)$ or $h_T = o_p(c_T)$ to indicate that all the elements of the vector or matrix individually satisfy the stated order in probability.

In many cases, our analysis involves the probability limits of functions of random vectors and so the following result is going to be very useful. For convenience the result is stated in terms of random vectors; however, the same result applies for random variables and random matrices.

Lemma 1.1 Slutsky's Theorem[32]
Let $\{h_T\}$ *be a sequence of random vectors which converges in probability to the random vector* h *and let* $f(.)$ *be a vector of continuous functions then* $plim f(h_T) = f(h)$.

In many cases $h_T = \hat{\theta}_T$, a GMM estimator of some unknown parameter vector θ_0, and so it is of interest to characterize the limiting relationship between estimator and estimand.

Definition 1.5 Consistency of an Estimator
Let $\{\hat{\theta}_T\}$ *be a sequence of estimators of the unknown parameter vector of constants* θ_0 *then* $\hat{\theta}_T$ *is said to be a consistent estimator of* θ_0 *if* $plim\, \hat{\theta}_T = \theta_0$.

[32] This theorem is named after Evgenii Slutsky (1880–1948), a Russian mathematician who first proved a version of this result. He made numerous other contributions to statistics including early work which helped to lay the foundations of stationary time series theory. He also made contributions to economics particularly in the area of demand analysis including the eponymous Slutsky effect and Slutsky matrix.

If $plim\,\hat{\theta}_T \neq \theta_0$ then the estimator is said to be *inconsistent*. Notice that the consistency of $\hat{\theta}_T$ for θ_0 implies $\hat{\theta}_T - \theta_0 = o_p(1)$. Consistency is a rather weak property because it merely states that as $T \to \infty$ the estimator converges in probability to the true value. It is perfectly reasonable to question how much comfort can be drawn from this property since it implies the true value is only recovered in the limit. However, earlier it was observed that convergence also implies a sense in which $\hat{\theta}_T$ becomes closer to θ_0 as T increases. This is a more intuitively appealing property; certainly we would be concerned if the estimator is inconsistent and so not converging in probability to the true value!

Convergence in probability implies that the difference between $\hat{\theta}_T$ and θ_0 disappears with probability one as $T \to \infty$. Therefore in the limit $\hat{\theta}_T$ and θ_0 are essentially identical. In deriving the asymptotic distribution of the GMM estimator, it will be convenient to appeal to the weaker notion of convergence in distribution. For this definition we revert to the more general notation because this concept is not just applied to estimators in our analysis.

Definition 1.6 Convergence in Distribution
The sequence of random vectors $\{h_T\}$ *with corresponding distribution functions* $\{F_T(c)\}$ *converges in distribution to the random vector* h *with distribution function* $F(c)$ *if and only if there exists* T_ϵ *for every* ϵ *such that* $|F_T(c) - F(c)| < \epsilon$ *for* $T > T_\epsilon$ *at all points of continuity* $\{c\}$. *This is denoted by* $h_T \xrightarrow{d} h$

The distribution of h is known as the *limiting (or asymptotic) distribution* of h_T. If h_T converges in distribution then $h_T = O_p(1)$. However, in practice, our focus is not just on establishing that h_T converges in distribution, but also on characterizing the exact nature of its limiting distribution. We now turn to various results which facilitate this type of analysis as well as the other aspects of asymptotic behaviour described above.

1.4.2 Stationary Time Series, the Weak Law of Large Numbers and the Central Limit Theorem

The asymptotic theory in this book revolves around analyses of the limiting behaviour of sums of random variables using the Weak Law of Large Numbers and Central Limit Theorem. For these results to apply, it is necessary to place restrictions on the nature of the random variables in the model. Various approaches can be taken but, throughout this book, we follow Hansen's (1982) original treatment involving stationary time series. In passing we note that this assumption is employed in nearly all the studies listed in Table 1.1.[33]

Definition 1.7 Strictly Stationary Processes
Let $\mathcal{N}(T) = \{1, 2, \ldots T\}$ *and* $\{v_t; t \in \mathcal{N}(T)\}$ *be a set of random vectors. Define* $\{t_1, t_2, \ldots, t_n\}$ *to be a subset of* $\mathcal{N}(T)$. *The set of random vectors are said to*

[33] See Appendix A for a brief discussion of the GMM framework under alternative assumptions about the data generation process.

be strictly stationary if the joint probability distribution function, F(.), of any subset of $\{v_t\}$ *satisfies:*

$$F(v_{t_1}, v_{t_2}, \ldots, v_{t_n}) = F(v_{t_1+c}, v_{t_2+c}, \ldots v_{t_n+c})$$

for any integer n and integer constant c such that $\{t_1 + c, t_2 + c, \ldots t_n + c\}$ *is a subset of* $\mathcal{N}(T)$.

One consequence of this definition is that all moments of the process are constant over time, provided they exist. The imposition of strict stationarity is insufficient by itself to permit the proof of Weak Law of Large Numbers and the Central Limit Theorem. In addition restrictions need to be placed on the dependence structure and certain higher moments of the series. Examples of such conditions on the dependency are *ergodicity* or various types of *mixing condition*. Both involve rather sophisticated mathematical ideas and so for the present, we just add the caveat "subject to certain regularity conditions" in the statement of the following results. However, we return to these conditions in Chapter 3.

Lemma 1.2 Weak Law of Large Numbers (WLLN)
Let $\{v_t; t = 1, 2, \ldots, T\}$ *be a sequence of strictly stationary random vectors with* $E[v_t] = \mu$ *then subject to certain regularity conditions*

$$T^{-1} \sum_{t=1}^{T} v_t \xrightarrow{p} \mu$$

Lemma 1.3 Central Limit Theorem (CLT)
Let $\{v_t; t = 1, 2, \ldots, T\}$ *be a sequence of strictly stationary* $(s \times 1)$ *random vectors with* $E[v_t] = \mu$ *then subject to certain regularity conditions*

$$T^{-1/2} \sum_{t=1}^{T} (v_t - \mu) \xrightarrow{d} N(0, \Sigma)$$

where $N(0, \Sigma)$ *denotes the s dimensional multivariate normal distribution with mean 0 and positive definite covariance matrix*

$$\Sigma = \lim_{T \to \infty} Var[T^{-1/2} \sum_{t=1}^{T} (v_t - \mu)]$$

The matrix Σ is known as the long run covariance matrix of v_t to distinguish it from the the contemporaneous covariance matrix $E[(v_t - \mu)(v_t - \mu)']$.

To conclude this section, it is useful to present one final result which is invoked frequently.[34]

[34] This result is proved in Fuller (1976, p.199).

Lemma 1.4 The Limiting Distribution of Random Linear Functions of Vectors Converging to a Normal Distribution
Let $\{M_T; t = 1, 2 \ldots, T\}$ *be a sequence of random matrices which converges in probability to* M*, a matrix of constants, and* $\{h_T, ; t = 1, 2 \ldots T\}$ *be a sequence of random vectors which converges to a* $N(0, \Sigma)$ *distribution then*

$$M_T h_T \xrightarrow{d} N(0, M\Sigma M')$$

1.5 Overview of Later Chapters

This chapter has provided the flavour of GMM and placed the technique in the context of both the econometrics and statistics literatures. In the next chapter, we introduce the key elements of the GMM framework using the IV estimator in the static linear model. This approach keeps the technical details to a minimum and allows the reader to appreciate more readily the main ideas and intuitions. The issues addressed here are: identification; the asymptotic properties of the estimator; the iterated GMM estimator; and a decomposition of the population moment conditions into identifying and over-identifying restrictions which leads to the overidentifying restrictions test amongst other things. The following chapters build from these foundations to present the GMM framework for estimation and inference which encompasses the majority of the models in Table 1.1.

Chapter 3 addresses GMM estimation and the asymptotic properties of the estimator in correctly specified nonlinear dynamic models. The topics covered are: identification, calculation of the estimator by numerical optimization routines, consistency, asymptotic normality, covariance matrix estimation and iterated GMM estimators. Formal proofs are presented for the main statistical results. However, the issues are also illustrated using the consumption based asset pricing model to provide guidance on the practical implementation of GMM as well. All this discussion takes the data, parameter vector and population moment condition as given. In some cases, the researcher may desire to impose a normalization on any one of these three features. Therefore, the impact of normalization is also discussed and this motivates a variant of GMM known as the continuous updating estimator. This chapter concludes with a more formal presentation of how many seemingly different estimators can be regarded as special cases of GMM.

Chapter 4 explores the consequence of misspecification for the statistical properties of the GMM estimator. Particular attention is focused on convergence in probability of the estimator, covariance matrix estimation and the the limiting distribution of the estimator. A comparison with the results in the previous chapter reveals that misspecification has a fundamental impact on the large sample behaviour of the GMM estimator and its associated statistics. These differences motivate the use of the model specification tests.

Chapter 5 examines a wide variety of hypothesis tests which have been proposed within the GMM framework. The main focus is on the following: the

overidentifying restrictions test, tests for the validity of a subset of population moment conditions, tests of whether the parameter vector satisfies a set of restrictions, and structural stability tests. However there is also some discussion of Hausman-type tests, non-nested hypothesis tests and conditional moment tests.

All the preceding analysis is based on asymptotic theory. Chapter 6 explores how well this theory approximates finite sample behaviour. If attention is reduced to a very specific class of models then it is possible to examine this question analytically. However, for more general specifications, it is necessary to resort to computer based simulation studies. Both approaches are reviewed in Chapter 6, and the results from each are synthesized to indicate what aspects of the specification appear to effect the quality of the asymptotic approximation to finite sample behaviour. This chapter begins with a discussion of the available asymptotic results on the consequences of increasing the number of the moment conditions upon which the estimation is based.

The asymptotic theory in Chapters 3 and 5 takes the population moment condition as given. However, the evidence reviewed in Chapter 6 indicates that the quality of the asymptotic approximation can be sensitive to the choice of moment condition. Chapter 7 reviews the literature on moment selection. The discussion falls into two parts. The first part summarizes available results on the optimal choice of instrument in the special case of GMM known as generalized instrumental variables (GIV). The second part describes a number of information criteria that have been proposed as a basis for moment selection.

In the face of evidence that the asymptotic theory from Chapters 3 and 5 can provide a poor approximation, it is natural to seek alternative approximations that permit more reliable inference. Three such approximations are reviewed in Chapter 8. These are: the use of the bootstrap, an asymptotic theory derived under the assumption that the population moment condition provides weak identification, and an asymptotic theory for the case in which the long run variance is estimated by a class of estimators that are random in the limit.

All the methods and issues described above are illustrated empirically using the consumption based asset pricing model in Section 1.3.1. Chapter 9 presents empirical results for the other four examples in Section 1.3 that illustrate various aspects of the GMM inference framework.

Finally, Chapter 10 briefly reviews some other estimation techniques that are closely related to GMM. These are Simulated Method of Moments, Indirect Inference, Efficient Method of Moments and the method of Empirical Likelihood.

2

The Instrumental Variable Estimator in the Linear Regression Model

One of the main advantages of GMM is that it can be used to perform inference about the parameters in nonlinear dynamic models. However, as might be anticipated, both nonlinearity and dynamics create a number of technical issues which need to be addressed in the statistical analysis. These issues can obscure the essential structure of the method for those readers less familiar with this type of analysis. Therefore, in this chapter, we introduce the key elements of the GMM framework using the IV estimator in the static linear model. This approach enables us to keep the technical details to a minimum and allows the reader to appreciate more readily the main ideas and intuitions. Those readers already familiar with the basic GMM framework may prefer to pass over this chapter.

Section 2.1 specifies the model and discusses the connections between the population moment condition and the condition for parameter identification. Section 2.2 derives the estimator and describes a fundamental decomposition of the population moment condition into "identifying" and "overidentifying" restrictions. Section 2.3 considers the asymptotic properties of the estimator and the estimated sample moment. In the course of this discussion, it emerges that a consistent estimator of the long run variance of the sample moment is required for inference procedures based on the parameters or estimated moments. Therefore, Section 2.3 also contains a brief discussion of how such a covariance matrix estimator can be constructed in this simple model. Section 2.4 examines the optimal way in which to use the information in the population moment conditions, and introduces the "two step" and iterated GMM estimators. Section 2.5 discusses the consequences of specification error, and introduces the overidentifying restrictions test statistic which is the standard diagnostic for the model

specification within the GMM framework. Section 2.6 contains a summary of the chapter.

2.1 The Population Moment Condition and Parameter Identification

Consider the linear regression model

$$y_t = x_t'\theta_0 + u_t, \qquad t = 1, 2, \ldots T \tag{2.1}$$

in which x_t is a $(p \times 1)$ vector of observed explanatory variables for the observed variable y_t, and u_t is the unobserved error term. The $(p \times 1)$ vector θ_0 is an element of the parameter space, Θ, a subset of the p-dimensional Euclidean space $\Re^p$. The instruments are contained in the $(q \times 1)$ vector z_t. To facilitate the discussion, it is useful to define: $u_t(\theta) = y_t - x_t'\theta$. Notice that $u_t(\theta_0) = u_t$. As the analysis progresses, certain restrictions need to be placed on the variables but these will only be imposed as they become necessary to emphasize their role. At this stage, we only require the following.

Assumption 2.1 Strict Stationarity
The random vector $v_t = (x_t', z_t', u_t)'$ is a strictly stationary process.

This assumption implies that any population moments of v_t are independent of t.

Estimation of θ_0 is based on the following population moment condition.

Assumption 2.2 Population Moment Condition
The $(q \times 1)$ vector z_t satisfies: $E[z_t u_t(\theta_0)] = 0$.

This type of condition is sometimes refered to as an "orthogonality condition" because it states that z_t is statistically orthogonal to u_t. At this stage, it may be useful to relate this structure back to one of the models encountered in Chapter 1.

Example: Wright's (1925) Demand Equation
It can be recalled from Section 1.2 that Wright (1925) proposed IV as a method for estimating the parameters of demand and supply equations. His original derivation was based on the Method of Moments principle and so its implementation only required the researcher to find one instrument z_t^D which satisfied the moment condition in (1.13). Two candidates were suggested: an input price, now denoted z_{1t}^D, and yield per acre, z_{2t}^D. However, rather than choose between these two instruments arbitrarily, intuition suggests that a far more appealing strategy is to base estimation on both. This leads to the (2×1) population moment condition

$$E[z_t(q_t - \alpha_0 p_t)] = 0$$

where $z_t = (z_{1t}^D, z_{2t}^D)'$. It can be recognized that this population moment condition fits within the framework of Assumption 2.2 once q_t, p_t and α_0 are substituted for y_t, x_t and θ_0 respectively in (2.1). $\diamond$

While Assumption 2.2 specifies the information upon which estimation is based, the resulting estimation is only going to be successful if this population moment condition provides enough information to determine θ_0 uniquely. In reality, this is not guaranteed to be the case. The parameter vector θ_0 is only uniquely determined by the moment condition if $E[z_t u_t(\theta)] \neq 0$ at all other values of θ. In this case θ_0 is said to be *identified* by the population moment condition. This condition is easily stated but, in this form, provides little guidance about the circumstances under which it holds. Fortunately, it is possible to obtain a more transparent version. With some simple rearrangement, it follows that

$$E[z_t u_t(\theta)] = E[z_t u_t(\theta_0)] + E[z_t x_t'](\theta_0 - \theta) \tag{2.2}$$

and this combined with the population moment condition implies

$$E[z_t u_t(\theta)] = E[z_t x_t'](\theta_0 - \theta) \tag{2.3}$$

Therefore θ_0 is identified if $E[z_t x_t'](\theta_0 - \theta) \neq 0$ for all $\theta \neq \theta_0$. Equation (2.3) is a system of linear equations in $\theta_0 - \theta$ and so this property is guaranteed if the rank of $E[z_t x_t']$ is p; for example see Strang (1988, p.96). This gives the following condition for identification.

Assumption 2.3 Identification Condition
$rank\{E[z_t x_t']\} = p$.

The population moment and identification conditions provide the essential information upon which estimation of θ_0 is based. In view of its fundamental importance, it is worth briefly pausing to reflect on the exact nature of this information. Assumptions 2.2 and 2.3 imply there is a unique value in the parameter space at which $E[z_t u_t(\theta)]$ equals zero. In our discussion we have denoted this value by θ_0 – however, nothing has been said about this value beyond its uniqueness.

Before proceeding to define the GMM estimator, it is worth briefly considering how parameter identification can fail. There are two basic scenarios. First, failure can occur because there are fewer moment conditions than parameters. In terms of the mathematics, this implies that $rank(E[z_t x_t']) \leq q < p$. Intuitively, the problem here is that it is impossible to extract the p pieces of information needed to determine θ_0 uniquely from less than p population moment conditions. Secondly, failure can occur even when $q \geq p$ because collectively the population moment conditions still do not provide enough information to uniquely determine θ_0. This second scenario is best understood by considering a simple example. Suppose $p = q = 2$; let $x_t = (x_{1,t}, x_{2,t})'$, $z_t = (z_{1,t}, z_{2,t})'$ and $\theta_i, \theta_{0,i}$ denote the i^{th} elements of θ, θ_0 respectively. In this case,

$$E[z_t u_t(\theta)] = \begin{bmatrix} E[z_{1,t}x_{1,t}] & E[z_{1,t}x_{2,t}] \\ E[z_{2,t}x_{1,t}] & E[z_{2,t}x_{2,t}] \end{bmatrix} \begin{bmatrix} \theta_{0,1} - \theta_1 \\ \theta_{0,2} - \theta_2 \end{bmatrix} \tag{2.4}$$

For this model, identification requires the rank of $E[z_t x_t']$ to be two. Failure can occur because either $E[z_t x_t']$ contains a row of zeros or because the first row is a multiple of the second. Each of these can be interpreted in terms of the statistical model as follows.

- *Case 1: $E[z_t x_t']$ contains a row of zeros*
  Suppose $E[z_{1,t} x_t'] = (0, 0)$ and $E[z_{2,t} x_t'] = (m_1, m_2)$. In this case $E[z_{1,t} u_t(\theta)] = 0$ regardless of the value of $\theta_0 - \theta$, and so it provides no information on θ_0. The other moment condition provides some information but not enough to uniquely determine θ_0. For example if $m_i \neq 0$ for $i = 1, 2$ then $E[z_{2,t} u_t(\theta)] = 0$ for any $\theta_0 - \theta$ of the form $(c, -m_1 c/m_2)$. Identification fails because an insufficient number of elements of $E[z_t u_t(\theta_0)] = 0$ provide information about θ_0.

- *Case 2: One row of $E[z_t x_t']$ is a multiple of the other*
  Suppose $E[z_{1,t} x_t'] = kE[z_{2,t} x_t'] = (m_1, m_2)$ for some constant k and for the sake of argument $m_i \neq 0$ for $i = 1, 2$. In this case $E[z_t u_t(\theta)] = (0, 0)'$ for any $\theta_0 - \theta$ of the form $(c, -m_1 c/m_2)$ and, once again, θ_0 is not uniquely determined by the population moment condition. So identification fails because both elements of $E[z_t u_t(\theta_0)] = 0$ provide exactly the same information about θ_0.

From this discussion, it is clear that parameter identification and the relationship between p and q are important. It is therefore useful to introduce the following terminology. If the identification condition fails then the parameter vector θ_0 is said to be *under-identified (or unidentified)* by the population moment condition. If the parameters are identified and $q = p$ then the parameters are said to be *just-identified* by the population moment condition. Notice in this case there are just p sources of the p pieces of information needed to identify θ_0. Finally, if the parameters are identified and $q > p$ then θ_0 is said to be *over-identified* by the population moment condition. In this case there are more than p sources of the p pieces of information needed to identify θ_0.

For the remainder of this chapter, it is assumed that the parameters are either just- or over-identified. In Section 8.2, we examine the kind of problems which can occur if the parameters are under-identified or close to being so, a scenario termed "weak identification".

2.2 The Estimator and a Fundamental Decomposition

Section 1.2 introduced the generic definition of the GMM estimator. To specialize this definition to our current context, it is most convenient to work with matrix notation rather than summations. Therefore we start by introducing the following definitions. Let y be the $(T \times 1)$ vector whose t^{th} element is y_t; X be the $(T \times p)$ matrix whose t^{th} row is x_t'; Z be the $(T \times q)$ matrix whose t^{th} row is z_t'; u be the $(T \times 1)$ vector whose t^{th} element is u_t; and $u(\theta) = y - X\theta$.

Using this notation to make the appropriate substitutions into (1.18), the GMM minimand for this model is:

$$Q_T(\theta) = \{T^{-1}u(\theta)'Z\}W_T\{T^{-1}Z'u(\theta)\} \tag{2.5}$$

Following Definition 1.2, the GMM estimator of θ_0 is defined as

$$\hat{\theta}_T = \text{argmin}_{\theta\in\Theta}\, Q_T(\theta) \tag{2.6}$$

where the notation "argmin" is a mathematical shorthand for the value of the *arg*ument – θ – which *min*imizes the function – $Q_T(\theta)$. Since,

$$Q_T(\theta) = T^{-2}\{y'ZW_TZ'y + \theta'X'ZW_TZ'X\theta - 2y'ZW_TZ'X\theta\}$$

the first order conditions for the minimization in (2.6) are[1]

$$(T^{-1}X'Z)W_T(T^{-1}Z'y) = (T^{-1}X'Z)W_T(T^{-1}Z'X)\hat{\theta}_T \tag{2.7}$$

So provided $(T^{-1}X'Z)W_T(T^{-1}Z'X)$ is nonsingular, the estimator is given by

$$\hat{\theta}_T = \{(T^{-1}X'Z)W_T(T^{-1}Z'X)\}^{-1}(T^{-1}X'Z)W_T(T^{-1}Z'y) \tag{2.8}$$

It can be recalled from Section 1.2 that GMM (or Minimum Chi-Square) were introduced to circumvent the problems encountered with Method of Moments. That earlier discussion emphasized the way in which GMM generalized the Method of Moments principle. However, the relationship between the two estimation principles is far more subtle. Although the GMM estimator is defined via the minimization in (2.6), it is actually the solution to the first order conditions in (2.7). With a simple rearrangement, these conditions can be rewritten as

$$(T^{-1}X'Z)W_TT^{-1}Z'u(\hat{\theta}_T) = 0 \tag{2.9}$$

This characterization of the first order conditions reveals that $\hat{\theta}_T$ is identical to the Method of Moments estimator based on,

$$E[x_tz_t']WE[z_tu_t(\theta_0)] = 0 \tag{2.10}$$

This Method of Moments interpretation is useful because it makes explicit the relationship between the estimator and the population moment condition in Assumption 2.2. Minimization of $Q_T(\theta)$ with respect to θ amounts to estimation based on the information that the p linear combinations of $E[z_tu_t(\theta_0)]$ given in (2.10) are zero. Notice that this interpretation implies that if $q = p$ then Method of Moments and GMM are equivalent because in this case $E[x_tz_t']W$ is nonsingular and so (2.10) implies $E[z_tu_t(\theta_0)] = 0$.[2] In this case, the weighting matrix plays no role and the GMM estimator is given by,[3]

$$\hat{\theta} = (T^{-1}Z'X)^{-1}(T^{-1}Z'y) \tag{2.11}$$

[1] See Dhrymes (1984)[Proposition 95 and Corollary 28, p.110–111].

[2] Recall that a similar observation is made in Section 1.2 regarding the equivalence of Method of Moments and Minimum Chi-Square.

[3] Notice that this solution is consistent with (2.8) because if $p = q$ then $\{(T^{-1}X'Z)W_T(T^{-1}Z'X)\}^{-1} = (T^{-1}Z'X)^{-1}W_T^{-1}(T^{-1}X'Z)^{-1}$ - subject to the existence of the stated inverses.

However if $q > p$ then no such reduction is possible, and the choice of weighting matrix is important because it determines the exact nature of the linear combinations of $E[z_t u_t(\theta_0)]$ set to zero in (2.10).

This Method of Moments interpretation also indicates that if $q > p$ then there is a difference between the information with which we began, Assumption 2.2, and the information actually used in estimation, equation (2.10). To characterize the relationship between the two, it is useful to develop an alternative representation for (2.10) which has the same dimension as the population moment condition. For this part of the analysis, it is more convenient to work with a nonsingular tranformation of the population moment condition, $W^{1/2}E[z_t u_t(\theta_0)]$, where $W^{1/2}$ satisfies $W = W^{1/2\prime}W^{1/2}$.[4] So we begin by rewriting (2.10) as

$$F'W^{1/2}E[z_t u_t(\theta_0)] \;=\; 0 \tag{2.12}$$

where $F' = E[x_t z_t']W^{1/2\prime}$. Equation (2.12) indicates that GMM estimation is based on the information that $W^{1/2}E[z_t u_t(\theta_0)]$ lies in the null space of the $(p \times q)$ matrix F'. Sowell (1996) observes that this condition is identical to the restriction that the least squares projection of $W^{1/2}E[z_t u_t]$ onto the column space of F is zero. By this logic, we obtain the following alternative representation of the information used in GMM estimation,

$$F(F'F)^{-1}F'W^{1/2}E[z_t u_t(\theta_0)] \;=\; 0 \tag{2.13}$$

While (2.13) consists of q equations in $E[z_t u_t(\theta_0)]$, not all of them are linearly independent because $rank\{F(F'F)^{-1}F'\} = rank\{F\} \leq p$. Notice that we have already assumed this rank equals p to ensure identification. The re-emergence of this quantity here provides an alternative perspective on the fundamental connection between identification and estimation: the p parameters are only identified if the estimation is based on p linearly independent equations. In view of this connection, Sowell (1996) refers to the elements of (2.13) as the *identifying restrictions* associated with GMM estimation. It follows immediately from (2.13) that the part of $W^{1/2}E[z_t u_t(\theta_0)]$ unused in estimation is given by

$$(I_q - F(F'F)^{-1}F')W^{1/2}E[z_t u_t(\theta_0)] = 0 \tag{2.14}$$

Equation (2.14) constitute a set of $rank\{I_q - F(F'F)^{-1}F'\} = q - p$ linearly independent equations in $W^{1/2}E[z_t u_t(\theta_0)]$. Hansen (1982) refered to the elements of (2.14) as the *overidentifying restrictions*.

This decomposition is fundamental to the analysis of GMM estimators of *overidentified* parameter vectors and so it is worth emphasizing its structure. The $(q \times 1)$ vector of population moment conditions is decomposed into p identifying restrictions and $q - p$ overidentifying restrictions. The identifying restrictions represent the part of the population moment condition used in estimation and the overidentifying restrictions are the remainder. Most importantly, these two components are linearly unrelated because $F(F'F)^{-1}F'\{I_q - F(F'F)^{-1}F'\} = 0$.

[4] There must be a $(q \times q)$ nonsingular matrix $W^{1/2}$ which satisfies this identity because W is positive definite from Definition 1.2; see Dhrymes (1984) [Corollary 14, p.73].

So far, these components have been defined in terms of population quantities. We now consider the extent to which this behaviour is mirrored by their sample counterparts. Since the identifying restrictions represent the information upon which estimation is based, it would be anticipated that their sample analog holds at $\hat{\theta}_T$. This is easily verified to be the case because the first order conditions in (2.9) imply

$$F_T(F_T'F_T)^{-1}F_T'W_T^{1/2}T^{-1}Z'u(\hat{\theta}_T) = 0 \tag{2.15}$$

where $F_T' = (T^{-1}X'Z)\,W_T^{1/2'}$ and $W_T = W_T^{1/2'}W_T^{1/2}$. In contrast, the overidentifying restrictions are ignored in estimation and so it would be anticipated that they do not generally hold in the sample. Again, this is the case. However, they do play a similar remainder role in the sample. From (2.15) it follows that

$$(I_q - F_T(F_T'F_T)^{-1}F_T')W_T^{1/2}\;T^{-1}Z'u(\hat{\theta}_T) = W_T^{1/2}T^{-1}Z'u(\hat{\theta}_T) \tag{2.16}$$

and so the estimated transformed sample moment is just the sample analog to the function of the data in the overidentifying restrictions. This leads to a useful interpretation of the GMM minimand. In Section 1.2, $Q_T(\theta)$ was introduced as a measure of how far the sample moment is from its expectation of zero. The substitution of (2.16) into (2.5) indicates that the minimized value, $Q_T(\hat{\theta}_T)$, measures how far the sample is from satisfying the overidentifying restrictions. This interpretation proves useful in the development of statistics for testing whether the model is correctly specified. However, before we can discuss such methods, it is necessary to consider the asymptotic properties of the parameter estimator and the estimated sample moment. So, we delay further discussion of methods for assessing the model specification until Section 2.5.

2.3 Asymptotic Properties

GMM estimation generates two important statistics which play a central role in inference about the underlying model; these are the parameter estimator and the estimated sample moment. Since the latter depends on the former, it makes most sense to begin our discussion of their asymptotic properties with the parameter estimator, and then to use these results to analyze the behaviour of the estimated sample moment. The asymptotic analysis of the parameter estimator focuses on the twin properties of consistency and asymptotic normality. The latter facilitates the construction of large sample confidence intervals for the elements of θ_0. As will emerge, these intervals involve a consistent estimator of the long run variance of the sample moment, and so we briefly consider how such an estimator can be calculated in our simple model. As mentioned in the previous section, the estimated sample moment plays an important role in the construction of hypothesis tests. In this capacity, it is the asymptotic normality of $T^{-1/2}Z'u(\hat{\theta}_T)$ which is important, and so it is this aspect of the statistic's behaviour upon which we concentrate.

The asymptotic analysis rests on applications of the Weak Law of Large Numbers (WLLN) and Central Limit Theorem (CLT) in Lemmas 1.2 and 1.3

respectively. It was noted in Section 1.4.2 that the assumption of strict stationarity is insufficient by itself for these theorems and so we must introduce an additional restriction. Our purpose here is to illustrate the basic ideas and so it is convenient to assume away any dependence structure in the data for the time being.

Assumption 2.4 Independence
The vector $v_t = (x_t', z_t', u_t)'$ *is independent of* v_{t+s} *for all* $s \neq 0$.

Together, assumptions 2.1 and 2.4 imply v_t is an independently and identically distributed process.

To begin with, it is most convenient to substitute for y in (2.8). Equation (2.1) implies $y = X\theta_0 + u$ and using this identity in (2.8) yields

$$\hat{\theta}_T = \theta_0 + \{(T^{-1}X'Z)W_T(T^{-1}Z'X)\}^{-1}(T^{-1}X'Z)W_T(T^{-1}Z'u) \tag{2.17}$$

The consistency and asymptotic normality of $\hat{\theta}_T$ can be deduced directly from (2.17). We start with consistency.

From (2.17), it follows that

$$plim\,\hat{\theta}_T = \theta_0 + plim\left[\{(T^{-1}X'Z)W_T(T^{-1}Z'X)\}^{-1}(T^{-1}X'Z)W_T(T^{-1}Z'u)\right] \tag{2.18}$$

Using Slutsky's Theorem (see Lemma 1.1), (2.18) can be rewritten as

$$\begin{aligned} plim\,\hat{\theta}_T = \theta_0 \quad + \quad & \{plim(T^{-1}X'Z)plim(W_T)plim(T^{-1}Z'X)\}^{-1} \\ & plim(T^{-1}X'Z)plim(W_T)plim(T^{-1}Z'u) \end{aligned} \tag{2.19}$$

From Definition 1.2, it follows immediately that $plim(W_T) = W$, a positive definite symmetric matrix. The limiting behaviour of the other matrices in (2.19) can be deduced from the WLLN. Since $z_t x_t'$ and $z_t u_t$ are contemporaneous functions of independent processes, they are themselves independent processes. Therefore the WLLN yields[5]

$$T^{-1}Z'X = T^{-1}\textstyle\sum_{t=1}^{T} z_t x_t' \xrightarrow{p} E[z_t x_t'] \tag{2.20}$$

$$T^{-1}Z'u = T^{-1}\textstyle\sum_{t=1}^{T} z_t u_t \xrightarrow{p} E[z_t u_t] \tag{2.21}$$

It is at this point that the population moment and identification conditions become important. The identification condition states that $E[z_t x_t']$ is of rank p and so the inverse of $E[x_t z_t']\,W\,E[z_t x_t']$ exists. The population moment condition states that $E[z_t u_t] = 0$. Using these two results in (2.19) yields

$$plim\,\hat{\theta}_T = \theta_0 + M\,E[z_t u_t] = \theta_0 \tag{2.22}$$

where $M = (F'F)^{-1}F'W^{1/2}$ and we have again put $F = W^{1/2}E[z_t x_t']$. Therefore, $\hat{\theta}_T$ is consistent for θ_0.

[5] Strictly, it must be assumed that all stated expectations exist. However, since the purpose of this chapter is purely expository, we suppress such details here.

The asymptotic distribution[6] of the estimator is derived by rewriting (2.17) as

$$T^{1/2}(\hat{\theta}_T - \theta_0) = \{(T^{-1}X'Z)W_T(T^{-1}Z'X)\}^{-1}(T^{-1}X'Z)W_T(T^{-1/2}Z'u) \quad (2.23)$$

and analyzing the behaviour of the components on the right hand side of (2.23). Since $z_t u_t$ is an independent process, the CLT can be invoked to deduce that

$$T^{-1/2}Z'u = T^{-1/2}\sum_{t=1}^{T} z_t u_t \xrightarrow{d} N(0, S) \quad (2.24)$$

where $S = \lim_{T\to\infty} Var[T^{-1/2}\sum_{t=1}^{T} z_t u_t]$ and the mean of this distribution follows from the population moment condition. Therefore, $T^{1/2}(\hat{\theta}_T - \theta_0) = M_T n_T$ where M_T converges in probability to the matrix of constants M and n_T converges in distribution to a normal random vector. Using Lemma 1.4, it follows that

$$T^{1/2}(\hat{\theta}_T - \theta_0) \xrightarrow{d} N(0, MSM') \quad (2.25)$$

where, as a reminder, $M = \{E[x_t z_t']\, W\, E[z_t x_t']\}^{-1}E[x_t z_t']\, W$. In the case where $p = q$ then M reduces to $E[z_t x_t']$ and so $MSM' = \{E[z_t x_t']\}^{-1}S\{E[x_t z_t']\}^{-1}$.

Equation (2.25) implies that an approximate large sample $100(1-\alpha)\%$ confidence interval for $\theta_{0,i}$ is

$$\hat{\theta}_{T,i} \pm z_{\alpha/2}\sqrt{\hat{V}_{T,ii}/T} \quad (2.26)$$

where $\hat{V}_{T,ii}$ is the (i,i) element of a consistent estimator of MSM' and $z_{\alpha/2}$ is the $100(1-\alpha/2)$ percentile of the standard normal distribution. A consistent estimator of MSM' can be obtained from consistent estimators of its components because by Slutsky's Theorem if $\hat{M}_T \xrightarrow{p} M$ and $\hat{S}_T \xrightarrow{p} S$ then $\hat{M}_T\hat{S}_T\hat{M}_T' \xrightarrow{p} MSM'$. The obvious choice of $\hat{M}_T$ is $\{(T^{-1}X'Z)W_T(T^{-1}Z'X)\}^{-1}(T^{-1}X'Z)W_T$ because it has already been shown this matrix converges in probability to M. To construct $\hat{S}_T$ it is necessary to be more specific about the form of the long run covariance matrix, S. Under our assumptions $z_t u_t$ is an independently and identically distributed process with a mean of zero. Together these restrictions imply

$$\begin{aligned} E[u_t u_s z_t z_s'] &= E[u^2 zz'],\ say \qquad for\ t = s \\ &= 0 \qquad for\ t \neq s \end{aligned}$$

and so

$$S = \lim_{T\to\infty} T^{-1}\sum_{t=1}^{T}\sum_{s=1}^{T} E[u_t u_s z_t z_s'] = E[u^2 zz'] \quad (2.27)$$

[6] There has been a vast literature on the finite sample properties of IV estimators in the linear model. Unfortunately, these results do not generalize to the nonlinear dynamic models which are the ultimate focus of this book. Therefore we concentrate on asymptotic results here. However, this finite sample theory is briefly reviewed in Section 6.2.

White (1984, Chapter 6) demonstrates that S can be consistently estimated by

$$\hat{S}_T \;=\; T^{-1}\sum_{t=1}^{T}\hat{u}_t^2 z_t z_t' \tag{2.28}$$

where $\hat{u}_t = y_t - x_t'\hat{\theta}_T$. In certain circumstances, more structure can be placed on $E[u^2 zz']$ which can be exploited in the construction of $\hat{S}_T$. For example, in most econometric textbooks IV is first encountered in the "classical" model in which u_t possesses the properties:

Assumption 2.5 Classical Assumptions about u_t
(i)$E[u_t] = 0$*; (ii)* $E[u_t^2] = \sigma_0^2$*; (iii)* u_t *and* z_t *are independent.*

Under these assumptions $E[u^2 zz'] = \sigma_0^2 E[z_t z_t']$ and this can be consistently estimated by

$$\hat{S}_{CIV} \;=\; \hat{\sigma}_T^2 T^{-1} Z'Z \tag{2.29}$$

where $\hat{\sigma}_T^2 = T^{-1}u(\hat{\theta}_T)'u(\hat{\theta}_T)$ and we have used the "CIV" subscript to emphasize the imposition of the Classical assumptions about u_t but suppressed the T subscript for notational simpicity.

Finally, we derive the asymptotic distribution of the estimated sample moment. For reasons that will become apparent, it is most convenient to consider the transformed version of this statistic obtained by premultiplying the original by $W_T^{1/2}$. First notice that

$$W_T^{1/2}T^{-1/2}Z'u(\hat{\theta}_T) \;=\; W_T^{1/2}T^{-1/2}Z'u \;-\; W_T^{1/2}T^{-1}Z'X\,T^{1/2}(\hat{\theta}_T - \theta_0) \tag{2.30}$$

and so it follows from (2.23) that

$$W_T^{1/2}T^{-1/2}Z'u(\hat{\theta}_T) \;=\; (I_q - P_T)W_T^{1/2}(T^{-1/2}Z'u) \tag{2.31}$$

where $P_T = F_T(F_T'F_T)^{-1}F_T'$ and – as in Section 2.3 – $F_T = W_T^{1/2}(T^{-1}Z'X)$. Inspection of (2.31) reveals that $T^{-1/2}Z'u(\hat{\theta}_T)$ has a similar structure to $T^{1/2}(\hat{\theta}_T - \theta_0)$ – that is, it takes the form $N_T n_T$ where N_T converges in probability to a matrix of constants and n_T converges to a vector of normal random variables. Therefore, we can once again use Lemma 1.4 to deduce the limiting distribution, namely

$$W_T^{1/2}T^{-1/2}Z'u(\hat{\theta}_T) \;\xrightarrow{d}\; N(0, NSN') \tag{2.32}$$

where $N = [I_q - P]W^{1/2}$. In Section 2.3, it is noted that the estimated sample moment is closely related to the overidentifying restrictions, and this connection also manifests itself in the asymptotic distribution. Equation (2.31) implies that

$$W_T^{1/2}T^{-1/2}Z'u(\hat{\theta}_T) \;=\; (I_q - P)W^{1/2}T^{-1/2}Z'u \;+\; o_p(1) \tag{2.33}$$

and so the asymptotic behaviour of $W_T^{1/2}T^{-1/2}Z'u(\hat{\theta}_T)$ is governed by the function of the data which appears in the overidentifying restrictions. Once this

relationship is recognized then it becomes apparent that the limiting distribution in (2.32) only has mean zero if the overidentifying restrictions are satisfied at θ_0. One other aspect of the limiting distribution should also be noted. The covariance matrix is

$$NSN' = (I_q - P)W^{1/2}SW^{1/2'}(I_q - P) \tag{2.34}$$

Since $W^{1/2}$ and S are nonsingular, it follows[7] from (2.34) that $rank(NSN') = rank(I_q - P) = q - p$, and hence that NSN' is singular.[8] Notice that this rank equals the degree of overidentification and so further emphasizes the connection between the estimated sample moment and the overidentifying restrictions.

2.4 The Optimal Choice of Weighting Matrix

So far, the analysis has taken the weighting matrix as given and only placed fairly mild restrictions on its composition in Definition 1.2. At the same time, it has been seen that this matrix plays a crucial role in the analysis because it determines the exact nature of the minimand. In this section, we characterize the optimal choice of weighting matrix and this leads us to a discussion of the two step or iterated GMM estimator.

To begin, we must consider what is meant by "optimality" in this context. An inspection of the previous analysis indicates that the weighting matrix only affects the asymptotic properties of the estimator via the covariance matrix in (2.25). This can be anticipated from the role of W_T in the estimation. The estimator will converge in probability to the true value as long as the population moment and identification conditions hold. Essentially, these conditions ensure there is sufficient information from which to estimate θ_0 and that this information is correct. The choice of weighting matrix determines how this information is used and so impacts directly upon the precision of the estimation. It is this feature which is captured by the variance of the asymptotic distribution. Therefore the optimal weighting matrix is defined to be the value which minimizes the asymptotic variance.

Inspection of (2.25) reveals that it is the probability limit of W_T, W, which affects the asymptotic variance of $\hat{\theta}_T$. Therefore, we begin by characterizing the optimal value of W and then consider the issues involved in constructing a matrix which converges to this limit. For this discussion it is useful to introduce the following notation for the asymptotic variance of $\hat{\theta}_T$ given in (2.25),

$$V(W) = \{E[x_t z_t']WE[z_t x_t']\}^{-1} E[x_t z_t']\, W\, S\, W\, E[z_t x_t']\{E[x_t z_t']WE[z_t x_t']\}^{-1} \tag{2.35}$$

The optimal value of W, W^0 say, is the value which minimizes $V(W)$ in a matrix sense and so satisfies

$$V(\tilde{W}) - V(W^0) = \textit{a positive semi-definite matrix}$$

[7] See Dhrymes (1984) [p.17].
[8] See Rao (1973) [Chapter 8] for a discussion of the singular normal distribution.

for any other valid choice of weighting matrix, $\tilde{W}$. Hansen (1982) shows that $W^0 = S^{-1}$. Substituting this value into (2.35) yields

$$V(S^{-1}) = \{E[x_t z_t'] \, S^{-1} \, E[z_t x_t']\}^{-1} \tag{2.36}$$

This matrix $V(S^{-1})$ represents an efficiency bound for GMM estimation of θ_0 based on the population moment condition $E[z_t u_t(\theta_0)] = 0$ because all other choices of W result in a variance which is at least as large.

To construct a GMM estimator which reaches this bound, it suffices to put W_T equal to $\hat{S}_T^{-1}$, where $\hat{S}_T$ is a consistent estimator of S. This appears to create a circularity because (2.28) indicates that $\hat{S}_T$ depends on $\hat{\theta}_T$; this is easily resolved, however. For the consistency of $\hat{S}_T$, it is only necessary that this matrix is constructed using a consistent estimator of θ_0 and not the optimal estimator. This leads us to Hansen's (1982) two step procedure for optimal GMM estimation. On the first step, a consistent estimator of θ_0 is obtained using GMM with a sub-optimal weighting matrix such as $W_T = I_q$ or $W_T = (T^{-1}Z'Z)^{-1}$. This estimator is used to construct $\hat{S}_T$. On the the second step, the model is re-estimated using $W_T = \hat{S}_T^{-1}$. These two steps are sufficient to obtain an estimator with asymptotic covariance matrix equal to $V(S^{-1})$. However, the estimator of S used in the second step estimation is based on a sub-optimal estimator of θ_0 and so there may be gains in finite sample performance from iterating this procedure. In some cases, iteration may be unnecessary. For example, in the Classical regression model setting (Assumptions 2.1-2.5) the optimal estimator can be constructed by setting W_T just equal to $(T^{-1}Z'Z)^{-1}$ instead of $\hat{S}_{CIV}^{-1}$ because the factors involving $\hat{\sigma}_T^2$, and so $\hat{\theta}_T$, cancel out. In this case the optimal estimator can be calculated in one step, and can be recognized as the Two Stage Least Squares (2SLS) estimator. In practice, this type of convenient cancellation is rare and so iteration is required in most cases of interest.

Finally, a matter of terminology should be addressed. The estimator described in this subsection is typically refered to as "the optimal two step (or iterated) GMM estimator". It is important to remember that this optimality only refers to the choice of weighting matrix and there is no implication that the population moment condition is optimal in any sense. It is possible to characterize the optimal set of population moment conditions to use in GMM estimation. However, this is an extremely complicated problem for the types of model in Table 1.1. Therefore, it serves no useful pedagogic value to explore this issue here but we return to it in Chapter 7.

2.5 Specification Error: Consequences and Detection

So far, it has been assumed that the underlying economic/statistical model is correctly specified. Unfortunately, this need not be the case, and so it is important to consider how specification error would impact on the asymptotic

properties of the estimator and the estimated sample moment. Intuition suggests that such an error renders all inferences suspect at best and completely invalid at worse. This is born out by the discussion below, and so motivates the development of statistical procedures to assess whether the model is correctly specified. In this section we introduce the overidentifying restrictions test which has become the standard diagnostic for model specification within the GMM framework. Other diagnostics are discussed in Chapter 5.

To facilitate the discussion, it is useful to recap briefly what aspects of the model impact on $\hat{\theta}_T$ and $T^{-1}Z'u(\hat{\theta}_T)$. To this end, it is useful to introduce the notation $\mathcal{M}$ to denote the underlying economic/statistical model. As we have seen, this model has the property

$$\mathcal{M} \implies E[z_t u_t(\theta_0)] = 0, \ \forall t \text{ for some unique } \theta_0 \in \Theta \tag{2.37}$$

The population moment condition in (2.37) implies the identifying restrictions are satisfied at θ_0 and so $\hat{\theta}_T$ both converges in probability to θ_0 and $T^{1/2}(\hat{\theta}_T - \theta_0)$ converges to a mean zero normal distribution. The population moment condition also implies the overidentifying restrictions are satisfied at θ_0 and so $T^{-1/2}Z'u(\hat{\theta}_T)$ converges to a mean zero normal distribution.

If $\mathcal{M}$ is no longer considered to be the truth, then there are two natural, alternative scenarios. First, the true model, $\mathcal{M}_A$ say, although different from $\mathcal{M}$, shares the property in (2.37) – that is

$$\mathcal{M}_A \implies E[z_t u_t(\theta_+)] = 0, \ \forall t \text{ for some unique } \theta_+ \in \Theta \tag{2.38}$$

Secondly, the true model, $\mathcal{M}_B$ say, implies the property in (2.37) does not hold – that is

$$\mathcal{M}_B \implies \nexists \ \theta \in \Theta \text{ such that } E[z_t u_t(\theta)] = 0, \ \forall t \tag{2.39}$$

Notice that (2.38) can hold for any $q \geq p$ but (2.39) can only hold for $q > p$. This follows because if $q = p$ then $E[z_t u_t(\theta)] = 0$ represents a set of p equations in p unknowns which must perforce have a solution – subject to the identification condition in Assumption 2.3. We now consider the behaviour of the estimator and estimated sample moment under $\mathcal{M}_A$ and $\mathcal{M}_B$.

First, consider the case where the true model is $\mathcal{M}_A$. Since $\mathcal{M}$ and $\mathcal{M}_A$ are different by definition, they must have different implications for some aspect of the distribution of v_t. However, a comparison of (2.37) and (2.38) indicates that $\mathcal{M}$ and $\mathcal{M}_A$ have the same implications for $E[z_t u_t(\theta)]$ – the only potential difference being in the parameter value at which the moment condition is satisfied. The population moment condition in (2.38) implies the identifying restrictions are satisfied at θ_+, and so the analysis in Section 2.3 can be replicated to show that $\hat{\theta}_T$ converges in probability to θ_+. Furthermore, this analysis can be continued as before to show that $T^{1/2}(\hat{\theta}_T - \theta_+)$ converges to a mean zero normal distribution. Equation (2.38) also implies the overidentifying restrictions are satisfied at θ_+ and so this in turn implies that the estimated sample moment converges to a mean zero normal random vector. So the only potential difference between $\mathcal{M}$ and $\mathcal{M}_A$ is in the value to which $\hat{\theta}_T$ converges. However, as stated

above, neither model implies anything about the value of θ which satisfies the population moment condition beyond its uniqueness. Therefore, $\mathcal{M}$ and $\mathcal{M}_A$ are *observationally equivalent* on the basis of $E[z_t u_t(\theta)]$ alone.

In contrast, $\mathcal{M}$ and $\mathcal{M}_B$ have very different implications for $E[z_t u_t(\theta)]$. Equation (2.39) states that there is no value of θ at which the population moment condition is satsified. In spite of this, there must be a solution to the identifying restrictions because they constitute a set of p equations in p unknowns.[9] If this solution is denoted θ_*, then it follows by the same logic as before that $\hat{\theta}_T$ converges in probability to θ_*. It is also possible to develop an asymptotic distribution theory for the estimator in this case, but the analysis is more complicated than under $\mathcal{M}$. However the most important difference emerges in the behaviour of the estimated sample moment. The analysis in Section 2.3 can be replicated to show that

$$W_T^{1/2} T^{-1/2} Z'u(\hat{\theta}_T) \;=\; (I_q - P) W^{1/2} T^{-1/2} Z'u(\theta_*) \;+\; o_p(1) \tag{2.40}$$

It is apparent from (2.40) that the asymptotic behaviour of $W_T^{1/2} T^{-1/2} Z'u(\hat{\theta}_T)$ is determined by whether or not the overidentifying restrictions are satisfied at θ_*. The answer to this question can be deduced from the properties of θ_*. By definition, θ_* satisfies the identifying restrictions and (2.39) implies that $E[z_t u_t(\theta_*)] \neq 0$. Since,

$$W^{1/2} E[z_t u_t(\theta_*)] \;=\; P W^{1/2} E[z_t u_t(\theta_*)] + (I_q - P) W^{1/2} E[z_t u_t(\theta_*)]$$

it must follow that

$$(I_q - P) W^{1/2} E[z_t u_t(\theta_*)] \neq 0 \tag{2.41}$$

Equations (2.40) and (2.41) imply that $W_T^{1/2} T^{-1/2} Z'u(\hat{\theta}_T)$ is not $O_p(1)$ – as it is under $\mathcal{M}$ or $\mathcal{M}_A$ – but diverges at rate $T^{1/2}$ and, in consequence, does not converge in distribution.[10]

Regardless of whether $\mathcal{M}_A$ or $\mathcal{M}_B$ is the truth, it is desirable to develop statistical tests which can indicate that the assumed model is incorrect. Clearly, it is impossible to discriminate between $\mathcal{M}$ and $\mathcal{M}_A$ on the basis of $T^{-1/2} Z'u(\hat{\theta}_T)$. This can only be achieved by deducing a different set of moment conditions from $\mathcal{M}$ and testing whether they are corroborated by the data.[11] On the other hand, $\mathcal{M}$ and $\mathcal{M}_B$ have different implications for the overidentifying restrictions and so it would be anticipated that it is possible to discriminate between these two models based on the estimated sample moment.

Sargan (1958) was the first person to introduce the idea of testing the overidentifying restrictions in a linear model estimated by IV, and Hansen (1982) extended the statistic to the GMM framework. It is natural to base the test on the GMM minimand, $Q_T(\hat{\theta}_T)$, since it is shown in Section 2.3 that this statistic measures how far the sample is from satisfying the overidentifying restrictions.

[9] Again, subject to the identification condition in Assumption 2.3.

[10] See Chapter 4.

[11] However, the same problem recurs because there is always more than one probability distribution which can generate a finite set of population moment conditions.

To develop the distribution theory, it is most convenient to focus on the optimal GMM estimator, and so we set $W_T = \hat{S}_T^{-1}$. Therefore, the *overidentifying restrictions test* statistic[12] is given by

$$J_T = TQ_T(\hat{\theta}_T) = T^{-1/2}u(\hat{\theta}_T)'Z\,\hat{S}_T^{-1}\,T^{-1/2}Z'u(\hat{\theta}_T) \tag{2.42}$$

Under the null hypotheses,

$$H_0 : E[z_t u_t(\theta_0)] = 0$$

J_T converges in distribution to a χ^2_{q-p}.[13] Notice that the degrees of freedom equal the number of overidentifying restrictions. Intuition suggests that J_T can detect when the true model is actually $\mathcal{M}_B$, and this is verified in Chapter 5.

2.6 Summary

The purpose of this chapter is to introduce the main elements of the GMM framework using the example of the IV estimator in the static linear regression model. This approach is feasible because the intrinsic information in IV estimation takes the form of a population moment condition. Specifically, IV rests crucially on the existence of a vector of instruments, z_t, that are uncorrelated with the regression error, $u_t(\theta_0)$, or equivalently that the instruments satisfy $E[z_t u_t(\theta_0)] = 0$. If this population moment condition is used as a basis for GMM estimation then the resulting GMM estimator is the IV estimator. The advantage of deriving IV in this way is that it enables us to highlight seven key features of the GMM framework:

- *Identification*: For the estimation to be succesful, the population moment condition must not only be valid but also provide sufficient information to identify the parameter vector.

- *Identifying and overidentifying restrictions*: GMM estimation in overidentified models involves a fundamental decomposition of the moment condition into identifying restrictions and overidentifying restrictions. The identifying restrictions contain the information that goes into the estimation, and the overidentifying restrictions are a remainder that manifests itself in the estimated sample moment.

- *Asymptotic properties*: The GMM estimator is consistent and, when appropriately scaled, has a limiting distribution that is normal.

[12] This is also sometimes refered to as the *J–test*.

[13] It can be recognized that the overidentifying restrictions test is a direct extension of Neyman and Pearson (1928) statistic $GF(\hat{\theta}_T)$ discussed in Section 1.2. At first glance, the degrees of freedom appear to be in conflict; however, there is a logical explanation. Only $k-1$ of the population moment conditions in (1.7) are free: the k^{th} condition, say, is implied by first $k-1$ plus the constraints $\sum_{i=1}^{k}[D_t(i) - h(i;\theta_0)] = 0$ which must hold because of the definitions of $D_t(i)$ and $h(i;\theta_0)$.

- *Estimated sample moment*: The estimated sample moment is shown to have a limiting normal distribution whose attributes depend directly on the function of the data in the overidentifying restrictions.
- *Long run covariance estimation*: To translate this asymptotic normality into practical inference procedures, it is necessary to estimate the long run variance of the sample moment consistently.
- *Optimal choice of weighting matrix*: The optimal choice of weighting matrix depends on the long run variance of the sample moment and so its use typically involves a two step or iterated estimation.
- *Model diagnostics*: The overidentifying restrictions provide a basis for testing the validity of the model specification via the estimated sample moment.

Subsequent chapters build from this foundation to present the GMM framework in nonlinear dynamic models. Chapter 3 focuses on estimation and, in its course, extends the discussion of the first five aspects highlighted above to the general setting. The statistical properties derived in Chapter 3 are premised on the assumption that the model is correctly specified. Chapter 4 considers the impact of misspecification on the limiting properties of the GMM estimator. Chapter 5 derives the large sample properties of both the overidentifying restrictions test and also a number of other hypothesis tests which have been proposed within the GMM framework.

3

GMM Estimation in Correctly Specified Models

The previous chapter has provided an introduction to the GMM framework and the types of inference issues which arise within it. Although many of the details reflected the static, linear nature of the model, the underlying intuition did not. The essential feature of the estimation is the minimization of a quadratic form in the sample analog to a population moment condition which provided sufficient information to identify the unknown parameters. In this chapter, we show this strategy can be successfully extended to nonlinear dynamic models. The focus here is on the estimator and the derivation of its statistical properties in correctly specified models. The impact of misspecification on these properties is examined in Chapter 4. Matters of inference are postponed until Chapter 5 when a variety of hypothesis testing procedures are reviewed. The level of the discussion is more rigorous than the previous chapter, and the main results are formally proved. However, the issues are also illustrated throughout with an empirical example to provide guidance on the practical implementation of the estimator as well. Here, we focus on Hansen and Singleton's (1982) consumption based asset pricing model which was described in Section 1.3.1. Chapter 9 reports empirical results for the other four models in Section 1.3.

Section 3.1 defines the population moment condition and presents conditions for parameter identification. Section 3.2 discusses the calculation of the estimator in practice and includes a brief review of numerical optimization techniques. Section 3.3 extends the fundamental decomposition of the population moment condition into identifying and overidentifying restrictions to the nonlinear model. Section 3.4 derives the asymptotic properties of the estimator and the estimated sample moment. Section 3.4.1 presents a proof of consistency and Section 3.4.2 derives the asymptotic distribution of the estimator, and also uses this analysis to provide further insights into the form of the identifying and overidentifying restrictions. Section 3.4.3 derives the asymptotic distribution of the estimated sample moment. Section 3.5 describes the construction

of consistent estimators of the long run variance under three scenarios for the dynamic structure of the sample moment. Section 3.5.1 covers the case where $f(v_t, \theta_0)$ is a serially uncorrelated process; Section 3.5.2 considers the case where $f(v_t, \theta_0)$ is generated by a vector autoregressive moving average process; and finally Section 3.5.3 considers the class of heteroscedasticity and autocorrelation covariance (HAC) matrix estimators whose properties only require the dependence structure to satisfy very mild restrictions. Section 3.6 derives the optimal choice of weighting matrix and this leads to a discussion of the two step and iterated GMM estimators. Section 3.7 examines the consequences of various transformations and normalizations on the GMM estimator, and this leads to a discussion of both the continuous updating GMM estimator and also the construction of confidence intervals based directly on the GMM minimand. The chapter concludes with a slight detour. In Chapter 1, it is stated that many estimators can be viewed as special cases of GMM. Although some simple examples were provided, it was not possible to elaborate on the point at that stage. However, this is possible after the material in the first five sections of this chapter. Section 3.8 shows formally how other estimators can be fit within the GMM framework. Section 3.9 contains a summary of the chapter.

3.1 Population Moment Condition and Parameter Identification

In Chapter 1, it was shown that a wide variety of econometric models lead to population moment conditions which involve nonlinear functions of the data and parameters. It is therefore desirable to adopt a very general framework which encompasses all these cases. This means that the analysis in this chapter begins with the population moment condition and no attempt is made to characterize the specific data generation process which lays behind it. This population moment condition involves a function $f(.,.)$ of the observable vector of random variables v_t and the unknown $(p \times 1)$ parameter vector, θ_0. As before the parameter space is denoted by $\Theta \subseteq \Re^p$. However, before we introduce the population moment and identification conditions, certain restrictions need to be placed on v_t and $f(.,.)$.

Assumption 3.1 Strict Stationarity
The $(r \times 1)$ *random vectors* $\{v_t; -\infty < t < \infty\}$ *form a strictly stationary process with sample space* $\mathbf{V} \subseteq \Re^r$.

Recall that this assumption implies all expectations of functions of v_t are independent of time.

Assumption 3.2 Regularity Conditions for $f(.,.)$
The function $f : \mathbf{V} \times \Theta \to \Re^q$, *where* $q < \infty$, *satisfies: (i) it is continuous on* Θ *for each* $v_t \in \mathbf{V}$; *(ii)* $E[f(v_t, \theta)]$ *exists and is finite for every* $\theta \in \Theta$; *(iii)* $E[f(v_t, \theta)]$ *is continuous on* Θ.

Formally, it is necessary to assume that $f(., \theta)$ is a measurable function but we suppress this type of condition throughout the text. All functions considered are assumed to be measurable; see Newey and McFadden (1994) for a discussion of circumstances in which this may not hold. Assumption 3.2 holds in most, if not all, of the models behind the studies listed in Table 1.1. However, this assumption excludes some cases of interest, such as step functions which are by their nature discontinuous. One further aspect of Assumption 3.2 should be noted. The function $f(.)$ is assumed to be finite dimensional. This assumption is standard and satisfied in all the applications listed in Table 1.1. However, there are circumstances in which it may be desirable to relax this assumption. In Section 6.1.3, we consider the limiting behaviour of the estimator when q tends to infinity with the sample size. It is also possible to generalize the GMM framework to a continuum of moment conditions but we do not pursue this extension. For the latter, the interested reader is refered to Carrasco and Florens (2000).

The analysis centers on the following population moment condition.

Assumption 3.3 Population Moment Condition
The random vector v_t and the parameter vector θ_0 satisfy the $(q \times 1)$ population moment condition: $E[f(v_t, \theta_0)] = 0$.

Just as in the linear model, the population moment condition can only be used as a basis for estimation if it provides enough information to uniquely identify the parameter vector θ_0. In the linear model, it is possible to relate parameter identification to a simple condition which only involved the data. In nonlinear models, the situation is more complicated. Identification can fail due to the properties of the data, v_t, or due to the properties of $f(.)$ as a function of θ or due to an interaction of the two. To characterize how these types of failure can occur in nonlinear models, it is necessary to introduce the concepts of *global* and *local* identification. The need for this distinction will become apparent below.

The basic condition for parameter identification is given by:

Assumption 3.4 Global Identification
$E[f(v_t, \bar{\theta})] \neq 0$ for all $\bar{\theta} \in \Theta$ such that $\bar{\theta} \neq \theta_0$.

The adjective "global" emphasizes that the population moment condition only holds at one value in the *entire* parameter space. This can be recognized as the concept of identification used in our discussion of the linear model in the previous chapter. Within that context, it was possible to derive a convenient condition for global identification. Unfortunately, this is rarely possible in nonlinear models. However, there is one type of identification failure in nonlinear models which can be diagnosed using the condition in Assumption 3.4. This is the case when failure occurs due to the nature of $f(.)$ as a function of θ. This type of problem is best understood by considering two examples: in the the first there are just two values of θ_0 which satisfy the population moment condition; in the second, there are an infinite number of values which do so.

Example: The Partial Adjustment Model
Suppose the data are generated by the model[1]

$$\begin{aligned} y_t - y_{t-1} &= \beta_0(y^* - y_{t-1}) + u_t \\ u_t &= \rho_0 u_{t-1} + e_t \end{aligned}$$

where y^* represents the desired level of the process y_t and e_t is an i.i.d. process with mean zero. Simple rearrangement yields

$$y_t = \beta_0(1-\rho_0)y^* + (1+\rho_0-\beta_0)y_{t-1} + (\beta_0-1)\rho_0 y_{t-2} + e_t \tag{3.1}$$

Now suppose there exists a set of variables z_t which satisfy the population moment condition $E[z_t e_t(\theta_0)] = 0$ where $e_t(\theta) = y_t - \beta(1-\rho)y^* - (1+\rho-\beta)y_{t-1} - (\beta-1)\rho y_{t-2}$ and $\theta = (\beta, \rho, y^*)'$. Although this is very similar to the population moment condition in Chapter 2, it is outside that framework because $e_t(\theta)$ is a nonlinear function of θ. Using the condition in Assumption 3.4, the parameter vector is identified if $E[z_t e_t(\theta)] = 0$ at only $\theta = \theta_0$. To see if this holds, it is useful to introduce the notation

$$e_t(\mu) = y_t - \mu_0 - \mu_1 y_{t-1} - \mu_2 y_{t-2} \tag{3.2}$$

where $\mu = (\mu_0, \mu_1, \mu_2)'$. Equation (3.2) can be viewed as a type of "reduced form" version of $e_t(\theta)$ because any value of θ implies a value for μ via the relationship,

$$\begin{aligned} \mu_0 &= \beta(1-\rho)y^* \\ \mu_1 &= 1+\rho-\beta \\ \mu_2 &= (\beta-1)\rho \end{aligned} \tag{3.3}$$

Using these definitions, the condition for identification can be restated as the requirement that each value of μ is implied by only one value of θ. However, inspection of (3.3) reveals this is not the case here. The problems arise because the bottom two equations imply a quadratic equation for ρ, namely $\rho^2 - \rho\mu_1 - \mu_2 = 0$, to which there are two solutions. Denote these by ρ_1 and ρ_2. Each of these solutions implies a value of β which satisfies the bottom two equations as well; denote these by $\beta_i = 1 + \mu_2/\rho_i$ for $i = 1, 2$. Finally let $y_i^* = \mu_0/\{\beta_i(1-\rho_i)\}$. Clearly $\theta_i = (\beta_i, \rho_i, y_i^*)'$ yields the same value of μ for both $i = 1, 2$ and so Assumption 3.4 is violated. $\diamond$

Example: Eichenbaum's (1989) Model for Inventory Holdings by Firms

[1] This type of model has been used to analyze a wide variety of economic series including money demand and inventory holdings. In these applications exogenous regressors are also included and formally this removes the identification problem. However, if the regressors only play a very marginal role then the same type of identification problems can emerge; see Blinder (1986), Hall and Rossana (1991).

It is shown in Section 1.3.4 that Eichenbaum's (1989) model for inventory holdings implies that the following population moment condition holds

$$E[\{h_{t+2}(\psi_0) - \rho_0 h_{t+1}(\psi_0)\}z_t] = 0 \tag{3.4}$$

where $h_{t+1}(\psi) = I_{t+1} - \{\lambda + (\lambda\beta)^{-1}\}I_t + \beta^{-1}I_{t-1} + S_{t+1} - \phi\beta^{-1}S_t$ and $\psi = (\lambda, \beta, \phi)'$. In our earlier discussion of this model, ϕ is treated as a parameter to be estimated rather than the three underlying parameters of which it is a function. It may have been wondered why (δ, γ, α) are not estimated directly and the answer is that they are not identified by the population moment condition. The problem arises because the elements of (δ, γ, α) only appear in a ratio form via $\phi = 1 - \delta\gamma/\alpha$. Therefore, for any non-zero constant k, both $(\bar{\delta}, \bar{\gamma}, \bar{\alpha})$ and $(k\bar{\delta}, \bar{\gamma}, k\bar{\alpha})$ yield the same value of ϕ. This would clearly cause a violation of Assumption 3.4. However, there is no such problem if only ϕ is estimated instead. $\diamond$

In both these examples, the identification failure arises because of the nature of $f(.)$ as a function of θ. As mentioned above, identification can fail for other reasons but these are harder, if not impossible, to diagnose by examining $E[f(v_t, \theta)]$ directly. In the linear model of the previous chapter, it is possible to deduce a relatively simple condition for global identification and it would clearly be desirable to develop something similar for nonlinear models. Unfortunately, this cannot be done because it is typically impossible to find a useful alternative representation for $f(v_t, \theta)$ which holds over all $\theta \in \Theta$. However such a representation can be found if attention is limited to some suitably defined neighbourhood of θ_0. The price of this approach is that we are now deriving conditions for identification only within this neighbourhood and these are refered to as conditions for *local* identification. As the names suggest, local identification does not guarantee global identification but global identification cannot hold without local identification. Therefore, a more transparent condition for local identification is useful because it provides insights into when identification can fail.

To derive the condition for local identification, it is necessary to introduce the following definition and assumption. An *ϵ–neighbourhood* of θ_0 is defined to be the set N_ϵ which satisfies $N_\epsilon = \{\theta;\ ||\theta - \theta_0|| < \epsilon\}$. The aforementioned alternative representation of $f(.)$ is based on a first order Taylor Series approximation for $f(v_t, \theta)$ over a neighbourhood of the form N_ϵ. For this to be valid, it is necessary that $N_\epsilon \subset \Theta$ and so θ_0 must be an interior point of Θ.[2] So this condition is included with certain other regularity conditions in the following assumption.

Assumption 3.5 Regularity Conditions on $\partial f(v_t, \theta)/\partial\theta'$
(i) The derivative matrix $\partial f(v_t, \theta)/\partial\theta'$ exists and is continuous on Θ for each $v_t \in \mathbf{V}$; (ii) θ_0 is an interior point of Θ; (iii) $E[\partial f(v_t, \theta_0)/\partial\theta']$ exists and is finite.

[2] In other words θ_0 must not lie on the boundary of Θ. See Apostol (1974) [p.49] for definition of the interior of a set.

Part (i) of this condition is satisfied by most, but not all, of the models behind the studies listed in Table 1.1. If violations occur they tend to stem from the presence of absolute values for which the derivative is not defined everywhere on Θ. For example, the stochastic volatility model in Section 1.3.5 leads to population moment conditions which involve absolute values.[3] It is possible to develop local identification conditions in these situations but the analysis becomes more complicated.[4] Since these cases tend to be the exception rather than the rule, we work here within the framework of Assumption 3.5. Notice that the other four models in Section 1.3 satisfy Assumption 3.5(i) and the other two parts of the assumption can reasonably be expected to hold as well.

The condition for local identification is derived by restricting attention to sufficiently small ϵ so that $f(.)$ is equal to the following first order Taylor series expansion [5] in N_ϵ

$$f(v_t, \theta) \;=\; f(v_t, \theta_0) + \{\partial f(v_t, \theta_0)/\partial \theta'\}(\theta - \theta_0) \tag{3.5}$$

The advantage of this approach is that (3.5) implies $f(v_t, \theta)$ is a a linear function of $\theta - \theta_0$ in this neighbourhood. Taking expectations on both sides of (3.5) and using Assumptions 3.3 and 3.5 yields

$$E[f(v_t, \theta)] \;=\; \{E[\partial f(v_t, \theta_0)/\partial \theta']\}(\theta - \theta_0) \tag{3.6}$$

Equation (3.6) is essentially the same structure as (2.3) and so we can appeal to our earlier analysis of the the linear model to deduce the following condition for *local* identification.

Assumption 3.6 Local Identification
$rank\{E[\partial f(v_t, \theta_0)/\partial \theta']\} = p.$

This condition can be recognized as the generalization of the identification condition for the linear model given in Assumption 2.3.[6] Notice the form of the condition immediately implies identification fails if there are fewer moment conditions than parameters, i.e. $q < p$. While this is no surprise given the discussion in Chapter 2, this restriction was not immediately apparent from the global identification condition in Assumption 3.4. As in the linear model, this type of condition can also fail if $q \geq p$. However, one important difference is that identification in nonlinear models may be sensitive to the value of θ_0 via $\partial f(v_t, \theta)/\partial \theta'$. This opens up the possibility that the population moment condition may provide enough information to identify the parameters at some values of θ_0 but not at others.

[3] Another example is encountered in Section 9.1 when we consider an extension of the mutual fund evaluation method described in Section 1.3.2.

[4] The interested reader is refered to Newey and McFadden (1994)[Section 7].

[5] See Apostol (1974)[p.361].

[6] In the linear model, $f(v_t, \theta) = z_t u_t(\theta)$ and so $\partial f(v_t, \theta_0)/\partial \theta' = -z_t x_t'$. The condition implies global identification in the linear model because (3.5) is then an identity which holds for all θ and not just in a neighbourhood of θ_0.

Clearly, the exact nature of the condition in Assumption 3.6 depends on the $f(.)$ in question. To illustrate the types of condition which can arise in practice, we now examine local identification in three examples. We begin with continuations of our earlier examples to illustrate the difference between global and local identification. We then derive the local identification condition for the consumption based asset pricing model in Section 1.3.1. Further examples can be found in Chapter 9.

Example: Partial Adjustment Model (Continued)
Recall that $f(v_t, \theta) = z_t e_t(\theta)$ and so $\partial f(v_t, \theta_0)/\partial \theta' = z_t \partial e_t(\theta_0)/\partial \theta'$. From the definition of $e_t(\theta)$ and θ it follows that

$$E[\partial f(v_t, \theta_0)/\partial \theta'] = E[z_t \tilde{x}_t'] M(\theta_0) \tag{3.7}$$

where $\tilde{x}_t = (1, y_{t-1}, y_{t-2})'$ and

$$M(\theta) = \begin{bmatrix} -(1-\rho)y^* & \beta y^* & -\beta(1-\rho) \\ 1 & -1 & 0 \\ -\rho & -(\beta - 1) & 0 \end{bmatrix}$$

Given this structure, it follows that[7]

$$rank\{E[\partial f(v_t, \theta_0)/\partial \theta']\} \leq min\{rank(E[z_t \tilde{x}_t']), rank(M(\theta_0))\}$$

Inspection of $M(\theta_0)$ indicates that in general this matrix is of full rank and so $rank\{E[\partial f(v_t, \theta_0)/\partial \theta']\} = rank(E[z_t \tilde{x}_t'])$.[8] Therefore local identification rests on the relationship between the instruments and $\tilde{x}_t$ in a similar way to our earlier analysis of the linear regression model. Assuming this rank condition holds, θ_0 is locally identified.

It is informative to relate this conclusion back to our earlier analysis of this model. It was shown there that the parameter vector is globally unidentified because there are two values of θ which satisfy the population moment condition. This failure arose because the solutions for ρ satisfy a quadratic equation to which the roots are

$$\rho = \mu_1 \pm (\sqrt{\mu_1^2 + 4\mu_2})/2$$

Notice that this structure suggests the two solutions are distinct values of θ_0 and not within an ϵ neighbourhood of each other for some suitably small value of ϵ. It is therefore consistent with the finding that the two solutions are locally identified even though θ_0 is globally unidentified. $\diamond$

Example: Eichenbaum's (1989) Model for Inventory Holdings by Firms (Continued)
We again reconsider the problem of estimating the augmented parameter vector

[7] See Dhrymes (1984) [Proposition 7, p.17].
[8] See Ibid [Proposition 6, p.16].

in which (δ, γ, α) are included in ψ instead of ϕ. To simplify the analysis it is convenient to set $\rho = 0$ but this does not effect the essence of the argument. This case maps into our generic notation with $f(v_t, \theta_0) = h_{t+2}(\psi_0)z_t$ where $\theta = (\lambda, \beta, \delta, \gamma, \alpha)'$. As in the previous example the nonlinearity only arises through the parameters and so the derivative matrix has a similar structure

$$E[\partial f(v_t, \theta_0)/\partial\theta'] = E[z_t\tilde{x}_t']M(\theta_0) \tag{3.8}$$

except this time $\tilde{x}_t = (I_{t+1}, I_t, S_{t+1})'$ and

$$M(\theta) = \begin{bmatrix} (\lambda^2\beta)^{-1} - 1 & (\lambda\beta^2)^{-1} & 0 & 0 & 0 \\ 0 & -\beta^{-2} & 0 & 0 & 0 \\ 0 & (1 - \delta\gamma/\alpha)\beta^{-2} & \gamma\beta^{-1}/\alpha & \delta\beta^{-1}/\alpha & -\delta\gamma\beta^{-1}/\alpha^2 \end{bmatrix}$$

However this time it is immediately apparent that $rank\{M(\theta)\} \leq 3$ and so $rank\{E[\partial f(v_t, \theta_0)/\partial\theta']\} \leq 3 < p$. Therefore θ_0 is locally unidentified in this model. Again this result ties in with our previous analysis of global identification. It was shown before that $(\bar{\delta}, \bar{\gamma}, \bar{\alpha})$ and $(k\bar{\delta}, \bar{\gamma}, k\bar{\alpha})$ yield the same value of ϕ for any nonzero constant k. Since k can be arbitrarily close to one, it follows that if $\theta_0 = (\lambda_0, \beta_0, \delta_0, \gamma_0, \alpha_0)'$ satisfies the population moment condition then there is always another value $\theta_* = (\lambda_0, \beta_0, k\delta_0, \gamma_0, k\alpha_0)'$ within an ϵ neighbourhood of θ_0 which also satisfies the population moment condition for any $\epsilon > 0$.

Finally, it should be noted that this problem disappears if ϕ is treated as a parameter to be estimated instead of (δ, γ, α). To see this, redefine the parameter vector to be $\theta = (\lambda, \beta, \phi)'$. In this case, $\partial f(v_t, \theta_0)/\partial\theta'$ is given by (3.8) with

$$M(\theta) = \begin{bmatrix} (\lambda^2\beta)^{-1} - 1 & (\lambda\beta^2)^{-1} & 0 \\ 0 & -\beta^{-2} & 0 \\ 0 & \phi\beta^{-2} & -\beta^{-1} \end{bmatrix}$$

It is immediately apparent that $rank\{M(\theta)\} = 3$ and so local identification depends on whether $rank\{E[z_t\tilde{x}_t']\} = 3$. $\diamond$

Example: Hansen and Singleton's (1982) Consumption Based Asset Pricing Model

It is shown in Section 1.3.1 that if the representative agent possesses a CRRA utility function then the data and parameter vector, $\theta = (\gamma, \delta)'$, satisfy the population moment condition in (1.23). For our purposes here, it is convenient to restrict attention to the case in which there is only one asset with a maturity of one period. The population moment condition is then $E[z_t u_t(\theta_0)] = 0$ where $u_t(\theta) = \delta x_{1,t+1}^{\gamma-1} x_{2,t+1} - 1$, and we have set $x_{1,t+1} = c_{t+1}/c_t$, $x_{2,t+1} = r_{t+1}/p_t$ with the j subscript being dropped as there is only one asset. In this model, we have

$$E[\partial f(v_t, \theta)/\partial\theta'] = E[z_t\delta log(x_{1,t+1})x_{1,t+1}^{\gamma-1}x_{2,t+1}\ ,\ z_t x_{1,t+1}^{\gamma-1}x_{2,t+1}] \tag{3.9}$$

For local identification this matrix must have rank two when evaluated at θ_0. Apart form requiring z_t to contain at least two elements, it is not easy to deduce

from (3.9) when this rank condition holds. ◇

These three examples illustrate how the rank condition can highlight what aspects of the model are important for identification. However, as we have also seen, it may be difficult to determine a priori whether these conditions are satisfied for the data in hand. In practice, failures in identification may only become apparent when estimation is attempted and so we return to this topic in that context in the next section.[9]

3.2 The Estimator and Numerical Optimization

It can be recalled from Definition 1.2 that the GMM minimand takes the form,

$$Q_T(\theta) = \{T^{-1}\sum_{t=1}^{T} f(v_t,\theta)\}' W_T \{T^{-1}\sum_{t=1}^{T} f(v_t,\theta)\} \tag{3.10}$$

For completeness we restate the properties of the weighting matrix here.

Assumption 3.7 Properties of the Weighting Matrix
W_T is a positive semi-definite matrix which converges in probability to the positive definite matrix of constants W.

By definition, the GMM estimator of θ_0 is

$$\hat{\theta}_T \;=\; argmin_{\theta\in\Theta}\, Q_T(\theta) \tag{3.11}$$

where "argmin" stands for value of the *argument* – θ – which *minimizes* the function – $Q_T(\theta)$. If Assumption 3.5 holds, and in most cases of interest it will, then the first order conditions for this minimization imply $\partial Q_T(\hat{\theta}_T)/\partial\theta = 0$. This condition yields[10]

$$0 \;=\; \{T^{-1}\sum_{t=1}^{T} \frac{\partial f(v_t,\hat{\theta}_T)}{\partial\theta'}\}' W_T \{T^{-1}\sum_{t=1}^{T} f(v_t,\hat{\theta}_T)\} \tag{3.12}$$

In the linear model of Chapter 2, these conditions could be solved to obtain a closed form solution for $\hat{\theta}_T$ as a function of the data. Unfortunately, in nonlinear models this is typically impossible. For example, the first order conditions for Hansen and Singleton's (1982) consumption based asset pricing model are

$$\begin{aligned} 0 \;=\;& \{T^{-1}\sum_{t=1}^{T}[z_t\hat{\delta}_T log(x_{1,t+1})x_{1,t+1}^{\hat{\gamma}_T-1}x_{2,t+1}\;,\; z_t x_{1,t+1}^{\hat{\gamma}_T-1}x_{2,t+1}]\}' W_T \\ & \times\{T^{-1}\sum_{t=1}^{T} z_t(\hat{\delta}_T x_{1,t+1}^{\hat{\gamma}_T-1}x_{2,t+1}-1)\} \end{aligned} \tag{3.13}$$

[9] Also see Section 3.6.
[10] See Dhrymes (1984) [Proposition 92, p.111].

Only a little trial and error is needed to verify that these cannot be solved to produce a closed form solution for $\hat{\theta}_T$.

Back in the days of Karl Pearson, the story would have stopped here. Fortunately, the advance of computer technology over the last forty years has enabled the development of a vast array of numerical optimization routines which can be used to calculate $\hat{\theta}_T$. These days, such optimization procedures can be implemented with just a few lines of code in most econometric or statistical software packages. In view of this, we do not provide a comprehensive review of these procedures here.[11] Instead we briefly discuss certain issues involved in their implementation.

These types of computer based routines essentially perform an "informed version" of trial and error to find the value of θ which minimizes $Q_T(\theta)$. The procedure begins with some trial value of θ, $\bar{\theta}(0)$ say. If this is the value which minimizes $Q_T(\theta)$ then it should not be possible to find a value of θ for which the minimand is smaller. So the computer uses some rule to see if it can find a value of θ, $\bar{\theta}(1)$ say, which satisfies $Q_T(\bar{\theta}(1)) < Q_T(\bar{\theta}(0))$. If it can, then $\bar{\theta}(1)$ becomes the new candidate value for $\hat{\theta}_T$ and the computer searches again to see it can find a value $\bar{\theta}(2)$ such that $Q_T(\bar{\theta}(2)) < Q_T(\bar{\theta}(1))$. This updating process continues until it is judged that the value of θ which minimizes $Q_T(\theta)$ has been found. It is useful to distinguish three important aspects of such routines.

- The *starting value* for θ, $\bar{\theta}(0)$.
- The *iterative search method* by which the candidate value of $\hat{\theta}_T$ is updated on the i^{th} step.
- The *convergence criterion* used to judge when the minimum has been reached.

The various numerical optimization routines differ in how the iterative search method is performed. In most problems it is computationally infeasible to perform a search over the entire parameter space[12] and so some rule is used to limit the calculations involved. For example, in a class known as *Gradient Methods*[13] the value of θ is updated on the i^{th} step by

$$\bar{\theta}(i) = \bar{\theta}(i-1) + \lambda_i D(\bar{\theta}(i-1))$$

where λ_i is a scalar known as the *step size* and $D(.)$ is a $(p \times 1)$ vector known as the *step direction*. The step direction vector is a function of the gradient, $\partial Q_T(\bar{\theta}(i-1))/\partial\theta$, and hence reflects the curvature of the function at $\bar{\theta}(i-1)$. As the names suggest, $D(\bar{\theta}(i-1))$ determines the direction in which to update $\bar{\theta}(i-1)$ and λ_i determines how far to go in that direction.

[11] Many excellent surveys already exist in the econometrics literature e.g. Quandt (1983), Judge, Griffiths, Hill, Lutkepohl, and Lee (1985)[Appendix B] Gallant (1987)[Chapter 2].

[12] Such a strategy is known as a *grid search*.

[13] For example, see Judge, Griffiths, Hill, Lutkepohl, and Lee (1985) [p.953].

Convergence can be assessed in a number of ways. For example, if $\bar{\theta}(i)$ is the value which minimizes $Q_T(\theta)$ then the updating routine should not move away from this point. This suggests that the minimum has been found if

$$||\bar{\theta}(i+1) - \bar{\theta}(i)|| < \epsilon \tag{3.14}$$

where ϵ is an arbitrarily small positive constant. A typical value for ϵ is 10^{-6} or less. This rule allows for the fact that the update $\lambda_{i+1}D(\bar{\theta}(i))$ is unlikely to be exactly zero even if $\bar{\theta}(i)$ is the minimum due to rounding errors in calculation. As stated in (3.14), the convergence criterion is independent of the magnitude of θ. In practice, this may be a problem if the latter is very small. Ideally, ϵ should be replaced by $\eta(||\bar{\theta}(i)|| + \tau)$ where η and τ are small positive constants in the order of 10^{-5} and 10^{-3} respectively. However, in some commercially available computer packages the rule is of the form in (3.14). If this is the case then the user must be sensitive to the order of magnitude of of θ when choosing ϵ. Alternatively, convergence can be assessed by examining the first order conditions. Once the minimum is reached then (3.12) should be satisfied and this leads to the criterion

$$||\partial Q_T(\bar{\theta}(i))/\partial\theta|| < \epsilon \tag{3.15}$$

where again allowance is made for rounding errors. Finally, if the minimum has been reached then the updating should not alter the value of the minimand and so

$$|Q_T(\bar{\theta}(i+1)) - Q_T(\bar{\theta}(i))| < \epsilon \tag{3.16}$$

Once again, it is desirable for the convergence criterion to reflect the size of the objective function and so a better version of the rule is obtained by substituting $\eta(Q_T(\bar{\theta}(i)) + \tau)$ for ϵ in (3.16). However, as above, the convergence criterion in some commercially available packages takes the form in (3.16) and if it does then the user must be sensitive to the values of the minimand in choosing ϵ. Which rule should be used? It is often prudent to check all three because anyone can be satisfied by itself without the minimum being reached; see Quandt (1983) [p.737–8], Gallant (1987) [p.29].

The choice of starting values is also important. Ideally, $\bar{\theta}(0)$ should be as close as possible to the value which minimizes $Q_T(\theta)$ because this reduces the number of iterations and hence the computational burden. Sometimes a preliminary estimate of θ_0 is available and this can be used as a starting value.[14] Whether this is the case or not, it is a wise precaution to run the routine with more than one set of starting values. In nonlinear models, the minimand may exhibit a less regular topology than in the linear model with the result that the numerical routine can have problems finding the minimum. The use of multiple starting values provides some safeguard against this problem because the routine can be restarted outside of the problem areas. However, if these problems

[14] This would be the case when calculating the two step or iterated GMM estimator; see Section 2.4 and 3.6. In other cases, various rules have been suggested for the calculation of starting values. We do not describe these here but refer the interested reader to Gallant (1987) [pp.29–30] and the references therein.

persist from different starting values then this may indicate the parameter vector is unidentified by the population moment condition upon which estimation is based.

To conclude this section, we provide an illustration of these issues.

Example: Hansen and Singleton's (1982) Consumption Based Asset Pricing Model

Hansen and Singleton (1982) estimate their model with various choices of assets. We concentrate here on just two of these choices; both are portfolios constructed from all the stocks on the New York Stock Exchange and the difference between them derives from the weights used in the portfolio. In one, all the assets receive an equal weight; this choice is refered to as "equally weighted returns" and denoted EWR. In the other, the weights on the assets reflect their relative values; this choice is refered to as "value weighted returns" and denoted VWR. In principle, the population moment condition in (1.23) holds jointly for both choices of assets but it is pedagogically more convenient to estimate the model separately for each choice of asset. Each asset has maturity $m = 1$ and so (1.23) implies each of the assets satisfies

$$E[z_t(\delta_0 x_{1,t+1}^{\gamma_0 - 1} x_{2,t+1} - 1)] = 0 \tag{3.17}$$

where $x_{1,t+1} = c_{t+1}/c_t$, $x_{2,t+1} = r_{t+1}/p_t$ and z_t is the vector of instruments.

To implement the model, it is necessary to specify z_t. In Section 1.3.1 it is shown that this moment condition holds for any $z_t \in \Omega_t$, and so the economic model leaves open a lot of possibilities. Our identification analysis indicated $\theta_0 = (\gamma_0, \delta_0)'$ is locally identified by (3.17) if the rank of the matrix in (3.9) is two. As remarked above, this is not particularly illuminating apart from the requirement that z_t has at least two elements. With so many options available, Hansen and Singleton (1982) estimate the model with a number of different choices of instrument. However, here we will focus on just one to simplify the presentation; this choice is $z_t = (1, x_{1,t}, x_{1,t-1}, x_{2,t}, x_{2,t-1})'$. It is also necessary to choose a value for the weighting matrix. We use two common choices $(T^{-1}\sum_{t=1}^{T} z_t z_t')^{-1}$ and cI_5 where c is a constant that is discussed below.

Hansen and Singleton (1982) estimate the model using monthly U.S. data for the period 1959:2–1978:12, but we take advantage of the march of time to use an extended sample covering 1959:1–1997:12. Once allowance is made for the two conditioning observations needed to construct z_t, this leaves a sample of size $T = 465$. The consumption of the representative agent in period t, c_t, is defined to be aggregate real consumption of nondurables and services in period t divided by total population in period t. Both consumption and population series are compiled by the U.S. Department of Commerce, and obtained from the *FRED* database constructed by the Federal Reserve Bank of St. Louis. The consumption figures are seasonally adjusted and expressed in billions of chained 1992 dollars. The nominal return on the assets is obtained from the CRSP tapes, and transformed into a real return using the implicit deflator associated

with the measure of consumption. Specifically, this gives

$$x_{2,t+1} = (1 + \textit{nominal return at time } t+1)\frac{\textit{deflator at time } t}{\textit{deflator at time } t+1}$$

where the *deflator at time t* is the ratio of aggregate real consumption of non-durables and services at time t to its nominal counterpart in period t. The latter is also seasonally adjusted and has the same source as the real data.

The estimations are performed by minimizing $TQ_T(\theta)$ using routines in the MATLAB version 6.0 Optimization Toolbox (Mathworks, 2000). This package provides a number of optimization procedures. All our estimations employ the procedure *fminu* which is a variant of the gradient method described above.[15] This estimation routine allows the researcher to specify constants which control the convergence criterion for the parameters and the minimand. In our estimation these two numbers are set equal and denoted ϵ_M. To illustrate their impact on the results, we perform the estimations using $\epsilon_M = 10^{-4}, 10^{-5}$ and 10^{-6}.

We begin with the estimation of the model for EWR. The results are presented in Table 3.1. Consider first the results for the case in which $W_T = 10^5 I_5$. The scaling factor of 10^5 is included because if $W_T = I_5$ then the value of the minimand is of the order 10^{-5} for parts of the parameter space and this made it difficult for *fminu* to find the minimum.[16] Even with this scaling, the minimand appears ill behaved. When $\epsilon_M = 10^{-4}$, all four starting values do not initiate procedures which converge to the same point. This behaviour could arise for two reasons. First, the minimand may have a well-defined local minimum at each of the two points to which the algorithm converged. In this case the parameters are locally identified at each point but obviously not globally identified. Secondly, the convergence criterion may be insufficiently tight and the iterative procedure is stopping before it reaches a local minimum. To assess which is the case here, we re–estimate with $\epsilon_M = 10^{-5}$ and then with $\epsilon_M = 10^{-6}$. As can be seen, this refinement causes the iterative procedure to converge to the same point for all the starting values. This diagnosis is confirmed by a plot of the minimands. Figures 3.1 contains a plot the minimand for the case in which $W_T = 10^5 I_5$. As can be seen, $Q_T(\theta)$ is very flat in the dimension of γ.

[15] See Section 9.1 for an empirical example in which this method does not work well and so an alternative routine is employed.

[16] This is an example of the problem noted above. The value of the objective function was of a lower order of magnitude than the convergence criteria.

Table 3.1
First step estimation results for the consumption-based asset pricing model with equally weighted returns

$W_T = 10^5 I_5$: *Starting values*	ϵ_M	$(\hat{\gamma}, \hat{\delta})$	$TQ_T(\hat{\theta})$
(0.5, 0.5)	$10^{-4}, 10^{-5}, 10^{-6}$	(-3.145, 0.999)	5.974
(-0.5, -0.5)	10^{-4}	(-0.334, 0.994)	6.064
	$10^{-5}, 10^{-6}$	(-3.145, 0.999)	5.974
(5.5, 5.5)	$10^{-4}, 10^{-5}, 10^{-6}$	(-3.145, 0.999)	5.974
(-5.5, -5.5)	$10^{-4}, 10^{-5}, 10^{-6}$	(-3.145, 0.999)	5.974
$W_T = (T^{-1}\sum_{t=1}^{T} z_t z_t')^{-1}$: *Starting values*	ϵ_M	$(\hat{\gamma}, \hat{\delta})$	$TQ_T(\hat{\theta})$
(0.5, 0.5)	10^{-4}	(0.500, 0.993)	0.031
	$10^{-5}, 10^{-6}$	(0.398, 0.993)	0.031
(-0.5,-0.5)	$10^{-4}, 10^{-5}, 10^{-6}$	(0.398, 0.993)	0.031
(5.5, 5.5)	$10^{-4}, 10^{-5}, 10^{-6}$	(0.398, 0.993)	0.031
(-5.5,-5.5)	$10^{-4}, 10^{-5}, 10^{-6}$	(0.398, 0.993)	0.031

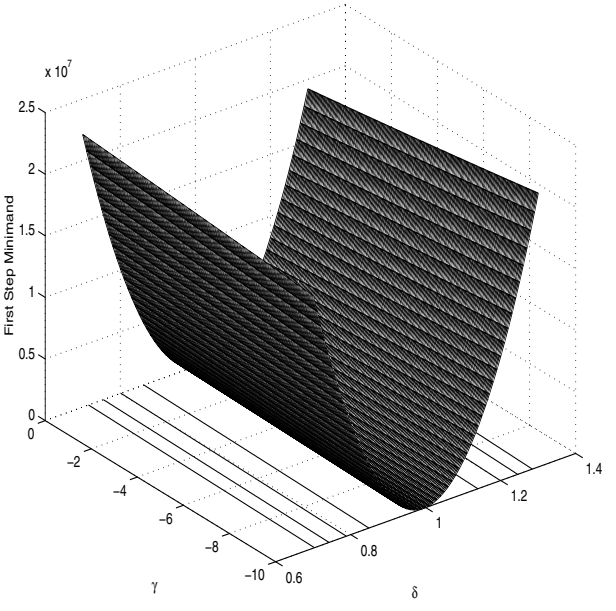

Figure 3.1: Minimand with $W_T = 10^5 I_5$ for the consumption-based asset pricing model with equally weighted returns

A similar problem emerges when $W_T = (T^{-1}\sum_{t=1}^{T} z_t z_t')^{-1}$, but again it disappears when the convergence criterion is tightened. The shape of the minimand is qualitatively similar to that in Figure 3.1 and so the plot is omitted. Although, we have convergence for each choice of weighting matrix, the parameter estimates are clearly very sensitive to this choice. In one case the estimated relative risk aversion of the representative agent $(1 - \hat{\gamma})$ is 0.602 and in the other it is 4.145. This discrepancy illustrates the motivation for estimation with the optimal weighting matrix. However, we must delay a presentation of those results until Section 3.6.

We now consider the estimation of the model with VWR. The results are presented in Table 3.2. From Table 3.2, it is clear that the same problems are encountered as before with $W_T = 10^5 I_5$. It can be seen from Figure 3.2 that the minimand has qualitatively the same shape with VWR as it did with EWR. Once again, the results are sensitive to the choice of weighting matrix.

Table 3.2
First step estimation results for the consumption-based asset pricing model with value weighted returns

$W_T = 10^5 I_5$: *Starting values*	ϵ_M	$(\hat{\gamma}, \hat{\delta})$	$TQ_T(\hat{\theta})$
(0.5, 0.5)	10^{-4}	(0.503, 0.994)	0.388
	$10^{-5}, 10^{-6}$	(-1.871, 0.998)	0.338
(-0.5,-0.5)	$10^{-4}, 10^{-5}$	(-0.348, 0.996)	0.359
	10^{-6}	(-1.871, 0.998)	0.338
(5.5, 5.5)	$10^{-4}, 10^{-5}, 10^{-6}$	(-1.871, 0.998)	0.338
(-5.5,-5.5)	$10^{-4}, 10^{-5}, 10^{-6}$	(-1.871, 0.998)	0.338
$W_T = (T^{-1}\sum_{t=1}^{T} z_t z_t')^{-1}$: *Starting values*	ϵ_M	$(\hat{\gamma}, \hat{\delta})$	$TQ_T(\hat{\theta})$
all *	$10^{-4}, 10^{-5}, 10^{-6}$	(0.698, 0.994)	0.003

Notes: * all = (0.5, 0.5), (-0.5,-0.5), (5.5,5.5), (-5.5,-5.5)

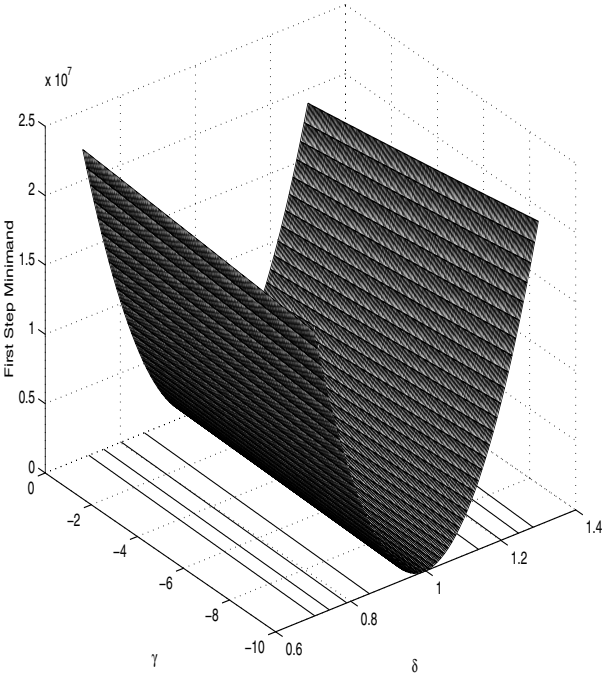

Figure 3.2: Minimand with $W_T = 10^5 I_5$ for the consumption-based asset pricing model with value weighted returns

3.3 The Identifying and Overidentifying Restrictions

The definition of the GMM estimator in (3.11) does not require $f(.)$ to be differentiable with respect to θ. In some cases this generality is useful, but it is unnecessary in nearly all the models in Table 1.1. When $f(.)$ is differentiable then the estimator can be defined equivalently as the solution to the first order equations in (3.12). This might appear a minor difference but it is important because it facilitates a Method of Moments interpretation for GMM. Just as in the linear model, this interpretation leads to a decomposition of the population moment condition into identifying and overidentifying restrictions. As shown in Chapter 2, this decomposition can be very useful for understanding the properties of GMM and it also plays an important role in the construction of diagnostics for the adequacy of the model specification. Similar dividends are reaped in the nonlinear model and so now we extend this decomposition to any models which satisfy the differentiablity conditions of Assumption 3.5.

An inspection of (3.12) reveals that the GMM estimator based on $E[f(v_t, \theta_0)]= 0$ can be interpreted as a Method of Moments estimator based on

$$F(\theta_0)'W^{1/2}E[f(v_t, \theta_0)] = 0 \tag{3.18}$$

where $F(\theta_0) = W^{1/2}E[\partial f(v_t, \theta_0)/\partial\theta']$. Equation (3.18) states that $W^{1/2}E[f(v_t, \theta_0)]$ lies in the null space of $F(\theta_0)'$, and implies $rank\{F(\theta_0)\}$ linear combinations of the transformed moment condition are set to zero. Assumption 3.6 guarantees this rank equals p and so, as in the linear model, the Method of Moments interpretation emphasizes the fundamental connection between identification and estimation. However, this time there is a slight difference. In the linear model, the concepts of local and global identification are identical but this is not the case in nonlinear models as seen in Section 3.1. The form of (3.18) indicates that it is the local version which is important here. The p parameters are only *locally* identified if the estimation is based on p linearly independent equations. The nature of this connection coincides with the our earlier definitions of the two types of identification. Local identification implies the population moment condition is satisfied uniquely at θ_0 in a suitably defined neighbourhood. In this case, (3.18) has a well-defined solution at θ_0. However, there may be other points in the parameter space at which (3.18) has well-defined solutions – this eventuality is only ruled out if θ_0 is globally identified.

If $p = q$ then (3.18) is equivalent to $E[f(v_t, \theta_0)] = 0$, and we note parenthetically that this means the weighting matrix plays no role in the analysis. However, if $q > p$ then there is a difference between information used in estimation and the original population moment condition. Since (3.18) is essentially the same structure as (2.10), we can repeat the same arguments here to show the population moment condition can be decomposed into identifying and overidentifying restrictions associated with GMM estimation. The *identifying restrictions* are[17]

$$F(\theta_0)[F(\theta_0)'F(\theta_0)]^{-1}F(\theta_0)'W^{1/2}E[f(v_t, \theta_0)] = 0 \tag{3.19}$$

These restrictions characterize the part of the transformed population moment condition used in estimation. Formally, (3.19) states that the least squares projection of $W^{1/2}E[f(v_t, \theta_0)]$ onto the column space of $F(\theta_0)$ is zero, and thereby places $rank\{F(\theta_0)[F(\theta_0)'F(\theta_0)]^{-1}F(\theta_0)'\} = p$ restrictions on the transformed population moment condition. The *overidentifying restrictions* represent the remainder and so by definition are[18]

$$\{I_q - F(\theta_0)[F(\theta_0)'F(\theta_0)]^{-1}F(\theta_0)'\}W^{1/2}E[f(v_t, \theta_0)] = 0 \tag{3.20}$$

Equation (3.20) states that the projection of $W^{1/2}E[f(v_t, \theta_0)]$ on to the orthogonal complement of $F(\theta_0)$ is zero, and thereby places $q - p$ restrictions on the transformed population moment condition. Notice that the identifying and overidentifying matrices have the same projection matrix structure encountered in the linear model, and so are orthogonal in nonlinear models as well.

The roles of the two sets of restrictions are reflected in their sample counterparts. Since the identifying restrictions represent the information used in estimation, their sample analogs are satisfied at $\hat{\theta}_T$ by construction. In contrast, the

[17] This terminology is introduced by Sowell (1996) who first characterized the identifying restrictions.

[18] This terminology is introduced by Hansen (1982) who first characterized the overidentifying restrictions in this context.

overidentifying restrictions are ignored in estimation and so their sample analog is not satisfied. However, they can be used to give a useful interpretation to the GMM minimand. From (3.12), it follows that

$$W_T^{1/2}T^{-1}\sum_{t=1}^{T} f(v_t,\hat{\theta}_T) = \{I_q - F_T(\hat{\theta}_T)[F_T(\hat{\theta}_T)'F_T(\hat{\theta}_T)]^{-1}F_T(\hat{\theta}_T)'\} \times W_T^{1/2}T^{-1}\sum_{t=1}^{T} f(v_t,\hat{\theta}_T) \quad (3.21)$$

where $F_T(\theta) = W_T^{1/2}T^{-1}\sum_{t=1}^{T}\partial f(v_t,\theta)/\partial\theta'$, and so the transformed estimated sample moment is the sample analog to the function of the data appearing in the overidentifying restrictions.[19] Therefore, $Q_T(\hat{\theta}_T)$ can be interpreted as a measure of how far the sample is from satisfying the overidentifying restrictions.

3.4 Asymptotic Properties

In the linear model, the asymptotic analysis rested crucially on a closed form expression for $\hat{\theta}_T$. However, as discussed in Section 3.2, such a representation typically does not exist in nonlinear models and so it is necessary to develop a different strategy of proof. As it turns out, the difference is most marked in the proof of consistency. Once consistency is established then it is possible to invoke the Mean Value Theorem to obtain a representation for $\hat{\theta}_T - \theta_0$ which facilitates the derivation of asymptotic normality along very similar lines to the argument used in the linear model. Hansen (1982) establishes these properties in his original article. Newey and McFadden (1994) and Wooldridge (1994) provide very useful treatments of the asymptotic analysis of a wide variety of econometric estimators. Our discussion takes advantage of their results and the reader is refered to these sources for some of the more technical details.

Before developing the asymptotic analysis it is necessary to place a further restriction on v_t. Recall from Section 1.4.2 that stationarity, by itself, is insufficient to allow the application of Laws of Large Numbers and Central Limit Theorem. Therefore we now impose the following.

Assumption 3.8 Ergodicity
The random process $\{v_t; -\infty < t < \infty\}$ *is ergodic.*

A formal definition of ergodicity involves rather sophisticated mathematical ideas and is beyond the scope of this book. Instead we refer the interested reader to Davidson (1994) [pp.199–203] or Spanos (1999) [pp.424–6]. It is sufficient for ergodicity that the dependence between v_t and v_{t-m} decreases at a certain rate to zero as $m \to \infty$. If v_t exhibits this behaviour then it is called a *mixing* process. This type of assumption has received a lot of attention in the econometrics literature because it can be used to underpin asymptotic analysis

[19] This assumes W_T is positive definite.

in either stationary or nonstationary environments, and so is more general than ergodicity which can only be used for stationary series. Further discussion of these issues here would constitute a major detour and would distract us from the main purpose of this chapter. Therefore, we provide a heuristic introduction to mixing processes in Appendix A. This appendix also contains a brief summary of the literature on GMM in a nonstationary environment.

3.4.1 Consistency of the Parameter Estimator

Even though there is no closed form expression for $\hat{\theta}_T$, it is clearly defined by (3.11). The key to a proof of consistency is the consideration of what happens if we perform a similar minimization on the population analog to $Q_T(\theta)$,

$$Q_0(\theta) = \{E[f(v_t, \theta)]\}' \, W \, \{E[f(v_t, \theta)]\} \tag{3.22}$$

The answer follows directly from our earlier assumptions. The population moment condition implies $Q_0(\theta_0) = 0$. The global identification condition and the positive definiteness of W, imply $Q_0(\theta) > 0$ for all $\theta \neq \theta_0$. Taken together these two properties imply $Q_0(\theta)$ has a unique minimum at $\theta = \theta_0$. Intuition suggests that if: (i) $\hat{\theta}_T$ minimizes $Q_T(\theta)$; and (ii) $Q_T(\theta)$ converges in probability to a function, $Q_0(\theta)$, whose unique minimum is at θ_0; then $\hat{\theta}_T$ must converge in probability to θ_0. In essence this intuition is correct but there is one mathematical detail which needs to be taken into account. It is not necessarily the case that the minimum of a sequence of functions converges to the minimum of the limit of the sequence of functions. For this to be the case, it is sufficient that $Q_T(\theta)$ converges *uniformly* to $Q_0(\theta)$.[20] This property is not guaranteed by Assumptions 3.1–3.8 and we must impose the following two additional restrictions.

Assumption 3.9 Compactness of Θ
Θ *is a compact set.*

This compactness assumption strictly requires the knowledge of bounds on θ_0 which is typically unavailable. However, this is often ignored in practice because these bounds can be assumed to be sufficiently large not to impact on the construction of the estimator.[21] The only other additional assumption is the requirement that $f(v_t, \theta)$ is bounded by a function with finite expectation for all θ.

Assumption 3.10 Domination of $f(v_t, \theta)$
$E[sup_{\theta \in \Theta} ||f(v_t, \theta)||] < \infty$

With these assumptions imposed, it is possible to deduce uniform convergence.[22]

[20] This property is not guaranteed by pointwise convergence of $Q_T(\theta)$. See Apostol (1974) [Chapter 9] for a useful discussion of the difference between pointwise and uniform convergence.

[21] Recall that a compact set is closed and bounded; see Apostol (1974) [Chapter 3]. Newey and McFadden (1994) discuss the potential for proving consistency without the imposition of compactness. Also see Pötscher and Prucha (1997) [Chapters 3 and 4].

[22] For example, see Newey and McFadden (1994) [Theorem 2.6], Wooldridge (1994) [Theorem 4.1] and the references therein.

Lemma 3.1 Uniform Convergence in Probability of $Q_T(\theta)$
If Assumptions 3.1, 3.2, 3.7-3.10 hold then $\sup_{\theta\in\Theta} |Q_T(\theta) - Q_0(\theta)| \xrightarrow{p} 0$.

Once uniform convergence is guaranteed, then consistency can be established.

Theorem 3.1 Consistency of the Parameter Estimator
If Assumptions 3.1–3.4 and 3.7–3.10 hold then $\hat{\theta}_T \xrightarrow{p} \theta_0$.

For completeness we now provide a more formal proof of this theorem. It is most convenient to break the proof down into two parts. First, it is shown that the conditions of the theorem imply:

$$\lim_{T\to\infty} P[\, 0 \leq Q_0(\hat{\theta}_T) < \epsilon \,] = 1 \text{ for any } \epsilon > 0 \tag{3.23}$$

This equation states that $\hat{\theta}_T$ minimizes $Q_0(\theta)$ with probability one as $T \to \infty$. The second part of the proof shows formally that this property implies consistency.

Part (i): Proof of (3.23).

This result is deduced from the following three statements about $Q_T(.)$ and $Q_0(.)$ implied by uniform convergence and the definition of the estimator.

(a): Lemma 3.1 states that the difference between $Q_T(\theta)$ and $Q_0(\theta)$ disappears with probability one as $T \to \infty$ at any value of $\theta \in \Theta$. Now, by definition $\hat{\theta}_T \in \Theta$, and so Lemma 3.1 implies $\lim_{T\to\infty} P[|\, Q_0(\hat{\theta}_T) - Q_T(\hat{\theta}_T)\,| < \epsilon/3] = 1$ for any constant $\epsilon > 0$.[23] This implies in turn that

$$\lim_{T\to\infty} P[Q_0(\hat{\theta}_T) < Q_T(\hat{\theta}_T) + \epsilon/3] = 1 \,.$$

(b): Since $\hat{\theta}_T$ minimizes $Q_T(\theta)$ it follows that

$$\lim_{T\to\infty} P[Q_T(\hat{\theta}_T) < Q_T(\theta_0) + \epsilon/3] = 1 \,.$$

(c): By similar reasoning to part (a), it follows that

$$\lim_{T\to\infty} P[Q_T(\theta_0) < Q_0(\theta_0) + \epsilon/3] = 1 \,.$$

A combination of the probability statements in (a) and (b) yields

$$\lim_{T\to\infty} P[Q_0(\hat{\theta}_T) < Q_T(\theta_0) + 2\epsilon/3] = 1$$

and this statement can be combined with (c) to deduce

$$\lim_{T\to\infty} P[Q_0(\hat{\theta}_T) < Q_0(\theta_0) + \epsilon] = 1$$

[23] The division of ϵ by three is for notational convenience below and has no substantive impact on the argument.

Equation (3.23) then follows immediately because Assumption 3.3 implies $Q_0(\theta_0) = 0$ and the positive definiteness of W implies $Q_0(\theta) \geq 0$.

Part (ii): (3.23) $\Rightarrow \hat{\theta}_T \xrightarrow{p} \theta_0$.

Let $\mathbf{N}$ be an open subset of Θ which contains θ_0 and $\mathbf{N}^c$ be the complement of $\mathbf{N}$ relative to Θ. By definition $\mathbf{N}^c$ is a closed subset of a compact set and so is itself compact.[24] Since $\mathbf{N}^c$ is compact and $Q_0(\theta)$ is a continuous function it follows that $Q_0(\theta)$ has an infimum on $\mathbf{N}^c$, which we denote by $\inf_{\theta \in \mathbf{N}^c} Q_0(\theta)$. From Assumption 3.4, it follows that this infimum is strictly positive. Therefore we can substitute $\epsilon = \inf_{\theta \in \mathbf{N}^c} Q_0(\theta)$ in (3.23) to deduce

$$\lim_{T \to \infty} P[Q_0(\hat{\theta}_T) < \inf_{\theta \in \mathbf{N}^c} Q_0(\theta)] = 1$$

This implies $\lim_{T\to\infty} P[\hat{\theta}_T \notin \mathbf{N}^c] = 1$ and hence that $\lim_{T\to\infty} P[\hat{\theta}_T \in \mathbf{N}] = 1$. Finally, since the above argument holds for any choice of $\mathbf{N}$ no matter how "small", it must follow that $\lim_{T\to\infty} P[\hat{\theta}_T = \theta_0] = 1$ which is the desired result. ◇

Notice that the conditions for Theorem 3.1 placed no restrictions on the derivative matrix $\partial f(v_t, \theta)/\partial\theta'$. It is true that we have refered to this derivative matrix in previous sections but its role has not been crucial. It was used to obtain a condition for local identification in models which satisfied Assumption 3.5; however the concept of global identification did not require its existence. The derivative matrix also played a role in the discussion of numerical optimization. However, as mentioned above, $Q_T(\theta)$ can be minimized by search methods which do not require the calculation of the gradient. As we shall see in the next sub-section, the derivative matrix plays a more central role in the proof of asymptotic normality of the estimator.

3.4.2 Asymptotic Normality of the Parameter Estimator

To develop the asymptotic distribution of the estimator, we require an asymptotically valid closed form representation for $T^{1/2}(\hat{\theta}_T - \theta_0)$. This representation comes from an application of the Mean Value Theorem.[25] This theorem relates $f(.)$ to its first derivatives $\partial f(v_t, \theta)/\partial\theta'$ and so it is necessary to impose Assumption 3.5.[26] To simplify the presentation, define $g_T(\theta) = T^{-1}\sum_{t=1}^{T} f(v_t, \theta)$ and $G_T(\theta) = T^{-1}\sum_{t=1}^{T} \partial f(v_t, \theta)/\partial\theta'$. The Mean Value Theorem implies that

$$g_T(\hat{\theta}_T) = g_T(\theta_0) + G_T(\hat{\theta}_T, \theta_0, \lambda_T)(\hat{\theta}_T - \theta_0) \tag{3.24}$$

where $G_T(\hat{\theta}_T, \theta_0, \lambda_T)$ is the $(q \times p)$ matrix whose i^{th} row is the corresponding row of $G_T(\bar{\theta}_T^{(i)})$ where $\bar{\theta}_T^{(i)} = \lambda_{T,i}\theta_0 + (1 - \lambda_{T,i})\hat{\theta}_T$ for some $0 \leq \lambda_{T,i} \leq 1$, and

[24] See Apostol (1974) [pp.50–3].
[25] See Apostol (1974) [p.355].
[26] Similar results can be developed for non-differentiable $f(v_t, \theta)$ in cases where $E[f(v_t, \theta)]$ is differentiable; see Newey and McFadden (1994) [Section 7].

λ_T is the $(q \times 1)$ vector with i^{th} element $\lambda_{T,i}$. Premultiplication of (3.24) by $G_T(\hat{\theta}_T)'W_T$ yields

$$G_T(\hat{\theta}_T)'W_T g_T(\hat{\theta}_T) = G_T(\hat{\theta}_T)'W_T g_T(\theta_0) + G_T(\hat{\theta}_T)'W_T G_T(\hat{\theta}_T, \theta_0, \lambda_T)(\hat{\theta}_T - \theta_0) \tag{3.25}$$

Now the first order conditions in (3.12) imply the left hand side of (3.25) is zero and so with some rearrangement it follows from (3.25) that

$$\begin{aligned} T^{1/2}(\hat{\theta}_T - \theta_0) &= -[G_T(\hat{\theta}_T)'W_T G_T(\hat{\theta}_T, \theta_0, \lambda_T)]^{-1} G_T(\hat{\theta}_T)'W_T T^{1/2} g_T(\theta_0) \\ &= -\bar{M}_T T^{1/2} g_T(\theta_0), \qquad \text{say.} \end{aligned} \tag{3.26}$$

Notice that this equation has the same basic structure as arose in the linear model at this stage: a random matrix, $-\bar{M}_T$, times a random vector, $T^{1/2}g_T(\theta_0)$. Just as in Section 2.3, we start by analyzing the limiting behaviour of these two components separately and then combine them to deduce the asymptotic distribution of the estimator. The asymptotic behaviour of $T^{1/2}g_T(\theta_0)$ is given by a version of the Central Limit Theorem. To apply the Central Limit Theorem, it is necessary to assume the second moment matrices of the sample moment satisfy certain restrictions.[27]

Assumption 3.11 Properties of the Variance of the Sample Moment
(i) $E[f(v_t, \theta_0)f(v_t, \theta_0)']$ *exists and is finite; (ii)* $lim_{T\to\infty} Var[T^{1/2}g_T(\theta_0)] = S$ *exists and is a finite valued positive definite matrix.*

The Central Limit Theorem is as follows.

Lemma 3.2 Central Limit Theorem for $T^{1/2}g_T(\theta_0)$
If Assumptions 3.1, 3.3, 3.8 and 3.11 hold then $T^{1/2}g_T(\theta_0) \xrightarrow{d} N(0, S)$.

The analysis of $\bar{M}_T$ is more complicated than in the linear model because it depends on $G_T(\hat{\theta}_T)$ and $G_T(\hat{\theta}_T, \theta_0, \lambda_T)$. Since $\hat{\theta}_T \xrightarrow{p} \theta_0$ and $\bar{\theta}_T^{(i)}$ lies on the line segment between $\hat{\theta}_T$ and θ_0, then it follows that $\bar{\theta}_T^{(i)} \xrightarrow{p} \theta_0$ for $i = 1, 2 \ldots p$. Intuition suggests that this should imply both $G_T(\hat{\theta}_T)$ and $G_T(\hat{\theta}_T, \theta_0, \lambda_T)$ converge in probability to $G_0 = E[\partial f(v_t, \theta_0)/\partial\theta']$. In essence this is correct, but the argument can only be formally justified if we impose two further restrictions on $\partial f(v_t, \theta)/\partial\theta'$.[28]

Assumption 3.12 Continuity of $E[\partial f(v_t, \theta)/\partial\theta']$
$E[\partial f(v_t, \theta)/\partial\theta']$ *is continuous on some neighbourhood* N_ϵ *of* θ_0.

Assumption 3.13 Uniform Convergence of $G_T(\theta)$
$\sup_{\theta\in N_\epsilon} ||G_T(\theta) - E[\partial f(v_t, \theta)/\partial\theta']|| \xrightarrow{p} 0$.[29]

[27] See Hansen (1982) for more primitive conditions for such an S to exist. We do not give these conditions here because they are superseded in the next section by the more restrictive conditions under which S can be consistently estimated.

[28] See Newey and McFadden (1994)[p.2145].

[29] For any matrix A, we define $\|A\| = [tr(A'A)]^{1/2}$.

With these assumptions imposed – and, of course, the conditions for the consistency of $\hat{\theta}_T$ – it is possible to deduce the following.

Lemma 3.3 Convergence of $G_T(\hat{\theta}_T)$ and $G_T(\hat{\theta}_T, \theta_0, \lambda_T)$
If Assumptions 3.1–3.5, 3.7–3.10, 3.12 and 3.13 hold then $G_T(\hat{\theta}_T) \xrightarrow{p} G_0$ and $G_T(\hat{\theta}_T, \theta_0, \lambda_T) \xrightarrow{p} G_0$.

Lemma 3.3 can be combined with Assumption 3.7 and Slutsky's Theorem to deduce that $\bar{M}_T \xrightarrow{p} (G_0' W G_0)^{-1} G_0' W$. Therefore just as in the linear model, $T^{1/2}(\hat{\theta}_T - \theta_0)$ is asymptotically the product of a random matrix which converges in probability to a constant, and a random vector which converges to a normal distribution. Therefore, the desired result follows once again from Lemma 1.4.

Theorem 3.2 Asymptotic Normality of the Parameter Estimator
If Assumptions 3.1–3.5 and 3.7–3.13 hold[30] *then: $T^{1/2}(\hat{\theta}_T - \theta_0) \xrightarrow{d} N(0, MSM')$ where $M = (G_0' W G_0)^{-1} G_0' W$.*

Theorem 3.2 implies that an approximate $100(1-\alpha)\%$ confidence interval for $\theta_{0,i}$ in large samples is given by

$$\hat{\theta}_{T,i} \pm z_{\alpha/2} \sqrt{\hat{V}_{T,ii}/T} \tag{3.27}$$

where $\hat{V}_{T,ii}$ is the $i - i^{th}$ element of a consistent estimator of MSM'. As in the linear model, a natural candidate is based on consistent estimators of the component matrices M and S. Notice that this time the matrix $\bar{M}_T$ cannot be used because although consistent, the values of $\{\bar{\theta}_T^{(i)}, i = 1, 2 \ldots p\}$ are unknown. However this problem is easily circumvented by replacing $\bar{\theta}_T^{(i)}$ with $\hat{\theta}_T$ and using $\hat{M}_T = [G_T(\hat{\theta}_T)' W_T G_T(\hat{\theta}_T)]^{-1} G_T(\hat{\theta}_T)' W_T$ to estimate M. However, the consistent estimation of S is more complicated and is the topic of Section 3.5.

As we have seen, Theorem 3.2 rests on an application of the Mean Value Theorem. The latter can only be applied if θ_0 is an interior point of Θ. It should be noted that if θ_0 is on the boundary then the limiting distribution theory is different. Since this situation is not common, we do not pursue it further here but refer the interested reader to Andrews (2002*a*).

To conclude this sub-section, we briefly return to the decomposition of the population moment condition into identifying and overidentifying restrictions. In Section 3.3, these components are defined and their role explained, but no intuition is offered for why they take these particular forms. It is now possible to remedy this omission because an intuition can be developed from the relationships used to deduce the asymptotic distribution.

The derivation of asymptotic normality began with (3.24). This equation is formally justified from the Mean Value Theorem and holds for any $\hat{\theta}_T$. However, an inspection of the subsequent analysis indicates that we would have

[30] Assumption 3.6 is only omitted because it is implied by Assumption 3.5.

obtained the same asymptotic distribution if instead we had confined attention to a sufficiently small neighbourhood around θ_0 for which

$$T^{1/2}g_T(\theta) = T^{1/2}g_T(\theta_0) + G_T(\theta_0)T^{1/2}(\theta - \theta_0) \tag{3.28}$$

In other words, for the purposes of the asymptotic distribution theory it is sufficient to concentrate on the behaviour of the sample moment in the neighbourhood of θ_0 for which $T^{1/2}g_T(\theta)$ is a linear function of $T^{1/2}(\theta-\theta_0)$. If we concentrate on this neighbourhood for the analysis of the minimization of $TQ_T(\theta)$ as well, then the identifying restrictions emerge naturally from the structure of the problem. Using (3.28), the GMM minimand in this neighbourhood can be rewritten as

$$TQ_T(\theta) = ||W_T^{1/2}T^{1/2}g_T(\theta)||^2 = ||W_T^{1/2}T^{1/2}g_T(\theta_0) + F_T(\theta_0)T^{1/2}(\theta - \theta_0)||^2 \tag{3.29}$$

where, as before, $F_T(\theta) = W_T^{1/2}T^{-1}\sum_{t=1}^{T}\partial f(v_t,\theta)/\partial\theta'$. Therefore if $\hat{\theta}_T$ minimizes $Q_T(\theta)$ in this neighbourhood then $T^{1/2}(\hat{\theta}_T - \theta_0)$ must also be the least squares solution to

$$W_T^{1/2}T^{1/2}g_T(\theta_0) + F_T(\theta_0)T^{1/2}(\theta - \theta_0) = 0 \tag{3.30}$$

The least squares solution to the inconsistent set of equations in (3.30) is found by solving the consistent set of equations[31]

$$P_T(\theta_0)W_T^{1/2}T^{1/2}g_T(\theta_0) + F_T(\theta_0)T^{1/2}(\theta - \theta_0) = 0 \tag{3.31}$$

where $P_T(\theta) = F_T(\theta)[F_T(\theta)'F_T(\theta)]^{-1}F_T(\theta)'$. Since the properties of the projection matrix and (3.28) in turn imply

$$\begin{aligned} P_T(\theta_0)W_T^{1/2}T^{1/2}g_T(\theta_0) + F_T(\theta_0)T^{1/2}(\theta - \theta_0) &= P_T(\theta_0)\{W_T^{1/2}T^{1/2}g_T(\theta_0) \\ &\quad + F_T(\theta_0)T^{1/2}(\theta - \theta_0)\} \\ &= P_T(\theta_0)W_T^{1/2}T^{1/2}g_T(\theta) \end{aligned}$$

it follows that the least squares solution to (3.30) must also set

$$||P_T(\theta_0)W_T^{1/2}T^{1/2}g_T(\theta)||^2 \tag{3.32}$$

to zero. Equations (3.28)–(3.32) show that the identifying restrictions possess their projection matrix form because, for the purposes of asymptotic distribution theory, the estimation can be considered as being based on a linearization of the sample moment condition in the neighbourhood of θ_0. Finally, note that the least squares solution to (3.30) is

$$T^{1/2}(\hat{\theta}_T - \theta_0) = -[F_T(\theta_0)'F_T(\theta_0)]^{-1}F_T(\theta_0)'W_T^{1/2}T^{1/2}g_T(\theta_0) \tag{3.33}$$

Equation (3.33) is easily verified to be asymptotically equivalent to the formula in (3.26) from which we deduced the asymptotic normality of the estimator.

[31] For example, see Strang (1988) [p.156].

3.4.3 Asymptotic Normality of the Estimated Sample Moment

It is shown in Section 3.3 that the estimated sample moment represents a source of information about whether the overidentifying restrictions are satisfied in the population. This property is exploited elsewhere to develop a test of the hypothesis that the model is correctly specified.[32] At this stage, we confine our attention to deriving the asymptotic distribution of $W_T^{1/2}T^{1/2}g_T(\hat{\theta}_T)$ in correctly specified models.

Equation (3.24) implies

$$W_T^{1/2}T^{1/2}g_T(\hat{\theta}_T) \;=\; W_T^{1/2}T^{1/2}g_T(\theta_0) \;+\; W_T^{1/2}G_T(\hat{\theta}_T,\theta_0,\lambda_T)T^{1/2}(\hat{\theta}_T-\theta_0) \tag{3.34}$$

If we substsitute for $T^{1/2}(\hat{\theta}_T-\theta_0)$ from (3.26) then (3.34) can be written as

$$W_T^{1/2}T^{1/2}g_T(\hat{\theta}_T) \;=\; N_T(\hat{\theta}_T)W_T^{1/2}T^{1/2}g_T(\theta_0) \tag{3.35}$$

where

$$N_T(\hat{\theta}_T)=I_q \;-\; W_T^{1/2}G_T(\hat{\theta}_T,\theta_0,\lambda_T)[G_T(\hat{\theta}_T)'W_TG_T(\hat{\theta}_T,\theta_0,\lambda_T)]^{-1}G_T(\hat{\theta}_T)'W_T^{1/2'}$$

Equation (3.35) implies $W_T^{1/2}T^{1/2}g_T(\hat{\theta}_T)$ has the same generic structure as the expression for $T^{1/2}(\hat{\theta}_T-\theta_0)$ in (3.26) namely: a random matrix times the random vector, $T^{1/2}g_T(\theta_0)$. Therefore we can use the same arguments as Section 3.4.2 to deduce the following result.

Theorem 3.3 Asymptotic Normality of the Estimated Sample Moment
If Assumptions 3.1–3.5 and 3.7–3.13 hold then: $W_T^{1/2}T^{1/2}g_T(\hat{\theta}_T)\xrightarrow{d}N(0,NW^{1/2}SW^{1/2'}N')$ *where* $N=[I_q-P(\theta_0)]$ *and* $P(\theta_0)=F(\theta_0)[F(\theta_0)'F(\theta_0)]^{-1}F(\theta_0)'$.

The connection between the estimated sample moment and the overidentifying restrictions manifests itself in the asymptotic distribution. Equation (3.35) implies that

$$W_T^{1/2}T^{1/2}g_T(\hat{\theta}_T) \;=\; [I_q \;-\; P(\theta_0)]W^{1/2}T^{1/2}g_T(\theta_0) \;+\; o_p(1) \tag{3.36}$$

Inspection of (3.36) reveals that the asymptotic behaviour of the estimated sample moment is governed by the function of the data which appears in the overidentifying restrictions. Therefore, the mean of the asymptotic distribution in Theorem 3.3 is zero because the overidentifying restrictions are satisfied at θ_0. This relationship also has an impact on the properties of the variance of the limiting distribution. Since $W^{1/2}$ and S are nonsingular, it follows that[33] $rank\{NSN'\} = rank\{I_q-P(\theta_0)\} = q-p$, and so the covariance matrix is singular.[34] This rank is easily recognized to be the number of overidentifying restrictions.

[32] See Section 2.5 and Chapter 5.
[33] See Dhrymes (1984) [p.17].
[34] See Rao (1973) [Chapter 8] for a discussion of the singular normal distribution.

3.5 Long Run Covariance Matrix Estimation

So far, very little has been said about S except that it exists and is positive definite. The latter is the matrix generalization of the requirement that a scalar variance be positive. It is important that the estimator also exhibits this property or is positive semi-definite at the very least; otherwise the estimated variances of the individual coefficient estimators can be negative. This is not always such a trivial property to impose and is one aspect of the various estimators upon which we focus below.

To understand more about the structure of S, it is useful to rewrite its definition as follows,

$$\begin{aligned} S &= \lim_{T\to\infty} Var[T^{-1/2}\sum_{t=1}^{T} f_t] \\ &= \lim_{T\to\infty} E\left[\left(T^{-1/2}\sum_{t=1}^{T} f_t - E[T^{-1/2}\sum_{t=1}^{T} f_t]\right) \times \right. \\ &\qquad \left. \left(T^{-1/2}\sum_{t=1}^{T} f_t - E[T^{-1/2}\sum_{t=1}^{T} f_t]\right)'\right] \end{aligned}$$

where to simplify notation we have set $f_t = f(v_t, \theta_0)$. Since

$$T^{-1/2}\sum_{t=1}^{T} f_t - E[T^{-1/2}\sum_{t=1}^{T} f_t] \;=\; T^{-1/2}\sum_{t=1}^{T}(f_t - E[f_t])$$

it follows that

$$\begin{aligned} S &= \lim_{T\to\infty} E[\{T^{-1/2}\sum_{t=1}^{T}(f_t - E[f_t])\}\{T^{-1/2}\sum_{t=1}^{T}(f_t - E[f_t])\}'] \\ &= \lim_{T\to\infty} E[T^{-1}\sum_{t=1}^{T}\sum_{s=1}^{T}(f_t - E[f_t])(f_s - E[f_s])'] \end{aligned} \tag{3.37}$$

The stationarity assumption implies that $E[(f_t - E[f_t])(f_{t-j} - E[f_{t-j}])'] = \Gamma_j$, say, for every t and so[35]

$$S \;=\; \Gamma_0 + \lim_{T\to\infty}\{\sum_{j=1}^{T-1}\left(\frac{T-j}{T}\right)(\Gamma_j + \Gamma_j')\} = \Gamma_0 + \sum_{i=1}^{\infty}(\Gamma_i + \Gamma_i') \tag{3.38}$$

The matrix Γ_j is known as the j^{th} autocovariance matrix of f_t.[36] From (3.38) it is clear that estimation of S is going to require assumptions about these autocovariance matrices.

[35] For example, see Hamilton (1994) [pp. 279–80].

[36] See Hamilton (1994) [pp.261–2] for a discussion of the properties of autocovariance matrices.

The long run covariance matrix estimation literature has focused on the ways to avoid any potential inconsistency caused by inappropriate assumptions about the dynamic specification of $\{f(v_t, \theta_0)\}$. Therefore, nearly all the contributions to this literature develop the properties of the estimator in question under the assumption that the model is correctly specified and so $E[f(v_t, \theta_0)] = 0$. We maintain this assumption throughout the section. However, it should be noted that if this assumption is inappropriate then all the estimators discussed below are inconsistent. In other words, the consistency of a covariance matrix estimator depends on the validity of the assumptions about *both* the mean and dynamic structure of $\{f(v_t, \theta_0)\}$. It might be felt that little concern need be attached to any inconsistency caused by $E[f(v_t, \theta_0)] \neq 0$ because once it is recognized that the model is misspecified then there is typically no interest in constructing confidence intervals for θ_0. However, the use of an inconsistent covariance matrix estimator has a detrimental effect on the properties of certain tests for misspecification, and may in turn affect the properties of moment selection procedures based upon these tests.[37] This motivates the use of covariance matrix estimators which are consistent even if the model is misspecified. Fortunately, there is a simple way to modify the estimators discussed here to achieve that end. However, we delay further discussion of this topic until Section 4.3.

In this section we describe estimators which have been proposed under three different sets of assumptions about the dynamic structure of f_t. The first is where $\{f_t\}$ forms a *serially uncorrelated sequence*. This type of restriction occurs in some of the models listed in Table 1.1 and so this case is treated separately in Section 3.5.1. The remainder of the section considers the more general case in which f_t is serially correlated. Two main approaches have been taken. The first assumes that f_t is generated by a *vector autoregressive moving average (VARMA)* process and is reviewed in Section 3.5.2. This approach has the advantage that the autocovariances can be estimated straightforwardly from the parameters of the VARMA model. The potential disadvantage is that if this model for f_t is incorrect then the resulting estimator of S may be inconsistent. The second approach uses a member of the class of *heteroscedasticity and autocorrelation covariance (HAC)* matrix estimators and these are described in Section 3.5.3. These estimators are consistent under the much weaker conditions on $\{f_t\}$. Unfortunately, these more general estimators can exhibit poor finite sample performance and this prompted the construction of *prewhitened and recoloured HAC* estimators. Initial evidence suggests this latter version performs better and so it is also described in Section 3.5.3.

Our discussion of covariance matrix estimation is less rigorous than the analysis in the previous sections. Instead we focus on the intuition behind the various methods and describing both their strengths and weaknesses. All the estimators can be established to be consistent under appropriate conditions but the reproduction of these very technical results is beyond the scope of this text. Instead, we refer the interested reader to the appropriate sources for a catalogue of the required regularity conditions and rigorous proofs of the stated

[37] See Chapters 5 and 7 respectively.

results. As we shall see, there is plenty to discuss even without this more formal analysis!

3.5.1 Serially Uncorrelated Sequences

If $\{f_t\}$ is a serially uncorrelated sequence then $\Gamma_j = 0$ for $j \neq 0$ and so it follows from (3.38) that S is given by

$$S = S_{SU} = E[f_t f_t'] \tag{3.39}$$

where we have used the SU subscript to distinguish this S from the cases considered below. The form of S_{SU} is essentially the same as the S matrix in (2.27) and a similar logic leads to the estimator[38]

$$\hat{S}_{SU} = T^{-1} \sum_{t=1}^{T} \hat{f}_t \hat{f}_t' \tag{3.40}$$

where $\hat{f}_t = f(v_t, \hat{\theta}_T)$. It can be shown that $\hat{S}_{SU} \xrightarrow{p} S_{SU}$; e.g. see White (1994) [Theorem 8.27, p.193]. Notice that this estimator is positive semi-definite by construction because

$$\hat{S}_{SU} = T^{-1} H' H \tag{3.41}$$

where H is the $(T \times q)$ matrix with t^{th} row $\hat{f}_t'$.

In the types of models in Table 1.1, this type of behaviour occurs because the underlying theory implies $\{f_t\}$ is a *martingale difference sequence* with respect to the information set $\Omega_{t-1} = \{f_{t-1}, f_{t-2}, \ldots f_1\}$. Such a process satisfies both $E[f_t] = 0$ for all t and also

$$E[f_t | \Omega_{t-1}] = 0 \qquad \text{for } t = 2, 3 \ldots \tag{3.42}$$

Consequently, for $t > s$, we have $E[f_t f_s' | \Omega_{t-1}] = E[f_t | \Omega_{t-1}] f_s' = 0$ which implies

$$E[f_t f_s'] = E[E[f_t | \Omega_{t-1}] f_s'] = 0 \tag{3.43}$$

3.5.2 VARMA Processes

If f_t is generated by a stationary and invertible vector autoregressive moving average (VARMA) model of order (m,n) and $E[f_t] = 0$, then it has the following representation[39]

$$\Psi(L) f_t \; = \; \Phi(L) e_t \tag{3.44}$$

in which $\{e_t\}$ is a sequence of independently and identically distributed random vectors with $E[e_t] = 0$ and $Var[e_t] = \Sigma$. The $(q \times q)$ matrix polynomials, $\Psi(L)$

[38] Also see Section 4.3.

[39] See Hamilton (1994) [Chapters 10 and 11] for an introduction to vector time series models and Reinsel (1993) for a more elaborate discussion of VARMA models.

and $\Phi(L)$ are respectively of orders m and n; $\Psi(L)$ contains the autoregressive parameters of the system and $\Phi(L)$, the moving average parameters. The restrictions on the parameters implied by the terms "stationary and invertible" are important for our discussion and so worth a brief explanation. A VARMA process is stationary if the roots, $\{s_i,\, i = 1, 2, \ldots m\}$, of the characteristic equation $det\{\Psi(s)\} = 0$ are all outside the unit circle. This implies that f_t has a VMA(∞) representation,[40]

$$f_t = \{\Psi(L)\}^{-1}\Phi(L)e_t \tag{3.45}$$

The process is invertible if the roots, $\{s_i^*,\, i = 1, 2, \ldots n\}$, of the characteristic equation $det\{\Phi(s^*)\} = 0$ are all outside the unit circle. This implies f_t has a VAR(∞) representation,[41]

$$\{\Phi(L)\}^{-1}\Psi(L)f_t \;=\; A(L)f_t \;=\; e_t \tag{3.46}$$

where $A(L) = I_q - A_1L - A_2L^2 - \ldots$ is a $(q \times q)$ matrix polynomial of infinite order.

Now let us return to the construction of a consistent estimator for S. From (3.45) it follows that[42]

$$S = S_{VARMA} = \{\Psi(1)\}^{-1}\Phi(1)\,\Sigma\,\Phi(1)'\{\Psi(1)'\}^{-1} \tag{3.47}$$

where $\Psi(1) = I_q + \sum_{i=1}^{m}\Psi_i$ and $\Phi(1) = I_q + \sum_{i=1}^{n}\Phi_i$. This matrix can be consistently estimated by

$$\hat{S}_{VARMA} = \{\hat{\Psi}(1)\}^{-1}\hat{\Phi}(1)\hat{\Sigma}\hat{\Phi}(1)'\{\hat{\Psi}(1)'\}^{-1} \tag{3.48}$$

where $\hat{\Psi}(1) = I_q + \sum_{i=1}^{m}\hat{\Psi}_i$, $\hat{\Phi}(1) = I_q + \sum_{i=1}^{n}\hat{\Phi}_i$ and $\{\hat{\Sigma}, \hat{\Psi}_i, \hat{\Phi}_j\,; i = 1, 2, \ldots m;\; j = 1, 2, \ldots n\}$ are consistent estimators of $\{\Sigma, \Psi_i, \Phi_j\,; i = 1, 2, \ldots m;\; j = 1, 2, \ldots n\}$. Since f_t is unobserved, these parameter estimates are obtained by estimating a VARMA model for $\hat{f}_t$. The estimator of Σ is of the form $\hat{\Sigma} = T^{-1}\sum_{t=1}^{T}\hat{e}_t\hat{e}_t'$ and so $\hat{S}_{VARMA}$ is positive semi-definite by construction. The estimation of VARMA models can be performed using generalized least squares or maximum likelihood; see Reinsel (1993) [Chapter 5]. However, it is computationally burdensome due to the presence of the MA terms. Various methods have been proposed for circumventing this problem in the context of covariance matrix estimation. Eichenbaum, Hansen, and Singleton (1988) and West (1997) suggest methods which can be employed if f_t follows a VARMA(0,n) process. Although, the absence of the autoregressive component can be justified in some of the models listed in Table 1.1, we do not review these procedures here. Instead, we focus on a more general method proposed by den Haan and Levin (1996) which can be applied when f_t follows a $VARMA(m, n)$ process.[43]

[40] See Hamilton (1994) [pp. 259–61].
[41] Ibid. [p. 263].
[42] Ibid. [pp. 276–84].
[43] Also see den Haan and Levin (1997).

To motivate den Haan and Levin's method, it is useful to rewrite S_{VARMA} in terms of the coefficients in the VAR(∞) representation. From (3.46) it follows that

$$S_{VARMA} = \{A(1)\}^{-1} \Sigma \{A(1)'\}^{-1} \tag{3.49}$$

This suggests an alternative approach is to estimate S using the coefficients from the VAR(∞) representation and thereby avoid the computational problems associated with the estimation of MA terms. There is just one snag: it is impossible to estimate an infinite order autoregressive model from a finite sample. To circumvent this problem, den Haan and Levin (1996) propose approximating (3.46) by a finite order VAR model whose order increases with the sample size. To implement this method in practice, it is necessary to choose the order of this approximation. Den Haan and Levin (1996) recommend this choice is made via a data-based model selection criterion. Specifically, they propose the following method for the estimation of S_{VARMA}.[44]

Den Haan and Levin's Method

1. *Calculate* $\hat{\Sigma}(0) = T^{-1} \sum_{t=1}^{T} \hat{f}_t \hat{f}_t'$.

2. *Estimate the model*

$$\hat{f}_t = A_1(k)\hat{f}_{t-1} + \ldots + A_k(k)\hat{f}_{t-k} + e_t(k) \tag{3.50}$$

for $k = 1, 2, \ldots K$ *and* $t = K+1, K+2 \ldots T$ *by least squares where* $\hat{f}_t = f(v_t, \hat{\theta}_T)$. *These estimates are given by*

$$\hat{A}(k) = \sum_{t=K+1}^{T} \hat{f}_t r_t' \{ \sum_{t=K+1}^{T} r_t r_t' \}^{-1}$$

where $A(k) = (A_1(k), A_2(k), \ldots A_k(k))$ *and* $r_t' = (\hat{f}_{t-1}', \hat{f}_{t-2}', \ldots \hat{f}_{t-k}')$. *Construct the forecast error* $\hat{e}_t(k) = \hat{f}_t - \hat{A}(k) r_t$ *and* $\hat{\Sigma}(k) = T^{-1} \sum_{t=K+1}^{T} \hat{e}_t(k) \hat{e}_t(k)'$.

3. *Let* $\hat{k}$ *be the value of* k *which minimizes Schwarz's (1978) information criterion*

$$SIC(k) = log\{det[\hat{\Sigma}(k)]\} + \frac{log(T) k q^2}{T} \tag{3.51}$$

over $k = 0, 1, \ldots K$.

4. *Estimate* S_{VARMA} *by*

$$\hat{S}_{VARMA} = \{I_q - \sum_{i=1}^{\hat{k}} \hat{A}_i(\hat{k})\}^{-1} \hat{\Sigma}(\hat{k}) \{I_q - \sum_{i=1}^{\hat{k}} \hat{A}_i(\hat{k})\}^{-1'} \tag{3.52}$$

[44] Also see Section 4.2

To implement this method, it is necessary to choose K. Den Haan and Levin (1996) show $\hat{S}_{VARMA}$ is consistent provided $K \to \infty$ as $T \to \infty$ and $K = O(T^{1/3})$ but an appropriate rule for picking K in finite samples remains an open question.[45] This choice has the advantage the lag selection procedure is consistent because if $n > 0$ then $\hat{k}$ tends in probability to ∞ as $T \to \infty$, but if $n = 0$ then $\hat{k} \xrightarrow{p} m$.[46] Finally, notice that once again this covariance matrix estimator is positive semi-definite by construction.

We have motivated this estimator by assuming f_t satisfies a VARMA model. However, inspection of den Haan and Levin's method indicates that it is consistent provided the autocovariance structure of f_t is equivalent to that of some infinite order autoregression. For this, it is only sufficient and not necessary that f_t be a VARMA process. Den Haan and Levin provide a set of more general conditions under which the estimator is consistent. These conditions are very similar to those employed in the next section and certain parallels will emerge between $\hat{S}_{VARMA}$ and some of the methods to which we now turn.

3.5.3 Heteroscedasticity and Autocorrelation Covariance Matrix Estimators

Unfortunately, VARMA processes may not be sufficiently general to capture the dependence structure of f_t in all cases of interest. This has prompted the development of the class of *heteroscedasticity and autocorrelation covariance* (HAC) matrices which are consistent under relatively weak assumptions on the dependence structure of the process. However, it is necessary to impose some further restrictions beyond those already assumed in Section 3.4 for the asymptotic analysis. The discussion in this section rests mostly on the work of Andrews (1991) and Newey and West (1994), and these authors catalogue the required regularity conditions.[47]

To motivate these estimators, it is useful to return to the definition of S given in (3.38), namely $S = \Gamma_0 + \sum_{i=1}^{\infty}(\Gamma_i + \Gamma_i')$. Given this structure, it is natural to estimate S by truncating this infinite sum and using the sample autocovariances, $\hat{\Gamma}_j = T^{-1}\sum_{t=j+1}^{T} \hat{f}_t \hat{f}_{t-j}'$, as estimates of their population analogs. This leads to the estimator

$$\hat{S}_{TR} = \hat{\Gamma}_0 + \sum_{i=1}^{\ell_T}(\hat{\Gamma}_i + \hat{\Gamma}_i') \tag{3.53}$$

where "TR" stands for truncated. White and Domowitz (1984) first proposed

[45] The asymptotic theory is satisfied by the closest integer to $cT^{1/3}$ for any finite positive constant c.

[46] Den Haan and Levin (1996) also consider using Akaike's (1973) information criterion, $AIC(k) = log\{det[\hat{\Omega}(k)]\} + \frac{2kq}{T}$ to pick the lag length. Their theoretical analysis suggests that SIC is a better choice because AIC is not a consistent method of lag selection; however their limited simulation evidence suggests that the two criteria perform comparably in this context.

[47] These include the conditions: (i) $\sup_{\theta \in N_\epsilon} E[|\partial^2 f(v_t, \theta)/\partial\theta_i \partial\theta_j|] < \infty$ for $i, , j = 1, 2, \ldots p$ and N_ϵ is some neighbourhood of θ_0; (ii) $(f(v_t, \theta_0)', vec(\partial f(v_t, \theta_0)/\partial\theta' - E[\partial f(v_t, \theta_0)/\partial\theta'])')$ has l–summable autocovariances and absolutely summable fourth order cumulants, where l is some positive constant.

this type of estimator and showed its consistency in certain least squares settings provided $\ell_T \to \infty$ as $T \to \infty$ and $\ell_T = o(T^{1/3})$. This would appear to solve the problem, but does not. While $\hat{S}_{TR}$ converges in probability to a positive definite matrix, it may be indefinite in finite samples.

The source of the trouble is not the truncation but the weights given to the sample autocovariances in (3.53). This is most readily seen by restricting attention to the case where f_t is a *ℓ–dependent process* so that $\Gamma_i = 0$ for all $i > \ell$, and $\ell_T = \ell$. In this case, the correct order of the process is being used in the estimator but the estimator is still not positive semi-definite. This failure is uncovered by rewriting $\hat{S}_{TR}$ as

$$\hat{S}_{TR} = T^{-1} H' D H$$

where H is the same matrix as in (3.41) and D is the $(T \times T)$ matrix whose only non-zero elements are $D_{i,j} = 1$ for $j = s_1(i), \dots s_2(i)$ for $i = 1, 2, \dots T$ and $s_1(i) = max(i - \ell, 1), s_2(i) = min(i + \ell, T)$. Since D is not positive semi-definite, neither is $\hat{S}_{TR}$. It is important to realize that the failure of positive semi-definiteness does not always imply negative sample variances. Rather it means that negative variances can occur for certain realizations of H. In the limit, the problem disappears because all realizations from the process must satisfy $\hat{S}_{TR} \xrightarrow{p} \Gamma_0 + \sum_{i=1}^{\ell} (\Gamma_i + \Gamma_i')$ which is positive definite by definition. One other important aspect of the problem can be learnt from this example. If $\ell = 0$ then $\hat{S}_{TR} = \hat{S}_{SU}$ and this estimator is positive semi-definite by construction. So the problem stems from the inclusion of the sample autocovariance matrices $\{\hat{\Gamma}_i; i = 1, 2, \dots \ell\}$.

The solution is to construct an estimator in which the contribution of the sample autocovariances matrices are weighted to downgrade their role sufficiently in finite samples to ensure positive semi-definiteness but have the weights tend to one as $T \to \infty$ to ensure consistency. This is the intuition behind the class of *heteroscedasticity autocorrelation covariance (HAC) matrices*. This class consists of estimators of the form

$$\hat{S}_{HAC} = \hat{\Gamma}_0 + \sum_{i=1}^{T-1} \omega_{i,T} (\hat{\Gamma}_i + \hat{\Gamma}_i') \tag{3.54}$$

where $\omega_{i,T}$ is known as the kernel (or weight). The kernel must be carefully chosen to ensure the twin properties of consistency and positive semi-definiteness. The three most popular choices in the econometrics literature are given in Table 3.3.

Table 3.3
Kernels for three common HAC estimators

Name	Author(s)	Kernel, $\omega_{i,T}$
Bartlett	Newey and West (1987a)	$1 - a_i$ for $a_i \leq 1$ 0 for $a_i > 1$
Parzen	Gallant (1987)	$1 - 6a_i^2 + 6a_i^3$ for $0 \leq a_i \leq 0.5$ $2(1 - a_i)^3$ for $0.5 \leq a_i \leq 1$ 0 for $a_i > 1$
Quadratic Spectral	Andrews(1991)	$\frac{25}{12\pi^2 d_i^2}\left[\frac{Sin(m_i)}{m_i} - Cos(m_i)\right]$

Note: $a_i = i/(b_T + 1)$; $d_i = i/b_T$; $m_i = 6\pi d_i/5$.

Here "name" refers to the term by which the particular choice of kernel is most commonly known, and is a reference back to an earlier literature on the estimation of the spectral density at frequency zero in which these types of problems were first solved.[48] The parameter b_T is known as the bandwidth, and must be non-negative. Notice that this parameter controls the number of autocovariances included in the HAC estimator when either the Bartlett or Parzen kernels are used. In these two cases, b_T must be an integer, but no such restriction is required for the quadratic spectral kernel. Which set of weights should be used? Andrews (1991) shows that the Quadratic Spectral weights are optimal in the sense that they minimize an asymptotic mean squared error criterion for the estimation of S. His results imply that this choice only marginally dominates the Parzen weights, but both should be much better than the Bartlett weights. This is mirrored to some extent by his simulation results for a linear model with some simple forms of autocorrelation and heteroscedasticity. However, although the Quadratic Spectral weights perform slightly better than the Parzen weights, neither dominate the Bartlett weights to the extent predicted by the theory. Newey and West (1994) report simulation evidence from two more general linear models; in one, their results corroborate Andrews's but in the other they find no clear ranking is possible. Newey and West (1994) conclude that the choice between the kernels is not particularly important; a view for which there is some precedent in the earlier spectral density estimation literature.[49]

The bandwidth is a much more important determinant of the finite sample properties of $\hat{S}_{HAC}$. For consistency, b_T must tend to infinity with T.[50] Andrews (1991) shows that the asymptotic mean square error is minimized by setting b_T equal to $O(T^{1/3})$ for the Bartlett weights and $O(T^{1/5})$ for both the

[48] See Priestley (1981) for a review of this earlier literature.

[49] See Priestley (1981) [p.574].

[50] Newey and West (1987a) and Gallant (1987) prove the consistency of their particular estimators under the assumption $b_T = o(T^{1/4})$. Andrews (1991) and Hansen (1992) prove the consistency of this general class of estimators under the assumption $b_T = o(T^{1/2})$.

Parzen and Quadratic Spectral weights. However, again, this type of condition provides little practical guidance because it only restricts the optimal bandwidth for the Bartlett weights, say, to be of the form $cT^{1/3}$ for any choice of finite $c > 0$. Andrews (1991) develops some procedures for picking the optimal c based on the assumption that f_t follows certain VARMA models. However, we do not pursue these here because if this specification is adopted then it seems more reasonable to use the $\hat{S}_{VARMA}$ described in the previous section.[51] Newey and West (1994) propose a nonparametric method for selecting the bandwidth and show it minimizes the asymptotic mean square error criterion. The mechanics of this approach are as follows; the parameters (h, n, c_γ, ν) are defined afterwards.

Newey and West's Method of Bandwidth Selection

1. *Use the* $(q \times 1)$ *vector* h *to construct the scalar random variable* $c_t = h' \hat{f}_t$.
2. *Construct* $\hat{\sigma}_j = T^{-1} \sum_{t=j+1}^{T} c_t c_{t-j}$ *for* $j = 0, 1, \ldots n$.
3. *Calculate* $\hat{s}^{(\nu)} = 2 \sum_{j=1}^{n} j^\nu \hat{\sigma}_j$ *and* $\hat{s}^{(0)} = \hat{\sigma}_0 + 2 \sum_{j=1}^{n} \hat{\sigma}_j$.
4. *Calculate* $\hat{\gamma} = c_\gamma \{ \{ \hat{s}^{(\nu)} / \hat{s}^{(0)} \}^2 \}^{1/(2\nu+1)}$.
5. *For the Bartlett and Parzen kernels, set* $b_T = int\{\hat{\gamma} T^{1/(2\nu+1)}\}$ *where* $int\{.\}$ *denotes the integer part of the number inside the brackets; for the Quadratic Spectral kernel, set* $b_T = \hat{\gamma} T^{1/(2\nu+1)}$.

It would be anticipated that the bandwidth depends on the autocovariances of $\hat{f}_t$ and close inspection of the above reveals this to be the case. However, there is no simple intuition for the exact nature of the calculations. The parameters (n, c_γ, ν) are given in Table 3.4.

Table 3.4
Parameter values for Newey and West's (1994) bandwidth selection method

Weight	ν	n	c_γ
Bartlett	1	$O(T^{2/9})$	1.4117
Parzen	2	$O(T^{4/25})$	2.6614
Quadratic Spectral	2	$O(T^{2/25})$	1.3221

Notice that the exact choice of n is not specified and so Newey and West's procedure does not completely solve the problem. They recommend that the calculations be repeated for different choices of n to ensure the resulting confidence intervals or hypothesis tests are not sensitive to the choice of this parameter.

[51] If f_t follows a VARMA process and so $S = S_{VARMA}$ then $\hat{S}_{VARMA}$ converges to this limit faster than $\hat{S}_{HAC}$; see Andrews (1991) and den Haan and Levin (1996).

To implement the method, the vector h must also be chosen. Newey and West (1994) focus on the case where $f_t = z_t u_t$ and suggest that if the first element of z_t is a constant then h can be set equal to $(0, 1, 1, \ldots 1)$. More generally, the choice of h can be data dependent subject to certain conditions; see Newey and West (1994) [p.636]. However to date, no further guidance is available about either how this choice should be made or its impact on the finite sample properties of the covariance matrix estimator.

In theory, the HAC estimators have solved the problem of constructing a consistent, positive semi-definite estimator of S under very weak conditions on f_t. However, in practice, they often do not work well in cases of interest. Simulation evidence suggets their use can lead to the confidence intervals in (3.27) which do not possess the anticipated coverage rates in finite samples; see Andrews (1991), Andrews and Monahan (1992) and Newey and West (1994). An examination of the estimation error indicates the types of circumstance in which this problem may be present. For ease of exposition, we restrict attention to HAC estimators for which $\omega_{i,T} = 0$ for $i > b_T$. From (3.38) and (3.54), the estimation error is

$$\begin{aligned} S - \hat{S}_{HAC} &= \Gamma_0 - \hat{\Gamma}_0 + \sum_{i=1}^{b_T} \omega_{i,T}\{(\Gamma_i - \hat{\Gamma}_i) + (\Gamma_i' - \hat{\Gamma}_i')\} \\ &\quad + \sum_{i=1}^{b_T}(1 - \omega_{i,T})(\Gamma_i + \Gamma_i') + \sum_{i=b_T+1}^{\infty}(\Gamma_i + \Gamma_i') \end{aligned} \tag{3.55}$$

So there are three sources of error: (i) error from the estimation of the autocovariances, $\{\Gamma_i - \hat{\Gamma}_i\}$; (ii) error due the weights on the estimated autocovariances, $1 - \omega_{i,T}$; (iii) approximation error due to the truncation of the sum, $\sum_{i=b_T+1}^{\infty}(\Gamma_i + \Gamma_i')$. The best way to appreciate when these errors are large is to start by describing a situation in which they should be relatively small. Suppose f_t is a ℓ–dependent process and ℓ is small relative to b_T. In this case it follows that: (i) the weights on the $\hat{\Gamma}_i$ for $i \leq \ell$ are very close to one; (ii) for $i > \ell$ the weights help to shrink the estimated covariance matrices towards their limiting value of zero; (iii) there is no approximation error. These three effects combine to produce an estimator that is reasonably accurate in finite samples. Now consider what happens as ℓ increases. Estimation error creeps in because the weights are substantially different from one for the longer lags less than or equal to ℓ and then once $b_T < \ell$, there is approximation error as well. This suggests that $\hat{S}_{HAC}$ is unlikely to perform well in finite samples if the population autocovariance matrices of f_t die out too slowly. Such behaviour would be observed if f_t is generated by a process with a substantial autoregressive component.

Autoregressive behaviour is a common feature of economic time series and so these problems motivated Andrews and Monahan (1992) to propose a modification to the HAC estimator based on a technique called *prewhitening and*

recolouring.[52] The basic idea is to filter $\hat{f}_t$ to reduce the size of its autoregressive component and hence to produce a series for which an HAC estimator works better. This is known as the "prewhitening" phase. The long run variance of the filtered series is estimated using a member of the class of HAC. Then in the "recolouring" phase, the long run variance of $\hat{f}_t$ is estimated from the HAC and the properties of the filter. Andrews and Monahan (1992) recommend using a VAR(m) process to filter the data and so their procedure is as follows.

Andrews and Monahan's Procedure

1. *Estimate the VAR(m) model for* $\hat{f}_t$,

$$\hat{f}_t = A_1(m)\hat{f}_{t-1} + \ldots + A_m(m)\hat{f}_{t-m} + e_t(m) \tag{3.56}$$

by least squares. These estimates are given by

$$\hat{A}(m) = \sum_{t=m+1}^{T} \hat{f}_t r_t' \{ \sum_{t=m+1}^{T} r_t r_t' \}^{-1}$$

where $A(m) = (A_1(m), A_2(m), \ldots A_m(m))$ *and* $r_t' = (\hat{f}_{t-1}', \hat{f}_{t-2}', \ldots \hat{f}_{t-m}')$. *Construct the forecast error* $\hat{e}_t(m) = \hat{f}_t - \hat{A}(m)r_t$.

2. *Construct the estimator* $\hat{\Sigma} = \hat{\Gamma}_0 + \sum_{i=1}^{T-1} \omega_{i,T}(\hat{\Gamma}_i + \hat{\Gamma}_i')$ *where* $\hat{\Gamma}_i = T^{-1} \sum_{t=m+i+1}^{T} \hat{e}_t(m)\hat{e}_{t-i}(m)'$.

3. *The estimator of* S *is*

$$\hat{S}_{PWRC} = \{I_q - \sum_{i=1}^{m} \hat{A}_i(m)\}^{-1} \hat{\Sigma} \{I_q - \sum_{i=1}^{m} \hat{A}_i(m)\}^{-1'} \tag{3.57}$$

Any value of m can be used; however, Newey and West (1994) recommend using $\hat{S}_{PWRC}$ with $m = 1$ and their method of bandwidth selection in step 2. This estimator is positive semi-definite by construction and Andrews and Monahan (1992) prove its consistency. There are clearly close parallels with den Haan and Levin's (1996) method: the main difference is that Andrews and Monahan use the autoregressive filter to remove *some* of the autocorrelation structure; whereas in den Haan and Levin's method the autoregressive filter must remove *all* the autocorrelation structure with the autoregression. This difference manifests itself in the consistency proofs. The consistency of $\hat{S}_{PWRC}$ depends mostly on the use of the HAC estimator in step 2, but the filter must also satisfy certain properties. In particular, if we write $plim_{T\to\infty} \hat{A}_i(m) = A_i(m)$, then $A(L) = I_q - \sum_{i=1}^{m} A_i(m)$ must satisfy the conditions for stationarity presented in the previous section. Since den Haan and Levin's method is based

[52] Like the HAC estimators, this technique has its origins in the literature on spectral density estimation where it was first proposed by Press and Tukey (1956).

essentially on the AR(∞) representation, this property is guaranteed in that case.[53] However, it may not hold if the AR polynomial is arbitrarily truncated at some finite lag as is done in Andrews and Monahan's procedure. However, since the AR filter is just a device to reduce the autocorrelation, and not to remove it, Andrews and Monahan propose modifying the filter as follows to ensure it satisfies the required "stationarity" condition.

To describe this modification, it is most convenient to set $m = 1$, which is the choice recommended by Newey and West (1994). Since there is now only one coefficient matrix, $\hat{A}_1(1)$, we denote this matrix by $\hat{A}$. Notice that in this case, the condition for stationarity reduces to the requirement that the eigenvalues of $plim_{T\to\infty}\hat{A} = A$ are less than one in absolute value.[54] In practice, problems may occur if the eigenvalues of $\hat{A}$ satisfy this condition but are close to one. Therefore, Andrews and Monahan (1992) propose modifying $\hat{A}$ to ensure its eigenvalues are less than 0.97 in absolute value. Their procedure is based on the Singular Value Decomposition of $\hat{A}$. This decomposition is $\hat{A} = \hat{B}\hat{\Delta}\hat{C}'$ where $\hat{\Delta}$ is a diagonal matrix whose elements are all non-negative.[55] Andrews and Monahan (1992) show that the eigenvalues of $\hat{A}$ are guaranteed to satisfy the required constraint if all the elements of $\hat{\Delta}$ are less than or equal to 0.97. If this is not the case, then Andrews and Monahan (1992) recommend the offending elements of $\hat{\Delta}$ are replaced by 0.97 to give a new matrix $\tilde{\Delta}$ and $\hat{A}$ is replaced by $\tilde{A} = \hat{B}\tilde{\Delta}\hat{C}'$. Simulation evidence in Andrews and Monahan (1992) and Newey and West (1994) suggests the use of prewhitening and recolouring improves the finite sample performance of the asymptotic confidence intervals in (3.27). So for completeness, we conclude this section by bringing together all these recommendations into a single procedure. Although, this was originally proposed by Newey and West (1994), we shall give it a more general name since it represents the synthesis of results and simulation evidence reported in all the papers cited above.[56]

Estimation of S when f_t is Stationary and Ergodic

1. *Estimate the model $\hat{f}_t = A\hat{f}_{t-1} + e_t$ by least squares to give $\hat{A}$. Let $\hat{A} = \hat{B}\hat{\Delta}\hat{C}'$ be the Singular Value Decomposition of $\hat{A}$. Define $\tilde{\Delta}$ to be the diagonal matrix whose $(i,i)^{th}$ element is given by $\tilde{\Delta}_{ii} = min\{\hat{\Delta}_{ii}, 0.97\}$ and $\tilde{A} = \hat{B}\tilde{\Delta}\hat{C}'$. Construct $\tilde{e}_t = \hat{f}_t - \tilde{A}\hat{f}_{t-1}$.*

[53] Of course, this statement is subject to certain regularity conditions being satisfied; see den Haan and Levin (1996).

[54] See Hamilton (1994) [p.259].

[55] This decomposition can be calculated straightforwardly in most computer packages for matrix analysis. It is defined as follows. First, note that $\hat{A}'\hat{A}$ and $\hat{A}\hat{A}'$ have exactly the same set of nonzero eigenvalues, which we denote by $\{\delta_i; i = 1, 2, \ldots r\}$. It is reasonable to assume in our context that $r = q$ and so both $\hat{A}'\hat{A}$ and $\hat{A}'\hat{A}$ are of full rank. The i^{th} diagonal element of $\hat{\Delta}$ is δ_i. The matrix $\hat{B}$ is the $(q \times q)$ matrix whose i^{th} column is the eigenvector of $\hat{A}\hat{A}'$ associated with the δ_i. The matrix $\hat{C}$ is the $(q \times q)$ matrix whose i^{th} column is the the eigenvector of $\hat{A}'\hat{A}$ associated with δ_i. For example, see Dhrymes (1984) [p.78] or Strang (1988) [Appendix A] for a more detailed discussion.

[56] Also see Section 4.3.

2. *Use an HAC estimator in conjunction with Newey and West's method of bandwidth selection given above to construct the matrix*

$$\hat{\Sigma} = \hat{\Gamma}_0 + \sum_{i=1}^{T-1} \omega_{i,T}(\hat{\Gamma}_i + \hat{\Gamma}_i')$$

where $\hat{\Gamma}_i = T^{-1}\sum_{t=i+1}^{T} \tilde{e}_t\tilde{e}_{t-i}'$.

3. *The estimator of* S *is*

$$\hat{S}_{SE} = \{I_q - \tilde{A}\}^{-1}\hat{\Sigma}\{I_q - \tilde{A}\}^{-1'} \tag{3.58}$$

where the subscript "SE" stands for "stationary and ergodic".

The choice between the covariance matrix estimators $\hat{S}_{SU}$, $\hat{S}_{VARMA}$ and $\hat{S}_{SE}$ depends on the model in question. Whichever estimator is appropriate, it can be used to calculate the approximate large sample confidence intervals for $\theta_{0,i}$ given in (3.27). This section concludes with an illustration of the various methods in the context of Hansen and Singleton's (1982) consumption based asset pricing model.

Example: Hansen and Singleton's (1982) Consumption Based Asset Pricing Model

Since $z_t \in \Omega_t$, it follows from (1.22) that $f(v_t, \theta_0) = z_t(\delta_0 x_{1,t+1}^{\gamma_0 - 1}x_{2,t+1} - 1)$ is martingale difference sequence. Therefore the economic model implies S can be consistently estimated by $\hat{S}_{SU}$ given in (3.40). In spite of this structure, we shall use this example to illustrate all the various methods discussed in this section. For den Haan and Levin's (1996) method K is set equal to $int\{T^{1/3}\} = 7$ but in each case the Schwarz criteria chooses $\hat{k} = 0$ and so indicates that f_t is serially uncorrelated. In this case, $\hat{S}_{VARMA}$ equals $\hat{S}_{SU}$. Three versions of $\hat{S}_{HAC}$ are calculated: one for each kernel in Table 3.3. In each case, we fix the bandwidth to $b_T = 7$. Finally, three versions of $\hat{S}_{SE}$ are calculated; again one for each kernel. The bandwidth for each is calculated using the parameters in Table 3.4, and so n equals $int\{T^{2/9}\} = 3$, $int\{T^{4/25}\} = 2$, $int\{T^{2/25}\} = 1$ respectively for the Bartlett, Parzen and Quadratic Spectral kernel. We arbitrarily chose to set $h = (1, 1, 1, 1, 1)'$. Clearly the width of the confidence intervals is determined by the standard error of the estimates,

$$s.e.(\hat{\theta}_i) = \sqrt{\hat{V}_{T,ii}/T} \tag{3.59}$$

where $\hat{V}_{T,ii}$ is the i^{th} main diagonal element of $\hat{V}_T = \hat{M}_T\hat{S}_T\hat{M}_T^{'}$ and $\hat{M}_T = [G_T(\hat{\theta})'W_TG_T(\hat{\theta})]^{-1}G_T(\hat{\theta})'W_T$. So, for brevity, only the standard errors of $\hat{\gamma}_T$ and $\hat{\delta}_T$ are reported. Table 3.5 contains these statistics for the case in which the model is estimated with equally weighted returns (EWR), and Table 3.6

presents the results for the case in which the model is estimated with value weighted returns (VWR).[57]

Certain features stand out. First, the different choice of covariance matrix estimator has some impact on the calculated standard errors. In principle, all the versions of $\hat{S}_T$ are consistent if the model is correctly specified because then $f(v_t, \theta_0)$ is a martingale difference sequence. Since den Haan and Levin's (1996) method confirms the absence of serial correlation in $f(v_t, \theta_0)$, these differences reflect inherent randomness or finite sample bias. Secondly, the estimates based on $W_T = (T^{-1}\sum_{t=1}^{T} z_t z_t')^{-1}$ give much smaller standard errors. Finally, no matter what the choice of asset or weighting matrix, δ_0 is far more precisely estimated than γ_0.

Table 3.5
Standard errors of the first step estimators for the consumption-based asset pricing model with EWR

W_T	$\hat{S}_T$	$s.e.(\hat{\gamma}_T)$	$s.e.(\hat{\delta}_T)$
$10^5 I_5$	SU, VARMA	6.844	1.210×10^{-2}
	HAC(B,7)	5.893	1.036×10^{-2}
	HAC(P,7)	6.458	1.132×10^{-2}
	HAC(Q,7)	5.549	9.720×10^{-3}
	SE(B,1)	7.670	1.360×10^{-2}
	SE(P,4)	7.148	1.254×10^{-2}
	SE(Q,2.2)	7.340	1.293×10^{-2}
$(T^{-1}\sum_{t=1}^{T} z_t z_t')^{-1}$	SU, VARMA	2.263	4.393×10^{-3}
	HAC(B,7)	2.134	4.544×10^{-3}
	HAC(P,7)	2.148	4.540×10^{-3}
	HAC(Q,7)	2.091	4.502×10^{-3}
	SE(B,0)	2.430	4.916×10^{-3}
	SE(P,1)	2.420	4.894×10^{-3}
	SE(Q,2.49)	2.308	4.726×10^{-3}

Notes: B, P, Q denote the Bartlett, Parzen, Quadratic Spectral kernel. For K=B,P or Q: HAC(K,7) denotes an HAC estimator kernel with K kernel and $b_T = 7$; SE(K,b) denotes $\hat{S}_{SE}$ with K kernel and estimated bandwidth b.

[57] It should be noted that evidence reported below indicates the model is misspecified for EWR and this renders the standard errors in (3.59) invalid. See Section 5.1.4 and Section 4.2 respectively.

Table 3.6
Standard errors of the first step estimators for the consumption-based asset pricing model with VWR

W_T	$\hat{S}_T$	$s.e.(\hat{\gamma}_T)$	$s.e.(\hat{\delta}_T)$
$10^5 I_5$	SU, VARMA	5.840	1.063×10^{-2}
	HAC(B,7)	4.559	8.447×10^{-3}
	HAC(P,7)	4.827	8.852×10^{-3}
	HAC(Q,7)	4.342	8.059×10^{-3}
	SE(B,1)	5.593	1.032×10^{-2}
	SE(P,5)	5.073	9.315×10^{-3}
	SE(Q,0.78)	5.632	1.038×10^{-2}
$(T^{-1}\sum_{t=1}^{T} z_t z_t')^{-1}$	SU, VARMA	1.867	3.761×10^{-3}
	HAC(B,7)	1.699	3.523×10^{-3}
	HAC(P,7)	1.722	3.548×10^{-3}
	HAC(Q,7)	1.674	3.489×10^{-3}
	SE(B,0)	1.850	3.765×10^{-3}
	SE(P,4)	1.761	3.626×10^{-3}
	SE(Q,1.62)	1.812	3.727×10^{-3}

Notes: See Table 3.5 for definitions.

3.6 The Optimal Choice of Weighting Matrix

In Section 3.3 it is shown that if $q = p$ then GMM is equivalent to the Method of Moments estimator based on $E[f(v_t, \theta_0)] = 0$ and so does not depend on the weighting matrix. However if $q > p$ then no such reduction is possible and it is clear from Theorem 3.2 that the asymptotic variance of $\hat{\theta}_T$ depends on W_T via W.[58] This opens up the possibility that inferences may be sensitive to W. Just as in the linear model, it is desirable to base inference on the most precise estimator and so the optimal choice of W is the one which yields the minimum variance in a matrix sense. Once again, this choice is S^{-1}; however this time we state the result more formally. Hansen (1982) proves this result but we note parenthetically that his argument is different from the one employed below.

Theorem 3.4 Optimal Choice of Weighting Matrix
If Assumptions 3.1–3.5, 3.7–3.13 hold then the minimum asymptotic variance of $\hat{\theta}_T$ *is* $(G_0' S^{-1} G_0)^{-1}$ *and this can be obtained by setting* $W = S^{-1}$.

Note that the regularity conditions are imposed to ensure that $\hat{\theta}_T$ has the asymptotic distribution given in Theorem 3.2.

Proof of Theorem 3.4:
Let $\hat{\theta}_T(W)$ be the GMM estimator based on Assumption 3.3 with weighting matrix W_T. It can be recalled from Section 2.4 that the result is established if it

[58] If $p = q$ then the asymptotic variance of $\hat{\theta}_T$ is $MSM' = (G_0' S^{-1} G_0)^{-1}$.

can be shown that $V(W)-V(S^{-1})$ equals a positive semi-definite matrix, where $V(W)$ denotes the variance of the limiting distribution of $T^{1/2}[\hat{\theta}_T(W) - \theta_0]$.

To begin the proof, it is useful to relate $T^{1/2}[\hat{\theta}_T(W) - \theta_0]$ to $T^{1/2}[\hat{\theta}_T(S^{-1}) - \theta_0]$. This is done quite simply by noting that

$$T^{1/2}[\hat{\theta}_T(W) - \theta_0] = T^{1/2}[\hat{\theta}_T(S^{-1}) - \theta_0] + T^{1/2}[\hat{\theta}_T(W) - \hat{\theta}_T(S^{-1})] \quad (3.60)$$

Now, from (3.33) it follows that

$$T^{1/2}[\hat{\theta}_T(W) - \theta_0] = -M(W)T^{1/2}g_T(\theta_0) + o_p(1) \quad (3.61)$$

where $M(W) = (G_0^{'}WG_0)^{-1}G_0^{'}W$, and so that

$$T^{1/2}[\hat{\theta}_T(W) - \hat{\theta}_T(S^{-1})] = -[M(W) - M(S^{-1})]T^{1/2}g_T(\theta_0) + o_p(1) \quad (3.62)$$

Therefore, if we substitute (3.61) and (3.62) into (3.60) and calculate the limiting variance of each side then it follows that

$$V(W) = V(S^{-1}) + V_1 + C + C' \quad (3.63)$$

where $V_1 = \lim_{T\to\infty} Var[\{M(W) - M(S^{-1})\}T^{1/2}g_T(\theta_0)]$ and

$$C = \lim_{T\to\infty} Cov\left[\{M(W) - M(S^{-1})\}T^{1/2}g_T(\theta_0), M(S^{-1})T^{1/2}g_T(\theta_0)\right] \quad (3.64)$$

Equation (3.63) is easily rearranged to give

$$V(W) - V(S^{-1}) = V_1 + C + C' \quad (3.65)$$

Now V_1 is positive semi-definite by construction, and so we focus attention on C. By definition, it follows that

$$\begin{aligned} C &= \lim_{T\to\infty} E\left[\{M(W) - M(S^{-1})\}T^{1/2}g_T(\theta_0)T^{1/2}g_T(\theta_0)'M(S^{-1})'\right] \\ &= \{M(W) - M(S^{-1})\}\lim_{T\to\infty} E\left[T^{1/2}g_T(\theta_0)T^{1/2}g_T(\theta_0)'\right]M(S^{-1})' \\ &= M(W)SM(S^{-1})' - M(S^{-1})SM(S^{-1})' = 0 \end{aligned} \quad (3.66)$$

Equations (3.65)–(3.66) and the definition of V_1 establish the desired result. ◇

The proof is derived by showing that $C = 0$. It can be recognized that C is the asymptotic covariance between $T^{1/2}[\hat{\theta}_T(S^{-1}) - \theta_0]$ and $T^{1/2}[\hat{\theta}_T(W) - \hat{\theta}_T(S^{-1})]$. Therefore, $C = 0$ implies that $T^{1/2}[\hat{\theta}_T(S^{-1}) - \theta_0]$ is asymptotically uncorrelated with $T^{1/2}[\hat{\theta}_T(W) - \hat{\theta}_T(S^{-1})]$ for any W.

Theorem 3.4 implies the optimal choice of W_T is $\hat{S}_T^{-1}$ where $\hat{S}_T$ is a consistent estimator of S. As in the linear model, the construction of this estimator requires at least two steps. On the first step a sub-optimal choice of W_T is used

to obtain a preliminary estimator, $\hat{\theta}_T(1)$. This estimator is used to obtain a consistent estimator of S, which is denoted $\hat{S}_T(1)$. On the second step θ_0 is re-estimated with $W_T = \hat{S}_T(1)^{-1}$. The resulting estimator, $\hat{\theta}_T(2)$, has the minimum asymptotic covariance matrix given in Theorem 3.4.[59] However, this *two step* estimator is based on a version of the optimal weighting matrix constructed using a sub-optimal estimator of θ_0. This suggests there may be finite sample gains from using $\hat{\theta}_T(2)$ to construct a new estimator of S, $\hat{S}_T(2)$ say, and then re-estimating θ_0 with $W_T = \hat{S}_T(2)^{-1}$. The resulting estimator, $\hat{\theta}_T(3)$, also has the same asymptotic distribution as $\hat{\theta}_T(2)$ but it is anticipated to be more efficient in finite samples. This potential finite sample gain in efficiency provides a justification for updating the estimate of S again and re-estimating θ_0. This process can be continued iteratively until the estimates converge; if this is done then it yields what has become known as the *iterated GMM estimator*. The i^{th} step of such an iterative procedure is as follows.

The i^{th} Step of Iterated GMM Estimation

- *If $i = 1$: Estimate θ_0 using GMM based on the population moment condition in Assumption 3.3 with a sub-optimal weighting matrix, such as $W_T = I_q$. Denote this estimator by $\hat{\theta}_T(1)$. Use this estimator to construct a consistent estimator of S by one of the methods described in Section 3.5.*[60] *Denote this estimator by $\hat{S}_T(1)$.*

- *If $i > 1$: Estimate θ_0 using GMM based on the population moment condition in Assumption 3.3 with $W_T = \hat{S}_T(i-1)^{-1}$ where $\hat{S}_T(i-1)$ is a consistent estimator of S based on $\hat{\theta}_T(i-1)$, the estimator of θ_0 from the $(i-1)^{th}$ step. If $\|\hat{\theta}_T(i) - \hat{\theta}_T(i-1)\| < \epsilon_\theta$ then the procedure has converged and the iterated GMM estimator is $\hat{\theta}_T = \hat{\theta}_T(i)$. If $\|\hat{\theta}_T(i) - \hat{\theta}_T(i-1)\| \geq \epsilon_\theta$ and $i < I_{max}$ then go to the $(i+1)^{th}$ step.*

Typically ϵ_θ is set equal to some small positive number such as 10^{-6}. Notice that a ceiling of I_{max} has been placed on the number of steps. This is needed because in practice there is no guarantee that this iterative procedure converges and so limiting the number of steps is a safeguard against putting the computer into an infinite loop! Regardless of whether convergence occurs before the chosen I_{max}, all $\{\hat{\theta}_T(i),\, i > 1\}$ have the same asymptotic distribution with the covariance matrix given in Theorem 3.4.

The choice of $W = S^{-1}$ has a second important implication for the asymptotic behaviour of the estimator which is presented in the following theorem.

Theorem 3.5 Asymptotic Independence of $T^{1/2}(\hat{\theta}_T - \theta_0)$ and $S^{-1/2}T^{1/2}g_T(\hat{\theta}_T)$

If (i) Assumptions 3.1–3.5, and 3.7–3.13 hold; (ii) $W = S^{-1}$; then $T^{1/2}(\hat{\theta}_T - \theta_0)$ and $S^{-1/2}T^{1/2}g_T(\hat{\theta}_T)$ are asymptotically independent.

[59] This estimator is sometimes refered to as Hansen's two step estimator because it is proposed in Hansen (1982).

[60] Also see Section 4.3.

Proof:
First recall that Theorems 3.2 and 3.3 establish that both statistics converge to normal distributions, and so a necessary and sufficient condition for asymptotic independence is that these two statistics are asymptotically uncorrelated. The latter can be deduced from (3.26) and (3.36). Using Lemma 3.3 and putting $W = S^{-1}$, it follows from (3.26) and (3.36) that

$$T^{1/2}(\hat{\theta}_T - \theta_0) = H_{1,T} + o_p(1) \tag{3.67}$$

$$W_T^{1/2} T^{1/2} g_T(\hat{\theta}_T) = H_{2,T} + o_p(1) \tag{3.68}$$

where

$$\begin{aligned} H_{1,T} &= -[F(\theta_0)'F(\theta_0)]^{-1}F(\theta_0)'S^{-1/2}T^{1/2}g_T(\theta_0) \\ H_{2,T} &= [I_q - P(\theta_0)]S^{-1/2}T^{1/2}g_T(\theta_0) \end{aligned}$$

If we let $C = \lim_{T\to\infty} Cov[H_{1,T}, H_{2,T}]$ then it follows from Theorems 3.2 and 3.3 that

$$C = \lim_{T\to\infty} E[H_{1,T}H_{2,T}^{'}] \tag{3.69}$$

Using (3.67) and (3.68) in (3.69), we obtain

$$\begin{aligned} C &= \lim_{T\to\infty} E\big[-[F(\theta_0)'F(\theta_0)]^{-1}F(\theta_0)'S^{-1/2}T^{1/2}g_T(\theta_0)T^{1/2}g_T(\theta_0)'S^{-1/2'} \times \\ &\qquad [I_q - P(\theta_0)]\big] \\ &= -[F(\theta_0)'F(\theta_0)]^{-1}F(\theta_0)'S^{-1/2}\left\{\lim_{T\to\infty} Var[T^{1/2}g_T(\theta_0)]\right\}S^{-1/2'} \times \\ &\qquad [I_q - P(\theta_0)] \\ &= -[F(\theta_0)'F(\theta_0)]^{-1}F(\theta_0)'S^{-1/2}\,S\,S^{-1/2'}[I_q - P(\theta_0)] \end{aligned}$$

Now, by definition, we have $S = S^{1/2'}S^{1/2}$ and $S^{-1} = S^{-1/2'}S^{-1/2}$ which together imply $S^{-1/2} = (S^{1/2'})^{-1}$. It therefore follows that $S^{-1/2}SS^{-1/2'} = I_q$. Using this identity C reduces to

$$C = -[F(\theta_0)'F(\theta_0)]^{-1}F(\theta_0)'[I_q - P(\theta_0)] = 0 \qquad \diamond$$

This independence property is exploited in the construction of certain test statistics described in Chapter 5. However, in our present context, it provides an interesting perspective on why this choice of W leads to an efficient estimator. First, notice that if we repeat the sequence of steps in the proof of Theorem 3.5 with any other choice of W then the end result is that $C \neq 0$. Therefore, $W = S^{-1}$ is the only choice of weighting matrix for which the estimator is statistically independent of the part of the moment condition unused in estimation. In other words, by making this choice of W, we have extracted all possible information about the parameters contained in the sample moment.

The estimators described in this section are often described as *"the optimal two step GMM"* or *"optimal iterated GMM"* estimator. It is important to realize that this optimality only refers to the choice of weighting matrix. These

are the most precise GMM estimators which can be constructed from the given population moment condition $E[f(v_t, \theta_0)] = 0$. It does not imply that there is anything optimal about the population moment condition itself. The optimal choice of moment condition is discussed in Chapter 7. We conclude this section with an empirical illustration of the two–step and iterated estimator.

Example: Hansen and Singleton's (1982) Consumption Based Asset Pricing Model

Table 3.7 contains the two step and iterated GMM estimation results for both equally weighted returns (EWR) and value weighted returns (VWR). Since the economic model implies $f(v_t, \theta_0)$ is a martingale difference sequence, the covariance matrix is estimated by $\hat{S}_{SU}$ at each step. The convergence criteria for the GMM iterative procedure is $\epsilon_\theta = 10^{-6}$. Convergence only took four iterations with VWR and five with EWR. After two steps the impact of the first-step weighting matrix is clearly diminishing. With iteration, the impact disappears completely.

Table 3.7
Two step and iterated GMM estimators for the consumption based asset pricing model with EWR and VWR

EWR: $W_T^{(1)}$	$(\hat{\gamma}_T, \hat{\delta}_T)$ for $i = 1$	$(\hat{\gamma}_T, \hat{\delta}_T)$ for $i = 2$	$(\hat{\gamma}_T, \hat{\delta}_T)$ after iteration
A	$(-3.145, 0.999)$	$(-0.328, 0.999)$	$(-0.343, 0.992)$
B	$(0.398, 0.993)$	$(-0.317, 0.992)$	$(-0.343, 0.992)$
VWR: $W_T^{(1)}$	$(\hat{\gamma}_T, \hat{\delta}_T)$ for $i = 1$	$(\hat{\gamma}_T, \hat{\delta}_T)$ for $i = 2$	$(\hat{\gamma}_T, \hat{\delta}_T)$ after iteration
A	$(-1.871, 0.998)$	$(0.706, 0.994)$	$(0.666, 0.994)$
B	$(0.698, 0.994)$	$(0.666, 0.994)$	$(0.666, 0.994)$

Notes: $W_T^{(1)}$ denotes the first-step weighting matrix, A denotes $W_T^{(1)} = 10^5 I_5$ and B denotes $W_T^{(1)} = (T^{-1} \sum_{t=1}^{T} z_t z_t')^{-1}$.

Table 3.8 reports the standard errors and 95% confidence intervals for the parameters. A comparison with the first step standard errors in Tables 3.5 and 3.6 indicates that iteration has increased the precision. As before, the discount factor is very precisely estimated, but the coefficient of relative risk aversion is not. In fact, the confidence intervals for γ_0 include values which exceed one. It

may be recalled from Section 1.3.1 that $\gamma_0 < 1$ was a necessary restriction for the representative agent to possess a concave utility function. However, this is not necessarily a concern since the confidence intervals are also consistent with the representative agent's utility function being concave.

Table 3.8
Approximate standard errors and 95% confidence intervals for the iterated GMM Estimators in the consumption based asset pricing model

Asset	$s.e.(\hat{\gamma}_T)$	$c.i.(\hat{\gamma}_T)$	$s.e.(\hat{\delta}_T)$	$c.i.(\hat{\delta}_T)$
EWR	2.215	$(-4.863, 4.000)$	0.004	$(0.983, 1.000)$
VWR	1.823	$(-2.916, 4.249)$	0.004	$(0.987, 1.001)$

Notes: $s.e.(.)$ denotes the standard error calculated using (3.59) with $W_T = \hat{S}_{SU}^{-1}$, $\hat{S}_T = \hat{S}_{SU}$ and $\hat{S}_{SU}$ is defined in (3.40). $c.i.(.)$ denotes the 95% confidence interval calculated using (3.27).

The imprecision of the estimates is a concern, however. The source of the problem can be traced to an interaction of the properties of the data and the nature of the nonlinearity in $f(v_t, \theta)$. The mean and standard deviations of real per capita consumption growth, $x_{1,t+1}$, are 1.002 and 0.004 respectively. The mean and standard deviation of the asset series, $x_{2,t+1}$, are 1.008 and 0.050 respectively for EWR and 1.006 and 0.042 for VWR. So clearly all the series fluctuate approximately around one; most importantly consumption growth deviates very little from this value. The nonlinearity enters through the Euler equation residual

$$u_t(\theta) \;=\; \delta x_{1,t+1}^{\gamma-1} x_{2,t+1} - 1 \tag{3.70}$$

Now, if we replace $x_{1,t+1}$ and $x_{2,t+1}$ in (3.70) by their approximate means of one then we have

$$u_t(\theta) \;\approx\; \delta 1^{\gamma-1} \;-\; 1$$

This approximation can be set to zero by putting $\delta = 1$ regardless of the value of γ. Of course, the data exhibit some variation so that the approximation does not hold exactly. However it is close enough to give the flavour of the problem here: the population moment condition provides very good information about δ_0 but poor information about γ_0. This is an example of the case in which a parameter is *weakly identified* by the population moment condition. This situation occurs sufficiently frequently to have generated its own branch of GMM theory, and this is reviewed in Section 8.2.

Although we return to this model to illustrate other aspects of the GMM framework, this nevertheless seems the most appropriate place to mention briefly subsequent developments in the empirical literature on this topic. Since Hansen and Singleton's (1982) study there have been a number of papers which have

estimated the consumption based asset pricing model with more sophisticated utility functions; see Kocherlakota (1996) for a survey. However, empirical success has been limited. Many studies encounter the same problem as we did above: aggregate consumption data exhibits far less variation than asset returns and so cannot possibly explain how these assets are priced. This could mean the economic model is fundamentally wrong or that we have the wrong measure of consumption. The latter explanation has recently received some attention. Mankiw and Zeldes (1991) document that stocks are owned by approximately only thirty percent of the U.S. population and therefore aggregate consumption is unlikely to be a good proxy for the consumption of asset holders. Unfortunately, aggregate data for stockholders are unavailable. Hagiwara and Herce (1997) circumvent this problem by using aggregate dividends to proxy the consumption of asset holders and find this subsitution leads to far more reasonable empirical results. ◇

3.7 Transformations, Normalizations and the Continuous Updating GMM Estimator

So far in this chapter, we have treated the data and parameter vector as given. However, in practice, a researcher may have to make decisions about the scale of the data or the parameterization of the model or whether to transform $f(.)$ in some fashion. In this section, we consider the extent to which the GMM estimator is invariant to such decisions. It emerges that the estimator can be sensitive to these types of transformations, and this motivates both a variant of GMM known as the continuous updating estimator and also an alternative method for the calculation of confidence intervals. Both these extensions are discussed in this section.

To begin, it is useful to distinguish five types of transformation which are considered below.

- *Units of measurement for v_t*: In some cases a researcher must decide what units in which to measure the data. For example, any nominal value can be measured in \$'s, 1000\$'s or 1,000,000\$'s. The choice between them determines whether a price of one thousand dollars is recorded as 1000, 1 or 0.001, and so determines the scale of the data.

- *Reparameterization*: Suppose θ_0 is globally identified and $\theta_0 = h(\gamma_0)$ where $h : \Re^p \to \Re^p$ is a continuous, differentiable bijective mapping. In this case, the population moment condition can be reparameterized as $E[f(v_t, h(\gamma_0))] = E[f_\gamma(v_t, \gamma_0)] = 0$, and GMM can be used to estimate γ_0 based on $E[f_\gamma(v_t, \gamma_0)] = 0$ instead of θ_0 based on $E[f(v_t, \theta_0)] = 0$.

- *Normalization of the parameter vector*: In some cases, θ_0 may only be identified up to some scaling factor and so it is necessary to impose some normalization on θ_0, such as $\theta_{0,1} = 1$, in order to achieve identification.

- *Curvature altering transformations of the population moment condition*:[61] In some cases, the objective function may be ill-behaved and researchers have found it advantageous to scale the population moment condition by some function of the θ_0. In other words, estimation of θ_0 is based on $c(\theta_0)E[f(v_t, \theta_0)] = 0$.

- *Stationarity inducing transformations*: In some cases, the underlying model may imply $E[h(\tilde{v}_t, \theta_0)] = 0$ in which $\tilde{v}_t$ is a vector of nonstationary variables. Such a specification is outside our framework because Assumption 3.1 is violated. However, it may be possible to find a nonsingular matrix $H(\tilde{v}_{t-1}, \theta_0)$ say, such that $H(\tilde{v}_{t-1}, \theta_0)h(\tilde{v}_t, \theta_0) = f(v_t, \theta_0)$ where v_t is a vector of stationary random variables, and $E[f(v_t, \theta_0)] = 0$. In this case, GMM estimation can be based on the population moment condition $E[f(v_t, \theta_0)] = 0$.

Below we consider the impact of each type of transformation on the GMM estimator in turn.

The GMM Estimator and the Units of Measurement for v_t

In general, the GMM estimator is not invariant to changes in the units of measurement of v_t. A simple example illustrates. Let v_t be a scalar random variable with unknown population mean θ_0. This definition implies that,

$$E[v_t] \; - \; \theta_0 \; = \; 0 \tag{3.71}$$

Since θ_0 is just identified by (3.71), the GMM estimator is just the Method of Moments estimator which, in turn, is $\hat{\theta}_T = T^{-1} \sum_{t=1}^{T} v_t$. Now suppose v_t is replaced by $x_t = cv_t$ in (3.71) for some non–zero, finite constant c. The resulting GMM estimator of θ_0 is $\tilde{\theta}_T = T^{-1} \sum_{t=1}^{T} x_t$. It is easily verified that $\tilde{\theta}_T = c\hat{\theta}_T$, and so the GMM estimator is not invariant to changes in the scale of the data. However, this lack of invariance is a strength rather than a weakness because the scaling of the data has changed the interpretation of the parameter θ_0. In one case, it is the population mean of v_t and in the other, it is the population mean of $x_t = cv_t$.

It is important to realize that the lack of invariance applies to scale changes in v_t, that is to the random variables which appear in the population moment condition. In some cases, v_t may itself be a function of a set of underlying variables and changes in the units of these variables may or may not have an impact on the scale of v_t. For example in Hansen and Singleton's (1982) consumption based asset pricing model, v_t is defined to be $(c_{t+1}/c_t, r_{t+1}/p_t)$. In this case, since the elements of v_t are ratios, changes in the units of c_t or asset prices (with commensurate changes in the returns) have no impact on v_t, and hence no impact on the GMM estimator. $\diamond$

[61] This type of transformation is sometimes refered to as "normalization" of the population moment condition. However, we eschew this terminology to avoid confusion with the concept of normalization of the parameter vector.

The GMM Estimator and Reparameterization

The GMM estimator is invariant to reparameterization in the sense that the two parameterizations yield logically consistent estimators. However, a similar result does not extend to the estimated asymptotic standard errors, and so inferences may be sensitive to the choice of parameterization. These two statements are now justified in turn.

Let $Q_{\gamma,T}(\gamma)$ be the GMM minimand associated with the reparameterized model, that is $Q_{\gamma,T}(\gamma) = Q_T(h(\gamma))$, and $\hat{\gamma}_T = argmin\, Q_{\gamma,T}(\gamma)$. Given the properties of $h(.)$ stated above, it is possible to calculate $\hat{\gamma}_T$ as follows. First, $Q_T(h(\gamma))$ can be minimized with respect to $h(\gamma)$ to yield $\hat{h}_T$, say. Then, $\hat{h}_T = h(\hat{\gamma}_T)$ can be solved to yield a unique value for $\hat{\gamma}_T$. It is easily recognized that $\hat{h}_T = \hat{\theta}_T$ and so by construction

$$\hat{\theta}_T \;=\; h(\hat{\gamma}_T) \tag{3.72}$$

Therefore the two estimators are logically consistent. However, the same cannot be said for inferences based on the estimator, as we now show.

It can be recalled from the discussion following (3.27) that the estimated asymptotic standard errors of $\hat{\theta}_T$ are the square roots of the diagonal elements of the matrix,

$$\begin{aligned}\hat{V}_{\theta,T} \;&=\; [G_T(\hat{\theta}_T)'W_T G_T(\hat{\theta}_T)]^{-1} G_T(\hat{\theta}_T)'W_T \hat{S}_T W_T G_T(\hat{\theta}_T) \\ &\quad \times [G_T(\hat{\theta}_T)'W_T G_T(\hat{\theta}_T)]^{-1}\end{aligned} \tag{3.73}$$

Similar arguments imply that the corresponding matrix for $\hat{\gamma}_T$ is given by

$$\begin{aligned}\hat{V}_{\gamma,T} \;&=\; [G_{\gamma,T}(\hat{\gamma}_T)'W_T G_{\gamma,T}(\hat{\gamma}_T)]^{-1} G_{\gamma,T}(\hat{\gamma}_T)'W_T \hat{S}_{\gamma,T} W_T G_{\gamma,T}(\hat{\gamma}_T) \\ &\quad \times [G_{\gamma,T}(\hat{\gamma}_T)'W_T G_{\gamma,T}(\hat{\gamma}_T)]^{-1}\end{aligned} \tag{3.74}$$

where $G_{\gamma,T}(.)$, and $S_{\gamma,T}$ are the analogs of $G_T(.)$, $\hat{S}_T$ only defined in terms of $f_\gamma(.)$ instead of $f(.)$. Intuition suggests that these two matrices should be related, and they are. To see how, note that (3.72) implies $f(v_t, \hat{\theta}_T) = f_\gamma(v_t, \hat{\gamma}_T)$, and hence that $\hat{S}_T = \hat{S}_{\gamma,T}$ – assuming the same generic covariance matrix estimator is used in each case, of course. Furthermore, by the Chain rule

$$\partial f_\gamma(.)/\partial\gamma' \;=\; \{\partial f(.)/\partial\theta'\}\, \partial h(.)/\partial\gamma'$$

and so, using (3.72), it follows that

$$G_{\gamma,T}(\hat{\gamma}_T) \;=\; G_T(\hat{\theta}_T) H(\hat{\gamma}_T) \tag{3.75}$$

where $H(.) = \partial h(.)/\partial\gamma'$. Collecting these results together and making the appropriate substitutions into (3.74), it can be shown that

$$\hat{V}_{\gamma,T} \;=\; [H(\hat{\gamma}_T)]^{-1} \hat{V}_{\theta,T} [H(\hat{\gamma}_T)']^{-1} \tag{3.76}$$

To illustrate how reparameterization may affect inferences, it suffices to take a simple example. Suppose $p = 1$ and $h(\gamma) = \gamma^3$. The asymptotic confidence interval for γ_0 is

$$\hat{\gamma}_T \;\pm\; z_{\alpha/2} \sqrt{\hat{V}_{\gamma,T}/T} \tag{3.77}$$

Since $\theta = \gamma^3$ and (3.76) holds with $H(\hat{\gamma}_T) = 3\hat{\gamma}_T^2$, it follows that (3.77) implies the following interval for θ_0

$$\left(\left\{ \hat{\gamma}_T - z_{\alpha/2}(3\hat{\gamma}_T^2)^{-1}\sqrt{\hat{V}_{\theta,T}/T} \right\}^3 , \left\{ \hat{\gamma}_T + z_{\alpha/2}(3\hat{\gamma}_T^2)^{-1}\sqrt{\hat{V}_{\theta,T}/T} \right\}^3 \right) \quad (3.78)$$

In contrast, the asymptotic confidence interval based upon $\hat{\theta}_T$ directly is

$$\hat{\theta}_T \pm z_{\alpha/2}\sqrt{\hat{V}_{\theta,T}/T} \quad (3.79)$$

In general, there is no reason why the intervals in (3.78) and (3.79) should be equal.

This sensitivity is a potential source of concern, and motivates an alternative method for the construction of confidence intervals that is discussed later in this section. However, it is worth noting one defence of the intervals described above. It can be argued that many economic models imply a "natural parameterization" and so this is the only parameterization of interest. For example, in Hansen and Singleton's (1982) consumption based asset pricing model, there are two aspects of the agents behaviour which are crucial for the model: his/her discount factor and coefficient of relative risk aversion. In our presentation in Section 1.3.1, these two aspects of the model are captured directly by unknown parameters (δ_0, γ_0). Alternatively, the model could have been parameterized so that the discount factor and risk aversion are captured by $h_1(\eta_1)$ and $h_2(\eta_2)$ say, for some prespecified functions $h_i(.)$ of unknown parameters (η_1, η_2). However, in this second approach the unknown parameters have no meaningful economic interpretation. So the first parameterization is argued to be the "natural" one for this model and the second, by implication, to be "unnatural". While this argument may not find universal favour, it is certainly the case that published studies tend to employ the natural parameterization. ◇

The GMM Estimator and Normalization of the Parameter Vector

In general, the GMM estimators associated with different normalizations of the parameter vector do not exhibit a logical consistency in finite samples. However, they do exhibit a logical consistency in the limit.

This particular issue has been the focus of some attention in the literature on the use of the linear quadratic model for inventory holdings, and this setting provides a convenient framework for our discussion. Several papers have contributed to this part of the literature but our discussion is based on Fuhrer, Moore, and Schuh (1995).[62] The model has essentially the same structure as the one described in Section 1.3.4 except that now the cost functions take the form,

$$\begin{aligned} C_{Q_t} &= (\theta_{0,1}/2)Q_t^2 + (\theta_{0,2}/2)(Q_t - Q_{t-1})^2 \\ C_{I_t} &= (\theta_{0,3}/2)(I_t - \omega_0 I_{t-1})^2 \end{aligned}$$

[62] The interested reader is refered to Fuhrer, Moore, and Schuh (1995) or Blinder and Maccini (1991) for the appropriate references.

With these definitions, the Euler equation becomes

$$\begin{aligned} E[\theta_{0,1}(Q_t - \beta_0 Q_{t+1}) &+ \theta_{0,2}(\Delta Q_t - 2\beta_0 \Delta Q_{t+1} + \beta_0^2 \Delta Q_{t+2}) \\ &+ \theta_{0,3} I_t + \theta_{0,4} S_t \,|\, \Omega_t] = 0 \end{aligned} \tag{3.80}$$

where β_0 and Ω_t denote the discount factor and information set at time t respectively (as in Section 1.3.4), Δ denotes the difference operator,[63] and we have set $\theta_{0,4} = \theta_{0,3}\omega_0$. It is common in the literature on this model to fix the value for β_0 *a priori* because then the Euler equation is linear in both the parameters and variables. We follow this practice, and so the Euler equation can be written more compactly as,

$$E[e_t(\theta_0) \,|\, \Omega_t] = 0 \tag{3.81}$$

where

$$e_t(\theta) = \theta_1 R_{1,t} + \theta_2 R_{2,t} + \theta_3 I_t + \theta_4 S_t \tag{3.82}$$

and we have set $R_{1,t} = (Q_t - \beta_0 Q_{t+1})$, $R_{2,t} = (\Delta Q_t - 2\beta_0 \Delta Q_{t+1} + \beta_0^2 \Delta Q_{t+2})$, and $\theta_0 = (\theta_{0,1}, \theta_{0,2}, \theta_{0,3}, \theta_{0,4})$. Using similar argument to (1.23), it follows from (3.81) that

$$E[z_t e_t(\theta_0)] = 0 \tag{3.83}$$

for any $z_t \in \Omega_t$.

Ideally (3.83) would form the basis for GMM estimation of θ_0. However, inspection of (3.82) reveals that θ_0 is not identified by this population moment condition: if (3.83) holds then so does $E[z_t e_t(\bar{\theta})] = 0$ for $\bar{\theta} = c\theta_0$ and any finite constant c. In other words, θ_0 is only identified up to a scaling factor. In the absence of any additional information about the parameters from the underlying economic theory, it is necessary to impose some arbitrary normalization on θ_0 in order to facilitate the estimation. For the purposes of exposition, we consider two such normalizations. First, suppose the elements of $e_t(\theta_0)$ are divided by $\theta_{0,1}$ to yield

$$\tilde{e}_t(\psi_0) = R_{1,t} + \psi_{0,1} R_{2,t} + \psi_{0,2} I_t + \psi_{0,3} S_t \tag{3.84}$$

where $\psi_{0,i} = \theta_{0,i+1}/\theta_{0,1}$. Secondly, suppose the elements of $e_t(\theta_0)$ are divided by $\theta_{0,4}$ to yield

$$\bar{e}_t(\phi_0) = \phi_{0,1} R_{1,t} + \phi_{0,2} R_{2,t} + \phi_{0,3} I_t + S_t \tag{3.85}$$

where $\phi_{0,i} = \theta_{0,i}/\theta_{0,4}$. Notice that both these normalizations of θ_0 are logically consistent in the sense that given ψ_0 it is possible to solve uniquely for ϕ_0 and *vice versa*.[64]

These normalizations lead to two different population moment conditions upon which estimation can be based,

$$E[z_t \tilde{e}_t(\psi_0)] = 0 \tag{3.86}$$

$$E[z_t \bar{e}_t(\phi_0)] = 0 \tag{3.87}$$

[63] That is $\Delta Q_t = Q_t - Q_{t-1}$.

[64] Specifically, the mapping between them is given by $\phi_1 = 1/\psi_3$, $\phi_2 = \psi_1/\psi_3$, $\phi_3 = \psi_2/\psi_3$ where it is assumed for simplicity that all coefficients are non-zero.

Since both population moment conditions have the linear structure considered in Chapter 2, we can appeal to that earlier analysis to deduce that ψ_0 is identified by (3.86) provided $rank\{E[z_t x_{1,t}^{'}]\} = 3$ where $x_{1,t} = (-R_{2,t}, -I_t, -S_t)'$, and ϕ_0 is identified by (3.87) provided $rank\{E[z_t x_{2,t}^{'}]\} = 3$ where $x_{2,t} = (-R_{1,t}, -R_{2,t}, -I_t)'$. The form of the estimators is given by (2.8), that is

$$\begin{aligned}\hat{\psi}_T &= \left[(T^{-1}\sum_{t=1}^{T} x_{1,t}z_t^{'})W_T(T^{-1}\sum_{t=1}^{T} z_t x_{1,t}^{'})\right]^{-1} \\ &\quad \times (T^{-1}\sum_{t=1}^{T} x_{1,t}z_t^{'})W_T(T^{-1}\sum_{t=1}^{T} z_t R_{1,t}) \qquad (3.88)\\ \hat{\phi}_T &= \left[(T^{-1}\sum_{t=1}^{T} x_{2,t}z_t^{'})W_T(T^{-1}\sum_{t=1}^{T} z_t x_{2,t}^{'})\right]^{-1} \\ &\quad \times (T^{-1}\sum_{t=1}^{T} x_{2,t}z_t^{'})W_T(T^{-1}\sum_{t=1}^{T} z_t S_t) \qquad (3.89)\end{aligned}$$

It is remarked above that the two normalizations of θ_0 are logically consistent. Since $\hat{\psi}_T \xrightarrow{p} \psi_0$ and $\hat{\phi}_T \xrightarrow{p} \phi_0$, the estimators must exhibit a similar logical consistency *in the limit*. However, there is no reason for $\hat{\psi}_T$ and $\hat{\phi}_T$ to exhibit this property in finite samples. For example, even though the model implies $\psi_{0,1}/\psi_{0,2} = \phi_{0,2}/\phi_{0,3}$, the corresponding estimators in (3.88)–(3.89) do not exhibit this property, that is $\hat{\psi}_{T,1}/\hat{\psi}_{T,2} \neq \hat{\phi}_{T,2}/\hat{\phi}_{T,3}$ in general. Fuhrer, Moore, and Schuh (1995) provide empirical evidence that the estimators of inventory models can be very sensitive to the choice of normalization. Further evidence is provided by the simulation study reported in West and Wilcox (1994). ◇

The GMM Estimator and Curvature Altering Transformations of the Population Moment Condition

The GMM estimator is invariant to curvature altering transformations of the population moment condition if the parameter vector is just identified; however, if the parameter vector is overidentified then it only exhibits this property in the limit.

We begin with the just identified case, that is $p = q$. Suppose that GMM estimation is to be based upon the transformed population moment condition,

$$c(\theta_0)E[f(v_t, \theta_0)] = 0 \qquad (3.90)$$

where $c(\theta_0)$ is a finite non–zero scalar.[65] Since $p = q$, the GMM estimator is just the Method of Moments estimator $\hat{\theta}_T$ obtained by solving the sample analog to

[65] For simplicity, we take $c(.)$ to be a scalar, but the same arguments go through if $c(\theta_0)$ is a $(p \times p)$ nonsingular matrix.

(3.90),

$$c(\hat{\theta}_T)T^{-1}\sum_{t=1}^{T} f(v_t, \hat{\theta}_T) \;=\; 0 \tag{3.91}$$

However, provided $c(\hat{\theta}_T)$ is finite and non-zero, (3.91) implies

$$T^{-1}\sum_{t=1}^{T} f(v_t, \hat{\theta}_T) \;=\; 0 \tag{3.92}$$

and so $\hat{\theta}_T$ is also the Method of Moments, and hence GMM, estimator of θ_0 based on $E[f(v_t, \theta_0)] = 0$.[66]

However, if $q > p$ then the above argument does not go through because the first order conditions do not set the sample moment to zero. Specifically, the GMM estimator based on (3.90) is now the solution to

$$\left\{\left[\frac{\partial c(\hat{\theta}_T)}{\partial \theta}\right] T^{-1}\sum_{t=1}^{T} f(v_t, \hat{\theta}_T)' \;+\; c(\hat{\theta}_T)G_T(\hat{\theta}_T)'\right\} W_T T^{-1}\sum_{t=1}^{T} f(v_t, \hat{\theta}_T) \;=\; 0 \tag{3.93}$$

In general, the solution to (3.93) does not satisfy the first order conditions associated with GMM estimation based on the untransformed population moment condition given in (3.12).[67] However, since (3.90) holds, the estimator is consistent for θ_0 and so the transformation does not affect the probability limit of the estimator.

As mentioned above, this type of transformation is employed when the minimand is ill-behaved making estimation difficult. Such a problem occurs in Eichenbaum's (1989) inventory model described in Section 1.3.3. and we illustrate this type of transformation in Section 9.3 as part of our empirical investigation of this model. ◇

Stationarity Inducing Transformations

If it is possible to find one stationarity inducing transformation of $f(.)$ then there are infinitely many such transformations. In general, the GMM estimator is sensitive to the choice of transformation in finite samples, but is consistent no matter which transformation is used.

These statements are most easily substantiated in the context of a specific example. To this end, we consider the consumption based asset pricing model described in Section 1.3.1 and, to simplify the discussion, focus on the specification used in our empirical implementation.[68]

[66] Note that if the entire population moment condition in (3.71) is scaled by c, instead of just scaling v_t, then the resulting GMM estimator is invariant to the choice of c.

[67] The reader should be alerted to an abuse of notation in making the comparison between these two equations. In Section 3.2, $\hat{\theta}_T$ is defined to be the solution to (3.12). In the current paragraph, $\hat{\theta}_T$ has been used to denote the solution to (3.93).

[68] That is with only one asset with a maturity of one period, and the constant relative risk aversion utility function given in (1.21).

To begin, it is useful to revisit the derivation of the population moment condition in Section 1.3.1 because the steps taken involve the implicit use of a stationarity inducing transformation. It can be recalled from this earlier discussion that the derivation of the population moment condition began with a characterization of the optimal path for consumption in (1.19). Under the conditions given above, this equation reduces to

$$p_t c_t^{\gamma_0 - 1} = \delta_0 E[r_{t+1} c_{t+1}^{\gamma_0 - 1} | \Omega_t] \tag{3.94}$$

From this starting point, we proceeded as follows. Since $p_t c_t^{\gamma_0 - 1} \in \Omega_t$, both sides of this equation were divided by $p_t c_t^{\gamma_0 - 1}$ to give

$$E[\delta_0 (r_{t+1}/p_t)(c_{t+1}/c_t)^{\gamma_0 - 1} - 1 | \Omega_t] = 0 \tag{3.95}$$

and then the population moment condition is deduced from (3.95) using an iterated expectations argument. However, we could have taken another approach. Equation (3.94) can be rewritten as

$$E[\delta_0 r_{t+1} c_{t+1}^{\gamma_0 - 1} - p_t c_t^{\gamma_0 - 1} | \Omega_t] = 0 \tag{3.96}$$

It is then possible to use the same iterated expectations argument to deduce population moment conditions based (3.96). Why use the first approach and not the second? The answer is simple. Population moment conditions based on (3.95) involve $x_{1,t+1} = c_{t+1}/c_t$ and $x_{2,t+1} = r_{t+1}/p_t$, both of which are stationary random variables. Whereas moment conditions deduced from (3.96) involve functions of the nonstationary variables (c_t, r_t, p_t).

While the choice between (3.95) and (3.96) may be clear cut. It should be noted that the stationarity inducing transformation used in (3.95) is not unique. For example, let $w_t \in \Omega_t$ be a stationary random variable, then division of (3.94) by $w_t p_t c_t^{\gamma_0 - 1}$ yields

$$E[\delta_0 w_t^{-1} (r_{t+1}/p_t)(c_{t+1}/c_t)^{\gamma_0 - 1} - w_t^{-1} | \Omega_t] = 0 \tag{3.97}$$

which can also form the basis for population moment conditions involving stationary random variables. It follows, therefore, that there are an infinite number of stationarity inducing transformations. The impact of the choice of w_t is most easily understood by considering the moment condition upon which estimation is ultimately based. It can be recalled from Section 1.3.1 that an iterated expectations argument is used to deduce $E[u_t(\theta_0) z_t] = 0$ where $u_t(\theta_0) = \delta_0 (x_{2,t+1} x_{1,t+1}^{\gamma_0 - 1} - 1)$. If the same argument is used starting from (3.97) then the resulting moment condition is simply $E[u_t(\theta_0) \tilde{z}_t] = 0$ where $\tilde{z}_t = w_t^{-1} z_t$. Therefore, the components in the stationarity inducing transformation play different roles: division by $p_t c_t^{\gamma_0 - 1}$ actually induces stationarity and w_t simply scales the instrument vector. From this perspective, it is immediately apparent that the resulting estimator is not invariant in finite samples to the choice of

stationarity inducing transformation, but is nevertheless consistent for θ_0 for any suitable choice of w_t.[69] $\diamond$

It is clear that either the estimator or subsequent inferences can be sensitive to the types of transformation considered above. The sensitivity of the estimator is particularly unappealing in the last three cases because there is clearly an arbitrariness to the specific normalization or transformation chosen. It is possible to modify the GMM minimand to produce an estimator which is invariant to curvature altering transformations. This version is known as the *continuous updating GMM estimator* and is considered below. The sensitivity of inferences to reparameterization may be viewed as a less serious problem because of the "natural parameterization" argument. However, the latter view is not universally accepted and so we explore an alternative method for the construction of confidence intervals. We now describe both these remedies in turn.

The continuous updating GMM estimator was introduced by Hansen, Heaton, and Yaron (1996). The motivation for this estimator is best understood by considering the population analog to the GMM minimand with the optimal weighting matrix. It can be recalled from Theorem 3.4 that the optimal choice of W is S^{-1}. For our purposes here, it is important to note that this choice of weighting matrix depends on θ_0 because $S = \lim_{T\to\infty} Var[T^{1/2} g_T(\theta_0)]$, and to emphasize this dependence we now write $S = S(\theta_0)$. Using this notation, the population analog to the GMM minimand is

$$Q_{pop}(\theta) \;=\; E[f(v_t,\theta)]' S(\theta)^{-1} E[f(v_t,\theta)] \tag{3.98}$$

Notice that both the population moment condition and weighting matrix depend on θ_0. However, since the consistency of the estimator depends crucially on $E[f(v_t,\theta_0)] = 0$ and not on $S(\theta_0)^{-1}$, we have treated the dependencies of $f(.)$ and $S(.)$ on θ differently so far. In the iterated estimation, a preliminary estimator of θ_0 is used to construct the weighting matrix and hence to eliminate the argument from the weighting matrix so that the minimand takes the form

$$Q_{iter,T}(\theta) \;=\; g_T(\theta)' \hat{S}_T(i-1)^{-1} g_T(\theta) \tag{3.99}$$

While this is approach is perfectly reasonable, it is not the only one possible. An alternative is to acknowlege the dependence of S on θ in the minimization and hence define the minmand to be

$$Q_{cont,T}(\theta) \;=\; g_T(\theta)' S_T(\theta)^{-1} g_T(\theta) \tag{3.100}$$

where

$$S_T(\theta) \;=\; \Gamma_{0,T}(\theta) \;+\; \sum_{i=1}^{T-1} \omega_{i,T} [\, \Gamma_{i,T}(\theta) \;+\; \Gamma_{i,T}(\theta)' \,] \tag{3.101}$$

[69] If the Euler equation is linear in the variables then it is possible to argue that an analogous conditional moment restriction is satisfied by the detrended variables. See Section 9.3 for further discussion of this approach to inducing stationarity.

where $\Gamma_{i,T}(\theta) = T^{-1}\sum_{t=i+1}^{T} f(v_t,\theta)f(v_{t-i},\theta)'$. Notice that $S_T(\theta)$ has the generic form of the HAC estimators discussed in Section 3.5.3 and so $S_T(\theta_0) \xrightarrow{p} S$ under appropriate conditions upon the dynamic structure of $f(v_t,\theta_0)$ and kernel, $\omega_{i,T}$. The *continuous updating GMM estimator* is defined to be,

$$\hat{\theta}_{cont,T} = argmin_{\theta\in\Theta} Q_{cont,T}(\theta) \tag{3.102}$$

Intuition suggests that the continuous updating estimator exhibits the same asymptotic properties as the two step or iterated estimator, and this is the case. This can be established using similar arguments to the proofs of Theorems 3.1 and 3.2 and is left to the reader. Although the iterated and continuous updating estimators have the same asymptotic distributions, they are typically different in finite samples. The first order conditions for the iterated estimation are given by (3.12) with $W_T = \hat{S}_T(i-1)^{-1}$, those for the continuous updating estimator are given by,[70]

$$\begin{aligned} 2G_T(\hat{\theta}_T)'S_T(\hat{\theta}_T)^{-1}g_T(\hat{\theta}_T) \quad &- \quad \left\{\frac{\partial vec[S_T(\hat{\theta}_T)]}{\partial\theta'}\right\}' [S_T(\hat{\theta}_T)^{-1}\otimes S_T(\hat{\theta}_T)^{-1}] \\ &\times \quad vec[g_T(\hat{\theta}_T)g_T(\hat{\theta}_T)'] = 0 \end{aligned} \tag{3.103}$$

A comparison of the two sets of equations indicates that the first order conditions for the continuous updating estimator contain an additional term due to the presence of the argument in the weighting matrix. To make this second term explicit, it is necessary to substitute in the appropriate formula for $\partial S(\theta)/\partial\theta'$. For our purposes here, it is sufficient to restrict attention to the case in which $\omega_{i,T} = 0$, that is in which the long run variance is estimated under the assumption that $f(v_t,\theta_0)$ is a serially uncorrelated process. In this case, it can be shown that

$$\frac{\partial vec[S_T(\theta)]}{\partial\theta'} = T^{-1}\sum_{t=1}^{T}\{[I_q\otimes f(v_t,\theta)] + [f(v_t,\theta)\otimes I_q]\}\frac{\partial f(v_t,\theta)}{\partial\theta'} \tag{3.104}$$

In general, there is no reason why the solutions to (3.12) and (3.103) should coincide for finite T. However, it can be verified that both sets of equations are satisfied by θ_0 in the limit.

The chief advantage of the continuous updating estimator is that it is invariant to curvature altering transformations of $f(v_t,\theta)$. To illustrate, consider again the situation described above in which the population moment condition is multiplied by $c(\theta_0)$, and so estimation is based on (3.90). The key difference now is that $Q_{cont,T}(\theta)$ depends on θ via both the sample moment *and* the inverse of the covariance matrix. After the transformation the sample moment is $c(\theta)g_T(\theta)$ and the inverse of the covariance matrix is $c(\theta)^{-2}S_T(\theta)^{-1}$. Once these terms are substituted into the minimand in (3.100), it is easily verified that the

[70] These equations can be derived using Dhrymes (1984) [Proposition 99, p.115; Proposition 106, p.124].

factors involving $c(\theta)$ cancel out, and so the estimator is unaffected by this type of transformation. In some cases, different elements of the population moment may be transformed by different functions of θ, and the previous argument is easily extended to cover this case and also the more general scenario in which $f(v_t, \theta)$ is premultiplied by any nonsingular matrix $C(\theta)$.

It is important to realize that the invariance of the continuous updating estimator is *only* with respect to curvature altering transformations of the population moment condition. However, there are cases in which the net effect of one of the other types of transformation is to premultiply the population moment by some nonsingular matrix $C(\theta)$, and so the continuous updating estimator is invariant to these types of transformation in such cases as well.

Example: Hansen and Singleton's (1982) Consumption Based Asset Pricing Model

Table 3.9 contains the continuous updating estimates, their standard errors and 95% confidence intervals for the parameters for both choices of assets. The starting values for the estimations are the iterated estimates reported in Table 3.7. This choice is made for two reasons. First, if the model is correctly specified, then the iterated estimator is consistent for θ_0 which is a solution to the first order conditions in (3.103) in the limit. Secondly, in this example, this choice of starting value initiates the minimization in an area within which the minimand is reasonably well behaved. In contrast to our experience with the iterated estimator, the results are very sensitive to the choice of starting value. In particular, for certain starting values, the numerical optimization routine diverges into parts of the parameter space clearly not in the neighbourhood of the global minimum of $TQ_{cont,T}(\theta)$. A similar experience is reported by Hansen, Heaton, and Yaron (1996) in the context of slightly more sophisticated versions of the consumption based asset pricing model. These differing experiences can be explained by considering the surface of the minimands with the VWR data. Figure 3.3 plots the second step minimand based on the first step estimates calculated using $W_T = 10^5 I_5$ with the VWR data. A comparison with Figure 3.2 indicates the minimand has the same valley like shape on both first and second steps. In contrast, the surface of the continuous updating minimand has a ravine in which the minimum is located as shown in Figure 3.4. The minimum is far harder to locate in the latter case particularly as the surface is relatively flat around the ravine.

With VWR, the results are very similar, although not identical, to those reported for the iterated estimator. With EWR, the only noticable difference between the estimation results is in the estimate of γ_0: continuous updating GMM yields 0.515 for this parameter, as opposed to -0.343 with iterated GMM. Notwithstanding this difference, the results are qualitatively the same from the iterated and continuous GMM estimations. In both cases, δ_0 is precisely

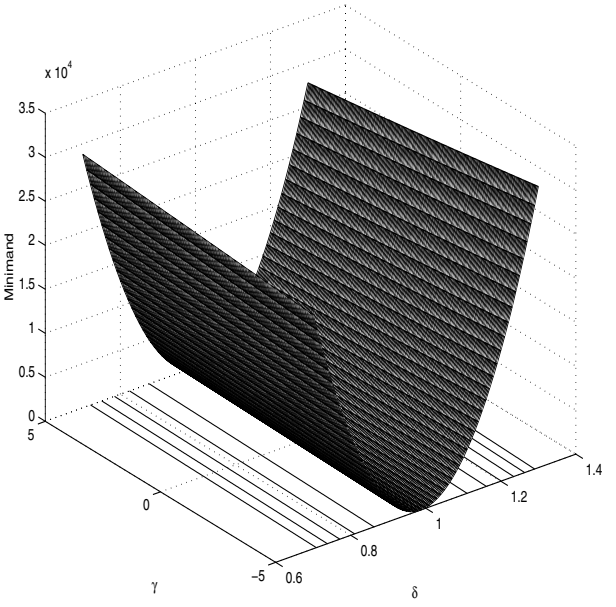

Figure 3.3: Second-step GMM minimand for the consumption based asset pricing model with value weighted returns

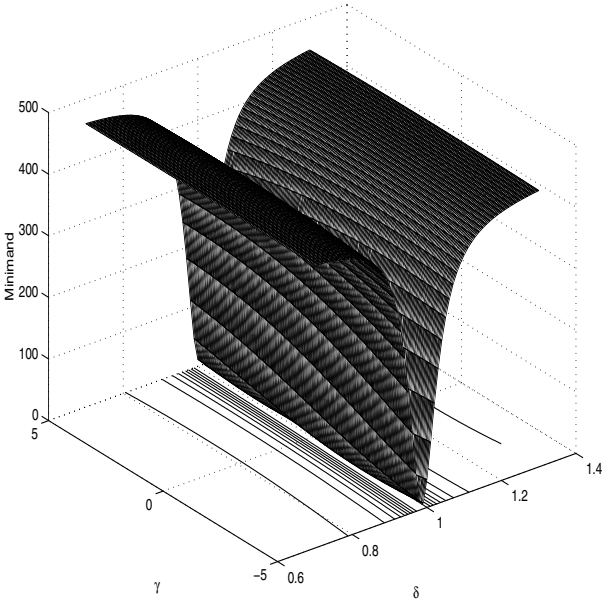

Figure 3.4: Continuous Updating GMM minimand for the consumption based asset pricing model with value weighted returns

estimated but γ_0 is not. Furthermore, the source of this imprecision is the same here as it is in the iterated estimation: γ_0 is weakly identified by the population moment condition associated with continuous updating GMM. $\diamond$

Table 3.9
Continuous Updating GMM estimation results for the consumption based asset pricing model

EWR: $(\hat{\gamma}_T, \hat{\delta}_T)$	$s.e.(\hat{\gamma}_T)$	$c.i.(\hat{\gamma}_T)$	$s.e.(\hat{\delta}_T)$	$c.i.(\hat{\delta}_T)$
$(0.515, 0.990)$	2.229	$(-3.853, 4.884)$	0.004	$(0.981, 0.998)$
VWR: $(\hat{\gamma}_T, \hat{\delta}_T)$	$s.e.(\hat{\gamma}_T)$	$c.i.(\hat{\gamma}_T)$	$s.e.(\hat{\delta}_T)$	$c.i.(\hat{\delta}_T)$
$(0.785, 0.993)$	1.829	$(-2.801, 4.370)$	0.004	$(0.986, 1.000)$

Notes: $s.e.(.)$ denotes the standard error calculated using (3.59) with $W_T = \hat{S}_{SU}^{-1}$, $\hat{S}_T = \hat{S}_{SU}^{-1}$ and $\hat{S}_{SU}$ is defined in (3.40). $c.i.(.)$ denotes the 95% confidence interval calculated using (3.27).

It is the form of the minimand that gives the continuous updating estimator its invariance to normalization of the population moment condition. It is this minimand which also provides the key to the construction of asymptotic confidence sets which are invariant to reparameterization. We use the term "confidence set" because the approach described below is based on a probability statement involving θ_0 rather than statements involving its individual elements. This approach was first introduced into the GMM literature by Stock and Wright (1995, 2000) although in the context of a different problem. Stock and Wright are concerned with the problem of inference in the presence of weakly identified parameters, and we discuss this approach to inference in that context in Section 8.2. For the present, we focus purely on the construction of confidence sets which are invariant to reparameterization.[71]

To derive these confidence sets, it is necessary to consider the limiting distribution of $TQ_{cont,T}(\theta_0)$. This distribution follows straightforwardly from the limiting behaviour of its components, $T^{1/2}g_T(\theta_0)$ and $S_T(\theta_0)^{-1}$. Under the conditions of Lemma 3.2, it follows that $T^{1/2}g_T(\theta_0) \xrightarrow{d} N(0, S)$. If it is also

[71] In the weak identification literature, these confidence sets are sometimes refered as S-sets, a terminology inpsired by the notation used by Stock and Wright (2000).

assumed that $S_T(\theta_0) \xrightarrow{p} S$, then $S_T(\theta_0)^{-1} \xrightarrow{p} S^{-1}$.[72] Combining these two results, it follows that

$$TQ_{cont,T}(\theta_0) \xrightarrow{d} \chi^2_q \tag{3.105}$$

An asymptotically valid $100(1-\alpha)\%$ confidence set for θ_0 is then given by

$$\{\, \theta : TQ_{cont,T}(\theta) < c_q(\alpha) \,\} \tag{3.106}$$

where $c_q(\alpha)$ is the $100(1-\alpha)\%$ percentile of the χ^2_q distribution. In other words, the confidence sets in (3.106) consist of all values of θ for which the minimand of the continuous updating GMM estimator does not exceed the appropriate percentile of the limiting distribution of $TQ_{cont,T}(\theta_0)$. It is easily recognized that our earlier arguments about the invariance of the estimator to reparameterization can be applied here to show that the confidence sets in (3.106) exhibit the same invariance property. This confidence set is illustrated below for our running example. In that particular case, the calculations are relatively straightforward because θ_0 is only a (2×1) vector. However, the computational burden increases rapidly with p and quickly becomes prohibitive. Therefore, this method of calculating confidence sets can be infeasible in many cases of interest.

Example: Hansen and Singleton's (1982) Consumption Based Asset Pricing Model

It can be recalled that the model has been estimated for two types of asset, VWR and EWR. As it turns out, these two cases provide a good illustration of a fundamental difference between the confidence sets and marginal intervals reported earlier. By construction the marginal intervals in (3.27) are non-empty. However, it is entirely possible for the confidence set in (3.106) to contain no elements, and this is exactly what happens when the model is estimated with EWR. Such a phenomenon provides evidence that the model is misspecified. We come to the same conclusion using model specification tests in Section 5.1, and delay further discussion of this outcome until then. Instead we focus here on the case in which the model is estimated with VWR. For this case, the 95% confidence set for (δ_0, γ_0) consists of all points within the ellipse plotted in Figure 3.5. This confidence set is clearly more informative than the marginal intervals reported in Tables 3.8 and 3.9 because it reveals a connection between the plausible values for γ_0 and δ_0. In general terms, higher values of γ_0 in the set are associated with smaller values of δ_0 and *vice versa*. However, in one sense the confidence set and marginal confidence intervals are similar: they both imply δ_0 is estimated very precisely but γ_0 is not. $\diamond$

[72] The reader is refered to the references given in Section 3.5.3 for appropriate regularity conditions for this result to hold.

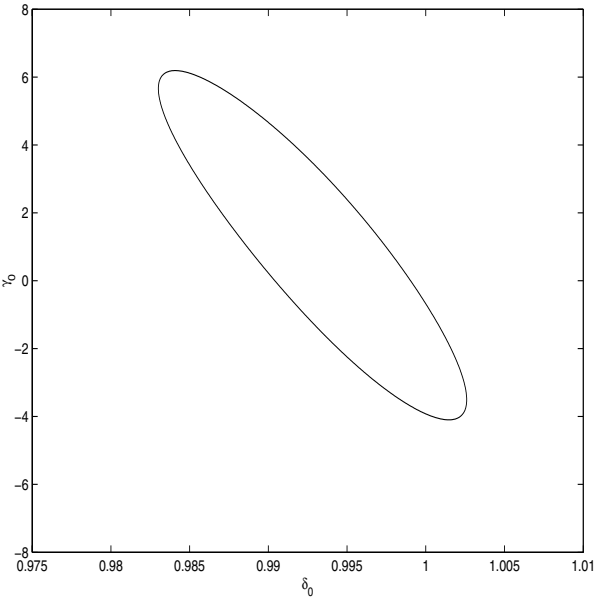

Figure 3.5: 95% confidence set for θ_0 in the consumption based asset pricing model with value weighted returns

3.8 GMM as a Unifying Principle of Estimation

It is stated in Chapter 1 that GMM provided a unifying framework for the analysis of many econometric estimators. At that point it was only possible to provide a few illustrations of this thesis but we are now in a position to elaborate further. So we conclude this chapter by describing how the GMM framework encompasses many other estimators derived using a seemingly different approach. This section covers material which is irrelevant to many of the applications listed in Table 1.1, and so some readers may wish to proceed to Chapter 4.

It is convenient to divide the discussion into two parts. First, we consider the case in which all the elements of θ_0 are estimated simultaneously. For reasons that will become apparent, we refer to such estimators as *single step*. This is the case upon which we have focused in the book so far. Then we consider the case of *sequential* estimators in which the elements of the parameter vector are estimated in stages.

3.8.1 Single Step Estimators

Many econometric estimators are obtained by optimizing a scalar of the form

$$\sum_{t=1}^{T} N_t(\theta) \tag{3.107}$$

Two leading examples are least squares and maximum likelihood, both of which we discuss in more detail below. If $N_t(\theta)$ is differentiable then the estimator, $\tilde{\theta}$, is the value which solves the associated first order conditions

$$\sum_{t=1}^{T} \partial N_t(\tilde{\theta})/\partial\theta \;=\; 0 \tag{3.108}$$

Equation (3.108) implies that $\tilde{\theta}$ is equivalent to the Method of Moments estimator based on the population moment condition

$$E[\partial N_t(\theta_0)/\partial\theta] \;=\; 0 \tag{3.109}$$

Since $\partial N_t(\theta)/\partial\theta$ is a $(p \times 1)$ vector it can be recalled from Section 3.3 that $\tilde{\theta}$ is also the GMM estimator based on (3.109).

As illustrations, we now derive the population moment condition implicit in the GMM interpretation of least squares and maximum likelihood estimation. A further example can be found in the next sub-section.

Example: Ordinary Least Squares Estimation in the Linear Regression Model
Suppose the static, linear regression model from Chapter 2 is estimated by ordinary least squares. Typically, this estimator is derived as the value of θ which minimizes the residual sum of squares. Within the terms of our discussion here, this involves

$$N_t(\theta) = (y_t - x_t'\theta)^2$$

Therefore, the OLS estimator can be interpreted as a GMM estimator based on the population moment condition

$$E[x_t(y_t - x_t'\theta)] \;=\; 0 \tag{3.110}$$

This condition states that the regressors and error are uncorrelated and is, of course, one of the assumptions of the "Classical regression model". $\diamond$

Example: Maximum Likelihood Estimation
Suppose the conditional probability density function of the continuous stationary random vector v_t given $\{v_{t-1}, v_{t-2}, \ldots\}$ is $p(v_t; \theta_0, V_{t-1})$ where $V_{t-1} = (v_{t-1}', v_{t-2}', \ldots v_{t-k}')$. The maximum likelihood estimator (MLE) of θ_0 based on the conditional log likelihood function is the value of θ which maximizes,

$$L_T(\theta) \;=\; \sum_{t=1}^{T} ln\{p(v_t; \theta, V_{t-1})\} \tag{3.111}$$

This fits within our framework with $N_t(\theta) = ln\{p(v_t;\theta,V_{t-1})\}$ and so the MLE can be interpreted as a GMM estimator based on the population moment condition

$$E\left[\partial ln\{\, p(v_t;\theta,V_{t-1})\,\}/\partial\theta\right] \;=\; 0 \tag{3.112}$$

$\diamond$

Since both OLS and MLE are derived from perfectly valid estimation principles in their own right, it is reasonable to question whether there is any value to this GMM interpretation. In fact there are two main advantages. First, the GMM interpretation focuses attention specifically on the information used in estimation; whereas this is often not apparent from the original derivation of the estimators. For example, the importance of (3.110) for OLS estimation only emerges in proofs of unbiasedness or consistency of the estimator. Secondly, this interpretation allows the asymptotic properties of a variety of seemingly different estimators to be deduced using the framework discussed in the previous sections. It is in this sense that we refer to GMM as a unifying principle of estimation. To illustrate both these advantages, we return to the case of Maximum Likelihood estimation.

Example: Maximum Likelihood Estimation (Continued)
It is argued in Chapter 1 that the dependence of MLE on the probability distribution was a major weakness in the types of nonlinear dynamic models in Table 1.1. This problem is more readily appreciated using the GMM interpretation of MLE. The above analysis indicates that the MLE is consistent if (3.112) is satisfied. In fact, this population moment condition is automatically satisfied if the distribution is correctly specified. It is useful to prove this result here because it provides a natural starting point for considering the consequences of misspecification.

By definition, a probability density function satisfies

$$\int_{\mathbf{V}} p(v_t;\theta_0,V_{t-1})dv_t' \;=\; 1 \tag{3.113}$$

where $\int_{\mathbf{V}}(.)dv_t'$ denotes integration with respect to v_t over the sample space $\mathbf{V}$. Differentiation of (3.113) yields

$$\frac{\partial}{\partial\theta}\left[\int_{\mathbf{V}} p(v_t;\theta_0,V_{t-1})dv_t'\right] \;=\; 0 \tag{3.114}$$

If $p(.)$ satisfies the relatively mild conditions for the reversal of the orders of differentiation and integration then (3.114) implies

$$\int_{\mathbf{V}} \{\partial p(v_t;\theta_0,V_{t-1})/\partial\theta\}dv_t' \;=\; 0$$

This equation can be rewritten as

$$\int_{\mathbf{V}} \frac{1}{p(v_t;\theta_0,V_{t-1})}\{\partial p(v_t;\theta_0,V_{t-1})/\partial\theta\}\; p(v_t;\theta_0,V_{t-1})\;\}dv_t' \;=\; 0 \qquad (3.115)$$

If the probability density function is correctly specified then (3.115) is identical to (3.112) because $\partial ln\{p(\theta)\}/\partial\theta = \{1/p(\theta)\}\partial p(\theta)/\partial\theta$, for any scalar function $p(.)$. However, notice that if $p(.)$ is no longer the true probability density function of v_t then (3.115) cannot be interpreted as an expectation and so does not imply (3.112).

Does this mean that (3.112) never holds if the distribution is misspecified? The answer is no, but once the possibility of misspecification is admitted then its theoretical justification disappears. This issue is best understood using an example. Consider again the consumption based asset pricing model we have used throughout this chapter. As mentioned in Section 1.3.1 the conditional distribution of $\tilde{x}_{t+1} = (\tilde{x}_{1,t+1},\tilde{x}_{2,t+1})' = (ln(x_{1,t+1}),ln(x_{2,t+1}))'$ is unknown but let us suppose it is assumed to be normal. To be consistent with the economic model, the likelihood must be maximized subject to the restriction $E[\delta x_{1,t+1}^{\gamma-1}x_{2,t+1}\,|\,\Omega_t] = 1$. Hansen and Singleton (1982) show that for this model one element of (3.112) is equivalent to the population moment condition

$$E\left[\;\left\{ln(\delta_0) + (\gamma_0-1)\tilde{x}_{1,t} + \tilde{x}_{2,t} + 0.5\{(\gamma_0-1)^2\sigma_{11} + \sigma_{22} + 2(\gamma_0-1)\sigma_{12}\}\right\}\right] \;=\; 0 \qquad (3.116)$$

where σ_{ij} is the $i-j^{th}$ element the conditional variance of $\tilde{x}_{t+1}$. Equation (3.116) holds if the conditional distribution of $\tilde{x}_{t+1}$ is normal. However, if the distribution has been misspecified then this condition can no longer be justified by the line of argument in (3.113)–(3.115). Furthermore, a comparison with (1.22) indicates that (3.116) is not implied by the Euler equation of the economic model. Therefore, if the distribution has been incorrectly specified then there is neither a statistical nor an economic justification for the moment condition upon which this Maximum Likelihood estimation is based. This motivates the use of GMM estimation based on population moment conditions implied by the economic model.

The problems here stem from the presence of nonlinear functions of endogenous variables in the population moment condition.[73] If this feature is not present, then (3.112) may hold for a wide class of plausible true probability distributions. So there are circumstances in which Maximum Likelihood is undertaken even though the distribution is unknown. In this case, it is refered to as *Quasi Maximum Likelihood estimation* (White, 1982) or *Pseudo Maximum Likelihood estimation* (Gourieroux, Monfort, and Trognon, 1984). Both these sets of authors derive the asymptotic distribution of the estimator. However, it can also be derived directly using the GMM framework. If it is assumed that (3.112) holds then Theorem 3.2 implies the suitably normalized *Quasi*–MLE converges to a normal distribution with mean zero and covariance matrix

[73] See Amemiya (1977) and Phillips (1982) for further discussion of this issue.

$(G_0'S^{-1}G_0)^{-1}$ where[74]

$$\begin{aligned} G_0 &= E[\partial^2 ln\{\, p(v_t;\theta_0, V_{t-1})\,\}/\partial\theta\partial\theta'] \\ S &= E[\{\partial ln[p(v_t;\theta_0, V_{t-1})]/\partial\theta\}\{\partial ln[p(v_t;\theta_0, V_{t-1})]/\partial\theta\}'] \end{aligned}$$

If the distribution of v_t is misspecified then no further reduction of the asymptotic covariance matrix is possible. However, if the distribution is correctly specified then the information matrix identity[75] implies $(G_0'S^{-1}G_0)^{-1}$ equals S^{-1} and so the GMM framework yields the familiar result from Maximum Likelihood theory. $\diamond$

3.8.2 Sequential Estimators

So far we have concentrated on the case in which all elements of θ_0 are estimated simultaneously. However, in some cases it is convenient to estimate θ_0 sequentially. In this section, it is shown that a class of sequential estimators are also special cases of GMM. We start with the general case and then illustrate the ideas using a model with generated regressors.

To introduce the basic idea, it is sufficient to focus on two step sequential estimation procedures. Accordingly, we partition the parameter vector into $\theta_0' = (\theta_{0,1}', \theta_{0,2}')$ where $\theta_{0,i}$ is $(p_i \times 1)$ vector. Suppose that in the first step, $\theta_{0,1}$ is estimated by GMM based on the population moment condition $E[f_1(v_t, \theta_{0,1})] = 0$ with weighting matrix $W_{1,T}$. Let this estimator be $\hat{\theta}_{1,T}$. Now suppose that in the second step, $\theta_{0,2}$ is estimated by GMM based on the $(p_2 \times 1)$ population moment condition $E[f_2(v_t, \theta_0)] = 0$ with $\hat{\theta}_{1,T}$ substituted for $\theta_{0,1}$. Notice that $\theta_{0,2}$ is just identified by $E[f_2(v_t, \theta_0)] = 0$ conditional on $\theta_{0,1}$ and so the weighting matrix plays no role in this estimation. Newey and McFadden (1994) show that this sequential estimation procedure is identical to the single step estimation of θ_0 via GMM based on $E[f(v_t, \theta_0)] = 0$ where

$$f(v_t, \theta_0) = \begin{bmatrix} f_1(v_t, \theta_{0,1}) \\ f_2(v_t, \theta_0) \end{bmatrix} \tag{3.117}$$

and the weighting matrix

$$W_T = \begin{bmatrix} W_{1,T} & 0 \\ 0 & W_{2,T} \end{bmatrix} \tag{3.118}$$

for any positive definite matrix $W_{2,T}$. At first glance this may seem surprising but there is a simple intuition behind the result. From (3.117) and (3.118) it follows that the minimand for the simultaneous estimation can be written as

$$\begin{aligned} Q_T(\theta) &= Q_{1,T}(\theta_1) + Q_{2,T}(\theta_1, \theta_2) \\ &= g_{1,T}(\theta_1)'W_{1,T}g_{1,T}(\theta_1) + g_{2,T}(\theta_1, \theta_2)'W_{2,T}g_{2,T}(\theta_1, \theta_2) \end{aligned}$$

[74] Notice that (3.115) implies $\partial ln[p(v_t;\theta_0, V_{t-1})]/\partial\theta$ is a martingale difference sequence with respect to V_{t-1}.

[75] For example, see White (1982).

where $g_{1,T}(\theta_1) = T^{-1}\sum_{t=1}^{T} f_1(v_t, \theta_1)$ and $g_{2,T}(\theta_1, \theta_2) = T^{-1}\sum_{t=1}^{T} f_2(v_t, \theta)$. Since $f_2(.)$ is $(p_2 \times 1)$, there always exists a value of θ_2 which sets $Q_{2,T}(\theta_1, \theta_2)$ to zero regardless of the value of θ_1. So the minimization of $Q_T(\theta)$ can be performed by first finding the value $\hat{\theta}_1$ which minimizes $Q_{1,T}(\theta_1)$ and then finding the value of θ_2 which sets $Q_{2,T}(\hat{\theta}_{1,T}, \theta_2)$ to zero. Clearly, this is just the sequential procedure described above.

It is important to notice that this argument only works for situations in which $f_2(.)$ is the same dimension as $\theta_{0,2}$. To illustrate what happens if this is violated, it is useful to denote the dimension of $f_2(.)$ by q_2. First consider the case where $q_2 < p_2$. This implies $\theta_{0,2}$ is unidentified by $E[f_2(v_t, \theta_0)] = 0$ conditional on $\theta_{0,1}$ and so θ_0 is unidentified. Now consider the case where $q_2 > p_2$. This time $\theta_{0,2}$ is over-identified by $E[f_2(v_t, \theta_0)] = 0$ conditional on $\theta_{0,1}$. Consequently, there is not generally a value of θ_2 which sets $Q_{2,T}(\theta_1, \theta_2)$ to zero for any value of θ_1. This means the value of θ_1 which minimizes $Q_T(\theta)$ is no longer the same as the value which minimizes $Q_{1,T}(\theta_1)$ alone.

In spite of this limitation, many sequential estimators are covered by these conditions. The main advantage of this GMM interpretation comes in the calculation of the correct asymptotic variance for $\hat{\theta}_{2,T}$. Since $\hat{\theta}_{2,T}$ is calculated conditional on $\hat{\theta}_{1,T}$, its asymptotic distribution must take account of the uncertainty inherent in the estimation of $\theta_{0,1}$. The correct distribution is typically not obvious when the estimator is viewed in its original sequential form. However, the GMM perspective allows the correct form of the distribution to be deduced immediately from Theorem 3.2. As an illustration, we consider a more general version of the partial adjustment model discussed in Section 3.1. Other examples can be found in Newey (1984) and Newey and McFadden (1994).

Example: A Partial Adjustment Model for Inventory Holdings
Hall and Rossana (1991) consider the following model for inventories

$$\begin{aligned} \Delta y_t &= \gamma_{0,0} + \gamma_{1,0} y_{t-1} + \gamma'_{2,0} x_{t-1} + \gamma_{3,0} w^e_{1,t} + \gamma_{4,0} w^e_{2,t} + u_t \\ u_t &= \rho_0 u_{t-1} + e_t \end{aligned}$$

where y_t are inventory holdings in period t, $\Delta y_t = y_t - y_{t-1}$, x_{t-1} is a vector containing the number of workers, the hours per production worker, materials, work in progress and unfilled orders in period $t-1$, $w^e_{1,t}$ is the expected new orders, and $w^e_{2,t}$ is expected material prices. All variables are in logs. The error term e_t is assumed to be independently and identically distributed with mean zero. If all the regressors are observed then the parameters can be estimated by nonlinear least squares. These estimators are defined to be

$$(\hat{\gamma}'_T, \hat{\rho}_T) = argmin_{(\gamma', \rho) \in \Gamma \times R} T^{-1} \sum_{t=1}^{T} \{e_t(\gamma, \rho)\}^2 \tag{3.119}$$

where $e_t(\gamma, \rho) = u_t(\gamma) - \rho u_{t-1}(\gamma)$, $u_t(\gamma) = \Delta y_t - (1, y_{t-1}, x'_{t-1}, w^e_{1,t}, w^e_{2,t})'\gamma$ and γ is the (9×1) vector of regression parameters. Unfortunately, neither of the

expected values are known at time t. To circumvent this problem, Hall and Rossana (1991) estimate these variables by their least squares predictions from the AR(12) models

$$w_{i,t} = \tilde{w}'_{i,t}\beta_{i,0} + e_{i,t}$$

where $\tilde{w}'_{i,t} = (1, w_{i,t-1}, w_{i,t-2}, \dots w_{i,t-12})$ and $\beta_{i,0}$ are the vector of regression parameters. These predictions are known as "generated regressors" because they are generated from a separate model. The need to predict $w^e_{i,t}$ creates a sequential estimation.[76] In the first step $\theta_{0,1} = (\beta'_{1,0}, \beta'_{2,0})$ are estimated. In the second step, $\theta_{0,2} = (\gamma'_0, \rho_0)$ are estimated conditional on $\hat{\theta}_{1,T}$. However, it is not obvious exactly how this structure would affect inference about the parameters of the inventory equation. As suggested above, the answer is found by interpreting the estimation from a GMM perspective. To achieve this, we must derive the population moment conditions which are being implicitly exploited in each step of the estimation. Since the univariate AR(12) models are linear regression models, it follows from the previous subsection[77] that $\hat{\beta}_{i,T}$ are the GMM estimators based on

$$E[f_1(v_t, \theta_{1,0})] = E\begin{bmatrix} \tilde{w}_{1,t}(w_{1,t} - \tilde{w}'_{1,t}\beta_{1,0}) \\ \tilde{w}_{2,t}(w_{2,t} - \tilde{w}'_{2,t}\beta_{2,0}) \end{bmatrix} = 0$$

The minimand for nonlinear least squares estimation also fits within the framework discussed in the previous sub-section. The minimand in (3.119) can be obtained from (3.107) by putting $N_t(\theta) = e_t(\gamma, \rho)^2$. Therefore it follows from (3.109) that the GMM interpretation of Hall and Rossana's (1991) estimator is completed by

$$E[f_2(v_t, \theta_0)] = E[\frac{\partial \tilde{e}_t(\theta_0)}{\partial \theta_2}\tilde{e}_t(\theta_0)] = 0$$

where

$$\begin{aligned} \tilde{e}_t(\theta) &= \tilde{u}_t(\gamma, \theta_2) - \rho\tilde{u}_{t-1}(\gamma, \theta_1) \\ u_t(\gamma, \theta_1) &= \Delta y_t - (1, y_{t-1}, x'_{t-1}, \tilde{w}'_{1,t}\beta_1, \tilde{w}_{2,t}\beta_2)'\gamma \end{aligned}$$

The correct form of the asymptotic distribution of the inventory equations can be deduced from Theorem 3.2. ◇

3.9 Summary

This chapter provides a comprehensive treatment of GMM estimation in correctly specified models. Building from the discussion in the previous chapter, it is shown that the basic approach to estimation employed in the linear static

[76] Pagan (1984) presents an in depth analysis of the problems caused by generated regressors. However, he does not exploit the GMM perspective described here. This approach was taken first by Newey (1984).

[77] Notice that the static nature of the variables in our earlier example played no role in the discussion.

model translates readily to nonlinear, dynamic models. The basic statistical framework also translates; although, inevitably, the presence of nonlinearity and dynamics complicates the analysis at various points. Seven key features emerge.

- *Identification*: For the estimation to be successful, the population moment condition must not only be valid but also provide sufficient information to identify the parameter vector. The intuition behind parameter identification is identical to the linear model, but nonlinearity considerably complicates its verification within a particular model. As a result, it is necessary to introduce the concepts of local and global identification.

- *Calculation of the estimator*: The presence of nonlinearity and, to a lesser extent, the dynamics means that the first order conditions do not yield a closed form solution for the estimator in general. Instead, the solution must be found via numerical optimization techniques.

- *Identifying and overidentifying restrictions*: GMM estimation in overidentified models involves a fundamental decomposition of the population moment condition into identifying and overidentifying restrictions. The identifiying restrictions contain the information that goes into the estimation, and the overidentifying restrictions are a remainder that manifests itself in the estimated sample moment.

- *Asymptotic properties*: The GMM estimator is consistent and, when appropriately scaled, has a limiting normal distribution. Here too, the absence of a closed form solution for the estimator, necessitates a different approach. This difference is most marked in the proof of consistency. However, once consistency is established, the Mean Value Theorem can be used to linearize the sample moment, and the proof of asymptotic normality can be viewed as a direct generalization of the arguments used in the linear model.

- *Estimated sample moment*: The estimated sample moment is shown to have a limiting normal distribution whose attributes depend directly on the function of the data in the overidentifying restrictions.

- *Long run covariance matrix estimation*: To translate the asymptotic normality into practical inference procedures, it is necessary to estimate the long run variance of the sample moment consistently. To construct a suitable estimator, it is necessary to make certain assumptions about the dependence structure of $f(v_t, \theta_0)$, the function of the data which appears in the population moment condition. Three cases are considered: $f(v_t, \theta_0)$ is a serially uncorrelated process; $f(v_t, \theta_0)$ is generated by a vector autoregressive moving average process; the class of heteroscedasticity and autocorrelation covariance (HAC) matrix estimators whose properties only require the dependence structure to satisfy very mild restrictions.

- *Optimal choice of weighting matrix*: The optimal choice of weighting matrix converges to the inverse of the long run covariance matrix of the

sample moment. Therefore, in general, its use necessitates a two step or iterated estimation.

In addition to the standard GMM estimation framework, this chapter also discusses certain important extensions. It is shown that both the estimator and/or subsequent inferences are sensitive to certain transformations of either the data, parameter vector or moment condition. These sensitivities motivate the discussion of the continuous updating GMM estimator and also an alternative method for the construction of confidence sets based on inverting the minimand.

It is also shown that GMM can be viewed as a unifying principle of estimation because it encompasses other methods such as Maximum Likelihood, Ordinary Least Sqares and certain Sequential estimation techniques.

A key assumption throughout is that the model is correctly specified. In the next chapter, we consider the consequences of misspecification for the asymptotic properties of the estimator and estimated sample moment. As would be anticipated, these consequences are not good and this motivates the use of specification tests, such as the overidentifying restrictions test. Such diagnostic tests are examined in Chapter 5 as part of a more general review of hypothesis testing within the GMM framework.

4

GMM Estimation in Misspecified Models

The previous chapter establishes the large sample properties of the estimator and its various associated statistics in correctly specified models. In practice, a researcher never knows whether his/her assumptions correspond to the real world, and so it is important to consider the impact of misspecification on the statistical properties derived in Chapter 3. Intuition suggests misspecification has a detrimental effect, and this is borne out by the analysis presented in this chapter.[1] In particular, it is shown that misspecification contaminates inferences about the parameter vector, and this pessimistic conclusion motivates the model specification tests presented in the next chapter. However, there is a secondary purpose to the presentation of a formal analysis of the GMM estimator under misspecification. Inspection of the empirical literature reveals that it is not uncommon to find cases in which the sample evidence suggests that the model is misspecified but inference about the parameters is still performed – either implicitly or explicitly – using the asymptotic theory appropriate for correctly specified models. The results presented here provide guidance on the interpretation of such inferences, and suggest that this approach to inference in misspecified models is invalid in general.

Before we proceed further, it is useful to consider exactly what is meant by the term "misspecification" in our context. As seen in Chapter 1, an economic/statistical model consists of a set of assumptions about the data generation process for v_t. For expositional convenience, we now denote this model by $\mathcal{M}$. This model implies a set of population moment conditions which can be used as a basis for GMM estimation of θ_0. This logical sequence can be represented by

$$\mathcal{M} \implies E[f(v_t, \theta_0)] = 0, \ \forall\, t, \text{ for some unique } \theta_0 \in \Theta \tag{4.1}$$

[1] Also see Section 2.5 for a heuristic discussion of the consequences of misspecification in the static linear model.

If $\mathcal{M}$ is no longer considered to be the truth, then there are two natural, alternative scenarios. First, the true model, $\mathcal{M}_A$, although different from $\mathcal{M}$, shares the property in (4.1); that is

$$\mathcal{M}_A \implies E[f(v_t, \theta_+)] = 0, \ \forall\, t, \text{ for some unique } \theta_+ \in \Theta \tag{4.2}$$

Secondly, the true model, $\mathcal{M}_B$, implies the property in (4.1) does not hold; that is

$$\mathcal{M}_B \implies \nexists\, \theta \in \Theta \text{ such that } E[f(v_t, \theta)] = 0, \ \forall\, t. \tag{4.3}$$

Clearly, $\mathcal{M}$ and $\mathcal{M}_A$ are observationally equivalent on the basis of $E[f(v_t, \theta)]$ alone.[2] Therefore, the estimator and the estimated sample moment have essentially the same large sample properties under $\mathcal{M}$ and $\mathcal{M}_A$ – the only difference is in the use of θ_0 or θ_+ to denote the value at which the population moment condition and other regularity conditions are satisfied. In contrast, $\mathcal{M}$ and $\mathcal{M}_B$ have different implications for $E[f(v_t, \theta)]$, and these manifest themselves in the behaviour of the estimator and the estimated sample moment. It is convenient to reserve the term "misspecification" to denote only this second situation. As it stands, (4.3) states only that $E[f(v_t, \theta)]$ is non-zero. For the analysis in this chapter, it is most convenient to retain the assumption that v_t is a stationary process, and so $E[f(v_t, \theta)]$ is independent of t. Therefore, we restrict attention to the following class of misspecified models.

Assumption 4.1 The Nature of the Misspecification
$E[f(v_t, \theta)] = \mu(\theta)$ *for all* t *and* $||\mu(\theta)|| > 0$ *for all* $\theta \in \Theta$.[3]

One immediate consequence of this assumption is that it excludes misspecification characterized by structural instability – that is, cases in which $E[f(v_t, \theta)] = \mu_t$. While this obviously limits the generality of the analysis, the price is worth paying because Assumption 4.1 smooths the passage from correctly to incorrectly specified models, and so enables us to highlight more simply the main differences between the two scenarios. However, in Section 5.4, we do return to the topic of structural instability in the context of hypothesis testing. There is one further consequence of Assumption 4.1 which should be noted. Taken together, Assumptions 4.1 and 3.1 (the stationarity of v_t) imply that $q > p$. This follows because if $p = q$ then the value which satisfies the identifying restrictions in (3.19), θ_+ say, must also satisfy the population moment condition.[4] In other words, if the parameter vector is just-identified then the true model must exhibit the properties of $\mathcal{M}_A$ above.[5]

[2] Notice this definition of observational equivalence depends crucially on $f(.)$. Since $\mathcal{M}$ and $\mathcal{M}_A$ are different models they will have different implications for other aspects of the distribution of v_t.

[3] For any vector a, $||a|| = (a'a)^{1/2}$.

[4] See Hall and Inoue (2003).

[5] This does not hold if v_t is non-stationary.

In practice, inference is typically based on the two step or iterated estimator. The key feature of such estimators is that the i^{th} step estimation employs a weighting matrix equal to the inverse of a covariance matrix estimator calculated using $\hat{\theta}_T(i-1)$. This structure means that the population analog to the minimand on the i^{th} step depends on the probability limit of $\hat{\theta}_T(i-1)$ via the weighting matrix. This construction provides a mechanism through which the consequences of misspecification are transmitted from one step to the next. This means that to deduce the impact of this misspecification on the iterated estimator, it is necessary to consider each step sequentially. Therefore, we begin our discussion with the first step estimator: Section 4.1 derives its probability limit, and Section 4.2 derives its limiting distribution. It emerges that misspecification considerably complicates the analysis of the limiting distribution. Specifically, the rate of convergence of $\hat{\theta}_T(1)$ to $\theta_*(1)$ depends on the rate of convergence of W_T to W. This means that in some cases $T^{1/2}[\hat{\theta}_T(1) - \theta_*(1)]$ does not converge in distribution. However, it is shown that this statistic has a limiting normal distribution under certain conditions which plausibly cover the most common choices of weighting matrix in practice. Section 4.3 considers the impact of misspecification on the long run covariance matrix estimators presented in Section 3.5. It is shown that none of these estimators are consistent if the model is misspecified. However, it is also shown that there is a simple way to modify all the estimators to ensure consistency regardless of whether or not the model is correctly specified. There are certain advantages to using one of these modified estimators in the construction of moment selection procedures. A formal justification of this statement is left until Chapter 7. Section 4.4 examines the limiting behaviour of the second-step estimator. Here too, the method of covariance matrix estimation is important because it determines the rate of convergence of the weighting matrix to its limit and hence the rate of convergence of the estimator. We concentrate on two cases. Section 4.4.1 presents the analysis when the covariance matrix estimator is constructed under the assumption that $f(v_t, \theta_*)$ is a serially uncorrelated process. Section 4.4.2 presents the same analysis when an HAC estimator is used. Section 4.5 considers the limiting behaviour of the estimated sample moment. Unlike the estimator, $T^{1/2} g_T(\hat{\theta}_T(i))$ diverges at rate $T^{1/2}$ regardless of the rate of convergence of the weighting matrix. Finally, Section 4.6 provides a summary of the consequences of misspecification on the GMM estimator.

Before we begin the analysis, it is necessary to address an item of notation. In the course of our discussion, it emerges that the $p\lim_{T\to\infty} \hat{\theta}_T(i)$ may be different for each i, and consequently, we use $\theta_*(i)$ to denote this limit. However, to avoid excessive repetition, we express assumptions in terms of θ_*, and then define θ_* in the appropriate theorem. In spite of the aforementioned dependence on i, there are times in which the analysis is generic to all steps and so we adopt the more economical notation of $\hat{\theta}_T$ for the estimator and θ_* for its probability limit.

4.1 Probability Limit of the First Step Estimator

By definition, the first step GMM estimator can be constructed with any weighting matrix which satisfies Assumption 3.7. In Section 3.4.1, it is shown that such an estimator converges in probability to θ_0 in correctly specified models *provided* certain regularity conditions are satisfied. So this earlier analysis provides the natural place to start our search for conditions under which the first step estimator converges in misspecified models. It can be recalled that the proof of Theorem 3.1 is broken down into two parts. Part (i) uses the uniform convergence property in Lemma 3.1 to establish that $\hat{\theta}_T$ minimizes $Q_0(\theta)$ with probability one as $T \to \infty$. Then part (ii) uses the population moment and identification conditions in Assumptions 3.3–3.4 to show that part (i) implies consistency. This overview suggests similar arguments can be used to establish the convergence of $\hat{\theta}_T$ in misspecified models *provided* a suitable replacement is found for Assumptions 3.3 and 3.4 in part *(ii)*. To this end, we now introduce the following assumption.

Assumption 4.2 Identification Condition
There exists $\theta_* \in \Theta$ *such that* $Q_0(\theta_*) < Q_0(\theta)$ *for all* $\theta \in \Theta \setminus \{\theta_*\}$.

Assumption 4.2 states that the population analog to the first step GMM minimand has a unique minimum at θ_*. This property defines $\theta_* = \theta_*(1)$ as the probability limit of $\hat{\theta}_T(1)$ in Theorem 4.1 below. Before we present that result, it is worth noting two ways in which Assumption 4.2 differs from the combination of Assumptions 3.3 and 3.4. First, Assumption 4.2 does imply a specific value for $E[f(v_t, \theta_*)]$ – although it does imply that $\| E[f(v_t, \theta_*)] \| < \infty$. Secondly, in misspecified models, there is no reason why the same parameter value should minimize $Q_0(\theta)$ for two different choices of W. Therefore, in general, θ_* is determined in part by W.[6]

Theorem 4.1 Convergence of $\hat{\theta}_T(1)$
If Assumptions 3.1–3.2, 3.7–3.10, 4.1 hold and 4.2 holds for $\theta_* = \theta_*(1)$ *then* $\hat{\theta}_T(1) \xrightarrow{p} \theta_*(1)$.

As anticipated above, the proof is split into two parts along similar lines to the proof of Theorem 3.1. Part (i) uses the definition of the estimator and Lemma 3.1 to deduce that

$$\lim_{T \to \infty} P[0 \leq Q_0(\hat{\theta}_T(1)) < Q_0(\theta_*(1)) + \epsilon] = 1 \text{ for any } \epsilon > 0 \qquad (4.4)$$

Part (ii) uses (4.4) and Assumption 4.1 to deduce that $\hat{\theta}_T(1) \xrightarrow{p} \theta_*(1)$. The details are left to the reader.

[6] See Section 4.4 for futher discussion of this issue in the context of the two step or iterated estimator.

4.2 Asymptotic Distribution Theory for the First Step Estimator

In Section 3.4.2 it is shown that $T^{1/2}(\hat{\theta}_T - \theta_0)$ converges to a normal distribution if the model is correctly specified. In this section, we develop an analogous limiting distribution theory for the first step estimator when the model is misspecified. It emerges that the weighting matrix plays a far more fundamental role in misspecified models, and this complicates the analysis. This dependence is present at each step of the GMM estimation, and so the first part of the analysis is not specific to the first step estimator. Therefore, we adopt the generic notation of $\hat{\theta}_T$ for the estimator and θ_* for its probability limit for most of this analysis and then specialize the results to $\hat{\theta}_T(1)$ at the end. This section is based on results in Hall and Inoue (2003) to which the reader is refered for rigorous proofs of the main results.[7]

As in the previous section, we need to determine appropriate conditions under which to perform the analysis. Once again, the logical starting place is the corresponding analysis in correctly specified models. Inspection of the regularity conditions in Theorem 3.2 reveals that many of them do not involve the specification of the model *per se*. In particular, Assumptions 3.1, 3.2, 3.7–3.10, 3.12–3.13 impose regularity conditions on v_t, Θ or the behaviour of $f(.), \partial f(.)/\partial \theta^{'}$ over Θ. Therefore, we can equally well impose those assumptions here. Obviously, Assumptions 3.3–3.4 depend on the model specification, and, as in the previous section, we replace them with Assumption 4.2. Once this is done, we can invoke Theorem 4.1 to deduce that $\hat{\theta}_T \xrightarrow{p} \theta_*$. The nature of this limit will have an impact on our analysis. It can be recalled from Section 3.4.2 that the analysis started with the Mean Value Theorem applied to $g_T(\hat{\theta}_T)$ around θ_0. We use a similar starting point below but take the linearization around θ_*. So we must replace Assumptions 3.5, 3.12 and 3.13 by the following assumption.[8]

Assumption 4.3 Regularity Conditions on $\partial f(v_t, \theta)/\partial \theta'$
(i) The derivative matrix $\partial f(v, \theta)/\partial \theta'$ exists and is continuous on Θ for each $v \in \mathbf{V}$; (ii) θ_* is an interior point of Θ; (iii) $E[\partial f(v_t, \theta_*)/\partial \theta']$ exists and is finite; (iv) $E[\partial f(v_t, \theta)/\partial \theta']$ is continuous on some ϵ-neighbourhood N_ϵ of θ_*; (v) $\sup_{\theta \in N_\epsilon} ||G_T(\theta) - E[\partial f(v_t, \theta)/\partial \theta']|| \xrightarrow{p} 0$.

Once the linearization is taken around θ_*, it is the behaviour of $T^{1/2} g_T(\theta_*)$ which becomes relevant. Accordingly, we define

$$E[f(v_t, \theta_*)] \; = \; \mu_* \tag{4.5}$$

Notice that Assumption 4.1 implies $\mu_* \neq 0$. We must also replace Assumption 3.11 and Lemma 3.2 by:

[7] Hall and Inoue's (2003) results subsume earlier work by Maasoumi and Phillips (1982); the latter paper presents the limiting distribution of the IV estimator in the linear regression model with W_T set equal to the inverse of the instrument cross product matrix.

[8] Part (i) is identical to Assumption 3.5(i). It is repeated here to simplify the presentation.

Assumption 4.4 Properties of the Variance of the Sample Moment
(i) $E[(f(v_t,\theta_*)-\mu_*)\,(f(v_t,\theta_*)-\mu_*)']$ *exists and is finite;*
(ii) $lim_{T\to\infty}Var[T^{1/2}g_T(\theta_*)] = S_*$ *exists and is a finite valued positive definite matrix.*

Lemma 4.1 Central Limit Theorem for $T^{-1/2}\sum_{t=1}^{T}[f(v_t,\theta_*)-\mu_*]$
If Assumptions 3.1, 3.8, 4.1, and 4.4 hold then $T^{-1/2}\sum_{t=1}^{T}[f(v_t,\theta_*)-\mu_*] \xrightarrow{d} N(0,S_*)$.

With all these assumptions imposed, we can now proceed to the analysis. As mentioned above, we begin by using the Mean Value Theorem to deduce that

$$g_T(\hat{\theta}_T) \;=\; g_T(\theta_*) + G_T(\hat{\theta}_T,\theta_*,\lambda_T)(\hat{\theta}_T-\theta_*) \tag{4.6}$$

where $G_T(\hat{\theta}_T,\theta_*,\lambda_T)$ is $(q\times p)$ matrix whose i^{th} row is equal to the i^{th} row of $G_T(\bar{\theta}_T^{(i)})$ where $\bar{\theta}_T^{(i)} = \lambda_T^{(i)}\theta_* + (1-\lambda_T^{(i)})\hat{\theta}_T$ for some $0 \le \lambda_T^{(i)} \le 1$, and $i = 1,2,\ldots q$. It is then possible to apply the same sequence of arguments as in Section 3.4.2 to show that (4.6) leads to

$$T^{1/2}(\hat{\theta}_T-\theta_*) \;=\; -[G_T(\hat{\theta}_T)'W_TG_T(\hat{\theta}_T,\theta_*,\lambda_T)]^{-1}G_T(\hat{\theta}_T)'W_TT^{1/2}g_T(\theta_*) \tag{4.7}$$

It is convenient to rewrite (4.7) as

$$T^{1/2}(\hat{\theta}_T-\theta_*) \;=\; H_{0,T}\{\,H_{1,T} \;+\; H_{2,T}\,\} \tag{4.8}$$

where

$$H_{0,T} \;=\; -[G_T(\hat{\theta}_T)'W_TG_T(\hat{\theta}_T,\theta_*,\lambda_T)]^{-1} \tag{4.9}$$

$$H_{1,T} \;=\; G_T(\hat{\theta}_T)'W_TT^{-1/2}\sum_{t=1}^{T}[f(v_t,\theta_*)-\mu_*] \tag{4.10}$$

$$H_{2,T} \;=\; T^{1/2}G_T(\hat{\theta}_T)'W_T\mu_* \tag{4.11}$$

It is instructive to compare (4.8) with the corresponding equation in our analysis of correctly specified models, (3.26). The term $H_{0,T}H_{1,T}$ can be recognized as the analog to the right hand side of (3.26), and so misspecification has introduced a second term, $H_{0,T}H_{2,T}$, into the equation.[9] To proceed further, it is useful to decompose $H_{2,T}$ as follows:

$$H_{2,T} \;=\; H_{2,T}(1) \;+\; H_{2,T}(2) \;+\; H_{2,T}(3) \;+\; H_{2,T}(4) \tag{4.12}$$

where

$$H_{2,T}(1) \;=\; T^{1/2}[G_T(\hat{\theta}_T)-G_T(\theta_*)]'W_T\mu_* \tag{4.13}$$

$$H_{2,T}(2) \;=\; T^{1/2}[G_T(\theta_*)-G_*]'W_T\mu_* \tag{4.14}$$

$$H_{2,T}(3) \;=\; G_*'T^{1/2}(W_T-W)\mu_* \tag{4.15}$$

$$H_{2,T}(4) \;=\; T^{1/2}G_*'W\mu_* \tag{4.16}$$

[9] Notice that if the model is correctly specified then $\mu_* = 0$ and so $H_{2,T} = 0$.

At this stage, we can take advantage of two simplifications. First, the population analog to the first order conditions imply $H_{2,T}(4) = 0$. Secondly, $H_{2,T}(1)$ can be written as[10]

$$\begin{aligned} H_{2,T}(1) &= (\mu_*' W_T \otimes I_p)\text{vec}\{T^{1/2}[G_T(\hat{\theta}_T) - G_T(\theta_*)]\} \\ &= (\mu_*' W_T \otimes I_p) G_T^{(2)}(\hat{\theta}_T, \theta_*, \phi_T) T^{1/2}(\hat{\theta}_T - \theta_*) \\ &= M_T T^{1/2}(\hat{\theta}_T - \theta_*), \qquad \text{say} \end{aligned}$$

where $G_T^{(2)}(\hat{\theta}_T, \theta_*, \phi_T)$ is the $pq \times p$ matrix whose i^{th} row is the corresponding row of $(\partial/\partial\theta')vec\{\partial f(v_t, \tilde{\theta}_T^{(i)})/\partial\theta'\}$ with $\tilde{\theta}_T^{(i)} = \phi_T^{(i)}\hat{\theta}_T + (1-\phi_T^{(i)})\theta_*$, $0 \leq \phi_T^{(i)} \leq 1$, and ϕ_T is the $pq \times 1$ vector with i^{th} element $\phi_T^{(i)}$.

Taking advantage of these two simplifications, (4.8)–(4.16) can be used to deduce that

$$T^{1/2}(\hat{\theta}_T - \theta_*) = [I_p - H_{0,T}M_T]^{-1} H_{0,T}\{H_{1,T} + H_{2,T}(2) + H_{2,T}(3)\} \quad (4.17)$$

Intuition suggests that $[I_p - H_{0,T}M_T]^{-1}H_{0,T}$ converges in probability to some matrix of constants and $H_{1,T}$ converges in distribution to normal vector under our conditions. It is also reasonable to assume that $T^{1/2}[G_T(\theta_*) - G_*]$ converges to a normal limiting distribution under certain conditions, and so $H_{2,T}(2)$ exhibits the same property. The key question is the limiting behaviour of $H_{2,T}(3)$. From (4.15), it is clear that the limiting behaviour of $H_{2,T}(3)$ depends on that of $T^{1/2}(W_T - W)$. In order for $T^{1/2}(\hat{\theta}_T - \theta_*)$ to converge in distribution, it is a necessary condition that $T^{1/2}(\hat{\theta}_T - \theta_*) = O_p(1)$. From (4.17), it is clear that such a condition can only be satisfied if $T^{1/2}(W_T - W) = O_p(1)$. Therefore, if W_T converges to W at a slower rate than $T^{1/2}$ then $T^{1/2}(\hat{\theta}_T - \theta_*)$ must diverge. This dependence of $T^{1/2}(\hat{\theta}_T - \theta_*)$ on $T^{1/2}(W_T - W)$ is in marked contrast to what is found in correctly specified models, and is directly attributable to the presence of $H_{2,T}$ in (4.8).

To make further progress, it is clearly necessary to make some assumption about the nature of the convergence of W_T to W. We focus on two particular scenarios which both satisfy $T^{1/2}(W_T - W) = O_p(1)$ and together cover the choices of first step estimator used in our empirical example in Chapter 3. The first scenario is where $W_T = W$, which obviously covers $W_T = I_q$, and the second is where $T^{1/2}(W_T - W)\mu_*$ converges to a normal distribution, which we show below covers $W_T = [T^{-1}\sum_{t=1}^{T} z_t z_t']^{-1}$ under plausible assumptions. However, before we present these results certain other conditions must be imposed. To ensure that $G_T^{(2)}(\hat{\theta}_T, \theta_*, \phi_T)$ converges to a well-defined limit, we impose:

Assumption 4.5 Regularity Conditions for $G_T^{(2)}(\theta)$
(i) $(\partial/\partial\theta')vec\{\partial f(v_t, \theta)/\partial\theta'\}$ exists and is continuous on Θ for each $v \in \mathcal{V}$;
(ii) $E[(\partial/\partial\theta')vec\{\partial f(v_t, \theta)/\partial\theta'\}]$ exists and is continuous on Θ;

[10] Dhrymes (1984)[Corollary 25, p.103] and the Mean Value Theorem applied to the $i-j^{th}$ element of $G_T(\hat{\theta}_T)$.

(iii) $\sup_{\theta \in N_\epsilon} ||G_T^{(2)}(\theta) - E[(\partial/\partial\theta')vec\{\partial f(v_t,\theta)/\partial\theta'\}]|| \xrightarrow{p} 0$ *where* N_ϵ *is an* ϵ *neighbourhood of* θ_*.

It is also necessary to ensure that the the inverse matrix in (4.17) is well defined in the limit. Therefore, we impose:

Assumption 4.6 Regularity Conditions on H_*
The $p \times p$ *matrix* $H_* = G_*'WG_* + (\mu_*'W \otimes I_p)G_*^{(2)}$ *is nonsingular where* $G_* = E[\partial f(v_t,\theta_*)/\partial\theta']$ *and* $G_*^{(2)} = E[(\partial/\partial\theta')vec\{\partial f(v_t,\theta_*)/\partial\theta'\}]$.

Assumption 4.6 is satisfied if $Q_0(\theta)$ satisfies the second-order sufficient condition for minimization at θ_*. It is also necessary to impose certain conditions in order for $H_{2,T}(2)$ to converge to a normal distribution. For ease of exposition, we impose those conditions implicitly in the statement of the following theorem.

Theorem 4.2 Limiting Distribution of the First Step Estimator
Let Assumptions 3.1, 3.2, 3.7–3.10, 4.1-4.6 (with $\theta_* = \theta_*(1)$*) hold.*

(i) If $W_T = W$ *and*

$$\begin{bmatrix} T^{-1/2}\sum_{t=1}^T [f(v_t,\theta_*) - \mu_*] \\ T^{1/2}[G_T(\theta_*) - G_*]'W\mu_* \end{bmatrix} \xrightarrow{d} N\left(0, \begin{pmatrix} S_* & V_{1,2} \\ V_{2,1} & V_{2,2} \end{pmatrix}\right).$$

then it follows that

$$T^{1/2}(\hat\theta_T - \theta_*) \xrightarrow{d} N(0, \Sigma_1)$$

where

$$\Sigma_1 = H_*^{-1}(G_*'WS_*WG_* + G_*'WV_{1,2} + V_{2,1}WG_* + V_{2,2})H_*^{'-1}$$

(ii) If

$$\begin{pmatrix} T^{-1/2}\sum_{t=1}^T [f(v_t,\theta_*) - \mu_*] \\ T^{1/2}[G_T(\theta_*) - G_*]'W\mu_* \\ T^{\frac{1}{2}}(W_T - W)\mu_* \end{pmatrix} \xrightarrow{d} N\left(0, \begin{pmatrix} S_* & V_{1,2} & V_{1,3} \\ V_{2,1} & V_{2,2} & V_{2,3} \\ V_{3,1} & V_{3,2} & V_{3,3} \end{pmatrix}\right)$$

then

$$T^{1/2}(\hat\theta_T - \theta_*) \xrightarrow{d} N(0, H_*^{-1}\Sigma_2 H_*^{'-1}),$$

where

$$\begin{aligned} \Sigma_2 \quad = \quad & G_*'WS_*WG_* + V_{2,2} + G_*'V_{3,3}G_* + G_*'WV_{1,2} \\ & + G_*'WV_{1,3}G_* + V_{2,1}WG_* + G_*'V_{3,1}WG_* + V_{2,3}G_* + G_*'V_{3,2} \end{aligned}$$

It is interesting to compare the results in parts (i)–(ii). First recall that $\theta_*(1)$ depends on W. Secondly, the structure of covariance matrices is different. So, in general, the limiting distributions in (i) and (ii) are different. However, there

is one obvious exception: if $\mu_* = 0$ – *i.e.* the model is correctly specified – then $\theta_*(1) = \theta_0$ and both variances reduce to[11]

$$V_C = (G_*'WG_*)^{-1}(G_*'WS_*WG_*)(G_*'WG_*)^{-1} \tag{4.18}$$

which can be recognized as the variance in Theorem 3.2 (if we put $\theta_* = \theta_0$). This comment also implies that, in general, the first step estimator has a different distribution in correctly specified and misspecified models.

It is remarked above that Theorem 4.2 covers the case in which the weighting matrix is the inverse of the instrument cross product matrix under plausible assumptions. To uncover the nature of these conditions, we let $W_T = [T^{-1}\sum_{t=1}^{T} z_t z_t']^{-1}$, $W = M_{zz}^{-1} = \{E[z_t z_t']\}^{-1}$, and rewrite $T^{1/2}(W_T - W)$ as follows

$$T^{1/2}(W_T - W) = -M_{zz}^{-1}\, T^{-1/2}[\sum_{t=1}^{T}(z_t z_t' - M_{zz})][T^{-1}\sum_{t=1}^{T} z_t z_t']^{-1} \tag{4.19}$$

From (4.19), it can be seen that this case is covered by Theorem 4.2(ii) provided that $vech\{T^{-1/2}[\sum_{t=1}^{T} z_t z_t' - M_{zz}]\}$ converges to a mean zero normal distribution.[12]

4.3 Long Run Covariance Matrix Estimation

Section 3.5 described various estimators of the long run variance of the sample moment. These estimators were grouped into three classes according to the assumption made about the dynamic structure of $f(v_t, \theta_*)$. However, all the estimators have one feature in common: they are constructed under the assumption that the model is correctly specified. Once we move into the world of misspecified models, *none* of the proposed estimators are consistent even if they are based on a correct assumption about the dynamic structure. This section describes the impact of misspecification on each of the covariance matrix estimators, and explains how they can be modified to ensure consistency in misspecified models. Gallant and White (1988)[Chapter 6] consider the impact of model misspecification on covariance matrix estimation under very general conditions, and some of our discussion represents a specialization of their results to stationary processes.

It is shown below that the exact impact of misspecification on each covariance matrix estimator is different. However, it is possible to gain a sense of both the problem and the solution by examining a single autocovariance matrix. By definition, the j^{th} autocovariance matrix of $f(v_t, \theta_*)$ is

$$\begin{aligned} \Gamma_j &= E[\{f(v_t, \theta_*) - \mu_*\}\{f(v_{t-j}, \theta_*) - \mu_*\}'] \\ &= E[f(v_t, \theta_*)f(v_{t-j}, \theta_*)'] - \mu_*\mu_*' \end{aligned} \tag{4.20}$$

[11] Notice that $\mu_* = 0$ implies $V_{i,j} = 0$.

[12] $vech\{.\}$ denotes the operator which stacks the lower triangular elements of a matrix into a vector.

Suppose we estimate Γ_j by

$$\hat{\Gamma}_j \;=\; T^{-1} \sum_{t=j+1}^{T} f(v_t, \hat{\theta}_T) f(v_{t-j}, \hat{\theta}_T)' \tag{4.21}$$

This statistic is a consistent estimator of the first term on the right-hand side of (4.20) but therefore an inconsistent estimator of Γ_j because $\mu_* \neq 0$. Given (4.20), the obvious solution is to estimate Γ_j by

$$\tilde{\Gamma}_j \;=\; T^{-1} \sum_{t=j+1}^{T} [f(v_t, \hat{\theta}_T) - g_T(\hat{\theta}_T)][f(v_{t-j}, \hat{\theta}_T) - g_T(\hat{\theta}_T)]' \tag{4.22}$$

As would be anticipated, this estimator is consistent for Γ_j. Also notice that if the model is correctly specifed then $\hat{\Gamma}_j = \tilde{\Gamma}_j + o_p(1)$, and so there is no cost asymptotically to using the mean correction when it is unnecessary.

It is useful to introduce a terminology to capture the difference between $\hat{\Gamma}_j$ and $\tilde{\Gamma}_j$. The key difference between them is that the data, $\{f(v_t, \hat{\theta}_T)\}$ are "centred" about their mean $g_T(\hat{\theta}_T)$ in $\tilde{\Gamma}_j$ but they are "uncentred" in $\hat{\Gamma}_j$. Therefore, we refer to $\tilde{\Gamma}_j$ as the *centred* version of the sample autocovariance and $\hat{\Gamma}_j$ as the *uncentred* version. These adjectives are similarly used to distinguish covariance matrices based on uncentred or centred autocovariances. We now examine the behaviour of each of the covariance matrix estimators from Section 3.5 in turn.

If $\{f(v_t, \theta_*)\}$ forms a serially uncorrelated sequence then $S_* = \Gamma_0$. Since $\hat{S}_{SU} = \hat{\Gamma}_0$, it follows from (4.20) that

$$\hat{S}_{SU} \;\xrightarrow{p}\; S_* \;+\; \mu_* \mu_*' \tag{4.23}$$

Equation (4.23) indicates that $\hat{S}_{SU}$ converges to a positive definite matrix of constants – but obviously not S_*. However, given the discussion above, it is clear that a consistent estimator for S_* is given by:

$$\hat{S}_{SU,\mu} \;=\; T^{-1} \sum_{t=1}^{T} [f(v_t, \hat{\theta}_T) - g_T(\hat{\theta}_T)][f(v_t, \hat{\theta}_T) - g_T(\hat{\theta}_T)]' \tag{4.24}$$

Now consider the impact of misspecification on den Haan and Levin's (1996) estimator. For this discussion, it is convenient to focus on the case where f_t is actually generated by

$$\Psi(L)(f_t - \mu_*) \;=\; \Phi(L) e_t \tag{4.25}$$

where the matrix polynomials satisfy the conditions for stationarity and invertibility in Section 3.5.2 and e_t satisfies the properties listed there as well. Starting from (4.25), it can be deduced along similar lines to Section 3.5.2 that f_t satisfies the autoregressive model

$$A(L)(f_t - \mu_*) \;=\; e_t \tag{4.26}$$

A comparison of (4.26) with (3.50) indicates that there are now two sources of misspecification in the autoregressive model used in *Step 2* of den Haan and Levin's (1996) method. Apart from the truncation error, there is the omission of the intercept. Unlike the truncation error, the problems caused by the omission of the intercept cannot be removed by letting the autoregressive lag length tend to infinity with the sample size. Intuition suggests this type of misspecification causes $\hat{S}_{VARMA}$ to be an inconsistent estimator of S_*. Unfortunately, a formal investigation of this question is complicated by the presence of the lag selection criterion in *Step 3* of den Haan and Levin's (1996) method. However, once again, consistency is restored by applying a mean correction. This time, the correction is implemented by applying den Haan and Levin's method to $f(v_t, \hat{\theta}_T)$ in mean deviation form.

Finally, we consider the impact of misspecification on the class of HAC estimators – both with and without the use of prewhitening and recolouring. We begin with the *uncentred* HAC estimator

$$\hat{S}_{HAC} = \hat{\Gamma}_0 + \sum_{i=1}^{T-1} \omega_{i,T} \left(\hat{\Gamma}_i + \hat{\Gamma}_i' \right) \tag{4.27}$$

In Section 3.5.3 it is observed that the kernel, $\omega_{i,T}$, and bandwidth, b_T, must be carefully chosen to ensure the estimator is consistent. However, this comment is conditional on the assumption that $\hat{\Gamma}_j$ is a consistent estimator of Γ_j. As we have seen, this premise is only valid if the model is correctly specified. This means inevitably that $\hat{S}_{HAC}$ is itself no longer a consistent estimator. While the source of the inconsistency is the same as with $\hat{S}_{SU}$, the consequences are more drastic because of the increasing bandwidth. Using results in Gallant and White (1988)[Chapter 6], it can be shown that

$$\hat{S}_{HAC} = S_* + B_T \mu_* \mu_*' + o_p(1) \tag{4.28}$$

where $B_T = 1 + 2\sum_{i=1}^{T-1} \omega_{i,T}$. It can be shown that B_T increases at rate b_T for either the Bartlett, Parzen or Quadratic Spectral kernels. So in these cases, $\hat{S}_{HAC}$ is asymptotically equivalent to the sum of two matrices: S_*, a positive definite matrix of constants, and $B_T \mu_* \mu_*'$, a rank one matrix of $O(b_T)$. While $S_* + B_T \mu_* \mu_*'$ is positive definite for finite T, it is clear that the rank one matrix dominates in the limit as $T \to \infty$. In the next section, it is shown that (4.28) has an important implication for the limiting behaviour of $\hat{S}_{HAC}^{-1}$ which in turn affects the limiting behaviour of the two step GMM estimator. For the present, we focus instead on how to modify the estimator to ensure consistency even if the model is misspecified. Once again, the answer is straightforward: replace $\hat{\Gamma}_i$ in (4.27) by $\tilde{\Gamma}_j$ from (4.22). This yields the *centred* HAC estimator,[13]

$$\hat{S}_{HAC,\mu} = \tilde{\Gamma}_0 + \sum_{i=1}^{T-1} \omega_{i,T} \left(\tilde{\Gamma}_i + \tilde{\Gamma}_i' \right) \tag{4.29}$$

[13] Hall (2000) proves this estimator is consistent with either the Bartlett, Parzen or Quadratic Spectral kernel and $b_T \to \infty$ with T but $b_T = o(T^{1/2})$.

4.4 The Two Step or Iterated GMM Estimator

In this section, we consider the implications of misspecification for the probability limit of the two step or iterated estimator. The exact nature of this transmission mechanism depends on which covariance matrix estimator is used. For reasons that emerge below, we split the analysis into two parts. Section 4.4.1 considers the case in which $f(v_t, \theta_*) - \mu_*$ is a serially uncorrelated process, and either $\hat{S}_{SU}$ or $\hat{S}_{SU,\mu}$ is used to construct the weighting matrix. Section 4.4.2 considers the case in which either an uncentred or centred HAC estimator is used to construct the weighting matrix.[14] It emerges that the behaviour in these two cases is very different, and also very different from the behaviour of the first step estimator. In this discussion, it is necessary to distinguish various functions of the parameter vector evaluated at different steps of the estimation. Therefore, we define $\mu_*(i) = \mu(\theta_*(i))$, $S_*(i) = \lim_{T\to\infty} Var[T^{-1/2}\sum_{t=1}^{T} f(v_t, \theta_*(i))]$, $\Gamma_0(i) = Var[f(v_t, \theta_*(i))]$, and let $\hat{\Gamma}_0(i)$, $\tilde{\Gamma}_0(i)$ denote respectively the uncentred and centred zero order sample autocovariance matrices evaluated at $\hat{\theta}_T(i)$.

4.4.1 Estimation with $W_T = \hat{S}_{SU}^{-1}$ or $W_T = \hat{S}_{SU,\mu}^{-1}$

It is most convenient to develop the analysis under the assumption that $f(v_t, \theta_*)$ is a serially uncorrelated sequence and so $S_*(1) = \Gamma_0(1)$. In this case, the inconsistency of $\hat{S}_{SU}$ stems solely from $\mu_* \neq 0$, and not from an incorrect assumption about the dynamic structure of $f(v_t, \theta_*) - \mu_*$. However, some of the results hold more generally and so we relax this assumption briefly at the end to consider the impact of dynamic misspecification.

We begin our discussion with the second step estimator, $\hat{\theta}_T(2)$. Recall from Section 3.6, that $\hat{\theta}_T(2)$ is calculated using $W_T = \hat{S}_T(1)^{-1}$ where $\hat{S}_T(1)$ is an estimator of the long run variance based on $\hat{\theta}_T(1)$. Therefore, the population analog to the second step minimand is given by:

$$Q_0^{(2)}(\theta) \;=\; E[f(v_t,\theta)]' W^{(2)} E[f(v_t,\theta)] \tag{4.30}$$

where $W^{(2)} = \{p\lim_{T\to\infty} \hat{S}_T(1)\}^{-1}$. Then from Theorem 4.1, (4.21) and (4.22) it follows that

$$\hat{S}_{SU} \;=\; \hat{\Gamma}_0 \;\xrightarrow{p}\; S_*(1) + \mu_*(1)\mu_*(1)' \tag{4.31}$$

$$\hat{S}_{SU,\mu} \;=\; \tilde{\Gamma}_0 \;\xrightarrow{p}\; S_*(1) \tag{4.32}$$

and so[15]

$$\begin{aligned} \hat{S}_{SU}^{-1} \;&\xrightarrow{p}\; S_*(1)^{-1} - c_*(1) S_*(1)^{-1}\mu_*(1)\mu_*(1)' S_*(1)^{-1} \qquad (4.33)\\ &=\; S_u(1)^{-1}, \qquad \text{say} \\ \hat{S}_{SU,\mu}^{-1} \;&\xrightarrow{p}\; S_*(1)^{-1} \qquad (4.34) \end{aligned}$$

[14] We do not explicitly consider the case in which den Haan and Levin's (1996) estimator is used because, as mentioned above, the presence of the lag selection method complicates the analysis.

[15] For example see Morrison (1976) [p.69].

where $c_*(1) = [1 + \mu_*(1)' S_*(1)^{-1}\mu_*(1)]^{-1}$. Inspection of (4.33)–(4.34) reveals that both $\hat{S}_{SU}^{-1}$ and $\hat{S}_{SU,\mu}^{-1}$ converge in probability to positive definite matrices of constants and so satisfy the conditions for a weighting matrix specified in Assumption 3.7. It is also apparent that $W^{(2)}$ is different in each case, and intuition suggests this difference should also manifest itself in the probability limits of the associated two step estimators. It is hard to confirm or disprove this intuition by looking at $Q_0^{(2)}$. However, more progress can be made by turning to the population analog of the first order conditions. To this end, let $\theta_*^{(u)}$ denote the unique minimizer of $Q_0^{(2)}(\theta)$ when $W^{(2)} = S_u(1)^{-1}$, and $\theta_*^{(c)}$ be the unique minimizer of $Q_0^{(2)}(\theta)$ when $W^{(2)} = S_*(1)^{-1}$.[16] In order to characterize these values by the first order conditions, it is necessary to assume that Assumption 4.3(ii)–(iii) hold at both $\theta_*^{(u)}$ and $\theta_*^{(c)}$. Once these conditions are imposed, it follows that $\theta_*^{(u)}$ is the solution to the first order conditions

$$G(\theta)' S_*(1)^{-1} E[f(v_t, \theta)] - c_*(1) G(\theta)' S_*(1)^{-1}\mu_*(1)\mu_*(1)' S_*(1)^{-1} E[f(v_t, \theta)] \; = \; 0 \tag{4.35}$$

and $\theta_*^{(c)}$ is the solution to

$$G(\theta)' S_*(1)^{-1} E[f(v_t, \theta)] \; = \; 0 \tag{4.36}$$

Inspection of (4.35) and (4.36) reveals two features of the probability limits: in general, $\theta_*^{(u)} \neq \theta_*^{(c)}$, and neither equals $\theta_*(1)$, the probability limit of $\hat{\theta}_T(1)$.[17] However, there is one exception which should be noted. If $\theta_*(1)$ satisfies (4.36), then it also satisfies (4.35), and so $\theta_*^{(u)} = \theta_*^{(c)} = \theta_*(1)$. Such a coincidence would occur if the first step weighting matrix is of the form $kS_*(1)^{-1}$ for some constant k, but this is unlikely to be the case in general. The equality between the two probability limits can also occur if the estimation is iterated beyond two steps. If both the iterated estimator based on $W_T = \hat{S}_{SU}^{-1}$ and the iterated estimator based on $W_T = \hat{S}_{SU,\mu}^{-1}$ individually converge then it can be shown using appropriately modified versions of (4.35) and (4.36) that both estimators have the same probability limit.

Now consider the limiting distribution of the second step estimator. Regardless of whether $W_T = \hat{S}_{SU}^{-1}$ or $W_T = \hat{S}_{SU,\mu}^{-1}$, it is possible to establish that the second step estimator has a limiting normal distribution under plausible conditions. For brevity, we focus on the case in which $W_T = \hat{S}_{SU,\mu}^{-1}$ but a similar argument applies for the case in which $W_T = \hat{S}_{SU}^{-1}$. The argument is based on an appeal to Theorem 4.2(ii). Using the same trick as (4.19), it can be shown that the limiting distribution of $\hat{\theta}_T(2)$ is given by Theorem 4.2(ii) provided $vech\{T^{1/2}(\tilde{\Gamma}_0(1) - \Gamma_0(1))\}$ converges to a normal distribution. However, this appeal to Theorem 4.2(ii) is not so benign as it at first appears. Using the Mean

[16] The superscript on θ_* reflects whether the covariance matrix is uncentred or centred, and the '(2)' argument is suppressed for ease of notation.

[17] Recall that if the model is correctly specified then the probability limit of all three estimators is θ_0.

Value Theorem, it can be shown that

$$\begin{aligned} vech\{T^{1/2}[\tilde{\Gamma}_0(1)-\Gamma_0(1)]\} &= vech\{T^{1/2}[\Gamma_{0,T}(1)-\Gamma_0(1)]\} \\ &\quad (\partial/\partial\theta')vech\{\Gamma_{0,T}(\theta_*(1))\}T^{1/2}[\hat{\theta}_T(1)-\theta_*(1)] \\ &\quad + o_p(1) \end{aligned} \tag{4.37}$$

where $\Gamma_{0,T}(\theta) = T^{-1}\sum_{t=1}^{T}[f(v_t,\theta)-\mu(\theta)][f(v_t,\theta)-\mu(\theta)]'$. Therefore, the large sample behaviour of $T^{1/2}[\hat{\theta}_T(2)-\theta_*(2)]$ depends on the large sample behaviour of $T^{1/2}[\hat{\theta}_T(1)-\theta_*(1)]$ *unless* $(\partial/\partial\theta')vech\{\Gamma_{0,T}(\theta_*(1))\} \xrightarrow{p} 0$. In general, there is no reason to suppose that this condition holds. A similar argument can be applied to the iterated versions of these estimators to deduce that the limiting distribution of $T^{1/2}[\hat{\theta}_T(i)-\theta_*(i)]$ depends on $\{T^{1/2}[\hat{\theta}_T(j)-\theta_*(j)],\ j=1,2,\ldots i-1\}$ in general. Needless to say, this recursive structure must be taken into account in the calculation of the asymptotic variance of the estimator. However, we do not pursue the form of this asymptotic variance further here.

So far, it has been assumed that $f(v_t,\theta_*)$ is a serially uncorrelated sequence. If this assumption is relaxed then $S_*(1)$ must be replaced by $\Gamma_0(1)$ in (4.35) and (4.36). However, this substitution has no qualitative impact on the foregoing analysis of the probability limits of the estimators, and so all the conclusions remain valid in this more general case. The assumption of no serial correlation also has no qualitative impact on the appeal to Theorem 4.2(ii) to deduce the asymptotic normality. However, its relaxation introduces a dynamic structure in $f(v_t,\theta_*)-\mu_*$ which must be accounted for in the definitions, and also the estimation, of the covariance matrices $V_{i,j}$ in Theorem 4.2(ii).

To conclude, this sub-section we examine the impact of using $\hat{S}_{SU,\mu}^{-1}$ in our empirical example.

Example: Hansen and Singleton's (1982) Consumption Based Asset Pricing Model

Table 3.7 in Section 3.6 reports the results from the two step and iterated estimations with $\hat{S}_{SU}^{-1}$ used as weighting matrix. Table 4.1 contains the analogous results when $\hat{S}_{SU,\mu}^{-1}$ is used. With equally weighted returns (EWR), convergence takes 5 and 4 iterations respectively with $W_T^{(1)} = 10^5 I_5$ and $W_T^{(1)} = (T^{-1}\sum_{t=1}^{T} z_t z_t')^{-1}$. With value weighted returns (VWR), one less iteration is needed in each case. If the model is correctly specified, then the probability limit of the estimator is the same on all steps. The results in Table 3.7 indicate the iterated estimator converges to the same values for a given asset irrespective of the the first step weighting matrix. Our analysis in this sub-section indicates that if convergence occurs then the probability limits of the iterated estimators should be the same regardless of whether the weighting matrix is either $\hat{S}_{SU}^{-1}$ or $\hat{S}_{SU,\mu}^{-1}$ even if the model is misspecified. These arguments lead us to expect that the corresponding estimates should be close in large finite samples *irrespective* of whether the model ultimately proves to be correctly or incorrectly specified. A comparison of Tables 3.7 and 4.1 indicates the iterated estimates are iden-

tical to three decimal places for VWR and to two decimal places for EWR. ◇

Table 4.1
Two step and iterated GMM estimators for the consumption based asset pricing model with EWR and VWR

EWR: $W_T^{(1)}$	$(\hat{\gamma}, \hat{\delta})$ for $i = 1$	$(\hat{\gamma}, \hat{\delta})$ for $i = 2$	$(\hat{\gamma}, \hat{\delta})$ after iteration
$10^5 I_5$	$(-3.145, 0.999)$	$(-0.253, 0.992)$	$(-0.344, 0.992)$
$(T^{-1}\sum_{t=1}^{T} z_t z_t')^{-1}$	$(0.398, 0.993)$	$(-0.335, 0.992)$	$(-0.344, 0.992)$
VWR: $W_T^{(1)}$	$(\hat{\gamma}, \hat{\delta})$ for $i = 1$	$(\hat{\gamma}, \hat{\delta})$ for $i = 2$	$(\hat{\gamma}, \hat{\delta})$ after iteration
$10^5 I_5$	$(-1.871, 0.998)$	$(0.716, 0.993)$	$(0.666, 0.994)$
$(T^{-1}\sum_{t=1}^{T} z_t z_t')^{-1}$	$(0.698, 0.994)$	$(0.666, 0.994)$	$(0.666, 0.994)$

Note: $W_T^{(1)}$ denotes the first-step weighting matrix.

4.4.2 Estimation with $W_T = \hat{S}_{HAC}^{-1}$ or $W_T = \hat{S}_{HAC,\mu}^{-1}$

Now let us consider the same questions in the cases where either an uncentred or centred HAC estimator is used to construct the weighting matrix. Although there are some similarities between the two cases, there are sufficient differences to necessitate a separate treatment for each. It emerges that the distribution theory is very different from the cases considered above and non-standard in the sense that the estimator no longer converges at rate $T^{-1/2}$. In this sub-section, we concentrate on explaining the sources of these differences and so only provide heuristic arguments to justify the stated results. A more rigorous treatment can be found in Hall (2000) and Hall and Inoue (2003).

4.4.2.1 Estimation with $W_T = \hat{S}_{HAC,\mu}^{-1}$

First notice that $\hat{S}_{HAC,\mu}^{-1}$ satisfies the conditions for a valid weighting matrix given in Assumption 3.7 because by construction, $\hat{S}_{HAC,\mu}$ is positive semi-definite for finite T and converges in probability to the positive definite matrix S_*. This suggests that we can appeal to similar arguments as in the proof of Theorem 4.1 in order to deduce that $\hat{\theta}_T(2)$ converges in probability to some value.

Corollary 4.1 Probability Limit of $\hat{\theta}_T(2)$
Let $W_T = \hat{S}_{HAC,\mu}^{-1}$ and $\hat{S}_{HAC,\mu} \xrightarrow{p} S_*(1)$, a positive definite matrix. If Assumptions 3.1–3.2, 3.8–3.10, 4.1 hold and Assumption 4.2 holds at $\theta_* = \theta_*(2)$ then $\hat{\theta}_T(2) \xrightarrow{p} \theta_*(2)$.

Notice that in general $\theta_*(2) \neq \theta_*(1)$, the probability limit of the first step estimator, unless the weighting matrix on the first step, $W_T(1)$, is proportional to $S_*(1)^{-1}$. In practice, there is no reason to suppose that $W_T(1) = kS_*(1)^{-1}$ except by coincidence and so the probability limits of the first and second step estimators are different in most circumstances.

Now let us consider the limiting distribution of $\hat{\theta}_T(2)$. In the previous section, it is shown that Theorem 4.2 can be invoked to deduce the asymptotic normality of the second step estimator when W_T equals $\hat{S}_{SU}^{-1}$ or $\hat{S}_{SU,\mu}^{-1}$. However, such a strategy does not work here. The key difference is in the rate of convergence of W_T to W. While $\hat{S}_{SU,\mu}^{-1}$ converges to $\Gamma_0(1)^{-1}$ at rate $T^{-1/2}$, $\hat{S}_{HAC,\mu}^{-1}$ converges to $S_*(1)^{-1}$ at a slower rate. This means that $T^{1/2}[\hat{S}_{HAC,\mu}^{-1} - S_*(1)^{-1}]$ diverges as $T \to \infty$, and hence $T^{1/2}[\hat{\theta}_T(2) - \theta_*(2)]$ does the same. Therefore, in order to derive the limiting distribution, we must scale $\hat{\theta}_T(2) - \theta_*(2)$ by some other function of T which increases at a slower rate than $T^{1/2}$.

Since a similar story is going to emerge when $W_T = \hat{S}_{HAC}^{-1}$ – albeit with a different rate of convergence – it is more convenient to develop the analysis at a general level and then specialize the derived result to deduce the limiting distribution when $W_T = \hat{S}_{HAC,\mu}^{-1}$. Accordingly, we consider the case in which W_T converges to W at rate c_T^{-1} where c_T is a sequence of constants with the properties $c_T \to \infty$ with $T \to \infty$ and $c_T = o(T^{1/2})$. We also return to the generic notation of $\hat{\theta}_T$ for the estimator and θ_* for its plim to facilitate comparsion with Section 4.2. Our starting point is (4.8) with c_T substituted for $T^{1/2}$, that is

$$\begin{aligned} c_T(\hat{\theta}_T - \theta_*) &= (c_T/T^{1/2})H_{0,T}\{\, H_{1,T} \,+\, H_{2,T}\,\} \\ &= (c_T/T^{1/2})H_{0,T}H_{1,T} \,+\, (c_T/T^{1/2})H_{0,T}H_{2,T} \end{aligned} \tag{4.38}$$

where $H_{0,T}$, $H_{1,T}$ and $H_{2,T}$ are defined in (4.9)–(4.11). We now consider the behaviour of the two terms on the right-hand side of (4.38) in turn. In Section 4.2 it is shown that $H_{0,T}H_{1,T} = O_p(1)$, and an inspection of the argument reveals that this conclusion did not depend on the rate at which W_T converges to W. Therefore, the same arguments can be used here. However, since this term is multiplied by $c_T/T^{1/2}$ and $c_T = o(T^{1/2})$, it follows that

$$(c_T/T^{1/2})H_{0,T}H_{1,T} \xrightarrow{p} 0 \tag{4.39}$$

Now consider $(c_T/T^{1/2})H_{0,T}H_{2,T}$. Using (4.12)–(4.16) and $H_{2,T}(4) = 0$, it follows that

$$(c_T/T^{1/2})H_{0,T}H_{2,T} = (c_T/T^{1/2})H_{0,T}\{\, H_{2,T}(1) + H_{2,T}(2) + H_{2,T}(3)\,\} \tag{4.40}$$

Using similar arguments to Section 4.2 to analyse $H_{0,T}H_{2,T}(i), i = 1, 2$, it can be shown that

$$(c_T/T^{1/2})H_{0,T}[H_{2,T}(1) + H_{2,T}(2)] = H_{0,T}M_T c_T(\hat{\theta}_T - \theta_*) + o_p(1) \quad (4.41)$$

Therefore, combining (4.38)–(4.41), it follows that

$$c_T(\hat{\theta}_T - \theta_*) \; = \; [I_p \; - \; H_{0,T}M_T]^{-1}H_{0,T}\,(c_T/T^{1/2})H_{2,T}(3) \; + \; o_p(1) \quad (4.42)$$

Just as in Section 4.2, $H_{0,T}$ and M_T converge in probability to matrices of constants (under certain conditions) and so (4.42) implies the limiting behaviour of $c_T(\hat{\theta}_T - \theta_*)$ is driven by

$$(c_T/T^{1/2})H_{2,T}(3) \; = \; G_*^{'}c_T(W_T - W)\mu_* \quad (4.43)$$

To proceed further, we must make some assumption about $c_T(W_T - W)$ and so we return to the specific example of interest here. Using a similar argument to (4.19),

$$c_T[\hat{S}_{HAC,\mu}^{-1} - S_*(1)^{-1}] \; = \; -\,\hat{S}_{HAC,\mu}^{-1}\{\, c_T[\hat{S}_{HAC,\mu} \; - \; S_*(1)]\,\}S_*(1)^{-1} \quad (4.44)$$

and so it suffices to consider $c_T[\hat{S}_{HAC,\mu} \; - \; S_*(1)]$ because $S_*(1)^{-1} = O(1)$ and $\hat{S}_{HAC,\mu}^{-1} = O_p(1)$. To this end, it is useful to introduce the following notation. We define

$$\begin{aligned} \bar{S}_{*,T} &= \Gamma_{0,T} + \sum_{i=1}^{T-1} \omega_{i,T} \left(\Gamma_{i,T} \; + \; \Gamma_{i,T}^{'} \right) \\ S_{*,T} &= \Gamma_0 + \sum_{i=1}^{T-1} \omega_{i,T} \left(\Gamma_i \; + \; \Gamma_i^{'} \right) \end{aligned}$$

where $\Gamma_{i,T} = T^{-1}\sum_{t=i+1}^{T}[f(v_t, \theta_*) - g_T(\theta_*)][f(v_{t-i}, \theta_*) - g_T(\theta_*)]'$. Using these definitions, $c_T(\hat{S}_{HAC,\mu} - S_*(1))$ can be decomposed into the sum of three terms as follows,

$$c_T(\hat{S}_{HAC,\mu} - S_*(1)) \; = \; c_T(\hat{S}_{HAC,\mu} - \bar{S}_{*,T}) + c_T(\bar{S}_{*,T} - S_{*,T}) + c_T(S_{*,T} - S_*(1)) \quad (4.45)$$

Notice that the first component, $\hat{S}_{HAC,\mu} \; - \; \bar{S}_{*,T}$, represents the difference between the HAC evaluated at $\hat{\theta}_T(1)$ and $\theta_*(1)$; the second component, $\bar{S}_{*,T} - S_{*,T}$ is the difference between the HAC evaluated at $\theta_*(1)$ and the corresponding function evaluated at population instead of sample autocovariances; and the third component, $S_{*,T} \; - \; S_*(1)$ is the difference between the population analog to the HAC and the long run covariance matrix. Notice that the sum of the first two components is $\hat{S}_{HAC,\mu} \; - \; S_{*,T}$, and so can be interpreted as the error inherent in using the HAC estimator to estimate its population analog. The third component can then be interpreted as the bias induced by estimating $S_{*,T}$ instead of $S_*(1)$.

Hall and Inoue (2003) verify that under a set of plausible regularity conditions the three components in (4.45) behave as follows.

Assumption 4.7 Limiting Behaviour of the Components of $\hat{S}_{HAC,\mu} - S_*(1)$

1. $(T/b_T)^{1/2} vech(\hat{S}_{HAC,\mu} - \bar{S}_{*,T}) = o_p(1)$.

2. $(T/b_T)^{1/2} vech(\bar{S}_{*,T} - S_{*,T}) \xrightarrow{d} N(0, \Omega_\omega)$ *where* Ω_ω *is a positive definite matrix depending on the kernel* $\omega(.)$.

3. $\lim_{T\to\infty} b_T^k (S_{*,T} - S_*(1)) = C$ *where* $k > 0$ *is known as the characteristic exponent of the kernel* $\omega(.)$,[18] *and*

$$C = -\lim_{x\to 0} \left(\frac{1-\omega(x)}{|x|^k} \right) \sum_{j=-\infty}^{\infty} |j|^k \Gamma_j < \infty.$$

Before proceeding, it is worth briefly commenting on certain aspects of this assumption. In all our previous invocations of asymptotic normality, such as Lemma 4.1, the rate of convergence has been $T^{-1/2}$. The key difference here is in the form of $\hat{S}_{HAC,\mu} - S_*(1)$. Recall that $\hat{S}_{HAC,\mu}$ is itself a weighted sum of $T-1$ autocovariances (and their transposes). While we can apply the Central Limit Theorem to deduce $T^{1/2} vech\{\tilde{\Gamma}_i - \Gamma_i\}$ converges to normal distribution for fixed i, the rate of increase in the number of autocovariances included in $\hat{S}_{HAC,\mu}$ slows down the rate of convergence.[19] Notice also that the rate of convergence of all three components depends on the bandwidth, and the behaviour of the second and third components also depends on the kernel.

Using equation (4.45) and Assumption 4.7, it follows that the limiting behaviour of $c_T(\hat{S}_{HAC,\mu} - S_*(1))$ depends on the bandwidth and the kernel:

- if $\lim_{T\to\infty} T^{1/2}/b_T^{1/2+k} = 0$ then $(T/b_T)^{1/2}(\hat{S}_{HAC,\mu} - S_*(1)) \xrightarrow{d} N(0, \Omega_\omega)$;
- if $\lim_{T\to\infty} T^{1/2}/b_T^{1/2+k} = \phi \in (0,\infty)$ then $(T/b_T)^{1/2}(\hat{S}_{HAC,\mu} - S_*(1)) \xrightarrow{d} N(\phi C, \Omega_\omega)$;
- if $\lim_{T\to\infty} T^{1/2}/b_T^{1/2+k} = \infty$ then $plim_{T\to\infty} b_T^k(\hat{S}_{HAC,\mu} - S_*(1)) = C$.

Notice that neither the rate of convergence nor the nature of the limiting behaviour is the same in all three cases. In particular, if $\lim_{T\to\infty} T^{1/2}/b_T^{1/2+k} = \infty$ then the bias term, $S_{*,T} - S_*(1)$, becomes dominant and this causes $b_T^k(\hat{S}_{HAC,\mu} - S_*(1))$ to converge to a constant. As would be anticipated, these differences also manifest themselves in the limiting behaviour of the estimator. Using (4.42)-(4.45) and Assumption 4.7, the following three possibilities emerge for the limiting behaviour of $\hat{\theta}_T(2)$.

[18] Anderson (1994)[Section 9.3.2] defines the characteristic exponent and discusses its properties.

[19] For further discussion see Andrews (1991) or Hall and Inoue (2003).

Lemma 4.2 Limiting Behaviour of $\hat{\theta}_T(2)$ When $W_T = \hat{S}^{-1}_{HAC,\mu}$
Assume that: (i) $W_T = \hat{S}^{-1}_{HAC,\mu}$ and $\hat{S}_{HAC,\mu} \xrightarrow{p} S_*(1)$, a positive definite matrix; (ii) Assumptions 3.1, 3.8 and 4.7, and certain other regularity conditions hold.[20] *The limiting distribution is as follows:*

- *if $lim_{T\to\infty} T^{1/2}/b_T^{1/2+k} = 0$ then $(T/b_T)^{1/2}[\hat{\theta}_T(2) - \theta_*(2)] \xrightarrow{d} N(0, \Sigma_3)$;*
- *if $lim_{T\to\infty} T^{1/2}/b_T^{1/2+k} = \phi \in (0,\infty)$ then $(T/b_T)^{1/2}[\hat{\theta}_T(2) - \theta_*(2)] \xrightarrow{d} N(\phi H_{**}^{-1} G'_{**} C \mu_{**}, \Sigma_3)$;*
- *if $lim_{T\to\infty} T^{1/2}/b_T^{1/2+k} = \infty$ then $b_T^k[\hat{\theta}_T(2) - \theta_*(2)] \xrightarrow{p} H_{**}^{-1} G'_{**} C \mu_{**}$;*

where $\Sigma_3 = H_{**}^{-1} D' B \Omega B' D H_{**}^{'-1}$, $D = -(\mu_*(2)' S_*(1)^{-1} \otimes G'_{**} S_*(1)^{-1})$, B is the selection matrix defined by $vec\{S_*\} = B vech\{S_*\}$, H_{**} and G_{**} are respectively H_* and G_* in Assumption 4.6 evaluated at $\theta_* = \theta_*(2)$ instead of $\theta_*(1)$.

It is interesting to contrast this result with the corresponding discussion in the case where $W_T = \hat{S}^{-1}_{SU}$ or $\hat{S}^{-1}_{SU,\mu}$. Notice that unlike those previous cases, the asymptotic distribution of the second step estimator does not depend on the first step estimator. The reason is that one of the regularity conditions behind Assumption 4.7 is the restriction that $\hat{\theta}_T(1) - \theta_*(1) = O_p(T^{-1/2})$.[21] This means that $(T/b_T)^{1/2}[\hat{\theta}_T(1) - \theta_*(1)] = o_p(1)$, and so can have no effect on the large sample behaviour of $(T/b_T)^{1/2}[\hat{\theta}_T(2) - \theta_*(2)]$.

We now consider the iterated estimator. It is straightforward to extend Corollary 4.1 to $\hat{\theta}_T(i)$. However, the limiting distribution of the iterated estimator is going to be very complicated in general. Using a similar argument to (4.37), it follows that if $lim_{T\to\infty} T^{1/2}/b_T^{1/2+k} \in [0,\infty)$ the limiting distribution of $(T/b_T)^{1/2}[\hat{\theta}_T(i) - \theta_*(i)]$ depends on $\{(T/b_T)^{1/2}[\hat{\theta}_T(j) - \theta_*(j)],\ j = 2, 3, \ldots i-1\}$ in general. Notice that, this time, the dependence only goes back to the second step for the reasons discussed above.

4.4.2.2 Estimation with $W_T = \hat{S}^{-1}_{HAC}$

We now consider the case in which the second step estimator is calculated using the uncentred HAC estimator based on $\hat{\theta}_T(1)$. To begin, we must consider whether $\hat{S}^{-1}_{HAC}$ satisfies the conditions for a valid weighting matrix given in Assumption 3.7. Since this part of the analysis is generic to all steps, we return to our more general notation of $\hat{\theta}_T$ for the estimator and θ_* for its limit. In Section 4.3.3, it is shown that the large sample behaviour of $\hat{S}_{HAC}$ is identical to $S_* + B_T \mu_* \mu_*'$. The following lemma characterizes the implications of this structure for the large sample behaviour of $\hat{S}^{-1}_{HAC}$.

[20] These include $\hat{\theta}_T(1) - \theta_*(1) = O_p(T^{-1/2})$. See Hall and Inoue (2003) [Theorem 3] for a complete list of regularity conditions and also a rigorous proof.

[21] This condition is "plausible" because it is implied by Theorem 4.2.

Lemma 4.3 Limiting Behaviour of $\hat{S}_{HAC}^{-1}$
If $\hat{S}_{HAC} = S_* + B_T\mu_*\mu_*' + o_p(1)$ where $B_T = 1 + 2\sum_{i=1}^{T-1}\omega_{i,T}$, $B_T = O(b_T)$ and the bandwidth satisfies $b_T \to \infty$, $b_T = o(T^{1/2})$ then: $\hat{S}_{HAC}^{-1} \xrightarrow{p} S^+$ where

$$S^+ = S_*^{-1} - \frac{1}{\mu_*' S_*^{-1}\mu_*} S_*^{-1}\mu_*\mu_*' S_*^{-1} \tag{4.46}$$

Since the structure of this inverse is non-standard and also important below, we present a heuristic proof.[22] Since

$$\hat{S}_{HAC} = S_* + B_T\mu_*\mu_*' + o_p(1) \tag{4.47}$$

and $S_* + B_T\mu_*\mu_*'$ is nonsingular for any finite T, it follows that the large sample behaviour of $\hat{S}_{HAC}^{-1}$ can be deduced from $(S_* + B_T\mu_*\mu_*')^{-1}$. For any T, we have[23]

$$(S_* + B_T\mu_*\mu_*')^{-1} = S_*^{-1} - \frac{B_T}{1 + B_T\mu_*' S_*^{-1}\mu_*} S_*^{-1}\mu_*\mu_*' S_*^{-1} \tag{4.48}$$

recall that $b_T \to \infty$ as $T \to \infty$, and so it follows from (4.47)–(4.48) that

$$\hat{S}_{HAC}^{-1} \xrightarrow{p} \lim_{T\to\infty}(S_* + B_T\mu_*\mu_*')^{-1} = S_*^{-1} - \frac{1}{\mu_*' S_*^{-1}\mu_*} S_*^{-1}\mu_*\mu_*' S_*^{-1} = S^+$$

The matrix S^+ has two properties which play an important role in the analysis.

Corollary 4.2 Properties of S^+
(i) $rank(S^+) = q - 1$; (ii) the nullspace of S^+ is spanned by μ_*.

Notice that part (i) implies that $\hat{S}_{HAC}^{-1}$ converges to a singular matrix and so does not satisfy the conditions for a weighting matrix laid down in Assumption 3.7.

With this in mind, now consider the population analog to the second-step minimand when $W_T = \hat{S}_{HAC}^{-1}$. From Lemma 4.3, this minimand is given by

$$Q_0^{(2)}(\theta) = E[f(v_t,\theta)]' S^+ E[f(v_t,\theta)] \tag{4.49}$$

Using Corollary 4.2(ii), it can be seen that $Q_0^{(2)}(\theta)$ attains its minimum possible value of zero at $\theta = \theta_*$. To explore the implications of this structure for the estimator, we must impose some form of identification condition. The simplest such condition is to assume that this minimum is unique or, in other words, that there is no other value of θ which generates a value of $\mu(\theta)$ in the nullspace of S_*.

Assumption 4.8 Identification Condition
$S^+ E[f(v_t,\theta)] \neq 0$ *for any* $\theta \in \Theta \setminus \{\theta_*\}$.

[22] See Hall (2000) for a rigorous proof.
[23] See the matrix inversion result in (4.33).

In some cases it is possible to verify that this assumption holds, but in other models its imposition is more an article of faith. Once identification is assumed we can use the same sequence of arguments as in Theorem 4.1 to deduce the following result.

Corollary 4.3 Probability Limit of $\hat{\theta}_T(2)$
Let $W_T = \hat{S}_{HAC}^{-1}$. If (i) Assumptions 3.1,3.2, 3.8–3.10, 4.1 and 4.8 hold; (ii) $\hat{\theta}_T(1) \xrightarrow{p} \theta_*(1)$; (iii) $\hat{S}_{HAC}^{-1} \xrightarrow{p} S^+$; then: $\hat{\theta}_T(2) \xrightarrow{p} \theta_*(1)$.

Corollary 4.3 states that the GMM estimator converges to the same probability limit in both the first- and second-step estimations. It is straightforward to extend this result to the iterated estimator as well. Therefore if $W_T = \hat{S}_{HAC}^{-1}$ then the probability limits of the estimators exhibit the same type of behaviour as they would in a correctly specified model. This also implies that the iterated estimation converges after just two steps with probability one. Therefore, the second step and iterated estimators are asymptotically identical. One other consequence of Corollary 4.3 is that there is no need to index population quantities, such as μ_* or S_*, by i, and so we drop this index for the rest of this section.

Now consider the limiting distribution of $\hat{\theta}_T(2)$. As mentioned in the previous section, the rate of convergence is slower than $T^{1/2}$, and so we must return to (4.42) in order to start the analysis. However, this time there are some additional simplications of which we can take advantage. Corollary 4.2(ii) implies $S^+\mu_* = 0$ and so both $p\lim_{T\to\infty} M_T = 0$ and $(W_T - W)\mu_* = W_T\mu_*$. Therefore, (4.42) reduces to

$$c_T(\hat{\theta}_T - \theta_*) \;=\; H_{0,T}G_*^{'}c_T\hat{S}_{HAC}^{-1}\mu_* \;+\; o_p(1) \tag{4.50}$$

The key question is what is the appropriate choice of c_T. To answer this question, it is convenient to rewrite (4.50) as

$$c_T(\hat{\theta}_T - \theta_*) \;=\; H_{0,T}G_*^{'}c_TS_T^{-1}\mu_* + H_{0,T}G_*^{'}c_T(\hat{S}_{HAC}^{-1} - S_T^{-1})\mu_* + o_p(1) \tag{4.51}$$

where $S_T = S_* + B_T\mu_*\mu_*^{'}$. Hall and Inoue (2003) establish that the following results hold under plausible regularity conditions.

$$\begin{aligned}
H_{0,T}G_*^{'} &= -(G_*^{'}S^+G_*)^{-1}G_*^{'} = O(1) && (4.52)\\
b_TS_T^{-1}\mu_* &= \frac{b_T}{1 + B_T\mu_*^{'}S_*^{-1}\mu_*}S_*^{-1}\mu_* = O(1) && (4.53)\\
(\hat{S}_{HAC}^{-1} - S_T^{-1})\mu_* &= \hat{S}_{HAC}^{-1}(S_T - \hat{S}_{HAC})S_T^{-1}\mu_* && (4.54)\\
&= O_p(1)O_p(b_T/T^{1/2})O_p(b_T^{-1}) = O_p(T^{-1/2}) && (4.55)
\end{aligned}$$

If these results are used in conjunction with (4.51) then a two-part answer emerges to our question.

Lemma 4.4 Rate of Convergence for $\hat{\theta}_T(2)$
Let $W_T = \hat{S}_{HAC}^{-1}$. If (a) Assumptions 3.1, 3.8,and 4.8 hold; (b) $\hat{S}_{HAC}^{-1} \xrightarrow{p} S^+$;

(c) Equations (4.52)–(4.55) hold; (d) certain other regularity conditions hold.[24] *Then (i)* $b_T[\hat{\theta}_T(2) - \theta_*] \xrightarrow{p} \beta(G_*'S^+G_*)^{-1}G_*'S_*^{-1}\mu_*$ *if* $G_*'S_*^{-1}\mu_* \neq 0$ *where* $\beta = -\lim_{T\to\infty}(b_T/B_T)/(\mu_*'S_*^{-1}\mu_*)$*; (ii)* $T^{1/2}[\hat{\theta}_T(2) - \theta_*] = O_p(1)$ *if* $G_*'S_*^{-1}\mu_* = 0$.

Notice that $G_*'S_*^{-1}\mu_* = 0$ implies that $p\lim_{T\to\infty}\hat{\theta}_T(1) = \theta_*$ is the solution to the population analog to the first order conditions when $W_T = \hat{S}_{HAC,\mu}^{-1}$. Therefore part (ii) is only relevant in the unlikely eventuality that the probability limit of the first step weighting matrix is proportional to the long run variance S_*.[25] So the most relevant part of the lemma in practice is likely to be part (i). Lemma 4.4(i) states that $b_T[\hat{\theta}_T(2) - \theta_*]$ converges to a degenerate distribution, or in other words a constant vector. This behaviour is similar to the case when $W_T = \hat{S}_{HAC,\mu}^{-1}$ and $lim_{T\to\infty}T^{1/2}/b_T^{1/2+k} = \infty$, and has a correspondingly similar explanation. However, this time it is the bias induced by the use of uncentred autocovariance matrices in the HAC that is dominant.

4.5 The Estimated Sample Moment

We now consider the large sample behaviour of the estimated sample moment. In contrast to the results derived for the estimator, this analysis is uncomplicated and independent of the weighting matrix.

The analysis rests in part on an application of the Weak Law of Large Numbers. This law has not yet been invoked in our discussion of the nonlinear dynamic model, and so we now state it formally.[26]

Lemma 4.5 Weak Law of Large Numbers
Let $\bar{\theta} \in \Theta$, $E[f(v_t, \bar{\theta})] = \mu(\bar{\theta})$ *and Assumptions 3.1, 3.2, 3.8 and 3.10 hold then* $T^{-1}\sum_{t=1}^{T} f(v_t, \bar{\theta}) \xrightarrow{p} \mu(\bar{\theta})$.

Let $\hat{\theta}_T$ be a GMM estimator and assume it converges to some point in the parameter space, θ_*. Notice that this definition is sufficiently broad to include all the choices of weighting matrix considered above. In this case, it is straightforward to establish the following result.

Theorem 4.3 Large Sample Behaviour of the Estimated Sample Moment
Let (i) Assumptions 3.1, 3.2, 3.8–3.10, 4.1 and 4.3 hold; (ii) $\hat{\theta}_T \xrightarrow{p} \theta_*$ *for some* $\theta_* \in \Theta$. *Then* $g_T(\hat{\theta}_T) \xrightarrow{p} \mu(\theta_*)$ *where* $||\mu(\theta_*)|| > 0$.

Proof:
Using the Mean Value Theorem, it follows that

$$g_T(\hat{\theta}_T) \;=\; g_T(\theta_*) \;+\; G_T(\hat{\theta}_T, \theta_*, \lambda_T)(\hat{\theta}_T \;-\; \theta_*)$$

[24] See Hall and Inoue (2003) [Theorem 4].
[25] It is for this reason that we do not characterize the nature of the limiting behaviour beyond the given order in probability statement.
[26] See Wooldridge (1994) for discussion of Laws of Large Numbers in dynamic models.

The result then follows directly because under the stated conditions $G_T(\hat{\theta}_T, \theta_*, \lambda_T) \xrightarrow{p} G_* = O(1)$, $\hat{\theta}_T - \theta_* \xrightarrow{p} 0$, $g_T(\theta_*) \xrightarrow{p} \mu(\theta_*)$ and $\|\mu(\theta)\| > 0$ for all $\theta \in \Theta$. $\diamond$

The important consequence of Theorem 4.3 is that $\|T^{1/2} g_T(\hat{\theta}_T)\|$ diverges to infinity at rate $T^{1/2}$. Therefore, taken together, Theorems 3.3 and 4.3 imply that $T^{1/2} g_T(\hat{\theta}_T)$ converges to a mean zero normal distribution if the model is correctly specified but diverges to infinity if the model is misspecified. This property is exploited in the construction of the model specification tests which are reviewed in the next chapter.

4.6 Summary of Consequences of Misspecification for GMM Estimation

It is useful to begin by recalling the properties of the GMM estimator in correctly specified models. Since Assumptions 4.1 and 3.1 imply $q > p$, we confine our attention here to the case in which the parameter vector is overidentified.

Properties of GMM in correctly specified models:

- $\hat{\theta}_T$ converges in probability to θ_0 for any choice of W_T which satisfies Assumption 3.7.
- $T^{1/2}(\hat{\theta}_T - \theta_0)$ converges to a normal distribution and the choice of weighting matrix only affects this distribution in the variance via W.
- The two step and iterated estimators have the same asymptotic properties.
- $T^{1/2} g_T(\hat{\theta}_T)$ converges to a mean zero normal distribution.

In contrast, it has been shown in this chapter that the following properties hold in misspecified models.

Properties of GMM in misspecified models:

- The probability limit of $\hat{\theta}_T$ depends on W in general.
- The rate of convergence of $\hat{\theta}_T$ to its limit, θ_*, depends on the rate of convergence of W_T to W, and the limiting distribution of $c_T(\hat{\theta}_T - \theta_*)$ depends on that of $c_T(W_T - W)$.
- The two step and iterated estimators have different asymptotic properties, and the asymptotic distribution of $\hat{\theta}_T(i)$ depends on the estimators from the previous steps.[27]

[27] This statement excludes the case in which $W_T = \hat{S}_{HAC}^{-1}$.

- $T^{1/2} g_T(\hat{\theta}_T)$ diverges.

So, basically, everything is different. Most importantly, misspecification means that in most cases we have not estimated what we anticipated. This is sufficient by itself to make all subsequent inferences misleading, and so provides a motivation for the model specification tests described in the next chapter.

5

Hypothesis Testing

The previous two chapters describe the behaviour of the estimator and its associated statistics in both correctly specified and misspecified models. The next step is to develop inference procedures through which the estimation results can be used to learn about the underlying model. There are three broad questions which naturally arise in this context – Is the model correctly specified? Does the model satisfy restrictions implied by economic/statistical theory? Which of two competing models is correct? Within the GMM framework, all these questions are addressed via hypothesis tests concerning either population moment conditions or the parameter vector or both. In practice, these inferences are most often – if not always – based on the two step or iterated estimator. Therefore, we focus attention *exclusively* on this case throughout the chapter.

Misspecification has the potential to make the estimator inconsistent, and so to render all subsequent inferences misleading. Therefore, it is prudent to begin by testing whether the model is correctly specified. Within our framework, the economic/statistical model implies that v_t satisfies the population moment condition $E[f(v_t, \theta_0)] = 0$. Since this is the starting point for our estimation, it is clearly desirable to test whether the sample are consistent with the hypothesis that this condition holds in the population. In most of the applications in Table 1.1, q is greater than p and so the overidentifying restrictions are available to form the basis for a test of the model specification. Section 5.1 extends the earlier discussion of the overidentifying restrictions test to nonlinear dynamic models. It also presents a formal analysis of the statistic's behaviour in both correctly and misspecified models. The latter involves two forms of misspecification: "non-local" and "local". It is most common in the literature to analyze the power properties of various statistics using local misspecification framework. This approach is particularly attractive in cases where more than one statistic is available to test a hypotheses because it facilitates a meaningful comparison of the candidates' power properties. However we include both here because it is only via a non-local analysis that it becomes possible to uncover the dependence of the limiting behaviour of the statistic on the method of covariance matrix estimation. This issue is only illustrated explicitly for the

overidentifying restrictions test but equally applies to the other tests of model specification described below.

In some cases, a priori information may indicate that the potential misspecification is confined to certain elements of the population moment condition. In certain circumstances, it is possible to exploit this information to construct a more powerful test than the overidentifying restrictions test. Section 5.2 describes when this is possible and presents statistics for testing so-called hypotheses about a subset of the moment conditions. If the model is validated by the previous test statistics, then it is reasonable to use the estimation results as a basis for inference about the phenomena captured by the model. In many economic models, these inferences reduce to hypotheses about restrictions on the parameter vector. Section 5.3 discusses methods for testing the hypothesis that the parameter vector satisfies a set of nonlinear restrictions of the form $r(\theta_0) = 0$. These types of restrictions naturally arise in many economic models and so test results can often provide useful insights about the underlying economic structure.

One of the main assumptions behind GMM is that the population moment condition holds throughout the entire sample; in other words the model is assumed to be "structurally stable". A natural concern is whether the population moment condition is only true for part of the sample in which case the model exhibits "structural instability". Section 5.4 describes various methods for testing structural stability. The differences between the tests are most easily understood by considering their sensitivity to instability of identifying and overidentifying restrictions separately. It is also shown how this decomposition can be exploited to develop tests which can distinguish between instability in the parameters alone and instability of a more general form.

The foregoing hypothesis tests are by far the most common in the types of applications in Table 1.1, and so merit detailed discussion. Section 5.5 provides a brief summary of certain other inference techniques which have been proposed in the literature. Section 5.5.1 discusses non-nested hypothesis tests, which have been proposed as a method of choosing between two competing specifications. In some cases, one competing model can be nested within the other and so it is possible to assess which is more appropriate using the types of procedure described in Sections 5.1 through 5.3. However, in other cases the competing models are not nested in this fashion, and so alternative procedures must be developed. As will be seen, this type of question is much harder to address within the majority of models listed in Table 1.1 without further restrictions. Section 5.5.2 describes so-called "Hausman" tests which involve the comparison of two estimators based on different sets of population moment conditions. Section 5.5.3 concludes the chapter with a discussion of "conditional moment" tests. These tests are commonly employed in models estimated by Maximum Likelihood to assess whether the assumed distribution is correct. Although Maximum Likelihood is not a focus of this book, these tests are included here because they have some important similarities and differences with the other procedures discussed above. Section 5.6 concludes with a brief summary of the chapter.

Finally, two omissions should be noted. First, this chapter focuses exclusively on the asymptotic properties of these tests. In most cases, the original articles did not provide simulation evidence on the finite sample properties of their proposed tests. Instead this type of evidence tends to be found in studies which sought to examine the finite sample behaviour of all aspects of GMM in the context of a particular model. We believe that it is more instructive to review these studies in a similar spirit, and so further discussion of this aspect of hypothesis testing can be found in Chapter 6. Secondly, it is beyond the scope of this book to provide an introduction to the general theory of statistical hypothesis testing; this material can be found in many other sources such as Lehmann (1959) or Cox and Hinckley (1974).

5.1 The Overidentifying Restrictions Test

Section 2.5 introduced the idea of using the overidentifying restrictions to test whether the model is correctly specified. Although this earlier discussion is in the context of the linear model, the underlying intuition is not specific to this structure. In this section we extend the overidentifying restrictions test to nonlinear dynamic models and formally analyse its properties in correctly specified and misspecified models. There are two main approaches to this type of analysis in misspecified models. The first employs the framework in Chapter 4, which it is now useful to refer to as *non-local* misspecification. The second is based on a *local* form of misspecification. The distinction between them is best motivated by briefly reconsidering the nature of Assumption 4.1. This assumption has two important implications. First, there is no value of θ for which $E[f(v_t, \theta)] = 0$ – that is, the model is misspecified. Secondly, $E[f(v_t, \theta)] = \mu(\theta)$ – that is, the "size" of the misspecification, $\mu(\theta)$, is the same for all t, regardless of the sample size. In other words, the model is wrong and the situation does not change as the sample size increases. This scenario contrasts with local misspecification in which the model is misspecified for finite T, but the size of the misspecification decreases with T so that in the limit the model is correct. This misspecification is "local" in the sense that the data are generated by a sequence of processes which become closer and closer to satisfying H_0 as T increases and in the limit do satisfy this hypothesis. As might be imagined, a different analysis is required for each type of misspecification. Therefore, we break our discussion down into three parts. Section 5.1.1 introduces the test statistic and derives its asymptotic distribution in correctly specified models. Section 5.1.2 considers the behaviour of the statistic in non-locally misspecified models, and Section 5.1.3 presents its local counterpart. As will be seen, the conclusions from these two types of analysis are couched in very different terms. Section 5.1.4 concludes the discussion with a demonstration that each form of analysis leads to the same *qualitative* conclusions about the properties of the test.

5.1.1 The Statistic and its Asymptotic Distribution in Correctly Specified Models

Section 2.5 introduces the idea of using the overidentifying restrictions test statistic to assess the adequacy of the model specification. It can be recalled that the idea behind the test is simple: if $E[z_t u_t(\theta_0)] = 0$ then the estimated sample moment, $T^{-1}Z'u(\hat{\theta}_T)$, should be zero once allowance is made for sampling error. The same logic can be applied equally in nonlinear dynamic models: if $E[f(v_t, \theta_0)] = 0$ then $g_T(\hat{\theta}_T)$ should be approximately zero. This insight motivated Hansen (1982) to propose testing the null hypothesis

$$H_0 : E[f(v_t, \theta_0)] \; = \; 0 \tag{5.1}$$

using the overidentifying restrictions test statistic

$$J_T \; = \; T\, g_T(\hat{\theta}_T)^{'}\, \hat{S}_T^{-1} g_T(\hat{\theta}_T) \tag{5.2}$$

where, as a reminder, $\hat{\theta}_T$ is the second step (or iterated) estimator. This statistic is easily recognized to be the generalization of Sargan's (1958) statistic (equation (2.42) above) to nonlinear dynamic models. Hansen (1982, Lemma 4.2) derived its limiting distribution under H_0, and this result is given in the following theorem.

Theorem 5.1 The Asymptotic Distribution of the Overidentifying Restrictions Test Statistic
If (i) Assumptions 3.1–3.5, 3.8–3.13 hold; (ii) $\hat{S}_T$ is positive semi-definite and converges in probability to S; then $J_T \xrightarrow{d} \chi^2_{q-p}$.

Proof:
Since $plim\, \hat{S}_T = S$ it follows from Slutsky's Theorem (Lemma 1.1) that $J_T - \tilde{J}_T \xrightarrow{p} 0$ where

$$\tilde{J}_T = T g_T(\hat{\theta}_T)' S^{-1} g_T(\hat{\theta}_T) \tag{5.3}$$

Therefore, the theorem can be established by proving that $\tilde{J}_T$ has the stated limiting distribution. Using Theorem 3.3 evaluated at $W = S^{-1}$, we obtain

$$\tilde{J}_T \xrightarrow{d} \; \|[I_q - P(\theta_0)]n_q\|^2 \; = \; n_q^{'}[I_q - P(\theta_0)]n_q \tag{5.4}$$

where n_q denotes a $(q \times 1)$ random vector with a standard normal distribution. Now $I_q - P(\theta_0)$ is a projection matrix whose rank is $q - p$ by Assumption 3.4.[1] The desired result then follows from (5.4) and Rao (1973, p.186). $\diamond$

Notice that Theorem 4.1 holds for any choice of covariance matrix estimator which is both positive semi-definite and consistent for S under the assumption that *the model is correctly specified*. This class includes any estimators in Sections 3.5 and 4.3 which adequately capture the dynamic structure of $f(v_t, \theta_0)$.

[1] Recall that Assumption 3.4 implies Assumption 3.6 and hence that $rank[F(\theta_0)] = p$.

Although we have stated Theorem 5.1 in terms of the two step or iterated GMM estimator, intuition suggests a similar result holds for minimand of the continuous updating estimator.[2] In fact, the proof of Theorem 5.1 is easily adapted to show that under H_0

$$J_{cont,T} \;=\; TQ_{cont,T}(\hat{\theta}_T) \;\xrightarrow{d}\; \chi^2_{q-p} \tag{5.5}$$

where - with an abuse of notation - $\hat{\theta}_T$ is the now the continuous updating estimator.[3] However, while the asymptotic distributions of $J_{cont,T}$ and J_T are the same, the numerical values differ in a predicatable way under certain circumstances. Specifically, if J_T is based on the iterated estimator and these iterations converge, then it follows from the definition of the continuous updating estimator that $J_{cont,T}$ cannot exceed J_T.[4]

This statistic has become a standard diagnostic for models estimated by GMM and is routinely calculated in most computer packages. In Section 2.5, we discussed the interpretation of this test in general terms. We now complement those earlier remarks with a more formal analysis of the statistic's behaviour in misspecified models in Sections 5.1.2 and 5.1.3.

5.1.2 Non-Local Misspecification

Our analysis of GMM in misspecified models is premised on Assumption 4.1.[5] As mentioned above, this misspecification is refered to as "non-local" because the "size" of the misspecification, $\mu(\theta)$, is the same for all observations and sample sizes. Intuition suggests that if the model is wrong for every observation then the evidence against it must mount up as the sample increases with the result that the model is rejected with probability one in the limit. In essence this intuition is correct, but there is an important caveat concerning the calculation of the covariance matrix. The analysis in this section is based on Hall (2000).

Before, we present the more formal analysis, it is useful to develop a heuristic understanding of the way in which the covariance matrix estimator can play such a crucial role. Recall that the overidentifying restrictions test is a quadratic form in $T^{1/2}g_T(\hat{\theta}_T)$ and $\hat{S}_T^{-1}$. Theorem 4.3 indicates that $T^{1/2}g_T(\hat{\theta}_T)$ diverges under non-local specification. Intuition suggests that this behaviour is inherited by J_T provided $\hat{S}_T^{-1}$ converges in probability to a positive definite matrix. However, it can be recalled from Section 4.4 that the inverse of certain covariance matrix estimators – $\hat{S}_{HAC}^{-1}$ in particular – only converge to a positive semi-definite matrix in misspecified models and in these cases it is no longer so obvious that J_T diverges. For this reason, it is most convenient to separate our analysis into two parts depending on the limiting behaviour of $\hat{S}_T^{-1}$. There is one other aspect of this heuristic discussion, which should be noted. We have made no mention

[2] See Section 3.7.

[3] We omit the details for brevity. See Hansen, Heaton, and Yaron (1996) for further discussion.

[4] See Section 6.3 for further discussion.

[5] See Chapter 4.

of whether $\hat{S}_T$ is a consistent estimator of S_*. The key issue is only whether or not $p\lim_{T\to\infty} \hat{S}_T^{-1}$ is positive definite for which consistency is sufficient but not necessary.

We begin with the more standard case in which $\hat{S}_T^{-1}$ converges to a positive definite limit. Inspection of Section 4.3 reveals that this case covers the estimators: $\hat{S}_{SU}$, and the versions of the covariance matrices based on $f(v_t,\theta)-g_T(\theta)$.[6] For this analysis, we require the second step estimator to converge in probability to some constant limit. Below we impose this condition directly for simplicity because more primitive conditions depend in part on the covariance matrix estimator; see Chapter 4.[7]

Theorem 5.2 Large Sample Behaviour of J_T: Part (i)
If (i) Assumptions 3.1, 3.2, 3.8–3.10, 4.1 and 4.3 hold; (ii) $\hat{S}_T^{-1}$ satisfies Assumption 3.7; (iii) $\hat{\theta}_T \xrightarrow{p} \theta_*$ for some $\theta_* \in \Theta$; then: $T^{-1}J_T \xrightarrow{p} c$ where $0 < c < \infty$ and so $\lim_{T\to\infty} P[J_T > c_\alpha] = 1$, where c_α is the $100(1-\alpha)^{th}$ percentile of the χ^2_{q-p} distribution.

The basic outline of the proof has been anticipated above, but for completeness we now fill in the details.
Proof:
Let W denote the probability limit of $\hat{S}_T^{-1}$ and $\mu_* = E[f(v_t,\theta_*)]$. From Theorem 4.3 and Slutsky's Theorem (Lemma 1.1) it follows that

$$T^{-1}J_T \;=\; \mu_*^{'}W\mu_* \;+\; o_p(1) \tag{5.6}$$

Since W is positive definite and $\mu_* \neq 0$ by Assumption 4.1, it follows from (5.6) that $T^{-1}J_T \xrightarrow{p} c = \mu_*^{'}W\mu_* > 0$. Therefore, $J_T = Tc + o_p(T)$ increases at rate T and so tends to ∞ in probability as $T \to \infty$, which gives the desired result. ◇

In statistical parlance, Theorem 5.2 states that J_T is a *consistent test* of $H_0 : E[f(v_t,\theta_0)] = 0$ against the alternative that the data satisfy Assumption 4.1.[8]

We now consider what happens if $\hat{S}_{HAC}^{-1}$ is used as the weighting matrix on the second step. It can be recalled from Lemma 4.3 that $\hat{S}_{HAC}^{-1}$ converges in probability to a positive semi-definite matrix and that the form of this limit has important implications for the two step estimator. We now establish that this limiting behaviour also has important consequences for the behaviour of the overidentifying restrictions test.

[6] $\hat{S}_{VARMA}$ is omitted from this list because, at time of writing, its limiting behaviour in misspecified models is unknown; see the discussion in Section 4.3.

[7] For the purposes of comparison with Chapter 4, note that here we suppress the (2) index on both $\hat{\theta}_T$ and θ_* for ease of notation.

[8] This is a somewhat unfortunate terminology since we have already used the term consistency to refer to a property of an estimator. However, the meaning should be obvious from the context.

Theorem 5.3 Large Sample Behaviour of J_T: Part (ii)
If: (i) Assumptions 3.1, 3.2, 3.8–3.10, 4.1, 4.3 and 4.4 hold; (ii) $\hat{S}_T = \hat{S}_{HAC} = S_* + B_T\mu_*\mu_*' + o_p(1)$ *where* $B_T = 1 + 2\sum_{i=1}^{T-1}\omega_{i,T}$, $B_T = O(b_T)$, $\lim_{T\to\infty} B_T/b_T \neq 0$ *and the bandwidth satisfies* $b_T \to \infty$, $b_T = o(T^{1/2})$; *then:* $J_T = O_p(T/b_T)$.

Proof:
Let the minimand on the second step GMM estimation be $Q_T^{(2)}(\theta)$. By definition $Q_T(\hat{\theta}_T(2)) \leq Q_T(\theta_*)$, and so it is sufficient to prove that $TQ_T(\theta_*) = O_p(T/b_T)$. By the Cauchy–Schwarz inequality[9] and condition (ii), we have

$$TQ_T^{(2)}(\theta_*) \;\leq\; |T/b_T||b_T/B_T||B_TQ_T^{(2)}(\theta_*)| \tag{5.7}$$

Since $T/b_T = O(T/b_T)$ and condition (ii) implies $b_T/B_T = O(1)$ we concentrate here on showing that $B_TQ_T^{(2)}(\theta_*) = O_p(1)$. Since

$$B_TQ_T^{(2)}(\theta_*) \;= B_T^{1/2}g_T(\theta_*)'\hat{S}_T^{-1}B_T^{1/2}g_T(\theta_*)$$

we first consider $B_T^{1/2}g_T(\theta_*)$. By definition, we have

$$B_T^{1/2}g_T(\theta_*) \;=\; B_T^{1/2}\mu_* \;+\; (B_T/T)^{1/2}T^{-1/2}\sum_{t=1}^{T}[f(v_t,\theta_*) - \mu_*] \tag{5.8}$$

Now Lemma 4.1 implies that $T^{-1/2}\sum_{t=1}^{T}[f(v_t,\theta_*) - \mu_*] = O_p(1)$. Furthermore we have assumed that $b_T = o(T^{1/2})$, and so (5.8) implies $B_T^{1/2}g_T(\theta_*) = B_T^{1/2}\mu_* + o_p(1)$. Therefore, it follows that

$$\begin{aligned} B_TQ_T(\theta_*) \;&=\; B_T\mu_*'\hat{S}_T^{-1}\mu_* \;+\; o_p(1) &(5.9)\\ &=\; B_T\mu_*'S_T^{-1}\mu_* \;+\; B_T\mu_*'(\hat{S}_T^{-1} - S_T^{-1})\mu_* \;+\; o_p(1) &(5.10)\end{aligned}$$

Using (4.53) it can be shown that

$$B_T\mu_*'S_T^{-1}\mu_* \;=\; \frac{B_T\mu_*'S_*^{-1}\mu_*}{1 + B_T\mu_*'S_*^{-1}\mu_*} \;=\; O(1) \tag{5.11}$$

Now consider the second term in (5.10), that is

$$B_T\mu_*'(\hat{S}_T^{-1} - S_T^{-1})\mu_* \;=\; B_T\mu_*'\hat{S}_T^{-1}(S_T \;-\; \hat{S}_T)S_T^{-1}\mu_* \;= n_{2,T}, \text{ say}$$

From (4.54)–(4.55), it follows that $n_{2,T} = o_p(1)$, and so, using this result with (5.11) in (5.10), we have $B_TQ_T(\theta_*) = O_p(1)$. The desired result then follows from (5.7). ◇

Theorem 5.3 indicates that J_T cannot increase at a faster rate than T/b_T when $W_T = \hat{S}_{HAC}^{-1}$. By itself, this result does not imply J_T increases at that

[9] See Apostol (1974) [p.294].

rate, although this is in fact the case. Therefore, the overidentifying restrictions test is still consistent.

Together Theorems 5.2 and 5.3 indicate there is a difference in the rate at which J_T diverges depending on how the HAC is calculated. If a centred HAC estimator is used then J_T increases at rate T, but if an uncentred HAC estimator is used then J_T increases at rate T/b_T. Notice if we use an uncentred HAC with the optimal bandwidth then there is also a difference in the rate of increase of J_T between the kernels.[10] With the Bartlett, J_T increases at rate $T^{2/3}$, whereas with the Parzen and Quadratic Spectral kernels, J_T increases at rate $T^{4/5}$. Hall (2000) provides simulation evidence which illustrates that the failure to centre the HAC can have a substantial impact on the magnitude of the statistic in finite samples as well. It is less clear whether this difference in rates also manifests itself in differing power properties for the two versions of the test. For power calculations at a fixed significance level, it is only the magnitude of the statistic relative to the critical point which matters. Intuition suggests that there may be circumstances in which the two versions of the tests have different finite sample power properties but this remains an open research question. However, the rate of increase is important for the construction of moment selection procedures based on the overidentifying restrictions test; see Section 7.3.1.

5.1.3 Local Misspecification

So far our analysis has considered the scenarios in which the model is either correctly specified or subject to non-local misspecification. The contrast between these two is stark. If the model is correct then the following holds: (a) the population moment condition is true for all t; (b) the parameter estimator is consistent; (c) $T^{1/2}g_T(\hat{\theta}_T)$ converges to a mean zero normal distribution; and (d) it is only necessary to capture the dynamic structure of f_t to construct a consistent estimator of the long run variance. In contrast if there is non-local misspecification then: (a) the population moment condition is invalid for all t; (b) the parameter estimator is likely to be inconsistent; (c) $T^{1/2}g_T(\hat{\theta}_T)$ diverges; and (d) the construction of a consistent covariance matrix estimator must account for both the non-zero mean and the dynamic structure of f_t. In this section, we move to a third scenario which lies between these two extremes. *Local* misspecification captures the case where the population moment condition is invalid for any finite T but the size of the violation is $O(T^{-1/2})$ and so disappears in the limit. This rate of decrease ensures the misspecification does not affect the probability limits of either the parameter or covariance matrix estimators, but does manifest itself in the mean of the limiting distribution of $T^{1/2}g_T(\theta_0)$ and consequently the asymptotic distributions of the estimator and estimated sample moment as well. Newey (1985*a*) was the first paper to present an analysis of the overidentifying restrictions test under local alternatives. However, we take a different approach to the construction of local misspecification which was

[10] See Table 3.4.

first exploited in this context by Hall (1999). The qualitative conclusions are the same as Newey's (1985*a*) but the route to them is slightly different.

To introduce the local misspecification framework, it is most convenient to begin with the transformed population moment condition introduced in Section 3.3. Since two–step estimation involves $W = S^{-1}$ on the final step, we express the hypothesis that Assumption 3.3 holds by

$$H_0 : \; S^{-1/2}\, E[f(v_t, \theta_0)] = 0 \tag{5.12}$$

The advantage of this approach is that it allows H_0 to be decomposed into hypotheses about the identifying and overidentifying restrictions. To this end, we once again set $P(\theta) = F(\theta)[F(\theta)^{'} F(\theta)]^{-1} F(\theta)^{'}$ where $F(\theta) = S^{-1/2} E[\partial f(v_t, \theta)/\partial \theta^{'}]$. It then follows from (3.19)-(3.20) that

$$\begin{aligned} H_0 : \quad & H_0^I \,\&\, H_0^O \\ H_0^I : \quad & P(\theta_0)\, S^{-1/2} E[f(v_t, \theta_0)] = 0 \\ H_0^O : \quad & [I_q - P(\theta_0)]\, S^{-1/2} E[f(v_t, \theta_0)] = 0 \end{aligned}$$

where H_0^I, H_0^O are respectively the hypotheses that the identifying and overidentifying restrictions hold at θ_0. Since the transformed population moment can always be decomposed into

$$S^{-1/2}\, E[f(v_t, \theta)] \;=\; P(\theta)\, S^{-1/2} E[f(v_t, \theta)] \;+\; [I_q - P(\theta)]\, S^{-1/2} E[f(v_t, \theta)] \tag{5.13}$$

we can characterize the local misspecification in terms of violations of the identifying and overidentifying restrictions. To this end, we introduce the following sequences of local alternatives to H_0^I and H_0^O

$$\begin{aligned} H_{A,T}^I : \quad & P(\theta_0)\, S^{-1/2} E_T[f(v_t, \theta_0)] = T^{-1/2} P(\theta_0)\, \eta_I = T^{-1/2} \mu_I \\ H_{A,T}^O : \quad & [\, I_q - P(\theta_0)\,] S^{-1/2} E_T[f(v_t, \theta_0)] = T^{-1/2} [\, I_q - P(\theta_0)\,] \eta_O = T^{-1/2} \mu_O \end{aligned}$$

in which $\mu_I \neq 0, \mu_O \neq 0$ and $E_T[.]$ denotes expectations with respect to the joint probability distribution of $\{v_t; t = 1, 2, \ldots T\}$. The reason for this subscript on the expectation operator is discussed below, but first we briefly consider the nature of these two alternatives. Notice that under $H_{A,T}^I$ the identifying restrictions are violated for finite T, but the "size" of this violation decreases as T increases and disappears in the limit as $T \to \infty$. Clearly, $H_{A,T}^O$ implies a similar pattern of violations of the overidentifying restrictions. This technical device for constructing local alternative hypotheses is known as *Pitman drift* after Pitman (1949) who first introduced it.[11] As mentioned above, equation (5.13) can be used to combine these two sequences into a sequence of local alternatives to H_0, that is $H_{A,T} = H_{A,T}^I \,\&\, H_{A,T}^O$.

[11] Edwin Pitman (1897–1993) was an Australian statistician who made a number of contributions to statistics including the eponymous efficiency measure. The 1949 reference is to a set of lecture notes prepared for a lecture series given at the University of North Carolina, Chapel Hill and also elsewhere in the U.S. Although not published at that time, the notes were widely circulated and played an influential role in the development of statistical theory.

One immediate consequence of local misspecification is that the data v_t cannot be a realization from a strictly stationary process because the value of $E[f(v_t, \theta_0)]$ changes with T. This means the probability distribution of the data depends on T, and so the sample is now a realization from a doubly indexed process $\{v_{t,T}, t = 1, \ldots T; T = 1, 2, \ldots\}$[12] and it is for this reason that we introduced above the subscript T for the expectation operator. It is possible to develop characterizations of the probability distribution of the data which lead to $H_{A,T}$; for example, see Newey (1985*a*). However we do not pursue that route here and choose instead to characterize the data generation process implicitly via the properties which play a part in the analysis. Intuition suggests that $H_{A,T}$ causes a relatively modest perturbation from stationarity, and it is reasonable to assume there are data generation processes which satisfy the following assumption.

Assumption 5.1 Data Generation Process under $H_{A,T}$
The observed data are assumed to be a realization from a stochastic process $\{v_t; t = 1, 2, \ldots\}$ *which satisfies the following conditions: (i)* $\hat{\theta}_T \xrightarrow{p} \theta_0$*; (ii)* $g_T(\hat{\theta}_T) \xrightarrow{p} 0$*; (iii)* $G_T(\hat{\theta}_T) \xrightarrow{p} G_0$*,* $G_T(\hat{\theta}_T, \theta_0, \lambda_T) \xrightarrow{p} G_0$*; (iv)* $\hat{S}_T \xrightarrow{p} S$*, a positive definite matrix; (v)* $S^{-1/2}T^{1/2}g_T(\theta_0) \xrightarrow{d} N(\mu_I + \mu_O, I_q)$*.*

So for our purposes, the only effective difference between the data generation processes under H_0 and $H_{A,T}$ is in the mean of the limiting distribution for $T^{1/2}g_T(\theta_0)$.

Before we analyze the behaviour of the overidentifying restrictions test, it is instructive to consider the impact of local misspecification on the asymptotic distribution of the parameter estimator. Since $\hat{\theta}_T \xrightarrow{p} \theta_0$, we can use (3.24)–(3.26) in order to establish the following result.

Lemma 5.1 The Asymptotic Behaviour of $T^{1/2}(\hat{\theta}_T - \theta_0)$ under $H_{A,T}$
If Assumption 5.1 holds then:

$$T^{1/2}(\hat{\theta}_T - \theta_0) \xrightarrow{d} N\left(-(G_0'S^{-1}G_0)^{-1}G_0'S^{-1/2'}\eta_I, (G_0'S^{-1}G_0)^{-1}\right)$$

There are two aspects of this distributional result which should be noted. First, a comparison with Theorem 3.2 reveals that local misspecification only impacts on the mean of the distribution. Secondly, this impact derives from $H_{A,T}^I$ alone. This conforms to our earlier comments about the different roles of these two sets of restrictions.[13] A local violation of the identifying restrictions causes a bias in the asymptotic distribution of $\hat{\theta}_T$ away from θ_0, but a local violation of the overidentifying restrictions has no impact.

With this in mind, we now characterize the behaviour of the overidentifying restrictions test under $H_{A,T}$.

[12] Such a process is called a triangular array; see Davidson (1994) [pp.34, 178]. However, for notational simplicity, we suppress the additional subscript on v.

[13] See Section 3.3.

Theorem 5.4 Large Sample Behaviour of J_T under $H_{A,T}$
If Assumption 5.1 holds then: $J_T \xrightarrow{d} \chi^2_{q-p}(\mu_O'\mu_O)$ where $\chi^2_a(b)$ denotes a χ^2 distribution with degrees of freedom a and non–centrality parameter b.[14]

Proof:
Once again, it suffices to consider the statistic $\tilde{J}_T = T g_T(\hat{\theta}_T)' S^{-1} g_T(\hat{\theta}_T)$. The first few steps of the argument are identical to the analysis of $T^{1/2} g_T(\hat{\theta}_T)$ in the proof of Theorem 5.1. The Mean Value Theorem can be used to deduce (3.34) and this in turn leads to (3.35). Since Assumption 5.1 implies the matrices in (3.35) converge to the same limits under H_0 and $H_{A,T}$, equation (3.35) is equivalent to

$$S^{-1/2}T^{1/2}g_T(\hat{\theta}_T) \;=\; [I_q \;-\; P(\theta_0)]S^{-1/2}T^{1/2}g_T(\theta_0) \;+\; o_p(1) \tag{5.14}$$

Using (5.14) and Assumption 5.1, it follows that

$$\tilde{J}_T \;\xrightarrow{d}\; \|[I_q - P(\theta_0)](n_q + \mu_I + \mu_O)\|^2 \;=\; (n_q + \mu_I + \mu_O)'[I_q - P(\theta_0)](n_q + \mu_I + \mu_O) \tag{5.15}$$

where $n_q \sim N(0, I_q)$. Equation (5.15) implies $\tilde{J}_T$ converges to a $\chi^2_{q-p}(b)$ distribution where $b = (\mu_I + \mu_O)'[I_q - P(\theta_0)](\mu_I + \mu_O)$. However, since $\mu_I = P(\theta_0)\eta_I$ and $\mu_O = [I_q - P(\theta_0)]\eta_O$, the non-centrality parameter reduces to $b = \mu_O'\mu_O$. ◇

Theorem 5.4 reveals that the non–centrality parameter depends on μ_O alone, and so the test only has power against local violations of the overidentifying restrictions. This implies that if the local misspecification is confined to the identifying restrictions then J_T converges to a central χ^2_{q-p} distribution. Therefore, the test has the same distribution under both H_0 and $H^I_{A,T}$ & H^O_0, and so cannot be used to discriminate between these two states of the world.

5.1.4 The Parallels Between Non-Local and Local Analysis

As we have just seen, very different techniques are required for the analysis of the test's behaviour in the presence of local and non–local misspecification. At first glance, it is not immediately obvious that they lead to the same conclusions about the interpretation of a significant statistic – but they do! Since the test is the standard diagnostic within the GMM framework, it is worthwhile briefly explaining the parallels between the two types of analysis.

In the preamble to Chapter 4, we introduced three models: the assumed model $\mathcal{M}$, and two alternative candidates for the true model $\mathcal{M}_A$ and $\mathcal{M}_B$. These models have the following properties:

$$\mathcal{M} \quad \Longrightarrow \quad E[f(v_t, \theta_0)] = 0 \text{ for some unique } \theta_0 \in \Theta$$

[14] See Johnson and Kotz (1970) [Chapter 28] for a review of the properties of the non-central χ^2 distribution.

$$\begin{aligned} \mathcal{M}_A &\implies E[f(v_t, \theta_+)] = 0 \text{ for some unique } \theta_+ \in \Theta \\ \mathcal{M}_B &\implies \nexists\, \theta \in \Theta \text{ such that } E[f(v_t, \theta)] = 0 \end{aligned}$$

If $\mathcal{M}$ is misspecified then whether or not we can detect this fact using J_T depends on whether $\mathcal{M}_A$ or $\mathcal{M}_B$ represents the truth. If $\mathcal{M}_B$ is true then the assumed population moment condition is subject to non-local misspecification. In this case, J_T is consistent against this alternative, and so leads to rejection of the model with probability one in the limit. Now suppose $\mathcal{M}_A$ represents the truth. Thus far, we have not explicitly considered this case, but it is easy to see what happens. Both $\mathcal{M}$ and $\mathcal{M}_A$ imply there is a unique value of θ at which the population moment condition is satisfied. Since neither model places any further restrictions on this unique value of θ, they are observationally equivalent on the basis of $E[f(v_t, \theta)]$. Therefore, the estimator and all its associated statistics behave exactly the same under both $\mathcal{M}$ and $\mathcal{M}_A$. So this type of model misspecification cannot be detected – for the good reason that the part of the model used in estimation is actually correct!

This behaviour is mirrored in the analysis under local misspecification. The local alternative $H^I_{A,T}\,\&\,H^O_0$ corresponds to a local version of $\mathcal{M}_A$. To bring out this connection, it is necessary to consider a local alternative to H_0 in which the population moment condition is satisfied at a sequence of parameter values that converges to θ_0, that is

$$H^P_{A,T}:\ S^{-1/2}E_T[f(v_t, \theta_T)] \ = \ 0 \tag{5.16}$$

where $\theta_T = \theta_0 + T^{-1/2}\eta_P$. Given the local nature of the alternative, it is possible to use a first order Taylor expansion in (5.16) to deduce that $H^P_{A,T}$ implies

$$S^{-1/2}E_T[f(v_t, \theta_0)] \ + \ T^{-1/2}F(\theta_0)\eta_P \ = \ 0 \tag{5.17}$$

Since $F(\theta_0) = P(\theta_0)F(\theta_0)$, equation (5.17) can be rewritten as

$$S^{-1/2}E_T[f(v_t, \theta_0)] \ = \ -T^{-1/2}F(\theta_0)\eta_P \ = \ T^{-1/2}P(\theta_0)\eta_I$$

where $\eta_I = -F(\theta_0)\eta_P$. It is then immediately apparent that $H^P_{A,T} = H^I_{A,T}\,\&\,H^O_0$. In other words, $H^I_{A,T}\,\&\,H^O_0$ can be characterized as a sequence of alternatives in which the population moment condition is satisfied at a unique parameter value for each T, and so each member of the sequence satisfies the definition of $\mathcal{M}_A$. Theorem 5.4 states that J_T has the same distribution under both H_0 and $H^I_{A,T}\,\&\,H^O_0$, and so can now be recognized as the precursor to our comments above about the statistic's behaviour under $\mathcal{M}_A$. Since $H^P_{A,T}$ implies that $S^{-1/2}E_T[f(v_t, \theta_0)]$ lies in the column space of $F(\theta_0)$, it follows that $H^O_{A,T}$ implies the data are generated by a sequence of models with the properties of $\mathcal{M}_B$. So, Theorems 5.2 and 5.4 represent two ways of saying that the test can be used to discriminate between $\mathcal{M}$ and $\mathcal{M}_B$.

To conclude this discussion, it is useful to bring one implicit assumption into the light. Throughout, it has been assumed that the estimation really did locate

the *global* maximum of $Q_T(\theta)$. While, this is a reasonable assumption to make for the theoretical analysis, it may not be such a trivial issue in practice as we discussed in Section 3.2. Andrews (1997) observes that a significant statistic may be attributable to the failure of the estimation routine to locate the global minimum. In fact, Andrews (1997) proposes a method based on J_T to determine whether the global minimum has been reached. However, we do not pursue the details here, because this approach confounds issues of numerical convergence and model specification. However, Andrews's observation does re-emphasize the importance of locating the global maximum.

Example: Hansen and Singleton's (1982) Consumption Based Asset Pricing Model
Table 5.1 reports the overidentifying restrictions test statistics based on both the two step and iterated GMM estimators. For brevity, we only report results for the case in which the first step weighting matrix is the inverse of the instrument cross product matrix. Two choices of covariance matrix estimator are used: $\hat{S}_{SU}$ and $\hat{S}_{SU,\mu}$. However, in this case, the conclusions are affected by neither iteration nor the choice of covariance matrix estimator. The model is rejected with equally weighted returns (EWR) but cannot be rejected with value weighted returns (VWR). For the record, we also note that the same conclusions are drawn using the continuous updating estimator described in Section 3.7. In each case, the J–statistic based on the continuous GMM estimator is only marginally smaller than its counterpart based on the iterated estimator. ◇

Table 5.1
Overidentifying restriction test statistics

Asset	$\hat{S}_T$	Statistic	*Two-step*	*Iterated*
EWR	$\hat{S}_{SU}$	J_T	11.645	11.810
		$p-value$	0.009	0.008
	$\hat{S}_{SU,\mu}$	J_T	11.945	12.116
		$p-value$	0.008	0.007
VWR	$\hat{S}_{SU}$	J_T	1.747	1.748
		$p-value$	0.626	0.626
	$\hat{S}_{SU,\mu}$	J_T	1.754	1.755
		$p-value$	0.625	0.625

Notes: $\hat{S}_{SU}$, $\hat{S}_{SU,\mu}$ are given in (3.40) and (4.24) respectively, J_T denotes the overidentifying restrictions test in (5.2) and *p-value* denotes the observed significance level of J_T.

5.2 Testing Hypotheses about Subsets of $E[f(v_t, \theta_0)]$

The vector of population moment conditions can often be partitioned into a set of sub-vectors each of which refer to a different aspect of the model. In some

cases, a priori information may indicate that if there is misspecification then it is confined to a particular part of the population moment condition so that the true model, $\mathcal{M}_{B,S}$, would have the property,

$$\mathcal{M}_{B,S} \implies E[f_1(v_t, \theta_0)] = 0 \text{ for some unique } \theta_0 \in \Theta \text{ but } E[f_2(v_t, \theta_0)] \neq 0 \tag{5.18}$$

for some partition $f(v_t, \theta) = [f_1(v_t, \theta)', f_2(v_t, \theta)']'$. Since $\mathcal{M}_{B,S} \subset \mathcal{M}_B$, the overidentifying restrictions test is consistent against this type of misspecification. However, it is possible to construct a more powerful test of the model specification by taking advantage of the a priori information on the likely source of the misspecification. In this section, we present this test and analyze its properties under local forms of misspecification.

To begin, it is necessary to define the partition of $f(.)$ more formally and also introduce a partition of θ_0. Let $\theta_0' = (\theta_{0,1}', \theta_{0,2}')$ where $\theta_{0,i}$ is $(p_i \times 1)$, and $f(v_t, \theta_0)' = [f_1(v_t, \theta_{0,1})', f_2(v_t, \theta_0)']$ where $f_i(.)$ is $(q_i \times 1)$. Without loss of generality we focus on the case in which it is desired to test the null hypothesis

$$H_0^S : E[f_1(v_t, \theta_{0,1})] = 0 \text{ and } E[f_2(v_t, \theta_0)] = 0 \tag{5.19}$$

against the alternative that

$$H_A^S : E[f_1(v_t, \theta_{0,1})] = 0 \text{ and } E[f_2(v_t, \theta_0)] \neq 0 \tag{5.20}$$

Two features of this specification should be noted. First, the veracity of $E[f_1(v_t, \theta_{0,1})] = 0$ is maintained under both null and alternative; so the potential misspecification is confined to $E[f_2(v_t, \theta_0)]$. Secondly, this framework allows for the possibility that the maintained moment conditions, $E[f_1(v_t, \theta_{0,1})] = 0$, only depend on part of the parameter vector.

Both Newey (1985*a*) and Eichenbaum, Hansen, and Singleton (1988) have proposed methods for discriminating between these two hypotheses. Although these authors take very different approaches, Ahn (1995) shows their resulting statistics are asymptotically equivalent under both H_0^S and local versions of H_A^S. From a practical perspective, Eichenbaum, Hansen, and Singleton's (1988) statistic is far easier to calculate, and so we concentrate exclusively on this test. Readers interested in the approach taken by Newey (1985*a*) are refered to his original paper or the discussion in the review article by Hall (1999).

Eichenbaum, Hansen, and Singleton's (1988) statistic is so convenient because it is simply the difference between two overidentifying restrictions tests. The first is the overidentifying restrictions test from GMM estimation based on the full set of population moment conditions, J_T in (5.2). The second is the overidentifying restrictions test associated with GMM estimation of $\theta_{0,1}$ based on the moment conditions maintained under both H_0^S and H_A^S, that is

$$J_{1,T} = T g_{1,T}(\tilde{\theta}_{1,T})' \tilde{S}_{1,1}^{-1} g_{1,T}(\tilde{\theta}_{1,T}) \tag{5.21}$$

where $\tilde{\theta}_{1,T}$ is the two step (or iterated) GMM estimator of $\theta_{0,1}$ based on $E[f(v_t, \theta_{0,1})] = 0$, $g_{1,T}(\theta_1) = T^{-1} \sum_{t=1}^{T} f_1(v_t, \theta_1)$, and $\tilde{S}_{1,1}$ is a consistent estimator of

$S_{1,1} = \lim_{T\to\infty} Var[T^{-1/2} \sum_{t=1}^{T} f_1(v_t, \theta_{0,1})]$. Eichenbaum, Hansen, and Singleton's (1988) statistic is then given by,

$$C_T \;=\; J_T \;-\; J_{1,T} \tag{5.22}$$

The intuition behind the test's construction is most readily appreciated after an exploration of its properties, and so we now proceed to the statistical analysis, but return to the intuition at the end of this section.

We begin with its limiting distribution under H_0. It is clear from the structure of the statistic that most of the regularity conditions are going to be the same as for the corresponding result for the overidentifying restrictions test in Theorem 5.1. However, C_T also depends on a second GMM estimation using $E[f_1(v_t, \theta_{0,1})] = 0$ alone, and so it is necessary to introduce the following identification condition.[15]

Assumption 5.2 Identification Condition for $\theta_{0,1}$
$E[\partial f_1(v_t, \theta_{0,1})/\partial \theta_1^{'}]$ has rank p_1.

Notice that this assumption implies $q_1 \geq p_1$. Clearly, if $q_1 = p_1$ then $J_{1,T} = 0$ and C_T reduces to J_T; therefore it is assumed below that $q_1 > p_1$. It must also be the case that $q_2 > p_2$ otherwise there be a value of $\theta_{0,2}$ which sets $E[f(v_t, \theta_0)]$ equal to zero for any given value of $\theta_{0,1}$.[16]

Theorem 5.5 The Asymptotic Distribution of Eichenbaum, Hansen, and Singleton's (1988) Statistic under H_0^S
If (i) Assumptions 3.1–3.5, 3.8–3.13 and 5.2 hold; (ii) $q_1 > p_1$, $q_2 > p_2$; (iii) $\tilde{S}_{1,1}$ is positive semi-definite and converges in probability to $S_{1,1}$; (iv) $\hat{S}_T$ is positive semi-definite and converges in probability to S; then $C_T \xrightarrow{d} \chi^2_{q_2-p_2}$.

The proof is somewhat involved and so is relegated to the technical details sub-section at the end of this section.

There is an interesting pattern to the degrees of freedom of J_T, $J_{1,T}$ and C_T. Theorem 5.1 implies that J_T and $J_{1,T}$ have $q - p$ and $q_1 - p_1$ degrees of freedom respectively. Theorem 5.5 implies that C_T has $q_2 - p_2$ degrees of freedom. Therefore, the subtraction of $J_{1,T}$ from J_T has created a statistic, C_T, with $(q_1 - p_1)$ fewer degrees of freedom. Notice that the resulting degrees of freedom equal the degree to which $\theta_{0,2}$ is overidentified by $E[f_2(v_t, \theta_0)] = 0$ *given* $\theta_{0,1}$.

At the beginning of this section, it is stated that it is possible to use information on the nature of the misspecification to construct a more powerful test than J_T. It is now time to show that C_T fulfils this promise. To do this, it is necessary to move into a setting in which the true model satisfies H_A^S. In Section 5.1, we introduced two frameworks for analyzing the behaviour of test statistics in misspecified models: a non-local and a local analysis. In that

[15] See Section 3.1 for a discussion of identification.

[16] This follows from the assumption of stationarity by the same logic used to deduce that Assumption 4.1 implies $q > p$; see the preamble to Chapter 4.

context, it is shown that either framework can be used to delineate the class of alternatives against which J_T has power. However, the non-local framework is not well suited to the question at hand here because the end result is that C_T, like J_T, rejects H_0^S with probability one in the limit.[17] While useful to know, this does not help us to characterize which is more powerful. In contrast, the local framework is carefully constructed so that the test statistics converge in distribution. Since the end product is a distribution, it is possible to compare the power properties of two statistics within this framework, and so this is the approach we take.

Section 5.1.3 presents a local power analysis of the overidentifying restrictions test. In that earlier context, it is instructive to set up local alternatives using the identifying and overidentifying restrictions. However, that approach is less convenient here. Instead, we consider the following sequence of local alternatives to H_0^S,

$$H_{A,T}^S: \; E_T \left[\begin{array}{c} f_1(v_t, \theta_{0,1}) \\ f_2(v_t, \theta_0) \end{array} \right] = \left[\begin{array}{c} 0_{q_1 \times 1} \\ T^{-1/2}\mu_2 \end{array} \right] = T^{-1/2}\mu_S \tag{5.23}$$

For brevity, we confine ourselves to a heuristic comparison of the distributions of J_T and C_T under $H_{A,T}^S$. First, recall from Section 5.1.3 that for our purposes there is only one important difference between the data generation processes under the null and local alternatives and that is in the limiting distribution of sample moment. Under $H_{A,T}^S$, we have

$$S^{-1/2}T^{1/2}g_T(\theta_0) \xrightarrow{d} N(\, S^{-1/2}\mu_S, \, I_q \,) \tag{5.24}$$

and it is shown in the technical details sub-section at the end of this section that this behaviour translates into

$$J_T \xrightarrow{d} \chi^2_{q-p}(\nu_J) \tag{5.25}$$

$$C_T \xrightarrow{d} \chi^2_{q_2-p_2}(\nu_J) \tag{5.26}$$

where $\nu_J = \mu_S^{'} S^{-1/2'}[I_q - P(\theta_0)]S^{-1/2}\mu_S$. Therefore, the only difference between the limiting distributions is in the degrees of freedom. If $\nu_J > 0$ then C_T is the more powerful test because it has fewer degrees of freedom.[18] [19]

The foregoing discussion gives a useful perspective on the construction of the test. We can think of the overidentifying restrictions based on $E[f(v_t, \theta_0)] = 0$ as being built up of two components. The first component is the set of

[17] C_T is a consistent test of H_0^S against H_A^S. The essence of the proof is quite simple. Since $\mathcal{M}_{B,S} \subset \mathcal{M}_B$ it follows from Theorem 5.2 that $T^{-1}J_T \xrightarrow{p} c_S > 0$ under H_A^S. Also, $E[f_1(v_t, \theta_{0,1})] = 0$ under H_A^S and so from Theorem 5.1 that $J_{1,T} = O_p(1)$. Taken together these two properties imply: $T^{-1}C_T \xrightarrow{p} c_S$, and hence is consistent.

[18] From the analysis in Section 5.1.3, it follows that $\nu_J > 0$ if the data are generated by a sequence of processes which satisfies both $H_{A,T}^S$ and $H_{A,T}^O$.

[19] See Johnson and Kotz (1970) [Chapter 28] for a discussion of the properties of the non-central χ^2 distribution.

$q_1 - p_1$ overidentifying restrictions for $\theta_{0,1}$ based on $E[f_1(v_t, \theta_{0,1})] = 0$; the second component is the set of $q_2 - p_2$ overidentifying restrictions for $\theta_{0,2}$ based on $E[f_2(v_t, \theta_0)] = 0$ *given* $\theta_{0,1}$. Each component contributes to the degrees of freedom of the test, but it is only the second which contributes to the non-centrality parameter under $H^S_{A,T}$. This structure is exploited in the construction of C_T because the statistic is effectively calculated by subtracting from J_T the part which is insensitive to the misspecification under $H_{A,T}$.

Although the statistic has been motivated as a test of a subset of the population moment condition, the null hypothesis, H^S_0, involves both $E[f_1(v_t, \theta_{0,1})] = 0$ and $E[f_2(v_t, \theta_0)] = 0$. Therefore, the test is potentially sensitive to misspecification of any part of the population moment condition. Therefore, the veracity of the a priori information is crucially important in the interpretation of a significant statistic.

Example: Hansen and Singleton's (1982) Consumption Based Asset Pricing Model

In Section 5.1 it is shown that the use of the overidentifying restrictions test leads to rejection of the model with equally weighted returns (EWR) but not with value weighted returns (VWR). We now investigate the specification of the model with VWR further using Eichenbaum, Hansen, and Singleton's (1988) statistic. It can be recalled that our estimation employs an instrument vector which contains an intercept and lagged values of both consumption growth and the asset return. It may be possible that the moment conditions associated with either of the latter two variables are incompatible with the data but this was not detected with the overidentifying restrictions for the types of reason described above. This possibility leads us to consider two versions of Eichenbaum, Hansen, and Singleton's (1988) statistic. To introduce the associated null and alternative, we set $z'_{1,t} = (c_t/c_{t-1}, c_{t-1}/c_{t-2})$ and $z'_{2,t} = (r_t/p_{t-1}, r_{t-1}/p_{t-2})$. The first version tests whether the moments associated with consumption growth are compatible with the data and so the null and alternative are given by (5.19)–(5.20) with

$$\begin{aligned} f_1(v_t, \theta_0) &= \begin{bmatrix} 1 \\ z_{2,t} \end{bmatrix} u_t(\theta_0) \\ f_2(v_t, \theta_0) &= z_{1,t} u_t(\theta_0) \end{aligned}$$

The second version tests whether the moment conditions associated with the asset return are compatible with data, that is H^S_0 and H^S_A in (5.19)-(5.20) with

$$\begin{aligned} f_1(v_t, \theta_0) &= \begin{bmatrix} 1 \\ z_{1,t} \end{bmatrix} u_t(\theta_0) \\ f_2(v_t, \theta_0) &= z_{2,t} u_t(\theta_0) \end{aligned}$$

The results are given in Table 5.2. In each case the long run variance is estimated using $\hat{S}_{SU}$ and the statistics are based on the iterated estimator. Clearly, neither test offers evidence against the specification in this case.

Table 5.2
Eichenbaum, Hansen, and Singleton's (1988) statistics for the consumption based asset pricing model

Statistic	*d.f.*	z_1	z_2
$J_{1,T}$	1	1.241	0.384
$p-value$		0.265	0.536
C_T	2	0.503	1.363
$p-value$		0.778	0.506

Notes: z_i denotes the choice of instrument in $f_2(v_t,\theta)$, $J_{1,T}$ denotes the overidentifying restrictions test in (5.21) C_T denotes the overidentifying restrictions test in (5.22) *d.f.* denotes degrees of freedom and *p-value* denotes the observed significance level.

5.2.1 Technical Details

I: Proof of Theorem 5.5:
Before we begin, it is useful to introduce the following partition of $G(\theta) = E[\partial f(v_t,\theta)/\partial\theta']$ into four blocks conforming to the partitions of $f(.)$ and θ,

$$G(\theta) \;=\; \left[\begin{array}{c} G_1(\theta) \\ G_2(\theta) \end{array}\right] \;=\; \left[\begin{array}{cc} G_{1,1}(\theta) & G_{1,2}(\theta) \\ G_{2,1}(\theta) & G_{2,2}(\theta) \end{array}\right]$$

where $G_{i,j} = E[\partial f_i(v_t,\theta)/\partial\theta_j^{'}]$.

In view of conditions $(iii) - (iv)$ of the the theorem, it suffices to consider

$$\bar{C}_T \;=\; T\,\{\, g_T(\hat{\theta}_T)' S^{-1} g_T(\hat{\theta}_T) \;-\; g_{1,T}(\tilde{\theta}_{1,T})' S_{1,1}^{-1} g_{1,T}(\tilde{\theta}_{1,T}) \,\} \tag{5.27}$$

Since the proof is quite long, it is useful to present an overview of the proof strategy. There are three main steps.
Step 1: It is shown that

$$S^{-1/2}T^{1/2} g_T(\hat{\theta}_T) \;=\; A_1 S^{-1/2}T^{1/2} g_T(\theta_0) \;+\; o_p(1) \tag{5.28}$$

$$S_{1,1}^{-1/2}T^{1/2} g_{1,T}(\tilde{\theta}_{1,T}) \;=\; A_2 S^{-1/2}T^{1/2} g_T(\theta_0) \;+\; o_p(1) \tag{5.29}$$

for certain matrices of constants A_1 and A_2, and hence that

$$\bar{C}_T \;=\; T g_T(\theta_0)' S^{-1/2}[A_1^{'}A_1 \;-\; A_2^{'}A_2] S^{-1/2} g_T(\theta_0) \;+\; o_p(1) \tag{5.30}$$

Step 2: It is shown that $A_1^{'}A_1 \;-\; A_2^{'}A_2$ is idempotent with rank $q_2 - p_2$.
Step 3: Steps 1 and 2 can be combined with the Central Limit Theorem to derive the stated result along similar lines to the proof of Theorem 5.1.

Since *Step 3* is straightforward, we concentrate purely on *Steps 1 and 2* below.
Proof of Step 1: The definition of A_1 in (5.28) is straightforward because (3.36) implies

$$S^{-1/2}T^{1/2} g_T(\hat{\theta}_T) \;=\; [I_q \;-\; P(\theta_0)] S^{-1/2}T^{1/2} g_T(\theta_0) \;+\; o_p(1) \tag{5.31}$$

and so $A_1 = [I_q - P(\theta_0)]$. The definition of A_2 in (5.29) requires a little more work. Since $T^{1/2} g_{1,T}(\tilde{\theta}_{1,T})$ is also an estimated sample moment – this time from the estimation of $\theta_{0,1}$ based on $E[f_1(v_t, \theta_{0,1})] = 0$ – we can appeal once again to (3.36) and deduce that

$$S_{1,1}^{-1/2} T^{1/2} g_{1,T}(\tilde{\theta}_{1,T}) \;=\; [I_{q_1} \;-\; P_1(\theta_{0,1})] S_{1,1}^{-1/2} T^{1/2} g_{1,T}(\theta_{0,1}) \;+\; o_p(1) \quad (5.32)$$

where $P_1(\theta_{0,1}) = F_{1,1}(\theta_{0,1})[F_{1,1}(\theta_{0,1})' F_{1,1}(\theta_{0,1})]^{-1} F_{1,1}(\theta_{0,1})'$ and $F_{1,1}(\theta_{0,1}) = S_{1,1}^{-1/2} G_{1,1}(\theta_{0,1})$. Now, since

$$S_{1,1}^{-1/2} T^{1/2} g_{1,T}(\theta_{0,1}) \;=\; S_{1,1}^{-1/2} [I_{q_1} \;:\; 0_{q_1 \times q_2}] S^{1/2} S^{-1/2} T^{1/2} g_T(\theta_0) \quad (5.33)$$

where $0_{q_1 \times q_2}$ is a $(q_1 \times q_2)$ null matrix, it follows that (5.32) can be rewritten as

$$S_{1,1}^{-1/2} T^{1/2} g_{1,T}(\tilde{\theta}_{1,T}) \;=\; [I_{q_1} \;-\; P_1(\theta_{0,1})] \Xi S^{-1/2} T^{1/2} g_T(\theta_0) \;+\; o_p(1) \quad (5.34)$$

where $\Xi = S_{1,1}^{-1/2} [I_{q_1} \;:\; 0_{q_1 \times q_2}] S^{1/2}$. A comparison of (5.29) and (5.34) indicates that $A_2 = [I_{q_1} \;-\; P_1(\theta_{0,1})] \Xi$.

Proof of Step 2: First notice that A_1 is idempotent and

$$A_2' A_2 \;=\; \Xi' [I_{q_1} \;-\; P_1(\theta_{0,1})] \Xi \;=\; B, \text{ say.} \quad (5.35)$$

So to complete this step of the proof, it is necessary to show that (i) $A_1 - B$ is idempotent, and (ii) $rank(A_1 - B) = q_2 - p_2$.

Consider (i) first. Using the idempotency of A_1, it follows that

$$(A_1 \;-\; B)(A_1 \;-\; B) \;=\; A_1 \;-\; BA_1 \;-\; A_1 B \;+\; BB \quad (5.36)$$

We now show that $BA_1 = A_1 B = BB = B$, and so that the right hand side of (5.36) reduces to $A_1 - B$ which is the desired result. First, notice that from the definition of A_1 we have that

$$BA_1 \;=\; B[I_q - P(\theta_0)] \;=\; B - BP(\theta_0)$$

So $BA_1 = B$ if $BP(\theta_0) = 0$. This latter result is established by observing that,

$$BP(\theta_0) \;=\; BF(\theta_0)[F(\theta_0)' F(\theta_0)]^{-1} F(\theta_0)'$$

and

$$\begin{aligned} BF(\theta_0) \;&=\; \Xi' [I_{q_1} - P_1(\theta_{0,1})] \Xi F(\theta_0) \\ &=\; \Xi' [I_{q_1} - P_1(\theta_{0,1})] F_{1,1}(\theta_{0,1}) \\ &=\; 0 \end{aligned}$$

A similar argument can be used for $A_1 B = B$, and so we now consider BB. By definition, it follows that

$$BB \;=\; \Xi' [I_{q_1} - P_1(\theta_{0,1})] \Xi \Xi' [I_{q_1} - P_1(\theta_{0,1})] \Xi \quad (5.37)$$

Since $I_{q_1} - P_1(\theta_{0,1})$ is idempotent, and

$$\Xi\Xi' = S_{1,1}^{-1/2} S_{1,1} S_{1,1}^{-1/2'} = I_{q_1}, \tag{5.38}$$

it follows from (5.37) that $BB = B$.

Now consider (ii). Since $A_1 - B$ is idempotent, it follows that $rank(A_1 - B) = trace(A_1 - B)$.[20] Furthermore, it can be shown that $trace(A_1 - B) = trace(A_1) - trace(B)$.[21] These two traces can be deduced as follows[22]:

$$\begin{aligned} trace(A_1) &= trace(I_q) - trace\{P(\theta_0)\} \\ &= q - trace\{F(\theta_0)'F(\theta_0)[F(\theta_0)'F(\theta_0)]^{-1}\} \\ &= q - trace(I_p) = q - p \end{aligned}$$

and[23]

$$\begin{aligned} trace(B) &= trace\{\Xi[I_{q_1} - P_1(\theta_0)]\Xi^{'}\} \\ &= trace\{\Xi^{'}\Xi[I_{q_1} - P_1(\theta_0)]\} \\ &= trace[I_{q_1} - P_1(\theta_0)] = q_1 - p_1 \end{aligned}$$

Taken together, these two results imply

$$rank(A_1 - B) = trace(A_1) - trace(B) = q - p - (q_1 - p_1) = q_2 - p_2$$

which completes *Step 2* of the proof. The theorem then follows by combining *Steps 1 – 3* in the manner described above. ◇

II: Derivation of Noncentrality Parameters for J_T and C_T

Equation (5.24) can be combined with (5.14) to show that

$$J_T \xrightarrow{d} \|[I_q - P(\theta_0)](n_q + S^{-1/2}\mu_S)\|^2 \tag{5.39}$$

where once again n_q denotes a random vector with a $N(0, I_q)$ distribution. Equation (5.39) implies

$$J_T \xrightarrow{d} \chi^2_{q-p}(\nu_J) \tag{5.40}$$

where $\nu_J = \mu_S^{'} S^{-1/2'}[I_q - P(\theta_0)]S^{-1/2}\mu_S$. Equation (5.24) can be combined with (5.30) and *Step 2* of the proof of Theorem 5.5 to show that

$$\bar{C}_T \xrightarrow{d} \|[A_1 - B](n_q + S^{-1/2}\mu_S)\|^2$$

[20] See Dhrymes (1984) [Proposition 55, p.66].

[21] See Dhrymes (1984) [Proposition 16, p.24].

[22] The arguments below use the property that $trace(D_1D_2) = trace(D_2D_1)$ for any conformable matrices D_1, D_2; see Dhrymes (1984) [Proposition 16, p.24].

[23] For the third step, note that(5.38) implies $\Xi^{'} = \Xi^{-1}$.

and hence that

$$C_T \xrightarrow{d} \chi^2_{q_2-p_2}(\nu_S) \tag{5.41}$$

where $\nu_S = \mu_S^{'} S^{-1/2'}[A_1 - B]S^{-1/2}\mu_S$. At first glance, ν_J and ν_S appear different, but closer inspection reveals that they are identical. This follows because: (i) $A_1 = [I_q - P(\theta_0)]$; and (ii) equations (5.35) and (5.23) can be combined to show that $BS^{-1/2}\mu_S = 0$. $\diamond$

5.3 Testing Hypotheses About the Parameter Vector

There are many cases in which a particular economic theory implies a set of restrictions on the parameter vector of the econometric model. This means it is possible to assess the veracity of the theory by testing whether the restrictions in question are satisfied by the data. This section describes various methods for performing this type of inference.

The structure of this testing problem is different from those described in the previous two sections. We now move into a world where the data are assumed to be generated by a model from the set $\mathcal{M}_A$ defined by[24]

$$\mathcal{M}_A \implies E[f(v_t, \theta_0)] = 0 \text{ for some unique } \theta_0 \in \Theta$$

The question of interest is whether the data are generated by the subset of $\mathcal{M}_A$ which satisfy

$$\mathcal{M}_{A,R} \implies E[f(v_t, \theta_0)] = 0 \text{ for some unique } \theta_0 \in \Theta_r = \{\theta : r(\theta) = 0\} \tag{5.42}$$

where $r(\theta_0)$ is a vector of nonlinear functions of θ_0. Notice that by definition $\Theta_r \subset \Theta$. Therefore, the issue is whether θ_0 lies in Θ_r or its complement in Θ, Θ_r^c. This type of problem is often refered to as a *nested hypothesis test* because Θ_r can be "nested" in Θ in the sense that Θ_r is a subset of Θ.

The vector $r(.)$ must satisfy certain conditions if the restrictions are to be meaningful.

Assumption 5.3 Regularity Conditions for $r(.)$
Let $r : \Re^p \to \Re^s$ *be a* $(s \times 1)$ *vector of real valued functions which satisfies: (i)* $r(.)$ *is a vector of continuous differentiable functions; (ii)* $rank\{R(\theta_0)\} = s$ *where* $R(\theta) = \partial r(\theta)/\partial\theta'$.

This assumption ensures that $r(\theta_0)$ form a coherent set of equations – that is, given $p - s$ elements of θ_0, it is possible to solve uniquely for the remaining s values using $r(\theta_0) = 0$.[25] Notice that this property automatically excludes redundant restrictions, and also that the rank condition necessarily implies $s \leq p$.

[24] Previously we used θ_+ to characterize $\mathcal{M}_A$ but we use θ_0 here for consistency with the specification of the hypotheses below.

[25] These conditions derive from the Implicit Function Theorem; for example, see Apostol (1974) [p.374].

Newey and West (1987*b*) develop the theory for testing

$$H_0^R:\ r(\theta_0) = 0 \qquad \text{versus} \qquad H_A^R:\ r(\theta_0) \neq 0$$

based on GMM estimators. They propose three main statistics which can be viewed as extensions to the GMM framework of the Wald, Lagrange Multiplier (LM) and Likelihood Ratio (LR) tests from Maximum Likelihood theory.[26] To facilitate the presentation, it is useful to define unrestricted and restricted estimators of θ_0. The unrestricted estimator is just $\hat{\theta}_T$ defined earlier. The restricted estimator is the value of θ which minimizes $Q_T(\theta)$ subject to $r(\theta) = 0$; this is denoted $\tilde{\theta}_T$. The asymptotic properties of the restricted estimator are derived in the technical details sub-section at the end of this section. It is assumed that both these minimizations use the weighting matrix $\hat{S}_T^{-1}$. We now introduce the three statistics in turn.

The *Wald* test examines whether the unrestricted estimator, $\hat{\theta}_T$, satisfies the restrictions with due allowance for sampling error. The statistic is

$$W_T = T\, r(\hat{\theta}_T)' \left[R(\hat{\theta}_T)\, [G_T(\hat{\theta}_T)' \hat{S}_T^{-1} G_T(\hat{\theta}_T)]^{-1} R(\hat{\theta}_T)' \right]^{-1} r(\hat{\theta}_T) \tag{5.43}$$

The *LM* test examines whether the restricted estimator, $\tilde{\theta}_T$, satisfies the first order conditions from the unrestricted estimation. This statistic is:

$$LM_T = T\, g_T(\tilde{\theta}_T)' \hat{S}_T^{-1} G_T(\tilde{\theta}_T) [G_T(\tilde{\theta}_T)' \hat{S}_T^{-1} G_T(\tilde{\theta}_T)]^{-1} G_T(\tilde{\theta}_T)' \hat{S}_T^{-1} g_T(\tilde{\theta}_T) \tag{5.44}$$

Finally, the *D* or *LR-type* test examines the impact on the GMM minimand of the imposition of the restrictions. This statistic is

$$D_T = T[Q_T(\tilde{\theta}_T) \ - \ Q_T(\hat{\theta}_T)] \tag{5.45}$$

In the context of Maximum Likelihood theory, it is well known that these three statistics are asymptotically equivalent under the null hypothesis. Newey and West (1987*b*) [Theorem 2] show that this equivalence extends to the GMM setting.

Theorem 5.6 Asymptotic Equivalence of W_T, LM_T and D_T under H_0^R *If (i) Assumptions 3.1–3.5, 3.7–3.13, and 5.3 hold; (ii) $\hat{S}_T^{-1} \xrightarrow{p} S^{-1}$; then under H_0^R: (a) $W_T = N_T + o_p(1)$; (b) $LM_T = N_T + o_p(1)$; (c) $D_T = N_T + o_p(1)$; where $N_T = n_T' V_n^{-1} n_T$, $n_T = R(G_0' S^{-1} G_0)^{-1} G_0' S^{-1} T^{1/2} g_T(\theta_0)$, and $V_n = R(G_0' S^{-1} G_0)^{-1} R'$.*

The proof is relegated to the technical details sub-section. One immediate consequence of Theorem 5.6 is that all three statistics share the limiting distribution of N_T under H_0^R. This distribution is easily deduced from the definition of N_T because under our conditions it follows that $n_T \xrightarrow{d} N(\,0,\, V_n\,)$. Therefore we obtain the following distributional result.

[26] It should be noted that there are a number of asymptotically equivalent versions of these tests. Our presentation focuses exclusively on the versions proposed by Newey and West (1987*b*). See Newey and McFadden (1994) [p.2222] for a discussion of the alternative versions.

Theorem 5.7 The Limiting Distribution of W_T, LM_T and D_T under H_0^R
If (i) Assumptions 3.1–3.5, 3.7–3.13, and 5.3 hold; (ii) $\hat{S}_T^{-1} \xrightarrow{p} S^{-1}$; then under H_0^R: $W_T \xrightarrow{d} \chi_s^2$, $LM_T \xrightarrow{d} \chi_s^2$ and $D_T \xrightarrow{d} \chi_s^2$ as $T \to \infty$.

There is one other consequence of Theorem 5.6 which should be noted. Using similar arguments to Theorem 3.5, it is possible to show that n_T is asymptotically independent of $S^{-1/2}T^{1/2}g_T(\hat{\theta}_T)$ under the conditions of the theorem. Since the large sample behaviour of W_T, LM_T and D_T are governed by n_T, it follows that these three statistics are also asymptotically independent of $S^{-1/2}T^{1/2}g_T(\hat{\theta}_T)$. This, in turn, implies that W_T, LM_T and D_T are asymptotically independent of the overidentifying restrictions test statistic, J_T, under the composite null hypothesis that $E[f(v_t, \theta_0)] = 0$ and $r(\theta_0) = 0$.

Newey and West (1987*b*) show that the asymptotic equivalence of the statistics extends to local alternatives characterized by

$$H_{A,T}^R : \; r(\theta_0) \; = \; T^{-1/2}\mu_R$$

Furthermore, they show that the statistics converge to a $\chi_s^2(\delta_R)$ where

$$\delta_R = \mu_R' \left[R(\theta_0)(G_0' S^{-1} G_0)^{-1} R(\theta_0)' \right]^{-1} \mu_R > 0$$

So the statistics have power against the alternative for which they are designed. In view of their equivalence, some other criteria must be used to choose between the three. One such criterion is computational burden, although this is less of a concern now than it once was. The D statistic is more burdensome because it requires two estimations, whereas the Wald and LM only require one. Sometimes the unrestricted estimation is easier and sometimes not – it all depends on the model in question and the nature of $r(.)$. However, the Wald test has two disadvantages which should be mentioned. First, it is not invariant to a reparameterization of the model or the restrictions. This means that it is possible to rewrite the model and restrictions in a logically consistent way, but end up with a different Wald statistic.[27] Neither of the other two tests have this problem.[28] The second disadvantage is that the Wald statistic tends to be less well approximated by the χ_s^2 distribution in finite samples than the other two statistics; for example, see the simulation evidence reported in Gallant (1987).

At this stage, it is useful to bring into the light one assumption that has been lurking in the shadows. Throughout the analysis in this section, it has been assumed that Assumption 3.3 holds and so $E[f(v_t, \theta_0)] = 0$. It is important to realize that a violation of this assumption can also lead to a significant statistic. In other words, H_0^R may be rejected because either Assumption 3.3 holds but $r(\theta_0) \neq 0$ – or it may be rejected because the model is misspecified. Hall and

[27] For example, the restriction $\theta_{0,i} = \bar{\theta}_i$ can also be rewritten as $\theta_{0,i}^k = \bar{\theta}_i^k$ for any finite positive integer k. This sensitivity of the Wald statistic derives from the sensitivity of the asymptotic standard errors to reparameterization; see Section 3.7.

[28] Davidson and MacKinnon (1993) [p.467–9] provide a useful discussion of this issue and some examples. Also see Critchley, Marriott, and Salmon (1996).

Inoue (2003) provide a formal justification for this statement within the framework of non-local misspecification employed in Chapter 4. Their results indicate that Wald, LM and D tests do not converge to limiting χ^2_s distributions in misspecified models even if the restrictions are satisfied. Furthermore, the limiting behaviour of the three test statistics depends crucially on the covariance matrix estimator employed. For example, Hall and Inoue (2003) show that W_T, LM_T and D_T diverge to infinity in the case where either a centred or uncentred HAC estimator is used. These results emphasize the importance of using the model specification tests, J_T or C_T, before undertaking inference about the parameters.

To conclude our discussion, it is useful to explore briefly a different perspective on H_0^R involving the identifying restrictions. It can be recalled from Section 3.3 that the identifying restrictions can be interpreted as the restriction that the projection of $S^{-1/2}E[f(v_t, \theta_0)]$ onto the column space of $F(\theta_0)$, $\mathcal{C}[F(\theta_0)]$, is zero.[29] We now show that the restrictions can be interpreted as a statement about the structure of $\mathcal{C}[F(\theta_0)]$. To do this, notice that if H_0^R is true then the Implicit Function Theorem implies that the population moment condition can be written as

$$E[f(v_t, g(\psi_0))] = 0 \tag{5.46}$$

where ψ_0 is a $p - s$ vector which satisfies $\theta_0 = g(\psi_0)$. Now, if (5.46) is treated as a basis for GMM estimation of ψ_0 then the associated identifying restrictions imply the projection of $S^{-1/2}E[f(v_t, \theta_0)]$ onto the column space of $F(\psi_0)$ is zero where $F(\psi_0) = S^{-1/2}E[\partial f(v_t, g(\psi_0))/\partial \psi']$. However, since

$$F(\psi_0) = F(\theta_0)\ \{\, \partial g(\psi_0)/\partial \psi' \,\}$$

it follows that the column space of $F(\psi_0)$, $\mathcal{C}[F(\psi_0)]$, is of dimension $p - s$ and $\mathcal{C}[F(\psi_0)] \subset \mathcal{C}[F(\theta_0)]$.

It is interesting to contrast this perspective on H_0^R with what we learned about testing $H_0 : E[f(v_t, \theta_0)] = 0$ in the course of our earlier analysis of the overidentifying restrictions test. The analyses in Sections 5.1.2 and 5.1.3 indicate that tests of the validity of the population moment condition revolve around the overidentifying restrictions which, it can be recalled from Section 3.3, involve the orthogonal complement of $F(\theta_0)$. Therefore, the fundamental decomposition inherent in GMM estimation reverberates into hypothesis testing based on the estimator: hypotheses about the parameters are equivalent to hypotheses about the columnspace of $F(\theta_0)$, and hypotheses about the population moment condition are equivalent to hypotheses about the orthogonal complement of $F(\theta_0)$.

Example: Hansen and Singleton's (1982) Consumption Based Asset Pricing Model

It can be recalled from Section 5.1 that the overidentifying restrictions test is significant when the asset is the index with equally weighted returns (EWR). We interpret this rejection as being indicative of misspecification, and so, in view

[29] Recall that in this chapter we focus exclusively on the two step or iterated estimator and so S^{-1} must be substituted for W in the Section 3.3.

of the remarks above, do not consider that version of the model here. Instead we concentrate purely on the value weighted returns (VWR) case for which the overidentifying restrictions test is insignificant.

Using L'Hopital's rule it can be shown that $\lim_{\gamma \to 0}(c^\gamma - 1)/\gamma = ln(c)$.[30] Therefore the restriction $\gamma_0 = 0$ reduces CRRA utility function to the log utility function. This restriction can be expressed in our general notation by putting $r(\theta_0) = \gamma_0$. If we define $\theta_0 = [\gamma_0, \delta_0]'$ then $R(\theta_0)$ is given by

$$R(\theta_0) \; = \; [1, \, 0]$$

It is immediately apparent that this choice of $r(.)$ satisfies the regularity conditions in Assumption 5.3. The restricted estimation is performed using the procedure *constr* in the MATLAB version 6.0 Optimization Toolbox (Mathworks, 2000).

Table 5.3 contains the W_T, LM_T and D_T statistics for the test of $H_0^R : \gamma_0 = 0$. All three statistics are calculated using $\hat{S}_T = \hat{S}_{SU}$. From Theorem 5.7, all three test statistics converge to a χ_1^2 under this null. Notice that for this case the Wald test has a very simple form. Since $r(\hat{\theta}_T) = \hat{\gamma}_T$ and $R(\hat{\theta}_T) = [1, 0]$ (5.43) reduces to

$$W_T \; = \; T\frac{\hat{\gamma}_T^2}{\hat{V}_{11}}$$

where $\hat{V}_{11}$ is the $1-1$ element of $[G_T(\hat{\theta}_T)'\hat{S}_T^{-1}G_T(\hat{\theta}_T)]^{-1}$. In other words, the Wald statistic is just the square of the "t–statistic" for $\gamma_0 = 0$.

In this particular example, the choice between the three statistics is of no consequence because they are identical to three decimal places. As can be seen, we fail to reject $H_0^R : \gamma_0 = 0$ at conventional levels of significance.

Table 5.3
Test statistics for $H_0^R : \gamma_0 = 0$

Test	*Statistic*	*p-value*
W_T	0.133	0.715
LM_T	0.133	0.715
D_T	0.133	0.715

Note: W_T, LM_T and D_T are defined in (5.43)–(5.45).

5.3.1 GMM Estimation Subject to Nonlinear Restrictions on θ_0 and Other Technical Details

I. The Asymptotic Properties of the Restricted GMM Estimator
The *restricted* two step GMM estimator is defined by

$$\tilde{\theta}_T \; = \; argmin_{\theta \in \Theta_r} \, Q_T(\theta) \tag{5.47}$$

where $\Theta_r = \{ \, \theta \text{ s.t. } \theta \in \Theta \text{ and } r(\theta) = 0 \, \}$ and $Q_T = g_T(\theta)'\hat{S}_T^{-1}g_T(\theta)$.

Throughout, it is assumed that θ_0 satisfies the restrictions.

[30] See Rudin (1976) [p.109].

Assumption 5.4 Restrictions on θ_0
$r(\theta_0) = 0$.

The analysis is split into two parts: the consistency of $\tilde{\theta}_T$ and the asymptotic distribution of $T^{1/2}(\tilde{\theta}_T - \theta_0)$. As in Chapter 3, the logical sequence is to begin with consistency.

A comparison of (3.11) and (5.47) indicates that the only difference between the restricted and unrestricted estimations stems from the set over which the minimization is taken. It is therefore straightforward to modify the proof of Theorem 3.1 to establish the following.

Lemma 5.2 Consistency of $\tilde{\theta}_T$
If Assumptions 3.1 – 3.4, 3.7 – 3.10, 5.3 and 5.4 hold then: $\tilde{\theta}_T \xrightarrow{p} \theta_0$

While the characterization in (5.47) can be used to establish consistency, it does not lend itself to the derivation of the asymptotic distribution of $T^{1/2}(\tilde{\theta}_T - \theta_0)$. For this question, it is more fruitful to define $\tilde{\theta}_T$ using Lagrange's method for constrained optimization. Accordingly, we introduce the Lagrangean function

$$\mathcal{L}_T(\theta, \rho) \;=\; Q_T(\theta) \;-\; 2r(\theta)'\rho \tag{5.48}$$

where 2ρ is the $(s \times 1)$ vector of Lagrange multipliers.[31] Subject to certain regularity conditions,[32] $\tilde{\theta}_T$ and the associated estimator of ρ, denoted $\tilde{\rho}_T$, satisfy the first order conditions, $\partial\mathcal{L}(\tilde{\theta}_T, \tilde{\rho}_T)/\partial\theta = 0$ and $\partial\mathcal{L}(\tilde{\theta}_T, \tilde{\rho}_T)/\partial\rho = 0$. In this case, these conditions yield

$$G_T(\tilde{\theta}_T)'\hat{S}_T^{-1}g_T(\tilde{\theta}_T) \;-\; R(\tilde{\theta}_T)'\tilde{\rho}_T \;=\; 0 \tag{5.49}$$

$$-\,r(\tilde{\theta}_T) \;=\; 0 \tag{5.50}$$

To derive the asymptotic distribution of $T^{1/2}(\tilde{\theta}_T - \theta_0)$, it is necessary to know the probability limits of $\tilde{\theta}_T$ and $\tilde{\rho}_T$; the former limit is provided by Lemma 5.2 above, and the latter is given in the following lemma.

Lemma 5.3 Probability Limit of $\tilde{\rho}_T$
If Assumptions 3.1 – 3.5, 3.7 – 3.10, 3.12–3.13, 5.3 and 5.4 hold then: $\tilde{\rho}_T \xrightarrow{p} 0$.

This result can be derived by considering the limiting behaviour of (5.49) as $T \to \infty$, however we leave the details to the reader.[33]

The asymptotic distribution of $T^{1/2}(\tilde{\theta}_T - \theta_0)$ is deduced from (5.49)–(5.50). However, before this can be done, each equation requires a certain amount of manipulation, and so we start by considering each equation individually. Equation (5.49) implies that

$$G_T(\tilde{\theta}_T)'\hat{S}_T^{-1}T^{1/2}g_T(\tilde{\theta}_T) \;-\; R(\tilde{\theta}_T)'T^{1/2}\tilde{\rho}_T \;=\; 0 \tag{5.51}$$

[31] The factor of 2 is introduced for ease of presentation below.
[32] See Intrilligator (1971) [Chapter 3].
[33] Or see Newey and McFadden (1994) [p.2218].

Under our conditions, $G_T(\tilde{\theta}_T) \xrightarrow{p} G_0$, and if we assume $\hat{S}_T \xrightarrow{p} S$ then (5.51) can be rewritten as

$$G_0' S^{-1} T^{1/2} g_T(\tilde{\theta}_T) \; - \; R(\theta_0)' T^{1/2} \tilde{\rho}_T \; + \; o_p(1) \; = \; 0 \tag{5.52}$$

The next step involves the use of the Mean Value Theorem to linearize $T^{1/2} g_T(\tilde{\theta}_T)$ around $T^{1/2} g_T(\theta_0)$. Under our assumptions, this linearized version implies

$$T^{1/2} g_T(\tilde{\theta}_T) \; = \; T^{1/2} g_T(\theta_0) \; + \; G_0 T^{1/2} (\tilde{\theta}_T - \theta_0) \; + \; o_p(1) \tag{5.53}$$

Finally, if (5.53) is substituted into (5.52) then we obtain

$$G_0' S^{-1} T^{1/2} g_T(\theta_0) \; + \; G_0' S^{-1} G_0 T^{1/2} (\tilde{\theta}_T - \theta_0) \; - \; R(\theta_0)' T^{1/2} \tilde{\rho}_T \; + \; o_p(1) \; = \; 0 \tag{5.54}$$

Now consider (5.50). The Mean Value Theorem and Lemma 5.2 can be used to deduce that

$$T^{1/2} r(\tilde{\theta}_T) \; = \; T^{1/2} r(\theta_0) \; + \; R(\theta_0) T^{1/2} (\tilde{\theta}_T - \theta_0) \; + \; o_p(1) \tag{5.55}$$

Using (5.55) and Assumption 5.4, it can be seen that (5.50) implies

$$R(\theta_0) T^{1/2} (\tilde{\theta}_T - \theta_0) \; + \; o_p(1) \; = \; 0 \tag{5.56}$$

Taken together, equations (5.54) and (5.56) imply that $T^{1/2}(\tilde{\theta}_T - \theta_0)$ satisfies the following set of equations,

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} \; = \; \begin{bmatrix} G_0' S^{-1} T^{1/2} g_T(\theta_0) \\ 0 \end{bmatrix} + \begin{bmatrix} G_0' S^{-1} G_0 & -R_0' \\ -R_0 & 0 \end{bmatrix} \begin{bmatrix} T^{1/2}(\tilde{\theta}_T - \theta_0) \\ T^{1/2} \tilde{\rho}_T \end{bmatrix} \; + \; o_p(1) \tag{5.57}$$

where for brevity we set $R_0 = R(\theta_0)$. Using the formulae for the inversion of a partitioned matrix,[34] it can be shown that (5.57) implies

$$T^{1/2}(\tilde{\theta}_T - \theta_0) \; = \; -\{V_U \; - \; V_U R' [R V_U R']^{-1} R V_U\} G_0' S^{-1} T^{1/2} g_T(\theta_0) \; + \; o_p(1) \tag{5.58}$$

where to simplify the formulae we have set $V_U = (G_0' S^{-1} G_0)^{-1}$ – this notation reflects the fact this matrix is the variance of the asymptotic distribution for the unrestricted estimator; see Theorem 3.2. Notice that (5.58) has essentially the same structure as appeared at this stage in the analysis of the unrestricted estimator in Section 3.4.2: a matrix of constants times the vector, $T^{1/2} g_T(\theta_0)$. So once again, the limiting distribution is normal.

Lemma 5.4 Asymptotic distribution of $T^{1/2}(\tilde{\theta}_T - \theta_0)$
If (i) Assumptions 3.1–3.5, 3.7–3.13, 5.3 and 5.4 hold; (ii) $\hat{S}_T \xrightarrow{p} S$*; then:* $T^{1/2}(\tilde{\theta}_T - \theta_0) \xrightarrow{d} N(0, V_R)$ *where* $V_R = V_U \; - \; V_U R' (R V_U R')^{-1} R V_U$.

[34] See Magnus and Neudecker (1991) [p.11].

Notice that $V_U - V_R$ is a positive semi-definite matrix and so the restricted estimator is at least as efficient as the unrestricted estimator – in other words, we are never worse off for imposing *valid* restrictions on the parameters, as would be anticipated.

A comparison of Lemma 5.4 and Theorem 3.2 suggests the limiting distributions of the unrestricted and restricted estimator have much in common. However, there is one key difference which needs to be brought into the light. The matrix V_R has rank $p - s$ and so the normal distribution in Lemma 5.4 is singular. Whereas, the limiting covariance matrix for the unrestricted estimator is nonsingular. This difference reflects the nature of the estimators. In the unrestricted estimation all the elements of $\hat{\theta}_T$ are "free". In contrast, only $p - s$ elements of $\tilde{\theta}_T$ are "free" because the remaining s elements are tied down by the restrictions.

This analysis has concentrated on the estimator under H_0^R. In Section 5.3, certain comments are made about the behaviour of the restricted estimator under local alternatives $H_{A,T}^R$. Newey and McFadden (1994) [p.2218–20] present a more general version of our analysis under this more general class of processes. Just as in Section 5.1.3, the only effective difference between H_0^R and $H_{A,T}^R$ appears in the mean of the asymptotic distribution. Therefore, both Lemmas 5.2 and 5.3 continue to hold under local alternatives.

II. Proof of Theorem 5.6

Part (a): Under the assumptions listed in condition (i) of the theorem, it follows that: $R(\hat{\theta}_T) \xrightarrow{p} R(\theta_0)$ and $G_T(\hat{\theta}_T) \xrightarrow{p} G_0$. These two results combined with condition (ii) of the theorem imply that

$$\tilde{W}_T \;=\; T^{1/2} r(\hat{\theta}_T)' V_n^{-1} T^{1/2} r(\hat{\theta}_T) \;+\; o_p(1)$$

Therefore, the result will be established if we can show that $T^{1/2} r(\hat{\theta}_T) = \pm n_T + o_p(1)$. To this end, we use the Mean Value Theorem to deduce that

$$T^{1/2} r(\hat{\theta}_T) \;=\; T^{1/2} r(\theta_0) \;+\; R(\hat{\theta}_T, \theta_0, \lambda_T) T^{1/2} (\hat{\theta}_T - \theta_0) \tag{5.59}$$

where $R(\hat{\theta}_T, \theta_0, \lambda_T)$ is an $(s \times p)$ matrix whose i^{th} row is the i^{th} row of $R(\bar{\theta}_T^{(i)})$ where $\bar{\theta}_T^{(i)} = \lambda_{T,i}\theta_0 + (1 - \lambda_{T,i})\hat{\theta}_T^{(i)}$ for some $0 \leq \lambda_{T,i} \leq 1$, and λ_T is the $(s \times 1)$ vector with i^{th} element $\lambda_{T,i}$. Since $\hat{\theta}_T \xrightarrow{p} \theta_0$, it follows that $\bar{\theta}_T^{(i)} \xrightarrow{p} \theta_0$ and so $R(\hat{\theta}_T, \theta_0, \lambda_T) \xrightarrow{p} R(\theta_0)$. Using this result and $r(\theta_0) = 0$ in (5.59), it follows that

$$T^{1/2} r(\hat{\theta}_T) \;=\; R(\theta_0) T^{1/2} (\hat{\theta}_T - \theta_0) \;+\; o_p(1) \tag{5.60}$$

Equation (3.26) implies that

$$T^{1/2} (\hat{\theta}_T - \theta_0) \;=\; -(G_0' S^{-1} G_0)^{-1} G_0' S^{-1} T^{1/2} g_T(\theta_0) \;+\; o_p(1) \tag{5.61}$$

Finally, the substitution of (5.61) into (5.60) yields $T^{1/2} r(\hat{\theta}_T) = -n_T + o_p(1)$, which completes the proof of (a).

Part (b): Lemma 5.2 establishes that $\tilde{\theta}_T \xrightarrow{p} \theta_0$, and under the stated conditions $G_T(\tilde{\theta}_T) \xrightarrow{p} G_0$. The second of these results can be combined with the consistency of $\hat{S}_T$ to deduce that $LM_T = L\tilde{M}_T + o_p(1)$ where

$$L\tilde{M}_T \;=\; T^{1/2} g_T(\tilde{\theta}_T)' S^{-1} G_0 V_U G_0' S^{-1} T^{1/2} g_T(\tilde{\theta}_T) \;+\; o_p(1) \tag{5.62}$$

where $V_U = (G_0' S^{-1} G_0)^{-1}$. So the desired result will be established if we can show that $L\tilde{M}_T = N_T + o_p(1)$. To this end, we now consider the limiting behaviour of $G_0' S^{-1} T^{1/2} g_T(\tilde{\theta}_T)$. Using (5.53), it follows that

$$G_0' S^{-1} T^{1/2} g_T(\tilde{\theta}_T) \;=\; G_0' S^{-1} T^{1/2} g_T(\theta_0) \;+\; G_0' S^{-1} G_0 T^{1/2} (\tilde{\theta}_T - \theta_0) \;+\; o_p(1) \tag{5.63}$$

Equation (5.58) provides an asymptotically equivalent expression for $T^{1/2}(\tilde{\theta}_T - \theta_0)$, and if this expression is subsitituted into (5.63) then we obtain

$$G_0' S^{-1} T^{1/2} g_T(\tilde{\theta}_T) \;=\; R'[R V_U R']^{-1} R V_U G_0' S^{-1} T^{1/2} g_T(\theta_0) \;+\; o_p(1) \tag{5.64}$$

If (5.64) is substituted into (5.63) then – with appropriate cancellations – we obtain $L\tilde{M}_T = N_T + o_p(1)$.

Part (c): Once again, the proof rests in part on an application of the Mean Value theorem to $T^{1/2} g_T(\tilde{\theta}_T)$ but this time it is taken around $T^{1/2} g_T(\hat{\theta}_T)$ to yield

$$T^{1/2} g_T(\tilde{\theta}_T) \;=\; T^{1/2} g_T(\hat{\theta}_T) \;+\; G_T(\tilde{\theta}_T, \hat{\theta}_T, \lambda_T) T^{1/2} (\tilde{\theta}_T - \hat{\theta}_T) \tag{5.65}$$

where $G_T(\tilde{\theta}_T, \hat{\theta}_T, \lambda_T)$ is the $(q \times p)$ matrix whose i^{th} row is the i^{th} row of $G_T(\bar{\theta}_T^{(i)})$ where (this time) $\bar{\theta}_T^{(i)} = \lambda_{T,i} \tilde{\theta}_T + (1 - \lambda_{T,i}) \hat{\theta}_T$ for some $0 \leq \lambda_{T,i} \leq 1$ and λ_T is the $(q \times 1)$ vector with i^{th} element $\lambda_{T,i}$. Since both $\hat{\theta}_T$ and $\tilde{\theta}_T$ are consistent, it follows that $\bar{\theta}_T^{(i)}$ must also converge in probability to θ_0 for $i = 1, 2, \ldots q$. This property can then be combined with Assumptions 3.5, 3.12–3.13 to deduce that $G_T(\tilde{\theta}_T, \hat{\theta}_T, \lambda_T) \xrightarrow{p} G_0$ and so that (5.65) implies

$$T^{1/2} g_T(\tilde{\theta}_T) \;=\; T^{1/2} g_T(\hat{\theta}_T) \;+\; G_0 T^{1/2} (\tilde{\theta}_T - \hat{\theta}_T) \;+\; o_p(1) \tag{5.66}$$

If (5.66) is used to substitute for $T^{1/2} g_T(\tilde{\theta}_T)$ in (5.45) then it emerges after a little rearrangement that

$$\begin{aligned} D_T \;&=\; 2T^{1/2} (\tilde{\theta}_T - \hat{\theta}_T)' G_0' \hat{S}_T^{-1} T^{1/2} g_T(\hat{\theta}_T) \\ &\quad + T^{1/2} (\tilde{\theta}_T - \hat{\theta}_T)' G_0' \hat{S}_T^{-1} G_0 T^{1/2} (\tilde{\theta}_T - \hat{\theta}_T) \;+\; o_p(1) \end{aligned} \tag{5.67}$$

Clearly to proceed further, we need an expression for $T^{1/2}(\tilde{\theta}_T - \hat{\theta}_T)$. Since

$$T^{1/2} (\tilde{\theta}_T - \hat{\theta}_T) \;=\; T^{1/2} (\tilde{\theta}_T - \theta_0) \;-\; T^{1/2} (\hat{\theta}_T - \theta_0)$$

it follows from (5.61) and (5.58) that

$$T^{1/2} (\tilde{\theta}_T - \hat{\theta}_T) \;=\; V_U R' [R V_U R']^{-1} R V_U G_0' S^{-1} T^{1/2} g_T(\theta_0) \;+\; o_p(1) \tag{5.68}$$

We note parenthetically that under our conditions (5.68) implies that

$$T^{1/2}(\tilde{\theta}_T - \hat{\theta}_T) \xrightarrow{d} N(0, V_U R'[RV_U R']^{-1} RV_U)$$

Using (5.68), we can now deduce the limiting behaviour of the terms on the right hand side of (5.67) in turn. First consider

$$D_{1,T} = 2T^{1/2}(\tilde{\theta}_T - \hat{\theta}_T)' G_0' \hat{S}_T^{-1} T^{1/2} g_T(\hat{\theta}_T)$$

The first order conditions for the unrestricted estimation, (3.12), imply that

$$G_T(\hat{\theta}_T)' \hat{S}_T^{-1} T^{1/2} g_T(\hat{\theta}_T) = 0 \tag{5.69}$$

Since $G_T(\hat{\theta}_T) \xrightarrow{p} G_0$, it follows from (5.69) that $G_0' \hat{S}_T^{-1} T^{1/2} g_T(\hat{\theta}_T) = o_p(1)$. Furthermore, (5.68) implies $T^{1/2}(\tilde{\theta}_T - \hat{\theta}_T) = O_p(1)$. Therefore we can combine combine these two order in probability statements to deduce $D_{1,T} = O_p(1)o_p(1) = o_p(1)$. Now consider the second term on the right hand side of (5.67), namely

$$D_{2,T} = T^{1/2}(\tilde{\theta}_T - \hat{\theta}_T)' G_0' \hat{S}_T^{-1} G_0 T^{1/2}(\tilde{\theta}_T - \hat{\theta}_T) \tag{5.70}$$

It follow from the consistency of $\hat{S}_T$ and (5.68) that $D_{2,T} = N_T + o_p(1)$. Therefore $D_T = D_{1,T} + D_{2,T} = N_T + o_p(1)$. ◇

5.4 Testing Hypotheses About Structural Stability

So far, it has been assumed that if Assumption 3.3 is violated then the value of $E[f(v_t, \theta_0)]$ is the same for all t (albeit for a given T in the case of local misspecification). This property is refered to as *structural stability*. However, Assumption 3.3 is also violated if $E[f(v_t, \theta_0)] \neq 0$ for only part of the sample; such behaviour is termed *structural instability*. This section reviews various methods for testing structural stability based on GMM estimators.

The null hypothesis for structural stability tests is very simple: it states that Assumption 3.3 holds throughout the sample. The alternative is more difficult, however, because it must specify how the model changes. In the GMM literature, attention has focused almost exclusively on the case where the instability involves a discrete change at a single point in the sample known as the "break point". So this scenario receives the most attention here. However, we briefly discuss other forms of instability at the end of the section. To present the null and alternative hypotheses, it is necessary to introduce the following notation. Let π be a constant defined on $(0, 1)$ and let πT denote the potential break point at which some aspect of the model changes. For our purposes here, it is convenient to divide the original sample into two sub-samples. Sub-sample 1 consists of the observations before the break point, namely $T_1 = \{1, 2, \ldots [\pi T]\}$,

where [.] denotes the integer part, and sub-sample 2 consists of the observations after the break point, $T_2 = \{[\pi T] + 1, \ldots T\}$. This break point may be treated as *known* or *unknown* in the construction of the tests. If it is known, then the break point is specified a priori by the researcher and it is only desired to test for instability at this point alone. For example, we investigate below whether the change in operating procedures by the Federal Reserve in October 1979 caused instability in Hansen and Singleton's (1982) consumption based asset pricing models. If the break point is unknown, then the null is the broader hypothesis that there is no instability at any point in the sample. It is easily imagined that tests for the two cases are closely related. We begin our discussion with the simpler case in which the break point is known because this provides a more convenient setting for introducing the null hypotheses and the test statistics. We then consider the extension of these techniques to the unknown break point case.

5.4.1 Known Break Point Case

As remarked above, the basic null hypothesis of structural stability is very straightforward, namely

$$H_0^{SS}(\pi): \; E[f(v_t, \theta_0)] \; = \; 0 \qquad \text{for all } t \in T_1 \& T_2$$

However, rather than work directly with $H_0^{SS}(\pi)$, it is useful to decompose this hypothesis into statements about the stability of the identifying and overidentifying restrictions. It can be recalled from Section 3.3 that these two sets of restrictions play different roles in the estimation, and we have already seen in this chapter that these roles are reflected in the types of inference question for which each is used. The identifying restrictions are imposed in estimation, and so underlie hypotheses about θ_0. The overidentifying restrictions are ignored in estimation, and so can form the basis for inference about the validity of the model specification. It emerges below that similar connections arise in the context of structural stability testing, and this leads to valuable model building information. It is therefore useful to decompose $H_0^{SS}(\pi)$ to reflect these two possible sources of instability, and develop a test for each.

To introduce these component null hypotheses, it is necessary to allow for the possibility that the data generation process for v_t is different in T_1 and T_2. Accordingly, let $E_i[.]$ and $Var_i[.]$ denote the expectation and variance operators relative to the data generation process for v_t in T_i. Furthermore, we define the following sub-sample analogs to $P(\theta)$, $F(\theta)$ and S: $P_i(\theta, \pi) = F_i(\theta, \pi)[F_i(\theta, \pi)' F_i(\theta, \pi)]^{-1} F_i(\theta, \pi)'$, $F_i(\theta, \pi) = S_i(\theta, \pi)^{-1/2} E_i[\partial f(v_t, \theta_i)/\partial \theta']$,

$$S_1(\theta_1, \pi) \; = \; \lim_{T \to \infty} Var_1[[\pi T]^{-1/2} \sum_{t=1}^{[\pi T]} f(v_t, \theta_1)]$$

$$S_2(\theta_2, \pi) \; = \; \lim_{T \to \infty} Var_2[(T - [\pi T])^{-1/2} \sum_{t=[\pi T]+1}^{T} f(v_t, \theta_2)]$$

Since the identifying restrictions are imposed in estimation, there are always parameter values which satisfy them in each of the two sub-samples. Therefore, the identifying restrictions are said to be structurally stable if they are satisfied by the same parameter value in each sub-sample. This null is formally stated as

$$\begin{aligned} H_0^I(\pi): \quad & P_1(\theta_0,\pi)\left\{S_1(\theta_0,\pi)\right\}^{-1/2}E_1[f(v_t,\theta_0)]=0, \qquad t \in T_1 \\ & P_2(\theta_0,\pi)\left\{S_2(\theta_0,\pi)\right\}^{-1/2}E_2[f(v_t,\theta_0)]=0, \qquad t \in T_2 \end{aligned}$$

In contrast, the overidentifying restrictions are ignored in estimation and so we can examine their stability directly. The overidentifying restrictions are said to be stable if they hold before and after the break point. This is formally stated as

$$H_0^O(\pi) \;=\; H_0^{O1}(\pi)\,\&\,H_0^{O2}(\pi)$$

where

$$\begin{aligned} H_0^{O1}(\pi): \quad & [I_q - P_1(\theta_1,\pi)]\left\{S_1(\theta_1,\pi)\right\}^{-1/2}E_1[f(v_t,\theta_1)]=0, \qquad t \in T_1 \\ H_0^{O2}(\pi): \quad & [I_q - P_2(\theta_2,\pi)]\left\{S_2(\theta_2,\pi)\right\}^{-1/2}E_2[f(v_t,\theta_2)]=0, \qquad t \in T_2 \end{aligned}$$

Notice that $H_0^{O1}(\pi)$ and $H_0^{O2}(\pi)$ allow for the possibility that the overidentifying restrictions are satisfied at different values in each sub-sample.

By the very nature of the decomposition, it is clear that any instability must be reflected in a violation of at least one of the hypotheses $H_0^I(\pi)$ and $H_0^O(\pi)$. Therefore it follows that

$$H_0^{SS}(\pi) \;=\; H_0^I(\pi)\,\&\,H_0^O(\pi)$$

The value of this decomposition is that it allows the researcher to discriminate between two scenarios of empirical interest. The first is one in which the instability is confined to the parameters alone; this case is consistent with a violation of $H_0^I(\pi)$ but the validity of $H_0^O(\pi)$. The second scenario is one in which the instability is not confined to the parameters alone but effects other aspects of the model; this would imply a violation of $H_0^O(\pi)$ and most likely $H_0^I(\pi)$ as well.

We now describe test statistics for each component, and then present their asymptotic properties. To this end, we introduce the following notation and an additional assumption. Let the sample moment in each sub-sample be

$$\begin{aligned} g_{1,T}(\theta;\pi) &= [\pi T]^{-1}\sum_{t=1}^{[\pi T]} f(v_t,\theta) \\ g_{2,T}(\theta;\pi) &= (T-[\pi T])^{-1}\sum_{t=[\pi T]+1}^{T} f(v_t,\theta) \end{aligned}$$

and $\hat{S}_{1,T}(\pi)$, $\hat{S}_{2,T}(\pi)$ be consistent estimators of $S_1(\theta_1,\pi)$, $S_2(\theta_2,\pi)$ respectively. With these definitions, the sub-sample two step GMM estimators are

$$\hat{\theta}_{i,T}(\pi) \;=\; argmin_{\theta\in\Theta}\, g_{i,T}(\theta;\pi)'\,\hat{S}_{i,T}(\pi)^{-1}g_{i,T}(\theta;\pi) \tag{5.71}$$

for $i = 1, 2$. We also need the sub-sample derivative matrices,

$$G_{1,T}(\theta;\pi) = [\pi T]^{-1} \sum_{t=1}^{[\pi T]} \partial f(v_t,\theta)/\partial\theta'$$

$$G_{2,T}(\theta;\pi) = (T - [\pi T])^{-1} \sum_{t=[\pi T]+1}^{T} \partial f(v_t,\theta)/\partial\theta'$$

The additional assumption governs the dependence – or, more appropriately, the lack of it – between the two sub-samples. Throughout the discussion we impose the following condition.

Assumption 5.5 Zero Covariance of Partial Sums
$\lim_{T\to\infty} Cov[T^{1/2} g_{1,T}(\theta_0;\pi), T^{1/2} g_{2,T}(\theta_0;\pi)] = 0.$

This assumption is not guaranteed under ergodicity but can be justified under certain mixing conditions; see Andrews (1993).

From our earlier discussion, it can be recognized that $H_0^I(\pi)$ is equivalent to a null hypothesis of no parameter variation. Andrews and Fair (1988) derive test statistics for the latter hypothesis and it is most convenient to follow their approach here. Therefore, we introduce the augmented population moment condition:

$$E[g(v_t,\phi_0)] = \begin{bmatrix} d_t(\pi) f(v_t,\theta_1) \\ (1 - d_t(\pi)) f(v_t,\theta_2) \end{bmatrix} = 0 \tag{5.72}$$

where $d_t(\pi)$ is a dummy variable which equals one when $t \leq \pi T$ and $\phi_0 = (\theta_1', \theta_2')'$. Notice that this population moment condition is more general than Assumption 3.3 because it allows for the possibility that $E[f(v_t,\theta)] = 0$ is satisfied at different parameter values before and after the break point. However, if ϕ_0 satisfies the restrictions

$$[I_p, -I_p]\phi_0 = 0_p \tag{5.73}$$

then $\theta_1 = \theta_2$ and so the moment condition is satisfied at the same parameter value throughout the sample. This structure suggests a straightforward method for testing $H_0^I(\pi)$: estimate ϕ_0 by GMM based on (5.72) and then use the Wald, LM or LR-type statistic from the previous section to test the restrictions in (5.73). This approach requires calculation of the unrestricted and restricted estimators of ϕ_0 denoted by $\hat{\phi}_{U,T}$ and $\hat{\phi}_{R,T}$ respectively. The unrestricted estimator is $\hat{\phi}_{U,T} = [\hat{\theta}_{1,T}(\pi)', \hat{\theta}_{2,T}(\pi)']'$. The restricted estimator is $\hat{\phi}_{R,T} = [\tilde{\theta}_T(\pi)', \tilde{\theta}_T(\pi)']'$ where

$$\tilde{\theta}_T(\pi) = argmin_{\theta\in\Theta} \sum_{i=1}^{2} g_{i,T}(\theta;\pi)' \hat{S}_{i,T}(\pi)^{-1} g_{i,T}(\theta;\pi) \tag{5.74}$$

However, Andrews (1993) shows that $sup_{\pi\in(0,1)} \|T^{1/2}(\hat{\theta}_T - \tilde{\theta}_T)\| = o_p(1)$ under the null hypothesis, where $\hat{\theta}_T$ is the "full sample" GMM estimator defined in

(3.11). As a consequence, the limiting distribution theory is unaffected by the use of the full sample GMM estimator in place of the restricted estimator. Since the full sample estimator has almost certainly been calculated prior to the implementation of the structural stability tests, there is some convenience to making this substitution. Therefore, Andrews proposes versions of the LM and D tests that are based on the full sample estimator and we follow this practice in our presentation as these versions have become common in practice. However, we note in passing that this substitution may have a considerable impact on the value of the statistic in practice; see Section 9.2 for further discussion in the context of an empirical example.

The Wald statistic is given by

$$W_T(\pi) \;=\; T\left[\hat{\theta}_{1,T}(\pi) \;-\; \hat{\theta}_{2,T}(\pi)\right]' \hat{V}_W(\pi)^{-1} \left[\hat{\theta}_{1,T}(\pi) \;-\; \hat{\theta}_{2,T}(\pi)\right] \tag{5.75}$$

where

$$\begin{aligned} \hat{V}_W(\pi) \;&=\; \frac{1}{\pi}[G_{1,T}(\hat{\theta}_{1,T}(\pi);\pi)'\hat{S}_{1,T}(\pi)^{-1}G_{1,T}(\hat{\theta}_{1,T}(\pi);\pi)]^{-1} \;+ \\ &\quad \frac{1}{1-\pi}[G_{2,T}(\hat{\theta}_{2,T}(\pi);\pi)'\hat{S}_{2,T}(\pi)^{-1}G_{2,T}(\hat{\theta}_{2,T}(\pi);\pi)]^{-1} \end{aligned} \tag{5.76}$$

and $\hat{S}_{i,T}(\pi)$ denotes a consistent estimator of $S_i(\pi)$ based on the unrestricted estimator $\hat{\theta}_{i,T}(\pi)$. The LM statistic is given by

$$\begin{aligned} LM_T(\pi) \;&=\; \frac{T\pi}{(1-\pi)} g_{1,T}(\hat{\theta}_T;\pi)\hat{S}_T^{-1}G_T(\hat{\theta}_T)[G_T(\hat{\theta}_T)'\hat{S}_T^{-1}G_T(\hat{\theta}_T)]^{-1} \times \\ &\quad G_T(\hat{\theta}_T)'\hat{S}_T^{-1}g_{1,T}(\hat{\theta}_T;\pi) \end{aligned} \tag{5.77}$$

The D statistic is given by,

$$D_T(\pi) \;=\; T[J(\hat{\theta}_T,\hat{\theta}_T;\pi) \;-\; J(\hat{\theta}_{1,T}(\pi),\hat{\theta}_{2,T}(\pi);\pi)] \tag{5.78}$$

where

$$\begin{aligned} J(\theta_1,\theta_2,\pi) \;&=\; \pi\, g_{1,T}(\theta_1;\pi)'\hat{S}_{1,T}(\pi)^{-1}g_{1,T}(\theta_1;\pi) \;+ \\ &\quad (1-\pi)\, g_{2,T}(\theta_2;\pi)'\hat{S}_{2,T}(\pi)^{-1}g_{2,T}(\theta_2;\pi) \end{aligned} \tag{5.79}$$

To test $H_0^O(\pi)$, Hall and Sen (1999) propose the statistic

$$O_T(\pi) \;=\; O_{1,T}(\pi) \;+\; O_{2,T}(\pi) \tag{5.80}$$

where $O_{1,T}(\pi)$ and $O_{2,T}(\pi)$ are the overidentifying restrictions tests based on the sub–samples T_1 and T_2 respectively, that is

$$O_{1,T}(\pi) \;=\; [\pi T]g_{1,T}(\hat{\theta}_{1,T}(\pi);\pi)'\hat{S}_{1,T}(\pi)^{-1}g_{1,T}(\hat{\theta}_{1,T}(\pi);\pi) \tag{5.81}$$

$$O_{2,T}(\pi) \;=\; (T-[\pi T])g_{2,T}(\hat{\theta}_{2,T}(\pi);\pi)'\hat{S}_{2,T}(\pi)^{-1}g_{2,T}(\hat{\theta}_{2,T}(\pi);\pi) \tag{5.82}$$

The following theorem gives the limiting distribution of these statistics. For brevity, we state the result in terms of the Wald test but the same results apply to either the LM or D statistics. For convenience, we also state the distributional results under the composite null $H_0^{SS}(\pi)$.

Theorem 5.8 Limiting Distributions of $W_T(\pi)$ and $O_T(\pi)$ under $H_0^{SS}(\pi)$ *If Assumptions 3.1–3.5, 3.8–3.13 and 5.5 hold then: (i) $W_T(\pi) \xrightarrow{p} \chi^2_p$; (ii) $O_T(\pi) \xrightarrow{p} \chi^2_{2(q-p)}$; (iii) $W_T(\pi)$ and $O_T(\pi)$ are asymptotically independent.*

Part (i) is first presented in Andrews and Fair (1988)[Theorem 4], and its proof has been anticipated in the derivation of the test statistic above. Parts (ii)–(iii) are presented in Hall and Sen (1999)[Theorem 2.1]. There is a simple intuition behind part (ii): Theorem 5.1 can be used to justify that $O_{1T}(\pi)$ and $O_{2,T}(\pi)$ are individually χ^2_{q-p} and then Assumption 5.5 implies their asymptotic independence which gives the stated result. Part (iii) derives from Assumption 5.5 as well as the arguments which underlie Theorm 3.5.[35]

It can be recalled that the decomposition of $H_0^{SS}(\pi)$ was motivated by the potential to uncover useful information about the source of the instability. To assess whether this potential is realized, we must explore the behaviour of the test statistics under an alternative hypothesis which allows for instability. Hall and Sen (1999) show that $W_T(\pi)$ has power against local alternatives to $H_0^I(\pi)$, denoted $H_A^I(\pi)$, but none against local alternatives to $H_0^O(\pi)$, denoted $H_A^O(\pi)$. Whereas, $O_T(\pi)$ has power against $H_A^O(\pi)$ but none against $H_A^I(\pi)$. Furthermore these two statistics are also asymptotically independent under the composite local alternative $H_A^I(\pi)\&H_A^O(\pi)$. These results suggest that the two statistics can be combined to discriminate between local instability which is due solely to parameter variation and local instability of a more general nature. Interestingly, Hall and Sen (1999) show that this conclusion holds even if the wrong break point is used in the calculation of the tests.[36] However, the same conclusion only holds for non-local alternatives if the correct break point is used. We return to this issue when we describe the extension of these statistics to the unknown break point case in the next sub-section.

At the conclusion of this sub-section, we illustrate the tests for the Hansen and Singleton's (1982) consumption based asset pricing model. However, before that, we briefly describe two other statistics which could be used to test for instability. These are the overidentifying restrictions test and the Predictive test.

Since the overidentifying restrictions test is the standard diagnostic for model specification, it is interesting to consider its properties against structural instability. Ghysels and Hall (1990*a*) show that J_T is insensitive to $H_0^I(\pi)$ and Sen (1997) shows that it has power against $H_0^O(\pi)$. The arguments behind each are essentially the same as those used to establish that J_T has power against H_A^O but none against H_A^I in Section 5.1.3. Hall, Inoue, and Peixe (2003) consider the limiting behaviour of J_T in the presence of non-local misspecification due to neglected structural instability. They provide conditions for the test to be consistent but show that these are not guaranteed to hold in all circumstances.

[35] It should be noted that Theorem 5.8 (i) only requires $H_0^I(\pi)$ to hold, and part (ii) only requires $H_0^O(\pi)$ to hold – provided also that the other regularity conditions are suitably modified; see Hall and Sen (1999).

[36] In other words, the test is calculated with $\pi = \pi_*$, say, but the true break point is $[\pi_0 T]$.

This is because while there may be no single value of θ that satisfies the population moment condition for every observation, there can be a value of θ that sets the average of these population moment conditions to zero. While noteworthy, such a scenario is likely to be the exception rather than the rule. So for practical purposes, it is reasonable to conclude that the overidentifying restrictions test can detect neglected structural instability in many settings. In spite of these properties, intuition suggests that $W_T(\pi)$ and $O_T(\pi)$ are likely to be more powerful tests than J_T against structural instability because they are specifically designed for that alternative. Simulation evidence reported in Sen (1997) supports this view.

Ghysels and Hall (1990*c*) proposed the Predictive test to discriminate between $H_0^{SS}(\pi)$ and the alternative hypothesis

$$H_A^{PR}(\pi): \; E_1[f(v_t, \theta_0)] = 0, \qquad t \in T_1 \qquad \text{and} \qquad E_2[f(v_t, \theta_0)] \neq 0, \quad t \in T_2$$

The statistic is based on evaluating the sample moments from T_2 at $\hat{\theta}_{1,T}(\pi)$. Under $H_0^{SS}(\pi)$, this estimated sample moment should converge in probablity to zero. This approach leads to the Predictive test statistic

$$PR_T(\pi) = T g_{2,T}(\hat{\theta}_{1,T}(\pi); \pi)' \, \hat{V}_{PR}^{-1} \, g_{2,T}(\hat{\theta}_{1,T}(\pi); \pi)$$

where $\hat{V}_{PR}$ is a covariance matrix defined in Ghysels and Hall (1990*c*). Ghysels and Hall (1990*c*) show that this statistic converges to a χ_q^2 distribution under $H_0^{SS}(\pi)$.[37] Ghysels, Guay, and Hall (1997) show that

$$H_A^{PR}(\pi) \; = \; H_A^I(\pi) \& H_0^{O1}(\pi) \& H_A^{O2}(\pi)$$

In other words, the Predictive test has no power against violations of $H_0^{O1}(\pi)$. This feature renders the Predictive test less attractive than the combined use of $W_T(\pi)$ – or $LM_T(\pi)$, $D_T(\pi)$ – and $O_T(\pi)$ described above and so we do not pursue it further here.[38]

Example: Hansen and Singleton's (1982) Consumption Based Asset Pricing Model

Since our sample spans five decades there are many events which may have caused structural instability in an asset pricing model. To illustrate the tests described above, we pick one such event: the change in the operating procedures of the Federal Reserve in October, 1979. During the 1960s and most of the 1970s, the Federal Reserve used the federal funds rate as its primary operating target for monetary policy.[39] In October 1979, it was decided to change this practice to one in which the level of non-borrowed reserves became the primary operating

[37] Ghysels and Hall (1990*a*) propose a structural stability test based along a similar principle to the Eichenbaum, Hansen, and Singleton (1988) statistic in Section 5.2, but Ahn (1995) shows this is asymptotically equivalent to the Predictive test – their finite sample properties may be different, however.

[38] Ghysels, Guay, and Hall (1997) also extend the Predictive test to the unknown break point case.

[39] The federal funds rate is the interest rate on funds loaned overnight between banks.

target.[40] It has been argued in the literature that this change in Fed policy may have had sufficient impact on the financial environment to cause instability in asset pricing models.[41]

The evidence from the overidentifying restrictions test suggests that this model may be correctly specified for value weighted returns (VWR), but misspecified for equally weighted returns (EWR). Both conclusions leave scope for the use of structural stability tests although for different reasons. It can be recalled above that the overidentifying restrictions test has power against structural instability but is anticipated to be less powerful than tests specifically designed for this alternative. So for VWR, the motivation is that the failure to reject with the overidentifying restrictions test may simply reflect the low power of the test against structural instability. Whereas for EWR, the motivation is to assess whether the significance of the overidentifying restrictions test can be attributed to structural instability. Table 5.4 reports the structural stability test statistics associated with the October 1979 break point. For brevity, we only report results based on $W_T^{(1)} = (T^{-1}\sum_{t=1}^{T} z_t z_t')^{-1}$ and $\hat{S}_T = \hat{S}_{SU}$ given in (3.40).

For VWR, the overidentifying restrictions based tests are all insignificant at the 10% level. However, the evidence from the parameter variation tests is mixed. The Wald and LM tests are insignificant but the D test is just significant at the 10% level. Unfortunately, there is no obvious way to interpret this discrepancy between the tests of parameter variation. Statistical theory tells us only that $W_T(\pi)$, $LM_T(\pi)$ and $D_T(\pi)$ are asymptotically equivalent under the null and local alternatives but this does not imply the tests need be numerically identical in finite samples. However, one possible explanation is that the $D_T(\pi)$ is calculated using the full sample GMM estimator as the "restricted estimator". While this substitution is asymptotically valid, it may inflate the value of the statistic because it follows from (5.74) that $J(\hat{\theta}_T, \hat{\theta}_T; \pi) \geq J(\tilde{\theta}_T, \tilde{\theta}_T; \pi)$.[42]

For EWR, the evidence is more clear cut. All the parameter variation tests are insignificant at the 10% level, but the overidentifying restrictions based tests indicate instability. Both $O_T(\pi)$ and $O_{2,T}(\pi)$ are significant at the 10% level, but $O_{1,T}(\pi)$ is insignificant. This pattern of results suggests the model specification is correct prior to 1979:9, but misspecified thereafter.[43] Provided we accept the general framework of the consumption based asset pricing model, the most logical source of this misspecification is the representative agent's utility function. So with this proviso, the evidence is consistent with the following scenario. The representative agent possesses a CRRA utility function for the period 1959:3–1979:9, but then the functional form of this utility function changes as

[40] See Mishkin (1995) for a historical review of the Federal Reserve's monetary policy.
[41] See *inter alia* Ghysels and Hall (1990*a*).
[42] See Section 9.2.
[43] This conclusion appears at odds with the results reported in Hansen and Singleton (1984) who report a significant overidentifying restrictions test for the model with EWR. However, the overidentifying restrictions test based on the pre break sample is sensitive to the choice of break point; see Section 5.4.2 for further details.

a result of some event in 1979.10.[44] However, there is one important caveat. Although we have selected this break point for a reason, all these results may be sensitive to the choice of break point and so it is important to conduct a more thorough investigation before drawing any definitive conclusions about the importance of this date. This is undertaken at the end of the next subsection.

Table 5.4
Structural stability tests associated with October 1979

Asset :	VWR		EWR	
Test	*Statistic*	*p-value*	*Statistic*	*p-value*
$W_T(\pi)$	2.640	0.267	1.810	0.405
$LM_T(\pi)$	3.382	0.184	0.888	0.641
$D_T(\pi)$	5.040	0.080	2.543	0.280
$O_T(\pi)$	4.135	0.658	12.031	0.061
$O_{1T}(\pi)$	1.535	0.674	4.288	0.232
$O_{2T}(\pi)$	2.601	0.457	7.743	0.052

Note: $W_T(\pi)$, $LM_T(\pi)$ and $D_T(\pi)$ are defined in (5.75), (5.77) and (5.78), $O_T(\pi)$, $O_{1T}(\pi)$ and $O_{2T}(\pi)$ are defined in (5.80), (5.81) and (5.82).

5.4.2 Unknown Break Point Case

If the break point is unknown, then it is desired to test whether there is evidence of instability at any point in the sample. However, in practice, it is necessary to limit attention to the null hypothesis:

$$H_0^{SS}(\Pi) \;=\; H_0^{SS}(\pi), \text{ for all } \pi \in \Pi \subset (0,1) \tag{5.83}$$

On one hand, it is desirable for Π to be as wide as possible so that the null is as broad as possible. On the other hand, it must not be so wide that asymptotic theory is a poor approximation in the sub-samples. In applications to models of economic time series, it has become customary to use $\Pi = [0.15, 0.85]$. As in the fixed break point case, we decompose the null into components involving the stability of the identifying and overidentifying restrictions, that is

$$H_0^{SS}(\Pi) \;=\; H_0^{I}(\Pi) \;\&\; H_0^{O}(\Pi) \tag{5.84}$$

where

$$H_0^{I}(\Pi) \;=\; H_0^{I}(\pi), \text{ for all } \pi \in \Pi \tag{5.85}$$
$$H_0^{O}(\Pi) \;=\; H_0^{O}(\pi), \text{ for all } \pi \in \Pi \tag{5.86}$$

[44] See Sen and Hall (1999) for further discussion.

We begin by describing statistics for testing $H_0^I(\Pi)$. The construction is a natural extension of the fixed break point methods. Now $W_T(\pi)$, say, is calculated for each possible π to produce a sequence of statistics indexed by π, and inference is based on some function of this sequence. This function is chosen to maximize power against a local alternative in which a weighting distribution is used to indicate the relative importance of departures from $H_0^I(\pi)$ in different directions at different break points. A general framework for the derivation of these optimal tests is provided by Andrews and Ploberger (1994) in the context of Maximum Likelihood estimators and this is extended to the GMM framework by Sowell (1996). One drawback with this approach is that a different choice of weighting distribution leads to a different optimal statistic; however, three choices have received particular attention. To facilitate their presentation, we define the following local alternative to $H_0^I(\pi)$,

$$\begin{aligned} H_{A,T}^I(\pi): \qquad & P_1(\theta_0;\pi)\{S_1(\theta_0,\pi)\}^{-1/2}E_{1,T}[f(v_t,\theta_0)] = T^{-1/2}\mu_{I,1}, \qquad t \in T_1 \\ & P_2(\theta_0;\pi)\{S_2(\theta_0,\pi)\}^{-1/2}E_{2,T}[f(v_t,\theta_0)] = T^{-1/2}\mu_{I,2}, \qquad t \in T_2 \end{aligned}$$

It is assumed that $\mu_{I,1} = 0$ and a weighting distribution is specified for $(\mu_{I,2}, \pi)$.[45] The aforementioned three choices are as follows:

Choice 1:

If the conditional weighting distribution of $\mu_{I,2}$ given π is of the form $rL(\pi)U$ where r is a scalar, $L(\pi)$ is a particular matrix and U is the uniform distribution on the unit sphere in $\Re^p$ then Andrews and Ploberger (1995) show that for r sufficiently large the optimal statistic is

$$SupW_T \;=\; \sup_{\pi\in\Pi}\{\, W_T(\pi)\,\}$$

Choices 2 and 3:

If the conditional weighting distribution of $\mu_{I,2}$ given π as $N(0, c\Sigma_\pi)$, for some constant c. Andrews and Ploberger (1994) and Sowell (1996) show that for a particular choice of Σ_π, the optimal statistic only depends on c and not Σ_π. So, for convenience, this choice is made and then attention has focused on two values of c. If $c = 0$ then the optimal statistic takes the form

$$AvW_T \;=\; \int_\Pi W_T(\pi)dJ(\pi)$$

where $J(\pi)$ defines the weighting distribution over π. If $c = \infty$ then the optimal statistic takes the form

$$ExpW_T \;=\; log\left\{\int_\Pi exp[0.5W_T(\pi)]dJ(\pi)\right\}$$

In principle, AvW_T and $ExpW_T$ can be calculated with any choice of marginal distribution for π. However, it has become customary to assume this distribution is uniform over Π.

[45] For these tests of parameter variation, the roles of $\mu_{I,1}, \mu_{I,2}$ can be interchanged.

As they stand these statistics are not operational because we have treated π as continuous, whereas in practice it is discrete. For a given sample size, the set of possible break points are $T_b = \{i/T;\ i = [\pi_L T], [\pi_L T] + 1, \ldots, [\pi_U T]\}$ where π_L and π_U are respectively the lower and upper endpoints of the closed interval Π. So in practice, inference is based on the discrete analogs to $SupW_T$, AvW_T and $ExpW_T$, that is

$$SupW_T = \sup_{i \in T_b} \{ W_T(i/T) \} \tag{5.87}$$

$$AvW_T = d(\pi_L, \pi_U, T)^{-1} \sum_{i=[\pi_L T]}^{[\pi_U T]} W_T(i/T) \tag{5.88}$$

$$ExpW_T = log \left\{ d(\pi_L, \pi_U, T)^{-1} \sum_{i=[\pi_L T]}^{[\pi_U T]} exp[0.5 W_T(i/T)] \right\} \tag{5.89}$$

where the last two statistics are specialized to the case in which the weighting distribution for π is uniform on Π, and $d(\pi_L, \pi_U, T) = [\pi_U T] - [\pi_L T] + 1$.

Andrews (1993, 2003) and Andrews and Ploberger (1994) derive and tabulate the limiting ditributions of $SupW_T$, Av_T and $ExpW_T$ under $H_0(\Pi)$. We delay a discussion of the theoretical arguments to the end of this section. Critical points for these distributions are reproduced here for $\Pi = [.15, .85]$ in Table 5.5.[46] These enable the researcher to ascertain whether the statistic is significant at preascribed level. Hansen (1997) reports response surfaces which can be used to calculate p-values for all three versions of these tests. As a reminder, all the previous remarks equally apply to the corresponding functionals of $LM_T(\pi)$ or $D_T(\pi)$.

[46] Table 5.5 only contains parts of the tabulations reported by Andrews (1993) and Andrews and Ploberger (1994). They report critical points for $p = 1, 2, \ldots 20$ and other choices of Π.

Table 5.5
Critical points for $SupW_T$, AvW_T and $ExpW_T$

Statistic: $SupW_T$

p	10%	5%	1%
1	7.12	8.68	12.16
2	10.00	11.72	15.56
3	12.28	14.13	18.07
4	14.34	16.36	20.47
5	16.30	18.32	22.66
6	18.11	20.24	24.74
7	19.87	22.06	26.72
8	21.55	23.82	28.55
9	23.20	25.54	30.42
10	24.80	27.13	32.31
11	26.38	28.81	33.96
12	27.90	30.43	35.67

Statistic: AvW_T

p	10%	5%	1%
1	2.16	2.88	4.72
2	3.75	4.61	6.73
3	5.10	6.07	8.21
4	6.50	7.67	10.18
5	7.76	9.01	11.32
6	9.02	10.19	12.93
7	10.28	11.47	14.34
8	11.54	12.94	16.14
9	12.71	14.16	17.30
10	13.77	15.29	18.72
11	15.00	16.46	19.44
12	16.31	17.85	21.03

Statistic: $ExpW_T$

p	10%	5%	1%
1	1.51	2.06	3.41
2	2.59	3.22	4.76
3	3.49	4.22	5.77
4	4.37	5.23	7.13
5	5.22	6.13	7.91
6	6.01	6.92	8.96
7	6.70	7.66	9.53
8	7.58	8.60	10.96
9	8.31	9.35	11.67
10	9.00	10.04	12.61
11	9.69	10.75	13.21
12	10.45	11.55	13.83

Source: Andrews (2003)[Table 1] and Andrews and Ploberger (1994) [Tables 1 and 2]. Copyright: The Econometric Society. Reproduced with permission.
Notes: the figures represent the critical points for the three tests at the 10%, 5% and 1% significance level for $\Pi = [0.15, 0.85]$.

The same ideas can be used to construct tests of the null hypothesis that $H_0^O(\pi)$ holds for all $\pi \in \Pi$ against the alternative that

$$H_{A,T}^O(\Pi) \;=\; H_{A,T}^{O1}(\pi) \;\&\; H_{A,T}^{O2}(\pi) \text{ for all } \pi \in \Pi$$

where

$$\begin{aligned} H_{A,T}^{O1}(\pi) \quad &: \quad [I_q \;-\; P_1(\theta_1,\pi)]\,\{S_1(\theta_1,\pi)\}^{-1/2} E_{1,T}[f(v_t,\theta_1)] \\ &\qquad = T^{-1/2}\mu_{O1}, \qquad t \;\in\; T_1 \end{aligned}$$

$$\begin{aligned} H_{A,T}^{O2}(\pi) \quad &: \quad [I_q \;-\; P_2(\theta_2,\pi)]\,\{S_2(\theta_2,\pi)\}^{-1/2} E_{2,T}[f(v_t,\theta_2)] \\ &\qquad = T^{-1/2}\mu_{O2}, \qquad t \;\in\; T_2 \end{aligned}$$

Hall and Sen (1999) propose using the following statistics:

$$SupO_T \;=\; \sup_{i \in T_b} \{\, O_T(i/T) \,\} \tag{5.90}$$

$$AvO_T \;=\; d(\pi_L,\pi_U,T)^{-1} \sum_{i=[\pi_L T]}^{[\pi_U T]} O_T(i/T) \tag{5.91}$$

$$ExpO_T \;=\; log\left\{ d(\pi_L,\pi_U,T)^{-1} \sum_{i=[\pi_L T]}^{[\pi_U T]} exp[0.5 O_T(i/T)] \right\} \tag{5.92}$$

However, although the functionals are the same it has proved impossible to date to deduce any optimality properties for the resulting tests along the lines described above.[47] Hall and Sen (1999) derive and tabulate the limiting distributions of these three statistics under $H_0^O(\Pi)$. Once again, we postpone a discussion of the derivation until the end of this sub-section. Critical points for these distributions are reproduced here in Table 5.6. Sen and Hall (1999) report response surfaces which can be used to calculate p-values for all three versions of these tests.

[47] It is possible to derive optimal tests against the more restrictive alternatives $H_0^{O1}(\pi)\&H_{A,T}^{O2}(\pi)$ for all $\pi \in \Pi$ or $H_{A,T}^{O1}(\pi)\&H_0^{O2}(\pi)$ for all $\pi \in \Pi$ but the statistics are different in each case; see the discussion in Hall and Sen (1999). However, notice that both these alternatives restrict the violation of the population moment condition to occur either after or before the break point. Whereas, in practice, a researcher typically lacks that kind of a priori information, and so we do not pursue those tests here.

Table 5.6
Critical points for $SupO_T$, AvO_T and $ExpO_T$

Statistic: $SupO_T$ $q-p$	10%	5%	1%
1	8.70	10.39	14.13
2	12.78	14.75	18.53
3	16.33	18.53	23.19
4	19.65	21.99	26.99
5	22.81	25.31	30.55
6	25.70	28.32	33.90
7	28.76	31.45	37.03
8	31.61	34.53	40.09
Statistic: AvO_T $q-p$	10%	5%	1%
1	4.17	5.37	8.11
2	7.21	8.60	11.96
3	9.91	11.54	15.40
4	12.51	14.32	18.28
5	15.01	17.01	21.39
6	17.52	19.68	24.02
7	19.91	22.24	27.00
8	22.44	24.78	29.52
Statistic: $ExpO_T$ $q-p$	10%	5%	1%
1	2.45	3.13	4.69
2	4.17	4.99	6.87
3	5.73	6.69	8.81
4	7.20	8.26	10.43
5	8.64	9.77	12.20
6	10.02	11.21	13.85
7	11.41	12.69	15.33
8	12.79	14.12	16.80

Source: Hall and Sen (1999)[Table 1]. Copyright 1999 by the American Statistical Association. Reprinted with permission from the *Journal of Business and Economic Statistics*.
Notes: the figures represent the critical points for the three tests at the 10%, 5% and 1% significance level for $\Pi = [0.15, 0.85]$.

Which functional should be used? Simulation evidence suggests that no one test dominates the others.[48] So, unless your priors happen to correspond to one of the weighting distributions underlying the statistics, it is probably best to calculate all three, and this seems to have become the most common practice. However, the *Sup* test does have one attractive feature not shared by the other two. If $SupW_T$, say, occurs at $t = t_B$ then $\hat{\pi}_W = t_B/T$ provides an estimate

[48] See Hall and Sen (1999).

of the break point fraction. To date, it is unknown whether this estimate is consistent for π under the alternative but there are grounds for conjecturing that this property holds in just-identified models at least.[49] This remains an interesting avenue for future research.

It can be recalled that the decomposition of the null hypothesis has been motivated by its potential to provide useful model building information. In the previous sub-section, it is argued that this potential is realized for local instability regardless of the true break point but is only realized for non-local instability if correct break point has been identified. The latter property is underscored by simulation evidence in Hall and Sen (1999) which shows that $SupO_T$, AvO_T and $ExpO_T$ have power against non-local parameter variation. These properties prompted Hall and Sen (1999) to propose the following strategy.

Hall and Sen's (1999) strategy for diagnosing the source of the instability.

Case 1: If all the unknown break point tests fail to reject then this is evidence that all aspects of the model are stable.

Case 2: If the parameter variation tests are significant and either the unknown break point overidentifying restriction based tests are insignificant or $O_T(\hat{\pi}_W)$ is insignificant, then this is evidence of parameter variation.

Case 3: In all other situations, the tests indicate that there is instability that involves more than just the parameters.

Two comments are in order. First, note that the method is premised on the assumption that $\hat{\pi}_W$ is consistent for π if the instability is confined to the parameters alone. Secondly, Hall and Sen (1999) propose evaluating the significance of $O_T(\hat{\pi}_W)$ using the appropriate critical point of the $\chi^2_{2(q-p)}$ distribution, and this ignores a sampling error associated with the estimation of the break point. However, they report simulation evidence which suggests this distributional approximation is reasonably accurate. Their simulation evidence as a whole suggests that the strategy provides a feasible method for discriminating between parameter variation alone and more general forms of instability.

One final point should be noted. Although, all these tests are designed against an alternative in which there is instability at a single point in the sample. All the tests have non-trivial power against other forms of instability. We do not reproduce the argument here but instead refer the reader to the papers already cited above.[50]

Example: Hansen and Singleton's (1982) Consumption Based Asset Pricing Model[51]

[49] For example, Nunes, Kuan, and Newbold (1995) prove its consistency in linear regression models estimated by quasi maximum likelihood.

[50] Also see Section 5.4.3.

[51] See Section 9.2 for another empirical example of these tests.

Table 5.7 reports the structural stability tests when the break point is treated as unknown. Following accepted practice, we set $\Pi = [0.15, 0.85]$ which means the potential break point is assumed to lie between 1965:1 and 1992:2. For brevity, we only report results based on using $W_T^{(1)} = (T^{-1}\sum_{t=1}^{T} z_t z_t')^{-1}$ and the covariance matrix estimator $\hat{S}_T = \hat{S}_{SU}$. Parenthetically, we note that if $W_T^{(1)} = 10^5 I_5$ is used then the sub-sample estimates diverge in a few cases.

For VWR, all the statistics are insignificant at the 10% level, and so these tests provide no evidence of misspecification in this case. For EWR, all the parameter variation statistics are insignificant at the 10% level, and all the overidentifying restrictions based tests are significant at the 5% level. This evidence clearly indicates misspecifcation, and so is consistent with our earlier findings based on the overidentifying restrictions test. However, the application of the structural stability tests provides further information about the nature of the misspecification. Using Hall and Sen's (1999) diagnostic strategy described above, the pattern of results suggests that the misspecification cannnot be attributed to parameter variation alone.

Table 5.7 also reports the dates associated with the supremum of each test. Two features of these results stand out. First, the supremum for the parameter variation tests occurs at virtually the same point for a given choice of asset – in spite of the insignificance of the tests concerned. Secondly, the supremum for the overidentifying restrictions test occurs at the second possible breakpoint, that is 1965:2, for each choice of asset. This could reflect instability, but there is another explanation which needs to be noted. It can be recalled that the statistical theory behind the tests relies on the applicability of asymptotic theory in each of the sub–samples. At either end of Π, one of the sub–samples consists of only seventy observations, and it may be that this is not sufficiently large for asymptotic theory to provide a good approximation. In that case, the supremum may occur close to $\pi = 0.15$ or $\pi = 0.85$ simply because the sequence of test statistics has not converged in distribution over the entire interval Π. Figures 5.1–5.2 plot the individual test statistics for $W_T(\pi)$ and $O_T(\pi)$ against π for each choice of asset. The plots for $D_T(\pi)$ and $LM_T(\pi)$ are qualitatively similar to those for $W_T(\pi)$ and so are omitted for brevity. $\diamond$

Table 5.7
Structural stability tests with unknown break point

VWR:				
Test	*Sup–*	*Date*	*Av–*	*Exp–*
W	4.899	1982:7	1.468	0.899
LM	5.603	1982:9	2.095	1.194
D	5.852	1982:7	2.239	1.427
O	12.759	1965:2	5.751	3.704
EWR:				
Test	*Sup–*	*Date*	*Av–*	*Exp–*
W	4.893	1975:1	1.035	0.623
LM	5.262	1975:1	0.896	0.571
D	6.909	1975:2	1.642	1.123
O	22.580	1965:2	13.712	8.093

Note: W, LM, D and O denote the tests based on $W_T(\pi)$, $LM_T(\pi)$, $D_T(\pi)$ and $O_T(\pi)$ defined in (5.75), (5.77), (5.78) and (5.80). *Date* denotes the date associated with the Supremum statistic.

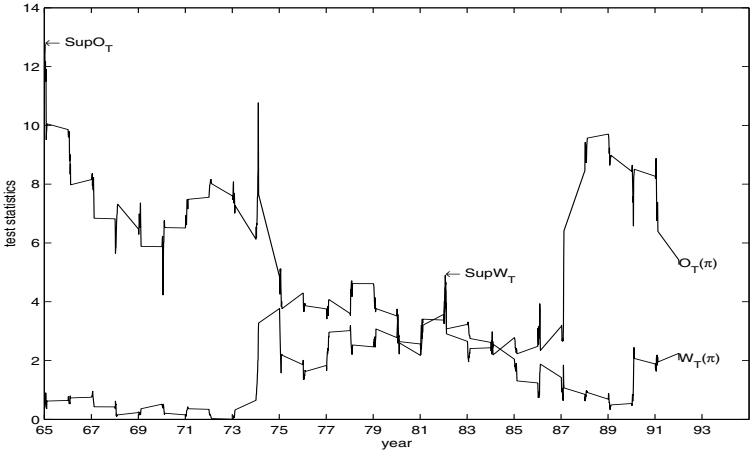

Figure 5.1: Wald and overidentifying restrictions tests for structural instability for the consumption based asset pricing model with value weighted returns

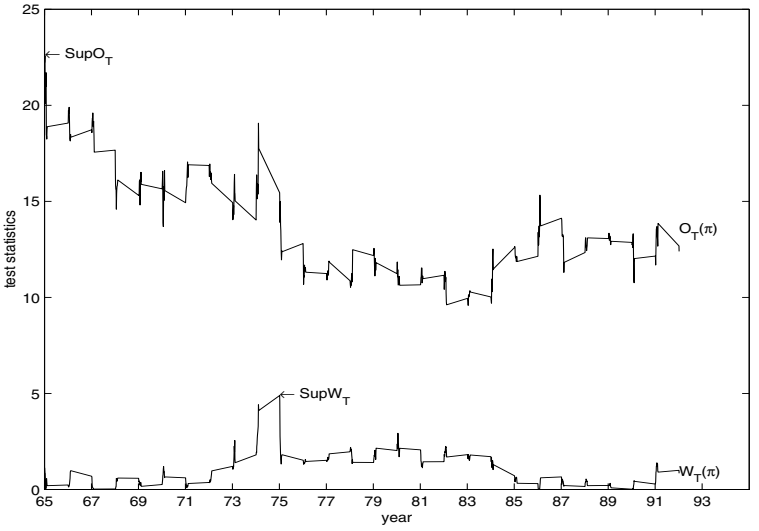

Figure 5.2: Wald and overidentifying restrictions tests for structural instability for the consumption based asset pricing model with equally weighted returns

5.4.2.1 Technical Details

There are two main steps to the analysis of structural stability tests derived above for the unknown break point case. First, it is necessary to characterize the limiting behaviour of individual members of the sequence of statistics $\{W_T(\pi); \pi \in \Pi\}$ and $\{O_T(\pi); \pi \in \Pi\}$. Secondly, these characterizations are used to deduce the limiting behaviour of the various functions of the sequences in which we are interested. The first part is closely related to our earlier analysis of the statistics for the fixed break point case. However, this time, the results must apply for all $\pi \in \Pi$, and this requires different techniques and assumptions. Below it is shown that the limiting distributions of the test statistics revolve around two continuous time processes known as a Brownian Motion and a Brownian Bridge. Therefore, we begin with definitions of these processes.

Definition 5.1 Brownian Motion
A n dimensional Brownian Motion $B_n(.)$ is a continuous time process associating each date $r \in [0,1]$ with the $(n \times 1)$ vector $B_n(r)$ satisfying the following properties:

(i) $B(0) = 0_n$ *where* 0_n *is a* $(n \times 1)$ *vector of zeros.*

(ii) For any dates $0 \leq r_1 \leq r_2 \leq \ldots \leq r_k \leq 1$ *the changes* $\{B_n(r_i) - B_n(r_{i-1}),\ i = 2, 3 \ldots k\}$ *are a set of mutually independent random vectors with* $B_n(r_i) - B_n(r_{i-1}) \sim N(0_n, (r_i - r_{i-1})I_n)$.

(iii) For any given realization, $B_n(r)$ *is continuous in* r *with probability one.*

A Brownian motion is the continuous time analog to a random walk, and is widely used in analyses of diffusion processes.[52]

Definition 5.2 Brownian Bridge
A n dimensional Brownian Bridge $BB_n(.)$ is a continuous time process associating each date $r \in [0,1]$ with the $(n \times 1)$ vector $BB_n(r) = B_n(r) - rB_n(1)$ where $B_n(.)$ is a Brownian motion.

Notice that a Brownian bridge both begins and ends at zero.

Below we establish that certain statistics converge in distribution to the distributions possessed by particular functions of Brownian Motions or Bridges. Such statements require an additional notation. Accordingly, we use $a_T \Rightarrow b$ to denote the statement a_T converges in distribution to the distribution possessed by the random variable b. More succinctly, a_T is said to *weakly converge* to b.

From (5.75) and (5.80), it is clear that the analysis of $W_T(\pi)$ and $O_T(\pi)$ is going to require assumptions about the partial sum, $T^{-1}\sum_{t=1}^{[\pi T]} f(v_t, \theta_0)$, its long run variance and the associated derivative matrix. To this end, we assume the data generation process satisfies the following assumptions.[53]

Assumption 5.6 Uniform Convergence of the Variance of the Partial Sums
$\sup_{\pi \in \Pi} \|Var[T^{-1/2}\sum_{t=1}^{[\pi T]} f(v_t, \theta_0)] - \pi S\| \xrightarrow{p} 0.$

Assumption 5.7 Uniform Convergence of the Partial Derivative Matrix
$\sup_{\pi \in \Pi} \|[T^{-1}\sum_{t=1}^{[\pi T]} \partial f(v_t, \theta_0)]/\partial \theta' - \pi G_0\| \xrightarrow{p} 0.$

Assumption 5.8 Functional Central Limit Theorem
$S^{-1/2}T^{-1/2}\sum_{t=1}^{[\pi T]} f(v_t, \theta_0) \Rightarrow B_q(\pi).$

Notice that Assumptions 5.6 implies both that $S_1(\pi) = S_2(\pi) = S$, and also, together with Assumption 5.7, that $F_1(\theta_0) = F_2(\theta_0) = F(\theta_0)$. The form of the distribution in Assumption 5.8 can be motivated from

$$T^{-1/2}\sum_{t=1}^{[\pi T]} f(v_t, \theta_0) \;=\; \left[\frac{[\pi T]}{T}\right]^{1/2} [\pi T]^{-1/2}\sum_{t=1}^{[\pi T]} f(v_t, \theta_0)$$

by noting that $([\pi T]/T)^{1/2} \approx \pi^{1/2}$, and that the CLT implies $[\pi T]^{-1/2}\sum_{t=1}^{[\pi T]} f(v_t, \theta_0) \xrightarrow{d} N(0, S)$. There is one consequence of Assumption 5.8 which is worth highlighting. Since,

$$T^{-1/2}\sum_{t=1}^{T} f(v_t, \theta_0) \;=\; T^{-1/2}\sum_{t=1}^{[\pi T]} f(v_t, \theta_0) \;+\; T^{-1/2}\sum_{t=[\pi T]+1}^{T} f(v_t, \theta_0)$$

[52] The name derives from the name of the first person to have recorded this type of uninterupted irregular motion in a natural phenomenon. R. Brown was a botanist and he observed the phenomenon when pollen dispersed on water. His results were published in 1828; see Brown (1828).

[53] Recall that for any matrix A, $\|A\| = [tr(A'A)]^{1/2}$.

and Assumption 5.8 implies

$$S^{-1/2}T^{-1/2}\sum_{t=1}^{T} f(v_t,\theta_0) \quad \Rightarrow \quad B_q(1)$$
$$S^{-1/2}T^{-1/2}\sum_{t=1}^{[\pi T]} f(v_t,\theta_0) \quad \Rightarrow \quad B_q(\pi)$$

then it must follow that

$$S^{-1/2}T^{-1/2}\sum_{t=[\pi T]+1}^{T} f(v_t,\theta_0) \;\Rightarrow\; B_q(1) \;-\; B_q(\pi) \tag{5.93}$$

Hamilton (1994)[Sections 17.1–17.3, 18.1] provides a very good introduction to Brownian Motions and the conditions behind Assumptions 5.6–5.8. Davidson (1994)[Part IV] provides a more comprehensive treatment.

With these assumptions in place, we now proceed to characterize the limiting behaviour of $W_T(\pi)$ and $O_T(\pi)$ in terms of Brownian Motions and Brownian Bridges. Consider first $W_T(\pi)$. The end result was first derived by Andrews (1993), but we follow the approach taken by Sowell (1996) which exploits the projection matrix structure inherent in the identifying restrictions.

Since $W_T(\pi)$ depends on $T^{1/2}[\hat{\theta}_{1,T}(\pi) - \hat{\theta}_{2,T}(\pi)]$ and

$$T^{1/2}[\hat{\theta}_{1,T}(\pi) - \hat{\theta}_{2,T}(\pi)] \;=\; T^{1/2}[\hat{\theta}_{1,T}(\pi) - \theta_0] \;-\; T^{1/2}[\hat{\theta}_{2,T}(\pi) - \theta_0] \tag{5.94}$$

we begin by deriving expressions for $T^{1/2}[\hat{\theta}_{i,T}(\pi) - \theta_0]$. To facilitate the analysis, we assume that the GMM estimators based on T_i are consistent for all π.

Assumption 5.9 Consistency of Sub-Sample Estimators
$sup_{\pi\in\Pi}\|\hat{\theta}_{i,T}(\pi) - \theta_0\| \xrightarrow{p} 0$.

We can now repeat exactly the same sequence of arguments as in Section 3.4.2 to obtain the following analogs to (3.26)

$$T^{1/2}[\hat{\theta}_{i,T}(\pi) - \theta_0] \;=\; -\bar{M}_{i,T}(\pi)T^{1/2}g_{i,T}(\theta_0;\pi) \tag{5.95}$$

where

$$\begin{aligned}\bar{M}_{i,T}(\pi) \;=\; & [G_{i,T}(\hat{\theta}_{i,T}(\pi);\pi)'\hat{S}_{i,T}(\pi)^{-1}G_{i,T}(\hat{\theta}_{i,T}(\pi),\theta_0,\lambda_T;\pi)]^{-1} \\ & \times G_{i,T}(\hat{\theta}_{i,T}(\pi);\pi)'\hat{S}_{i,T}(\pi)^{-1}\end{aligned}$$

and $G_{i,T}(\hat{\theta}_{i,T}(\pi),\theta_0,\lambda_T;\pi)$ is defined in an analogous fashion to $G_T(\hat{\theta}_T,\theta_0,\lambda_T)$. To proceed we adopt the following high level assumption.[54]

[54] Andrews (1993) or Ghysels, Guay, and Hall (1997) for more primitive conditions under which Assumption 5.10 holds.

Assumption 5.10 Uniform Convergence of $\bar{M}_{i,T}(\pi)$
$sup_{\pi\in\Pi}\|\bar{M}_{i,T}(\pi) - M_0\| \xrightarrow{p} 0$ *where* $M_0 = (G_0^{'}S^{-1}G_0)^{-1}G_0^{'}S^{-1}$.

It then follows from (5.95), Assumptions 5.6–5.10 and (5.93) that

$$T^{1/2}(\hat{\theta}_{1,T}(\pi) - \theta_0) \Rightarrow -\frac{1}{\pi}[F(\theta_0)'F(\theta_0)]^{-1}F(\theta_0)^{'}B_q(\pi) \tag{5.96}$$

$$T^{1/2}(\hat{\theta}_{2,T}(\pi) - \theta_0) \Rightarrow -\frac{1}{1-\pi}[F(\theta_0)'F(\theta_0)]^{-1}F(\theta_0)^{'} \times (B_q(1) - B_q(\pi)) \tag{5.97}$$

where once again we have set $F(\theta_0) = S^{-1/2}G_0$. The combination of (5.94) and (5.96)–(5.97) yields

$$T^{1/2}[\hat{\theta}_{1,T}(\pi) - \hat{\theta}_{2,T}(\pi)] \Rightarrow -\frac{1}{\pi(1-\pi)}[F(\theta_0)'F(\theta_0)]^{-1}F(\theta_0)^{'}BB_q(\pi) \tag{5.98}$$

Now consider $\hat{V}_W(\pi)$. By similar arguments to above, it can be shown that

$$\begin{aligned}\hat{V}_W(\pi) &\xrightarrow{p} \frac{1}{\pi}[F(\theta_0)'F(\theta_0)]^{-1} + \frac{1}{1-\pi}[F(\theta_0)'F(\theta_0)]^{-1} \\ &= \frac{1}{\pi(1-\pi)}[F(\theta_0)'F(\theta_0)]^{-1}\end{aligned} \tag{5.99}$$

The combination of (5.98)–(5.99) implies

$$W_T(\pi) \Rightarrow \frac{1}{\pi(1-\pi)}BB_q(\pi)'P(\theta_0)BB_q(\pi) \tag{5.100}$$

where once again $P(\theta_0) = F(\theta_0)[F(\theta_0)'F(\theta_0)]^{-1}F(\theta_0)'$. Now $P(\theta_0)$ is a projection matrix with rank equal to p by Assumption 3.6. Therefore $P(\theta_0)$ has p eigenvalues equal to one, $q-p$ eigenvalues equal to zero, and there exists an orthogonal matrix H such that[55]

$$P(\theta_0) = H^{'}\Lambda H \tag{5.101}$$

where $\Lambda = diag(1_p, 0_{q-p})$ and 1_p is a $(p \times 1)$ vector of ones. If we partition H into $[H_1, H_2]$, where H_1 is $q \times p$ then (5.101) implies that

$$P(\theta_0) = \begin{bmatrix} H_1^{'}H_1 & 0 \\ 0 & 0 \end{bmatrix} \tag{5.102}$$

If (5.102) is substituted into (5.100) then it follows that

$$W_T(\pi) \xrightarrow{d} \frac{1}{\pi(1-\pi)}BB_p(\pi)'H_1^{'}H_1BB_p(\pi) \tag{5.103}$$

where $BB_p(\pi)$ denotes the first p elements of $BB_q(\pi)$. Now by definition, H_i are orthogonal matrices and so $H_1H_1^{'} = I_p$. Therefore, $H_1B_p(\pi) \sim N(0_p, \pi I_p)$ and so it follows that $H_1B_p(\pi) \Rightarrow B_p(\pi)$ and hence that $H_1BB_p(\pi) \Rightarrow BB_p(\pi)$. This gives us the following result.

[55] See Dhrymes (1984) [Propositions 52 and 55, pp.61 and 65].

Theorem 5.9 Limiting Distribution of $W_T(\pi)$
If Assumptions 3.1–3.5, 3.8–3.9, 5.6–5.10 hold then: $W_T(.) \Rightarrow W(.)$ where $W(.)$ is a continuous time process associating each date $\pi \in \Pi$ with the scalar $W(\pi) = \frac{1}{\pi(1-\pi)} BB_p(\pi)' BB_p(\pi)$.

Now consider $O_T(\pi)$. By definition, we have

$$O_{1,T} = \|\hat{S}_{1,T}(\pi)^{-1/2}[\pi T]^{1/2} g_{1,T}(\hat{\theta}_{1,T}(\pi);\pi)\|^2 \quad (5.104)$$

$$O_{2,T} = \|\hat{S}_{2,T}(\pi)^{-1/2}(T-[\pi T])^{1/2} g_{2,T}(\hat{\theta}_{2,T}(\pi);\pi)\|^2 \quad (5.105)$$

and we can repeat the same sequence of arguments as in Section 3.4.3 to deduce the following sub-sample analogs to (3.35),

$$\begin{aligned} \hat{S}_{1,T}(\pi)^{-1/2}[\pi T]^{1/2} g_{1,T}(\hat{\theta}_{1,T}(\pi);\pi) &= N_{1,T}(\pi)\hat{S}_{1,T}(\pi)^{-1/2} \\ &\quad \times [\pi T]^{1/2} g_{1,T}(\theta_0;\pi) \quad (5.106) \\ \hat{S}_{2,T}(\pi)^{-1/2}(T-[\pi T])^{1/2} g_{2,T}(\hat{\theta}_{2,T}(\pi);\pi) &= N_{2,T}(\pi)\hat{S}_{2,T}(\pi)^{-1/2}(T-[\pi T])^{1/2} \\ &\quad \times g_{2,T}(\theta_0;\pi) \quad (5.107) \end{aligned}$$

where

$$\begin{aligned} N_{i,T}(\pi) = I_q - \hat{S}_{i,T}^{-1/2} G_{i,T}(\hat{\theta}_{i,T}(\pi),\theta_0,\lambda_T;\pi)[G_{i,T}(\hat{\theta}_{i,T}(\pi);\pi)'\hat{S}_{i,T}(\pi)^{-1} \\ \times G_{i,T}(\hat{\theta}_{i,T}(\pi),\theta_0,\lambda_T;\pi)]^{-1} G_{i,T}(\hat{\theta}_{i,T}(\pi);\pi)'\hat{S}_{i,T}(\pi)^{-1/2'} \end{aligned}$$

for $i = 1, 2$. As with $\bar{M}_{i,T}$, we must assume this matrix converges uniformly in π.

Assumption 5.11 Uniform Convergence of $N_{i,T}(\pi)$
$sup_{\pi\in\Pi}\|N_{i,T}(\pi)\hat{S}_{i,T}(\pi)^{-1/2} - N_0 S^{-1/2}\| \xrightarrow{p} 0$ where $N_0 = [I_q - P(\theta_0)]$.

To illustrate the argument from here on, it is most convenient to focus on only $O_{1,T}(\pi)$, and then to state the corresponding result for $O_{2,T}(\pi)$ afterwards. Assumptions 5.8 and 5.11 together with (5.106) imply that

$$O_{1,T}(\pi) \Rightarrow \|\frac{1}{\pi^{1/2}}[I_q - P(\theta_0)]B_q(\pi)\|^2 \quad (5.108)$$

$$= \frac{1}{\pi} B_q(\pi)'[I_q - P(\theta_0)]B_q(\pi) \quad (5.109)$$

Now, using (5.101)and $H'H = I_q$, we have

$$\begin{aligned} I_q - P(\theta_0) &= H'H - H'\Lambda H = H'[I_q - \Lambda]H \\ &= \begin{bmatrix} 0 & 0 \\ 0 & H_2' H_2 \end{bmatrix} \end{aligned}$$

This result can be combined with (5.108)–(5.109) to deduce that

$$O_{1,T}(\pi) \Rightarrow \frac{1}{\pi} B_{q-p}(\pi)' H_2' H_2 B_{q-p}(\pi)$$

where $B_{q-p}(\pi)$ is the vector consisting of the last $q-p$ elements of $B_q(\pi)$. Since H_2 is an orthogonal matrix, we can use the same reasoning as above to deduce that $H_2 B_{q-p}(\pi) \Rightarrow B_{q-p}(\pi)$ and hence that

$$O_{1,T}(\pi) \Rightarrow \frac{1}{\pi} B_{q-p}(\pi)' B_{q-p}(\pi)$$

Similar reasoning yields

$$O_{2,T}(\pi) \Rightarrow \frac{1}{1-\pi}[B_{q-p}(1) - B_{q-p}(\pi)]'[B_{q-p}(1) - B_{q-p}(\pi)]$$

So finally we obtain the following result for $O_T(\pi)$.

Theorem 5.10 Limiting Distribution of $O_T(\pi)$
If Assumptions 3.1–3.5, 3.8–3.9, 5.6–5.9, 5.11 hold then: $O_T(.) \Rightarrow O(.)$ where $O(.)$ is a continuous time process associating each date $\pi \in [0,1]$ with the scalar $O(\pi) = \frac{1}{\pi} B_{q-p}(\pi)' B_{q-p}(\pi) + \frac{1}{1-\pi}[B_{q-p}(1) - B_{q-p}(\pi)]'[B_{q-p}(1) - B_{q-p}(\pi)]$.

Theorems 5.9 and 5.10 give the limiting behaviour of $W_T(\pi)$ and $O_T(\pi)$ for $\pi \in \Pi$. The limiting distribution of the test statistics then follows directly from the Continuous Mapping Theorem. This theorem states that if $Z_T(.) \Rightarrow Z(.)$ and $h(.)$ is a continuous functional, then $h(Z_T(.)) \Rightarrow h(Z(.))$.[56] Since $Sup-$, $Av-$ and $Exp-$ versions of the statistics involve continuous functionals of $\{W_T(\pi)\}$ or $\{O_T(\pi)\}$, we can use the Continuous Mapping theorem to deduce the following corollary to Theorems 5.9 and 5.10.

Corollary 5.1 Limiting Distributions of Structural Stability Tests for the Unknown Break Point Case
If Assumptions 3.1–3.5, 3.8–3.9, 5.6–5.11 hold then: (i) $SupW_T \Rightarrow Sup_{\pi\in\Pi} W(\pi)$; (ii) $AvW_T \Rightarrow \int_\Pi W(\pi) dJ(\pi)$; (iii) $ExpW_T \Rightarrow log\{\int_\Pi exp[0.5W(\pi)] dJ(\pi)$; (iv) $SupO_T \Rightarrow Sup_{\pi\in\Pi} O(\pi)$; (v) $AvO_T \Rightarrow \int_\Pi O(\pi) dJ(\pi)$;
(vi), $ExpO_T \Rightarrow log\{\int_\Pi exp[0.5O(\pi)] dJ(\pi)$.

These results are presented in the following places: (i) Andrews (1993) [Theorem 3]; (ii)–(iii) Sowell (1996) [Theorem 3];[57] (iv)–(vi) Hall and Sen (1999) [Theorem 3.1]. It is the critical points from these distributions with $J(\pi)$ equal to the uniform distribution on Π which are reproduced in Tables 5.5 and 5.6.

One final comment is in order. It can be recalled from Theorem 3.5 that the parameter estimator (identifying restrictions) and the estimated sample moment (overidentifying restrictions) are asymptotically independent if the model is correctly specified. This independence has already manifested itself in various other inference procedures discussed earlier in this chapter, and it is also present here. The sequence of statistics $\{W_T(\pi)\}$ are functions of the first p elements of $B_q(.)$, and the $\{O_T(\pi)\}$ are functions of the last $q-p$ elements. Since, by definition, the elements of a Brownian motion are mutually independent, it follows that the tests of parameter variation are asymptotically independent of the tests based on the overidentifying restrictions under $H_0^{SS}(\Pi)$.

[56] For example, see Hamilton (1994) [p.482] and the discussion therein.
[57] Also see Andrews and Ploberger (1994).

5.4.3 Other Types of Structural Instability

As mentioned in the preamble to this section, the single break point case has received by far the most attention within the GMM literature on structural stability. However, other types have also been considered and we now provide a brief review of these alternatives.

An obvious extension of the single break point case is to allow for the prescence of multiple break points. To date this approach has not been developed in the context of GMM estimators. However, Bai and Perron (1998) have developed methods in the context of linear regression models. One aspect of their results is particularly interesting. They show that if it is assumed that there is a single break point then the estimated fraction $\hat{\pi}$ is consistent for the fraction associated with one of the multiple break points. This enables them to propose an iterative procedure in which the researcher gradually increases the number of break points until the structural stability tests are no longer significant. To date, it is unknown whether this type of sequential estimation procedure works in the more general GMM framework.

Hansen (1990) considers tests for $H_0^{SS}([0,1])$ against the alternative $E[f(v_t,\theta_t)] = 0$ and

$$\theta_t \;=\; \theta_{t-1} \;+\; \eta_t$$

where $\eta_t \sim i.i.d.(0,\tau^2 H_t)$. Notice that if $\tau^2 = 0$ then this model reduces to the null hypothesis. Interestingly, Hansen (1990) shows that the LM statistic against this alternative is well approximated by AvW_T and so this statistic is likely to have good power properties against this alternative as well.[58]

More generally, Sowell (1996) provides a framework for the construction of optimal tests for parameter variation based on GMM estimators. His results provide a generic approach which can be specialized to the form of instability of interest.

Finally, it should be noted that all these procedures rely on asymptotically large samples and so are unlikely to have good power properties against instability at the very beginning or end of the sample. Dufour, Ghysels, and Hall (1994) propose a Generalized Predictive test which can be applied in this situation. The null and alternative hypothesis are the same as the Predictive test except this time only T_1 need be asymptotically large and T_2 may be as small as one observation. The statistic is based on $\{f(v_t,\hat{\theta}_{1,T}(\pi)),\, t \in T_2\}$ and not the sub-sample average. Since the focus is now the individual observations, it is not possible to use a conventional asymptotic analysis to deduce the distribution. One solution is to make a distributional assumption, but this is unattractive in most GMM settings. Therefore Dufour, Ghysels, and Hall (1994) consider various distribution free methods of approximating or bounding the p-value of their statistics.

[58] Hansen (1990) analysis is motivated by earlier work due to Nyblom (1989) in the context of Maximum Likelihood estimators.

5.5 Other Hypothesis Tests

The foregoing tests are by far the most commonly used in the types of application listed in Table 1.1. However, certain other tests have been proposed and in this section we provide a brief review of these methods. The discussion covers non-nested hypothesis tests (Section 5.5.1), Hausman tests (Section 5.5.2) and conditional moment tests (Section 5.5.3).

5.5.1 Non-Nested Hypothesis Tests

So far, we have concentrated on methods for testing hypotheses about population moment conditions or parameters within a particular model. However, in many cases more than one model has been advanced to explain a particular economic phenomenon and so it may become necessary to choose between them. Sometimes, one model is nested within the other in the sense that it can be obtained by imposing certain parameter restrictions. In this case the choice between them amounts to testing whether the data support the restrictions in question using the methods described in Section 5.3. Other times, one model is not a special case of the other and so they are said to be *non-nested*. There have been two main approaches to developing tests between non-nested models. One is based on creating a more general model which nests both candidate models as a special case; the other examines whether one model is capable of explaining the results in the other. Most of this literature has focused on regression models or models estimated by maximum likelihood. While these situations technically fall within the GMM framework, they do not possess its distinctive features and so are not covered here.[59] Instead, we focus on methods for discriminating between two non-nested Euler equation models. These models involve partially specified systems and so involve aspects unique to the GMM in its most general form.

We consider the case where there are two competing models denoted $M1$ and $M2$. If $M1$ is true then the parameter vector θ_1 and the data satisfy the Euler equation

$$E_1[u_1(v_t, \theta_1)|\Omega_{t-1}] = 0 \tag{5.110}$$

where Ω_{t-1} is the information available at time $t-1$ and $E_1[.]$ denotes expectations under the assumption that $M1$ is correct. For our purposes, it is sufficient to assume the Euler equation residual $u_1(v_t, \theta_1)$ is a scalar. From (5.110) it follows that the residual is orthogonal to any $(q_1 \times 1)$ vector $z_{1,t} \in \Omega_{t-1}$, and this yields the population moment condition

$$E_1[z_{1,t}u_1(v_t, \theta_1)] = 0 \tag{5.111}$$

Using analogous definitions, $M2$ leads to the $(q_2 \times 1)$ population moment condition

$$E_2[z_{2,t}u_2(v_t, \theta_2)] = 0 \tag{5.112}$$

[59] These techniques are well described in the recent comprehensive review by Gourieroux and Monfort (1994).

where again the Euler equation residual is taken to be a scalar. It is assumed that the two models are globally non-nested in the sense that one model is not a special case of the other.[60] Since both models can be subjected to the tests in Sections 5.1–5.4, there can only be a need to discriminate between them if both models pass all these diagnostics; so we assume this to be the case.

As mentioned above there are two main strategies to developing non-nested hypothesis tests and each has been applied within the context of Euler equation models. Singleton (1985) proposes nesting the Euler equations of $M1$ and $M2$ within the Euler equation of a more general model. Ghysels and Hall (1990*b*) propose tests of whether one model can explain the results in another. We now describe these in turn.

Singleton's (1985) analysis begins with the observation that if $M1$ is false and its overidentifying restrictions test is insignificant then it must be because the test has poor power properties when $M2$ is true. Therefore, he proposes choosing the linear combination of the overidentifying restrictions which has the most power in the direction of $M2$. The problem is how to characterize this direction. Singleton (1985) solves this issue by introducing a more general Euler condition which is the following convex combination of those from $M1$ and $M2$,

$$E_G[e_t(\theta_1, \theta_2, \omega)|\Omega_{t-1}] = 0 \tag{5.113}$$

where

$$e_t(\theta_1, \theta_2, \omega) = \omega u_1(v_t, \theta_1) \; + \; (1 - \omega)\, u_2(v_t, \theta_2)$$

where $0 \leq \omega \leq 1$ and $E_G[.]$ taken with respect to the true distribution of the data under this more general model. Notice that $\omega = 1$ implies $M1$ is correct, and $\omega = 0$ implies $M2$ is correct. The other values of ω imply a continuum of residual processes which lie between those implied by $M1$ and $M2$ in some sense. If ω is replaced by a suitably defined sequence ω_T which converges to one from below at rate $T^{1/2}$ and $z_{1,t} = z_{2,t} = z_t$, then

$$E_G[z_t e_t(\theta_1, \theta_2, \omega)] = 0$$

defines a sequence of local alternatives to (5.111) in the direction of (5.112). Singleton (1985) shows that the linear combination of the overidentifying restrictions in $M1$ which maximizes power against this local alternative is the transpose of

$$A_T = S_{1,T}^{-1} \left(g_{1,T}(\hat{\theta}_{1,T}) \; - \; g_{2,T}(\hat{\theta}_{2,T}) \right)$$

where $g_{i,T}(\hat{\theta}_{i,T}) = T^{-1} \sum_{t=1}^{T} z_t u_i(v_t, \hat{\theta}_{i,T})$, $\hat{S}_{1,T}$ is a consistent estimator of $\lim_{T\to\infty} Var[T^{1/2} g_{1,T}(\theta_1)]$ and $\hat{\theta}_{i,T}$ is the GMM estimator of θ_i. This leads to the test statistic

$$NN_T(1,2) = T\, g_{1,T}(\hat{\theta}_{1,T})' A_T \left(A_T' \Sigma_{1,T} A_T \right)^{-1} A_T' g_{1,T}(\hat{\theta}_{1,T})$$

[60] See Pesaran (1987) for a formal definition of nested, partially non-nested and globally non-nested models. The distinction between the last two can be important but need not concern us here.

where $\Sigma_{1,T} = \hat{S}_{1,T} - \hat{G}_{1,T}(\hat{G}'_{1,T}\hat{S}^{-1}_{1,T}\hat{G}_{1,T})^{-1}\hat{G}'_{1,T}$ and $\hat{G}_{1,T} = \partial g_{1,T}(\hat{\theta}_{1,T})/\partial\theta'$. Singleton (1985) shows that if $M1$ is correct then $NN_T(1,2)$ converges to a χ^2_1 distribution. The roles of $M1$ and $M2$ can be reversed to produce the analogous statistic $NN_T(2,1)$ which would be asymptotically χ^2_1 if $M2$ is correct. In fact, the test should be performed both ways and so there are four possible outcomes: $NN_T(1,2)$ is significant but $NN_T(2,1)$ is not and so $M2$ is chosen; $NN_T(2,1)$ is significant but $NN_T(1,2)$ is not and so $M1$ is chosen; both $NN_T(1,2)$ and $NN_T(2,1)$ are significant and so both models can be rejected; both $NN_T(1,2)$ and $NN_T(2,1)$ are insignificant and so it is not possible to choose between them in this way.

This approach is relatively simple to implement because it does not require any additional assumptions or computations beyond those already involved for the estimation of $M1$ and $M2$. Its weakness is that the convex combination of the Euler equations from $M1$ and $M2$ may not be the Euler equation of a well defined economic model.[61] In such cases, it is unclear how a significant statistic should be interpreted. The only way to avoid this problem is to consider sequences of local alternatives to the data generation process implied by $M1$ which are in the direction of the data generation process implied by $M2$. However, this involves making the type of distributional assumption which the use of GMM was designed to avoid.

Ghysels and Hall (1990*b*) propose an alternative approach to testing based on whether one model can explain the results in the other.[62] More specifically, the data are said to support $M1$ if

$$T^{-1}\sum_{t=1}^{T} z_{2,t}u_2(v_t,\hat{\theta}_{2,T}) \;-\; E_1[z_{2,t}u_2(v_t,\hat{\theta}_{2,T})] \tag{5.114}$$

is zero allowing for sampling error. To implement the test it is necessary to know or be able to estimate the expectation term in (5.114). Unfortunately, this typically involves specifying the conditional distribution of v_t and so is unattractive for the reason mentioned above.[63] Ghysels and Hall (1990*b*) develop a test based on approximating the expectation using quadrature based methods, but we omit the details here.

Both these statistics are clearly focusing on the overidentifying restrictions alone. It is possible to extend Ghysels and Hall's (1990*b*) approach to tests of whether $M1$ can explain the identifying restrictions in $M2$. Such a test would focus on whether the solution to the identifying restrictions in $M2$ is equal to the value predicted by $M1$. In other words, it would examine

$$\hat{\theta}_{2,T} \;-\; E_1[\hat{\theta}_{2,T}]$$

[61] For example, Ghysels and Hall (1990*b*) show that a model constructed by taking a convex combination of the data generating processes for v_t implied by $M1$ and $M2$ does not typically possess an Euler equation of the form in (5.113).

[62] This general approach is often refered to as the encompassing test principle; see Mizon and Richard (1986).

[63] Furthermore Ghysels and Hall (1990*b*) show that a misspecification of this distribution can cause their statistic to be significant.

However, it would suffer from the same drawbacks as mentioned above and so we do not pursue such a test here.

Neither of these approaches is really satisfactory. Singleton's (1985) test is only appropriate in the limited setting where (5.113) is the Euler condition of a meaningful model. Ghysels and Hall's (1990*b*) test is always appropriate but requires additional assumptions about the distribution, and once these are made, it is more efficient to use Maximum Likelihood estimation.[64] This contrasts with the more successful treatments of the hypotheses in Sections 5.1–5.4. In these earlier cases, the partial specification caused no problems, but it clearly does so for non-nested hypotheses. In one sense, these results are more important because they illustrate the potential limits to inference based on a partially specified model.

5.5.2 Hausman Tests

Hausman (1978) proposes testing a hypothesis on the basis of a comparsion of two estimators of the parameter vector. One estimator must be consistent under the null hypothesis but inconsistent under the alternative. The other must be consistent under both null and the alternative. The simplest illustration is one of the examples used by Hausman, and which is also the most common application of this approach to testing. Suppose we have a linear regression model and are suspicious that one of the regressors, $x_{i,t}$ say, is endogenous. The null hypothesis that $x_{i,t}$ is exogenous can be tested via a Hausman test which compares the OLS estimator with an IV estimator. The former is consistent only if x_{it} is exogenous; the latter is consistent regardless. Clearly the difference between them converges to zero under the null, but some non-zero value under the alternative.[65]

It is readily recognized that this basic principle can be applied in a wide variety of settings. It is often applied in the context of Maximum Likelihood estimation to test if the specification is correct. To present this version of the statistic, let $\hat{\theta}_T$ denote the MLE and $\tilde{\theta}_T$ be a GMM estimator of θ_0 based on some population moment condition $E[f(v_t, \theta_0)] = 0$. The Hausman test statistic is then given by

$$H_T \;=\; T\left(\hat{\theta}_T \;-\; \tilde{\theta}_T\right)\left(\tilde{V}_T \;-\; \hat{V}_T\right)^{-1}\left(\hat{\theta}_T \;-\; \tilde{\theta}_T\right)$$

where $\hat{V}_T$ and $\tilde{V}_T$ are consistent estimators of the asymptotic covariance of $\hat{\theta}_T$ and $\tilde{\theta}_T$ respectively. Under the joint null hypotheses that the Maximum Likelihood estimation is based on the correct model and $E[f(v_t, \theta_0)] = 0$ is valid

[64] Although, full information maximum likelihood may be more computationally burdensome; see Ghysels and Hall (1990*b*).

[65] This statistic is often refered to as the *Wu–Hausman* test because – to quote Nakamura and Nakamura (1998) – "it was Hausman [(1978)] who presented it in the form that led to its widespread use but Wu [(1973)] who presented it first." [p.220]. See Nakamura and Nakamura (1998) for further discussion of the literature on these types of endogeneity test.

then Hausman (1978) shows that H_T converges to a χ^2_p distribution. The alternative hypothesis is that the model is not correctly specified but nevertheless $E[f(v_t, \theta_0)] = 0$.

Newey (1985*a*) extends this test principle to models estimated by GMM. Newey (1985*a*) derives a Hausman statistic based on the difference between GMM estimators obtained from two sets of moment conditions which may contain elements in common. Interestingly, he shows that if one estimator is obtained with the optimal weighting matrix then the asymptotic variance of the difference of the estimators has the same difference structure as in the Maximum Likelihood case. However, this asymptotic variance may also be singular. In principle, this matter is easily fixed by using a generalized inverse in the construction of the quadratic form, and then comparing the statistic to the critical point from a χ^2 distribution with degrees of freedom equal to the rank of $\tilde{V}_T - \hat{V}_T$. However, in practice, this adjustment is not so straightforward for two reasons. First, $rank\{p\lim_{T\to\infty}(\tilde{V}_T - \hat{V}_T)\}$ may be difficult to deduce a priori. Secondly, and unlike inverses, generalized inverses are not necessarily continuous functions of the elements, and so additional conditions are needed to ensure that the generalized inverse of $\tilde{V}_T - \hat{V}_T$ converges in probability to the generalized inverse of $p\lim_{T\to\infty} \tilde{V}_T - \hat{V}_T$; see Andrews (1987). Both these problems may explain the infrequent use of this test in the types of applications listed in Table 1.1.

5.5.3 Conditional Moment Tests

All the statistics presented in Section 5.1, 5.2 and 5.4 test hypotheses about the the population moment conditions upon which estimation is based. This mirrors the majority of empirical applications mentioned in the introduction. In these types of application, the model is only partially specified and so it is desirable to base estimation on as much relevant information as possible.[66] Therefore all available moment conditions tend to be used in estimation.[67] However, if the distribution of the data is known then the most asymptotically efficient estimates are obtained by using Maximum Likelihood. As shown in Section 3.7.1, maximum likelihood amounts to GMM estimation based on the score function of the data. So, in this case there is no advantage to including any other moment conditions implied by the model. These other moment conditions can, however, be used to test whether the specification of the model is correct. This generic approach yields what have become known as *conditional moment tests*.

Newey (1985*b*) and Tauchen (1985*a*) independently introduce a general framework for conditional moment testing based on Maximum Likelihood estimators. To illustrate this framework, suppose that the conditional probability density

[66] This statement is formally justified in Chapter 6.

[67] The choice of moment conditions may be limited by other factors such as data availability or computational constraints.

of v_t given $\{v_{t-1}, v_{t-2}, \ldots v_1\}$ is $p_t(v_t; \theta_0)$, and so the score function is

$$E[L_t(\theta_0)] = 0$$

where $L_t(\theta_0) = \partial log(p_t(v_t; \theta_0))/\partial\theta$. As mentioned above this is the moment condition upon which estimation is based. Now assume that if this model is correctly specfied then the data also satisfy the $(q \times 1)$ population moment condition $E[g(v_t, \theta_0)] = 0$. Therefore one way to assess the validity of the model is to test

$$H_0 : \ E[g(v_t, \theta_0)] = 0$$

against the alternative

$$H_A : \ E[g(v_t, \theta_0)] \neq 0$$

This hypothesis can be tested using the statistic

$$CM_T = T^{-1} \sum_{t=1}^{T} g_t(\hat{\theta}_T)' Q_T^{-1} \sum_{t=1}^{T} g_t(\hat{\theta}_T) \tag{5.115}$$

where $\hat{\theta}_T$ is the maximum likelihood estimator, Q_T is a consistent estimator of $\lim_{T\to\infty} Var[T^{-1/2} \sum_{t=1}^{T} c_t(\theta_0)]$, and

$$c_t(\theta_0) \ = \ g_t(\theta_0) \ - \ E[\partial g_t(\theta_0)/\partial\theta'] \{E[\partial L_t(\theta_0)/\partial\theta']\}^{-1} L_t(\theta_0)$$

Under H_0, CM_T converges to a χ^2_q distribution. The statistic has a similar structure to the overidentifying restrictions test but there is an important difference. Since $E[g(v_t, \theta_0)] = 0$ is not used in estimation, the statistic has power against any violations of H_0; see Newey (1985*b*). In spite of this, some caution is needed in the interpretation of the results. While a rejection of H_0 implies the model is misspecified, a failure to reject only implies that the assumed distribution exhibits this particular characteristic of the true distribution.

The choice of $g(.)$ varies from model to model. For example, in the normal linear regression model, $g(.)$ often involves the third and fourth moments of the error process; see Bowman and Shenton (1975). White (1982) suggests that one generally applicable choice is to base $g(.)$ on the information matrix identity,

$$E[L_t(\theta_0) L_t(\theta_0)'] \ = \ - E[\partial L_t(\theta_0)/\partial\theta']$$

because if the the null hypothesis cannot be rejected then conventional formulae for Wald, LR and LM statistics are valid. Consequently, this approach has been explored in many settings; for example, see Chesher (1984) and Hall (1987*a*). Various other examples are provided by Newey (1985*b*) and Tauchen (1985*a*).

5.6 Summary

This chapter has presented a number of inference procedures that can be used to learn about the underlying model. The discussion focused on four main types of hypothesis test within the GMM framework: the overidentifying restrictions test; an overidentifying restrictions based test for the validity of a subset of the population moment condition; Wald, D and LM tests for testing whether

the parameter vector satisfies a set of nonlinear restrictions; structural stability tests based on both the identifying restrictions and also the overidentifying restrictions. The limiting behaviour of these test statistics is derived under the appropriate null hypothesis. The power properties are analyzed using either local or non-local alternatives, and these two approaches are contrasted. A brief review is also provided of other less common hypothesis tests within the GMM framework such as non-nested tests, Hausman tests and conditional moment tests.

In the preamble to this chapter, it is observed that three types of inference questions arise in practice. These are – Is the model correctly specified? Does the model satisfy restrictions implied by economic/statistical theory? Which of two competing models is correct? We now briefly summarize what has been learnt about how these questions can be addressed.

- *Is the model correctly specified?* Misspecification can take two basic forms. First, the model can be misspecified in the sense described in Chapter 4 that is, $E[f(v_t, \theta)]$ is the same for all t but there is no value of θ that makes this expectation zero. Secondly, the model can be structurally unstable so that $E[f(v_t, \theta_0)] = 0$ for some part of the sample but not for all of it. The overidentifying restrictions test is designed to test against the first of these types of misspecification, and is consistent against this type of alternative. The overidentifying restrictions test has power against certain types of misspecification due to structural instability but is not consistent against all forms of structural instability. This type of misspecification can be detected using specially designed structural stability tests. The latter tests can be based on either the identifying restrictions, in which case they amount to tests for parameter variation, or the overidentifying restrictions.

- *Does the model satisfy restrictions implied by economic/statistical theory?* In many cases of interest, the restrictions implied by economic theory take the form of a set nonlinear restrictions on the parameter vector. Such restrictions can be tested using Wald, D or LM tests.

- *Which of two competing models is correct?* Assuming that both models appear correctly specified on the basis of diagnostics decribed above, the answer then depends on the relationship between the two models. If they are nested, in the sense that one is obtained by imposing a set of parameter restrictions on the other, then the choice between them can be based on the Wald, D or LM statistics for testing the validity of the restrictions in question. However, if the models are non-nested then this becomes a far harder question to address within the the types of model in Table 1.1 without the specification of the probability distribution of the data.

All the inference procedures described above are based on asymptotic theory. However, as noted at the outset, asymptotic theory is only used as an

approximation to large sample behaviour. It is therefore important to investigate how good this asymptotic approximation is to finite sample behaviour in the kinds of circumstance encountered in practice. This topic is addressed in the next chapter.

6

Asymptotic Theory and Finite Sample Behaviour

So far, all the analysis has rested on asymptotic theory. This approach has been taken for two good reasons. First, to date, it has proved impossible to develop a finite sample distribution theory for GMM estimators in nonlinear dynamic models. Secondly, as we have seen, asymptotic analysis delivers a very powerful inference framework. However, there is inevitably a price to be paid. All the asymptotic results are only strictly valid in the limit as $T \to \infty$, and so represent an approximation to finite sample behaviour. The question to which we now turn is: how good is this approximation? Intuition suggests that the answer varies from case to case, and so one goal of this chapter is to identify what aspects of the specification determine the quality of the asymptotic approximation.

Since finite sample distribution theory is intractable for nonlinear dynamic models, this question has been addressed in this context via computer based simulation studies calibrated to match models of particular interest. These studies form the main focus of this chapter and are reviewed in Section 6.3. However, we precede our review of these simulation studies with a discussion of two relevant aspects of the theoretical literature.

First, since many of the simulation studies examine the consequences of increasing the number of moment conditions, it is useful to consider what can be learnt about these consequences from asymptotic analysis. Section 6.1.1 considers the case in which there is a finite increase in the degree of overidentification. It emerges from this analysis that such an increase can never have a detrimental effect on the asymptotic distribution of the estimator. However, there are some circumstances in which there is no effect, and so the additional moment conditions are said to be *redundant*. This scenario turns out to be pertinent to our discussion of the aforementioned simulation studies, and so a formal definition of redundancy is provided in Section 6.1.2. Given the potential asymptotic benefits from increasing the degree of overidentification, it is natural to consider an

estimation strategy in which this degree is allowed to increase with the sample size. In Section 6.1.3, it is shown that there are potential gains from such a strategy but these can only be reaped if the degree of overidentification does not increase too quickly with the sample size.

The second relevant aspect is the theoretical literature on the finite sample behaviour of the GMM estimator in static models. While it is true that finite sample distribution theory has proved intractable for nonlinear dynamic models to date, this is not the case for the IV estimator in the linear regression model discussed in Chapter 2. Although the exact finite sample distribution is not easily interpreted, its form does reveal the aspects of the specification upon which it depends, and these are summarized in Section 6.2.1. Further insights are gained by considering higher order approximations such as Edgeworth expansions for the distribution of the estimator or so called Nagar expansions for the finite sample bias and mean squared error of the GMM estimator. Both methods have been applied in the context of the linear simultaneous equations model, but the second has recently been employed very fruitfully to examine the bias of GMM estimators in nonlinear static models. Section 6.2.2 summarizes the main insights gained from both these analyses. Although these results only apply to static models, intuition suggests that if a factor of the specification effects the quality of the asymptotic approximation in static models then the analogous factor has a corresponding effect in dynamic models. At the same time, it would be anticipated that the presence of dynamics introduces additional complications.

As mentioned above, Section 6.3 reviews the insights gained from a number of simulations studies calibrated to the types of models underlying the applications in Table 1.1. Finally, Section 6.4 pulls together the evidence from the preceding three sections to provide an overview of what factors appear to affect the quality of the asymptotic approximation. These factors are also used to motivate the topics addressed in the following two chapters.

6.1 The Impact of the Degree of Overidentification on the Asymptotic Behaviour of the Estimator

Theorems 3.1 and 3.2 establish the consistency and asymptotic normality of $\hat{\theta}_T$. Inspection of these results reveals that they hold for any population moment condition satisfying certain regularity conditions of which the most important, for our purposes here, are the orthogonality condition in Assumption 3.3 and the identification condition in Assumption 3.4. In most cases, the underlying model implies multiple choices of $f(v_t, \theta_0)$ which satisfy these conditions. Therefore, it is important to consider how these asymptotic properties are affected by the expansion of the set of population moment conditions upon which estimation is based. We split the analysis into three parts. Section 6.1.1 considers the case in which there is a finite increase in the number of moment conditions, Section

6.1.2 introduces the concept of redundant moment conditions and Section 6.1.3 considers the case in which the number of moments increases with T.

6.1.1 Finite Increase in the Degree of Overidentification

To facilitate the analysis, it is necessary to introduce the following notation. We partition $f(.)$ into $f(v_t,\theta)' = [f_1(v_t,\theta)', f_2(v_t,\theta)']$ where $f_i(.)$ is $(q_i \times 1)$ and $q = q_1 + q_2$ is finite. Now let $\hat{\theta}_{1,T}$ be the (optimal) two step GMM estimator based on

$$E[f_1(v_t,\theta_0)] \;=\; 0 \tag{6.1}$$

It is assumed that θ_0 is identified by (6.1) and hence that $q_1 \geq p$. Finally, let $\hat{\theta}_T$ denote the (optimal) two step estimator based on

$$E[f(v_t,\theta_0)] \;=\; 0 \tag{6.2}$$

It is straightforward to invoke Theorems 3.1 and 3.2 in order to deduce that both estimators are consistent for θ_0 and

$$T^{1/2}(\hat{\theta}_{1,T} - \theta_0) \;\xrightarrow{d}\; N(0, V_1) \tag{6.3}$$

$$T^{1/2}(\hat{\theta}_T - \theta_0) \;\xrightarrow{d}\; N(0, V) \tag{6.4}$$

where $V = (G_0' S^{-1} G_0)^{-1}$, $V_1 = (G_{1,0}' S_{1,1}^{-1} G_{1,0})^{-1}$, S and G_0 are defined as before,[1] $S_{1,1}$ is the $(q_1 \times q_1)$ upper left hand block of S, and $G_{1,0}$ is the $(q_1 \times p)$ matrix comprising the first q_1 rows of G_0. Therefore the only difference between the two limiting distributions lies in their variance. The following theorem establishes the relationship between V and V_1.

Theorem 6.1 Asymptotic Efficiency and the Inclusion of Additional Population Moment Conditions
If (i) Assumptions 3.1–3.5 and 3.7–3.13 hold; (ii) rank$(G_{1,0}) = p$ (iii) $\hat{\theta}_{1,T}$ is the (optimal) two step GMM estimator based on (6.1); (iv) $\hat{\theta}_T$ be the (optimal) two step GMM estimator based on (6.2); then $V_1 - V$ is positive semi-definite and so $\hat{\theta}_T$ is asymptotically at least as efficient as $\hat{\theta}_{1,T}$.

The regularity conditions are needed to ensure that (6.3)–(6.4) hold. The proof rests purely on showing that $V_1 - V$ is positive semi-definite.
Proof:
Since V and V_1 are positive definite, $V_1 - V$ is positive semi-definite if $V^{-1} - V_1^{-1}$ is also positive semi-definite.[2] The latter difference is more convenient to work with, and is our focus here. To this end, partition G_0 and S into

$$G_0 \;=\; \begin{bmatrix} G_{1,0} \\ G_{2,0} \end{bmatrix}, \qquad S \;=\; \begin{bmatrix} S_{1,1} & S_{1,2} \\ S_{2,1} & S_{2,2} \end{bmatrix}$$

[1] See Section 3.4.2.

[2] See Dhrymes (1984) [Proposition 65, p.76]. Strictly, Dhrymes only establishes the result for the case in which the difference in positive definite, but his proof is easily amended to cover the case in which the difference is positive semi-definite.

Using the partitioned matrix inversion formula,[3] it follows that

$$S^{-1} = \begin{bmatrix} S_{1,1}^{-1}(I_{q_1} + S_{1,2}AS_{2,1}S_{1,1}^{-1}) & -S_{1,1}^{-1}S_{1,2}A \\ -AS_{2,1}S_{1,1}^{-1} & A \end{bmatrix} \tag{6.5}$$

where $A = (S_{2,2} - S_{2,1}S_{1,1}^{-1}S_{1,2})^{-1}$. The substitution of both (6.5) and the partition of G_0 into $D = V^{-1} - V_1^{-1}$ yields

$$\begin{aligned} D &= G_{1,0}'S_{1,1}^{-1}(I_{q_1} + S_{1,2}AS_{2,1}S_{1,1}^{-1})G_{1,0} - G_{2,0}'AS_{2,1}S_{1,1}^{-1}G_{1,0} \\ &\quad -G_{1,0}'S_{1,1}^{-1}S_{1,2}AG_{2,0} + G_{2,0}^{'}AG_{2,0} - G_{1,0}^{'}S_{1,1}^{-1}G_{1,0} \end{aligned}$$

Multiplying out this expression, it can be verified that $D = B'AB$ where $B = S_{2,1}S_{1,1}^{-1}G_{1,0} - G_{2,0}$. Now S is positive definite (p.d) by assumption, and so S^{-1} shares this property, which in turn implies A is p.d. Therefore, D is positive semi-definite by construction, and so we have established the desired result. ◇

This result makes intuitive sense. The elements of the population moment condition can be viewed as pieces of information about θ_0 and, from this perspective, Theorem 6.1 can be paraphrased as saying that more correct information never hurts. For the puposes of our later discussion, it is useful to examine the circumstances under which it does not help either. This is the topic of the next sub-section.

6.1.2 Redundant Moment Conditions

Breusch, Qian, Schmidt, and Wyhowski (1999) use the term redundancy to describe the situation in which the augmentation of the population moment condition has no effect on the asymptotic variance of the estimator. This idea can be expressed formally as follows.

Definition 6.1 Redundant Moment Condition
Let V denote the asymptotic variance of the GMM estimator based on $E[f_1(v_t, \theta_0)] = 0, E[f_2(v_t, \theta_0)] = 0$, and let V_1 be the corresponding variance when estimation is based on $E[f_1(v_t, \theta_0)] = 0$ alone. The population moment condition $E[f_2(v_t, \theta_0)] = 0$ is said to be redundant for θ_0 given $E[f_1(v_t, \theta_0)] = 0$ if $V_1 = V$.

Intuition suggests that $E[f_2(v_t, \theta_0)] = 0$ is redundant given $E[f_1(v_t, \theta_0)] = 0$ if it provides no information about θ_0 beyond that already in $E[f_1(v_t, \theta_0)] = 0$. To formalize this intuition, it is necessary to first characterize the part of $f_2(v_t, \theta_0)$ which cannot be explained by $f_1(v_t, \theta_0)$. To this end, it is assumed that the Central Limit Theorem can be applied to deduce that

$$\begin{bmatrix} T^{1/2}g_{1,T}(\theta_0) \\ T^{1/2}g_{2,T}(\theta_0) \end{bmatrix} \xrightarrow{d} N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} S_{1,1} & S_{1,2} \\ S_{2,1} & S_{2,2} \end{bmatrix} \right) \tag{6.6}$$

[3] See Magnus and Neudecker (1991) [p.11].

where $T^{1/2}g_{i,T}(\theta_0) = T^{-1/2}\sum_{t=1}^{T} f_i(v_t, \theta_0)$ for $i = 1, 2$. It follows from (6.6) that the conditional distribution of $T^{1/2}g_{2,T}(\theta_0)$ given $T^{1/2}g_{1,T}(\theta_0)$ is given by

$$N\left(S_{2,1}S_{1,1}^{-1}T^{1/2}g_{1,T}(\theta_0),\ S_{2,2} - S_{2,1}S_{1,1}^{-1}S_{1,2}\right)$$

Given the form of this conditional distribution, the unexplained part of $T^{1/2}g_{2,T}(\theta_0)$ is given by

$$T^{1/2}g_{2,T}(\theta_0) \ - \ S_{2,1}S_{1,1}^{-1}T^{1/2}g_{1,T}(\theta_0) \ = \ T^{-1/2}\sum_{t=1}^{T} r(v_t, \theta_0), \text{ say,}$$

where $r(v_t, \theta_0) = f_2(v_t, \theta_0) - S_{2,1}S_{1,1}^{-1}f_1(v_t, \theta_0)$. Therefore, we now focus on the residual, $r(v_t, \theta_0)$. At this stage, it is useful to recall three aspects of our discussion of local identification in Section 3.1. First, the local information about θ_0 contained in a moment condition is captured by the expectation of its derivative with respect to θ. Secondly, the moment condition uniquely determines θ_0 if this expected derivative is full rank. Thirdly, if this expected derivative is rank deficient then the moment condition provides *some* information about θ_0 but not enough to determine it uniquely. Taken together these three points imply that the moment condition is only completely uninformative if the expected derivative is zero. Therefore, $E[f_2(v_t, \theta_0)] = 0$ provides no local information about θ_0 beyond that in $E[f_1(v_t, \theta_0)] = 0$ if and only if

$$E[\partial r(v_t, \theta_0)/\partial\theta'] \ = \ 0$$

This condition is one of three for redundancy provided by Breusch, Qian, Schmidt, and Wyhowski (1999). The other two are less intuitive but may be easier to verify in practice. For completeness, we reproduce all three here, but omit the proof.[4]

Lemma 6.1 Conditions for Redundancy
The following statements are equivalent. (A): $E[f_2(v_t, \theta_0)] = 0$ *is redundant given* $E[f_1(v_t, \theta_0)] = 0$. *(B):* $E[\partial r(v_t, \theta_0)/\partial\theta'] = 0$. *(C):* $E[\partial f_2(v_t, \theta_0)/\partial\theta'] = S_{2,1}S_{1,1}^{-1}E[\partial f_1(v_t, \theta_0)/\partial\theta']$. *(D): There exists a* $(q_1 \times p)$ *matrix* A *such that* $E[\partial f_1(v_t, \theta_0)/\partial\theta'] = S_{1,1}A$ *and* $E[\partial f_2(v_t, \theta_0)/\partial\theta'] = S_{2,1}A$.

The concept of redundancy proves useful in understanding some of the simulation results described in Section 6.3.

6.1.3 The Degree of Overidentification Increases with the Sample Size

If we follow Theorem 6.1 to its logical conclusion, then it leads us to an estimation strategy in which we include as many population moment conditions as possible. For a given sample, q must be less than T. However, as T increases,

[4] Also see Section 7.1.

Theorem 6.1 appears to suggest that it may be advantageous to allow q to increase as well – in other words, to adopt a strategy in which the number of population moment conditions is q_T and $q_T \to \infty$ with T. In spite of its intuitive appeal, this logical step must be taken with caution because Theorem 6.1 is premised on Theorem 3.2, and the latter only holds for fixed q. To date, there have been only a few studies which shed light on the asymptotic behaviour of the estimator when p is fixed but $q_T \to \infty$ with $T \to \infty$. This evidence suggests that the asymptotic theory derived in Chapter 3 may be valid if $q_T - p$ increases fairly slowly, but is unlikely to be so if $q_T - p$ increases too rapidly. It should be noted, however, that all these studies consider the issue in the context of i.i.d. data. It is left to future research to consider whether these rates of increase for q_T translate to dependent data. We now briefly summarize the main results on this issue.

Newey (1990) examines the limiting behaviour of the IV estimator in the context of a nonlinear simultaneous equations model under the assumption that the error, $u_t(\theta_0)$, is conditionally homoscedastic given z_t. This restriction is important because it implies that the optimal weighting matrix is $\hat{S}_{CIV}^{-1}$ in (2.29), and so proportional to $(T^{-1}Z'Z)^{-1}$, the choice assumed in Newey's (1990) analysis.[5] He shows that Theorem 3.2 continues to hold provided $q_T = o(T^{1/2})$. Koenker and Machado (1999) consider only linear models but allow for the possibility that $u_t(\theta_0)$ may be conditionally heteroscedastic and so S is estimated by $\hat{S}_{SU}$ in (3.40).[6] They show that $q_T \to \infty$ and $q_T = o(T^{1/3})$ are sufficient conditions for Theorem 3.2. This rate is rather slow, and so implies a more limited scope for an estimation strategy based on an expanding set of moment conditions. Interestingly, this slow rate appears to stem directly from the behaviour of $\hat{S}_{SU}$. However, this rate is sufficient and not necessary, and as such is a lower bound on the possible rate of increase for q_T.

If q_T increases faster than the rates given above then this impacts on the limiting behaviour of the estimator in some way. Morimune (1983) considers the limiting behaviour of the 2SLS (IV) estimator in the context of the linear simultaneous equation model. He shows that if q_T increases at rate $T^{1/2}$ then the estimator is consistent, $T^{1/2}(\hat{\theta}_T - \theta_0)$ has a limiting normal distribution but the mean of this distribution is not zero. He further shows that if q_T increases at rate T then the estimator is inconsistent. Bekker (1994) derives the limiting behaviour of the IV estimator in the case where the equation of interest is linear in the parameters and q_T increases with T. Using $\bar{\theta}$ to denote the probability limit of $\hat{\theta}_T$, Bekker (1994) shows that $(T-p)^{1/2}(\hat{\theta}_T - \bar{\theta})$ converges to a normal distribution with mean zero but a different variance than the one in Theorem 3.2.

[5] The main focus of Newey's (1990) study is actually the construction of optimal instruments, a topic that is considered in Chapter 7.

[6] Note that the dimension of $\hat{S}_T$ is q_T and so increases with T. This case is outside the settings reviewed in Section 3.5 for which $q_T = q$.

6.2 Finite Sample Theory for Static Models

This section describes the insights gained from two theoretical frameworks for learning about the finite sample behaviour of GMM estimators in static models. Section 6.2.1 describes the available exact finite sample results for GMM estimators. Section 6.2.2 summarizes results derived using higher order approximations based on Edgeworth and Nagar expansions.

6.2.1 Exact Results for the IV Estimator in the Linear Simultaneous Equations Models

There has been a considerable literature on the finite sample distributions of estimators in the static linear simultaneous equations model.[7] Here we focus exclusively on the results for the IV estimator described in Chapter 2.

For our purposes in Chapter 2, it suffices to specify just the equation of interest and make certain broad assumptions about the interrelationship between the variables. Here it is necessary to be more specific. Accordingly, we now assume the equation of interest is a member of the simultaneous system

$$YB \; + \; N\Gamma \; = \; U \tag{6.7}$$

in which Y is the $(T \times J)$ matrix of of observations on the J endogenous variables, N is the $(T \times K)$ matrix of observations of the K exogenous variables and U is the $(T \times J)$ matrix of errors. It is assumed that the t^{th} row of U, $U_{t,.}$ is a vector of independent random variables with zero mean and covariance matrix Σ whose typical element is $\sigma_{i,j,0}$, and which is independent of $U_{s,.}$ for all $s \neq t$. Without loss of generality, we focus attention on the the IV estimator of the parameters in the first equation of the system. To this end, we partition Y and N as follows: $Y = [y_1, Y_1]$ and $N = [N_1, N_2]$, where y_1 is the $(T \times 1)$ vector of observations on the first endogenous variable in the system, and N_i is $(T \times K_i)$ with t^{th} row $N'_{i,t}$, and $K_1 + K_2 = K$. The first equation of the system can then be written as

$$y_1 = Y_1\beta_0 + N_1\gamma_0 + u_1 \tag{6.8}$$

where $u_1 = U_{.,1}$ is the first column of U. The reduced form of the system in (6.7) is given by

$$Y \; = \; N\Pi \; + \; A$$

where $\Pi = -\Gamma B^{-1}$ and $A = UB^{-1}$. It is convenient to write this reduced form as

$$[y_1, \, Y_1] \; = \; \left[\begin{array}{cc} N_1 & N_2 \end{array} \right] \left[\begin{array}{cc} \Pi_{1,1} & \Pi_{1,2} \\ \Pi_{2,1} & \Pi_{2,2} \end{array} \right] \; + \; [a_1, \, A_1] \tag{6.9}$$

Below it is necessary to refer to the reduced form error variance, and so we let $\omega_{i,j,0}$ denote the $i - j^{th}$ element of $\Omega_0 = Var[a_t]$ where a_t is the t^{th} row of $[a_1, A_1]$.

[7] See Phillips (1983) or Bowden and Turkington (1984, pp.137–44) for a survey of these results.

If we set $X_1 = [Y_1,\ N_1]$ and $\theta^{'} = [\beta^{'},\ \gamma^{'}]$ then (6.8) can be written as

$$y_1 = X_1\theta_0 + u_1 \tag{6.10}$$

Equation (6.10) can be recognized to be of the same generic form as the model in Chapter 2 with $y = y_1$, $X = X_1$ and $p = J + K_1 - 1$. As in that earlier setting, the observations on the instruments are contained in the $T \times q$ matrix Z where

$$Z \ = \ [N_1,\ N_2C_2] \ = \ NC_z$$

where $K \geq q \geq p$, and C_z, C_2 are selection matrices. Two aspects of this instrument choice should be noted. First, the instruments taken from the set of exogenous variables that appear in the system. Secondly, the instrument matrix always includes the exogenous variables from the equation being estimated. The results discussed below are based on the assumption that N is fixed in repeated samples. In this case, it is easily verified that u_1 satisfies the "Classical assumptions" listed in Assumption 2.5, and so the "optimal" two step estimator is just the two stage least squares estimator.[8] We therefore focus on this version of the IV estimator, that is

$$\hat{\theta}_T \ = \ \left[\begin{array}{c} \hat{\beta}_T \\ \hat{\gamma}_T \end{array} \right] \ = \ \left(X_1^{'} P_z X_1 \right)^{-1} X_1^{'} P_z y_1 \tag{6.11}$$

where $P_z = Z(Z^{'}Z)^{-1}Z^{'}$.

Phillips (1980) derives the exact distribution of $\hat{\beta}_T$ in the case where $U_{t,\cdot}$ possesses a normal distribution. The resulting expression is extremely complicated and – to quote Phillips himself – "not as easy to interpret as we would like".[9] Therefore we do not present the precise details here. Instead, we abstract to more general level and use Phillips's (1980) result to examine what aspects of the specification affect this distribution. To simplify this discussion, we restrict attention to the case in which $J = 2$ and $C_z = I_k$ – therefore the system consists of only two equations and all the exogenous variables are used as instruments. It is worth noting that prior to Phillips's work, the finite sample distribution of $\hat{\beta}_T$ had been derived for certain special cases, and one of these is for $J = 2$. So, since we limit attention to this case, our discussion can take advantage of insights gained from the earlier studies by Richardson (1963), Sawa (1969) and Anderson and Sawa (1973, 1979).[10]

Using the aforementioned results, it can be shown that the finite sample distribution of $\hat{\beta}_T$ depends on the following aspects of the specification:

- β_0, the true parameter value.
- $q - p$, the degree of overidentification.

[8] See Section 2.4.

[9] See Phillips (1980) [p.870].

[10] Notice that Richardson (1963) and Phillips (1980) employ normalizations so that variance of the reduced form error and instrument cross product matrix are both identity matrices. These restrictions facilitate the analysis but also must be borne in mind when considering how the properties of the instruments effect the distribution.

- μ^2, the concentration parameter,

$$\mu^2 \;=\; \omega_{2,2,0}^{-1}\Pi_{2,2}^{'}[N_2^{'}N_2 - N_2^{'}N_1(N_1^{'}N_1)^{-1}N_1^{'}N_2]\Pi_{2,2} \tag{6.12}$$

- Σ_0, the covariance matrix of the errors.

It is not surprising that the distribution depends on both β_0 and Σ_0, but for our purposes here, there is little to be learnt from exploring the nature of their impact on the distribution. It is the roles of $q-p$ and μ^2 which provide the most useful insights. Most of these insights have been revealed by numerical calculations, but there is one interesting facet which can be deduced directly from the form of the distribution. Phillips's (1980) analytical result reveals that the finite sample moments of $\hat{\beta}_T$ only exist up to the order $q-p$.[11] Anderson and Sawa (1979) evaluate the distribution of $\hat{\beta}_T$ numerically for a wide variety of parameter settings. In general terms, their results suggest the following conclusions *ceteris paribus*. As $q-p$ increases the finite sample distribution tends to be negatively skewed and to exhibit less variation than would be predicted by the asymptotic distribution. In other words, as $q-p$ increases the distribution becomes increasingly concentrated about some point away from the true value. In contrast, increases in μ^2 tend to offset both these effects, although the distribution still exhibits less variation than would be anticipated from the asymptotic approximation.

Since all our statistical theory is based on $T \to \infty$, two questions naturally arise: – at what sample size does asymptotic theory provide a good approximation? – and on what aspects of the specification does this depend? It can be recalled from Theorem 3.1 that $\hat{\beta}_T$ is consistent and so as the sample size increases the distribution of $\hat{\beta}_T$ must collapse onto β_0. Whereas Theorem 3.2 states that $T^{1/2}(\hat{\beta}_T - \beta_0)$ converges in distribution to a normal random vector. In fact, both behaviours only occur if $\mu^2 \to \infty$.[12] Therefore this is the route through which T affects the distribution, and this relationship can be made explicit by rewriting (6.12) as

$$\mu^2 \;=\; T\omega_{2,2,0}^{-1}\Pi_{2,2}^{\prime}(M_{2,2} - M_{2,1}M_{1,1}^{-1}M_{1,2})\Pi_{2,2} \;=\; T\tilde{\mu}^2, \text{ say} \tag{6.13}$$

where $M_{i,j} = T^{-1}N_i^{'}N_j$. Equation (6.13) reveals an interesting feature of the passage from finite sample to asymptotic behaviour: it is not T *per se* that matters, but $T\tilde{\mu}^2$. Therefore, $\tilde{\mu}^2$ effects the sample size at which asymptotic theory manifests itself. In particular, notice that if $\tilde{\mu}^2$ is very close to zero then the passage from finite sample to asymptotic behaviour is likely to be slow. Since all our inference rests on asymptotic theory, it is important to gain a better understanding of what $\tilde{\mu}^2 \approx 0$ implies about the specification. This is most readily achieved by considering the extreme case in which $\tilde{\mu}^2 = 0$. Clearly

[11] Such a relationship had previously been conjectured by Basmann (1961, 1963).

[12] See Anderson and Sawa (1979) [p.174] or Phillips (1983) [footnote 10, p.470]. It is this behaviour which gives μ^2 its name : as $\mu^2 \to \infty$ the distribution of $\hat{\beta}_T$ becomes increasingly *concentrated* around β_0 and collapses onto this point in the limit.

this condition holds if either $\Pi_{2,2} = 0$ or $M_{2,2} - M_{2,1}M_{1,1}^{-1}M_{1,2} = 0$. However, a more instructive answer can be obtained by relating these two conditions back to the condition for identification in this model. It can be recalled from Section 2.1 that the condition for identification is $rank\{E[z_t x_t^{'}]\} = p$. For our model here,

$$\begin{aligned} E[z_t x_t^{'}] &= E\left[\begin{pmatrix} N_{1,t} \\ N_{2,t} \end{pmatrix} (y_{2,t},\, N_{1,t}^{'}) \right] \\ &= E\begin{bmatrix} N_{1,t}y_{2,t} & N_{1,t}N_{1,t}^{'} \\ N_{2,t}y_{2,t} & N_{2,t}N_{1,t}^{'} \end{bmatrix} \end{aligned}$$

where – since $J = 2$ – we set $Y_1 = y_2$ and y_2 is the $(T \times 1)$ vector with t^{th} element $y_{2,t}$. Using this substititution in (6.9) and the properties of a_t, it follows that

$$E[z_t x_t^{'}] = E\begin{bmatrix} N_{1,t}N_{1,t}^{'}\Pi_{1,2} + N_{1,t}N_{2,t}^{'}\Pi_{2,2}, & N_{1,t}N_{1,t}^{'} \\ N_{2,t}N_{1,t}^{'}\Pi_{1,2} + N_{2,t}N_{2,t}^{'}\Pi_{2,2}, & N_{2,t}N_{1,t}^{'} \end{bmatrix}$$

Inspection reveals that this matrix has rank less than p if either $\Pi_{2,2} = 0$ or $N_{2,t}$ is an exact linear function of $N_{1,t}$. Notice that the second condition implies that $N_2 = N_1 H$ and so

$$\begin{aligned} R_{2,1} &= M_{2,2} - M_{2,1}M_{1,1}^{-1}M_{1,2} = T^{-1}[H'N_1'N_1H - H'N_1'N_1(N_1^{'}N_1)^{-1}N_1^{'}N_1H] \\ &= 0 \end{aligned}$$

Therefore, the conditions for θ_0 to be unidentified are exactly the same as those for $\tilde{\mu}^2 = 0$.[13] The re-emergence of the condition for identification here is not surprising, because it is fundamental to our ability to estimate θ_0 from the population moment condition. However, this analysis also adds a new facet to our understanding of the relationship between the two. If either $\Pi_{2,2}$ or $R_{2,1}$ is very close to zero then $E[z_t u_t(\theta)]$ may be very close to zero for $\theta \neq \theta_0$. In this case θ_0 is said to be "weakly identified" by $E[z_t u_t(\theta_0)] = 0$. Under these conditions, $\tilde{\mu}^2$ is also likely to be small and so the estimator converges slowly toward the behaviour predicted by asymptotic theory.[14] Anderson and Sawa (1979) report evidence that this convergence is further slowed down by increases in $q - p$. They conclude that "the desirable asymptotic properties of the 2SLS estimator are not necessarily expected to be relevant to the cases that appear in practice, that is, the sample size being at least 50 but less than 100 and the number of excluded exogenous variables" – $q-p$ in our notation – "being more than 10 but less than 50"(Anderson and Sawa (1979) [p.175]). It is important to remember their time of writing when interpreting what sample sizes are "relevant" in practice. Nevertheless, their conclusions give us an indication of circumstances in which the asymptotic approximation may not be accurate.

The above discussion provides insights into the nature of the finite sample distribution. It is also useful to have similar insights for specific features of the distribution such as the mean and variance. Hillier, Kinal, and Srivastava

[13] This assumes $\omega_{2,2,0} < \infty$.
[14] Also see Section 8.2.

(1984) derive exact formulae for the moments of the IV estimator under normality. These formulae can be used to calculate the bias and mean squared error but are sufficiently complicated to be uninterpretable.[15] However, it is possible to develop more revealing expressions if we are prepared to settle for approximations to the finite sample moments. This topic is discussed in the next sub-section.

6.2.2 Higher Order Approximations

The asymptotic analysis in Chapters 2 through 5 is often refered to as "first order" asymptotics. This terminology originates from the idea of expressing the statistic of interest, c_T say, as a polynomial expansion in negative powers of T such as

$$c_T \;=\; c_0 \;+\; c_1 T^{-1/2} \;+\; c_2 T^{-1} \;+\; c_3 T^{-3/2} + \ldots$$

The limiting behaviour of c_T is governed by the lead or first term of the expansion c_0, and this gives rise to the terminology. As mentioned above, these first order asymptotics only provide an approximation to finite sample behaviour. Intuition suggests that a better approximation can be obtained by including higher order terms from this expansion. In this section, we review the literature on two types of higher order expansions for GMM estimators: Edgeworth expansions for the distribution function, and Nagar expansions for the bias and mean square error.

Edgeworth expansions provide a bridge between the finite sample and limiting distributions, and by examining their lead terms it is possible to uncover what factors affect the passage to the limiting distribution. Sargan and Mikhail (1971) and Sargan (1975) derive the Edgeworth expansion for the IV estimator in the static linear simultaneous model in (6.7) with normal errors.[16] For our purposes here, it is sufficient to focus on the case in which $J = 2$ and so there are only two endogenous variables. In this case, Sargan and Mikhail (1971) show that

$$P\left[\frac{T^{1/2}(\hat{\beta}_T \;-\; \beta_0)}{\sqrt{AVar(\hat{\beta}_T)}} \;\leq r\right] \;=\; \Phi(r) \;+\; \frac{1}{\sqrt{T}} D_1(r) \;+\; \frac{1}{T} D_2(r) \;+\; O_p(T^{-3/2}) \tag{6.14}$$

where $\hat{\beta}_T$ is the estimator defined in (6.11), $AVar(\hat{\beta}_T)$ is the asymptotic variance of $\hat{\beta}_T$, $\Phi(.)$ is the cumulative distrbution function of the standard normal distribution, and $D_i(.), i = 1, 2$ are constants that depend on the model.[17] It can be recognized that the first term on the right hand side of (6.14) is the probability

[15] Knight (1986) derives exact formulae for the moments of the 2SLS estimator when the errors follow an Edgeworth type distribution but these expressions possess the same advantages and disadvantages as their counterparts when the error has a normal distribution.

[16] Also see Morimune (1983).

[17] Sargan (1975) extends the analysis to the case where $\hat{\beta}_T - \beta_0$ is standardized by the square root of the estimated asymptotic variance.

of the particular event based on the asymptotic distribution of the estimator. It therefore follows that the use of the limiting distribution to calculate such probabilities involves an error of order $O_p(T^{-1/2})$. More can be learned about this error by examining the determinants of $D_i(.)$. Sargan and Mikhail (1971) report calculations that indicate the asymptotic approximation tends to deteriorate as the degree of overidentification increases. This leads them to conclude that:

> "by using an intelligent choice of instrumental variables, little may be lost in the asymptotic variance of the estimator and a good deal may be gained in the decreased error of the asymptotic approximations."[Sargan and Mikhail, 1971, p.158]

In terms of more modern terminology, this conclusion can be stated as saying that the inclusion of redundant or nearly redundant instruments tends to lead to a deterioration in the quality of the asymptotic approximation.

Nagar (1959) develops expansions for the first two moments of the Two Stage Least Squares estimator in the linear simultaneous equations model with normal errors. Although the approximations have been generalized subsequently to certain other distributions,[18] it is most convenient to maintain normality and also to continue to restrict attention to the case in which $J = 2$ – all notation is the same as in the previous sub-section.

Nagar (1959) derives a random vector, b_z, and a random matrix, M_z, such that:

$$\hat{\theta}_T - \theta_0 = b_z + o_p(T^{-1}) \tag{6.15}$$

$$(\hat{\theta}_T - \theta_0)(\hat{\theta}_T - \theta_0)' = M_z + o_p(T^{-2}) \tag{6.16}$$

He then approximates the bias (first moment) of $\hat{\theta}_T$ by $E[b_z]$ and the mean square error matrix of $\hat{\theta}_T$ by $E[M_z]$.[19] This leads to the following approximations: for the bias (up to the order of T^{-1})

$$E[b_z] = (q - p - 1)Q_z s \tag{6.17}$$

where $Q_z = (X'P_zX)^{-1}$,

$$s = \begin{bmatrix} B_1'\sigma_1 \\ 0_{p-1} \end{bmatrix}$$

B_1 is the matrix satisfying $A_1 = UB_1$, σ_1 is the first column of Σ and 0_r is the $r \times 1$ null vector; and for the mean squared error (up to order T^{-2})

$$E[M_z] = \sigma_{11}Q_z\,(I + A^*) \tag{6.18}$$

where

$$\begin{aligned} A^* &= [-2(q - p - 1)tr\,(Q_zH_\sigma) + tr\,(Q_zH_\Sigma)] \cdot I_p \\ &+ \left\{\left[(q - p)^2 - 3(q - p) + 4\right]H_\sigma - (q - p - 2)H_\Sigma\right\}Q_z, \end{aligned}$$

[18] See Buse (1992), Donald and Newey (2001) and Peixe and Hall (2000).

[19] While this step has an obvious intuitive appeal, it is not valid in all circumstances; see Srinivasan (1970). However, Sargan (1974) establishes a set of conditions under which it is valid in the context here.

$H_\sigma = \sigma_{11}^{-1} ss'$ and

$$H_\Sigma = \begin{bmatrix} B_1' \Sigma B_1 & 0_{p-1}' \\ 0_{p-1} & 0_{(p-1)\times(p-1)} \end{bmatrix}$$

and $0_{r\times r}$ is the $r \times r$ null matrix.

These two approximations can be used to explore the impact of the inclusion of additional instruments upon the first two moments of the estimator. Inspection of (6.17) reveals that the approximate bias depends on Z via $q - p$ and Q_z. Therefore, the bias is sensitive to both the number of instruments and their relationship to y_2. The bias is also different for each element of $\hat{\theta}_T$. To gain a better understanding, it is convenient to focus on $m_z = \|E[b_z]\|$ which can be interpreted as an aggregate measure of bias in the estimation of θ_0. Buse (1992) derives a relatively simple condition for m_z to increase when additional instruments are included in the estimation. To present this condition, define Z_1 and Z_2 to be respectively $(T \times q_1)$ and $(T \times q_2)$ matrices of instruments and assume Z_1 represent the first q_1 columns of Z_2. This means $q_1 < q_2$. For what follows, it is also important to recall that by assumption the first K_1 columns of any instrument matrix contain the explanatory variables, N_1, which appear in the equation being estimated. If m_i equals the value of m_z associated with $Z = Z_i$ then Buse (1992) shows that

$$\frac{m_2}{m_1} \left\{ \begin{array}{c} \geq \\ = \\ \leq \end{array} \right\} 1 \quad \text{if} \quad \frac{q_2 - p - 1}{q_1 - p - 1} \left\{ \begin{array}{c} \geq \\ = \\ \leq \end{array} \right\} \frac{R_2^2 - R_0^2}{R_1^2 - R_0^2} \tag{6.19}$$

where R_i^2 represents the uncentred R^2 from the regression of y_2 on Z_i, and R_0^2 represents the R^2 from the regression of y_2 on N_1. Therefore the approximate bias will increase with the number of excess instrumental variables only if the proportional increase in the number of instruments is faster than the rate of increase in R^2 measured relative to the fit of Y_1 on N_1. This means that the potential impact of additional instruments depends on the explanatory power of those already included.

While it is desirable to avoid bias, an increase in bias may be tolerated if the mean squared error is reduced. Unfortunately, the formula in (6.18) is not so amenable to interpretation. However, it can be used to numerically evaluate how the inclusion of new instruments affects the approximate mean squared error. Peixe and Hall (2000) report this type of calculation for the special case of the model described above in which $J = 2$, $K_1 = 1$, $K_2 = 8$. More specifically, they consider the system

$$y_1 = y_2\beta + n_1\gamma_{1,1} + u_1 \tag{6.20}$$

$$y_2 = N\gamma_2 + u_2 \tag{6.21}$$

in which $\beta = \gamma_{1,1} = 1$, $\gamma_{2,1} = \ldots = \gamma_{2,5} = .03$, $\gamma_{2,6} = \ldots = \gamma_{2,9} = .33$. These choices imply the first five columns of N have only a marginal contribution to the explanation of y_2, but the last four variables have a more significant impact so that the population R^2 for (6.21) is around 30%. To reflect this dichotomy,

we refer to $\{n_i, i = 1, \ldots 5\}$ as "bad" instruments (for y_2), and $\{n_i, i = 6, \ldots 9\}$ as "good" instruments (for y_2). Strictly none of these instruments are redundant but there is clearly a sense in which the "bad" instruments can be viewed as "nearly" redundant given the "good" ones.[20] The error specification is as follows: letting $u_{i,t}$ denote the t^{th} element of u_i, then $u_t = (u_{1,t}, u_{2,t})'$ is independently and identically distributed as a normal random vector with mean zero and a variance–covariance matrix Σ_0 whose diagonal elements are one and off-diagonal elements are 0.8. The sample size is set at $T = 30$. Table 6.1 reproduces the calculated values of the approximate bias and mean squared error for various instrument combinations reported in Peixe and Hall (2000). There are five cases each involving four instruments.[21] The only difference between the five cases lies in the number of good and bad instruments included. As would be expected, the approximate bias and MSE decrease every time a bad instrument is *replaced* by a good one. Table 6.1 also reports the percentage change in bias and MSE if an *additional* instrument is included. The results reveal that if the additional instrument is "bad" then both the bias and mean squared error increase. However, if the additional instrument is "good" then the impact on the bias is more subtle. If only one of the four instruments is good then the inclusion of another good one reduces the bias. Whereas, if at least two of the four are good then the inclusion of another good one increases the bias. Also the size of this increase is an increasing function of the number of good instruments. In spite of this, the inclusion of an additional good instrument always decreases the mean squared error. While caution must be exercised in generalizing the specific results to more general settings, one conclusion is clear. There is a far more complex relationship between the behaviour of the estimator and the properties of the instrument vector in finite samples than is predicted by asymptotic theory.

Table 6.1
Impact of an additional instrument

Inst.	*Bias*	$\%+1G$	$\%+1B$	*MSE*	$\%+1G$	$\%+1B$
3B1G	0.478	-24.08	48.80	0.546	-27.36	3.66
2B2G	0.243	0.25	49.36	0.390	-17.09	1.90
1B3G	0.163	12.59	49.57	0.319	-12.45	1.25
4G	0.122		49.75	0.277		0.98

Source: Peixe and Hall (2000).
Notes: *Inst* denotes the composition of the benchmark set of instruments: e.g. 3B1G denotes three bad instruments and one good one. *Bias* and *MSE* are calculated using (6.17) and (6.18) respectively. $\%+1G$ ($\%+1B$) denotes the percentage change in either the bias or MSE as a result on the inclusion of an additional good (bad) instrument.

[20] For ease of expression here, we attribute the property of redundancy directly to the instrument and not the associated population moment condition.

[21] Notice that this is the smallest number of instruments for which the second moment of the estimator exists within this model. See the comments made earlier in this section.

Newey and Smith (2004) develop Nagar type expansions for the bias of both the two step GMM and continuous updating GMM estimators in nonlinear static models. To present these results, we return to our generic notation and so assume that estimation of θ_0 is based on $E[f(v_t, \theta_0)] = 0$. The data vector, v_t, is assumed to be a realization from some independently and identically distributed process. Let $\hat{\theta}_T$ denote the two step GMM estimator and $\tilde{\theta}_T$ denote the continuous updating GMM estimator.[22] Also define $G_t = \partial f(v_t, \theta)/\partial\theta'|_{\theta=\theta_0}$, $f_t(\theta) = f(v_t, \theta)$ and let $f_{t,i}(\theta)$ denote the i^{th} element of $f_t(\theta)$. Newey and Smith (2004) show that the approximate bias of the GMM estimator is given by

$$E[\hat{\theta}_T] - \theta_0 = T^{-1}\{B_I + B_G + B_S + B_W\} + o(T^{-1}) \tag{6.22}$$

where

$$\begin{aligned}
B_I &= M(S^{-1})\{E[G_t M(S^{-1}) f_t(\theta_0)] - a\} \\
B_G &= -(G_0^{'} S^{-1} G_0)^{-1} E[G_t^{'} S^{-1/2'}\{I_q - P(\theta_0)\} S^{-1/2} f_t(\theta_0)] \\
B_S &= M(S^{-1}) E[f_t(\theta_0) f_t(\theta_0)' S^{-1/2'}\{I_q - P(\theta_0)\} S^{-1/2} f_t(\theta_0)] \\
B_W &= -M(S^{-1}) \sum_{j=1}^{p} E\left[\frac{\partial f_t(\theta) f_t(\theta)'}{\partial \theta_j}\Big|_{\theta=\theta_0} \right] [M(W) - M(S^{-1})]' e_j
\end{aligned}$$

a is $(q \times 1)$ vector with i^{th} element,

$$a_i = 0.5\, tr\left\{ (G_0^{'} S^{-1} G_0)^{-1} E\left[\frac{\partial^2 f_{t,i}(\theta_0)}{\partial\theta\partial\theta'} \right] \right\}$$

$M(W) = (G_0' W G_0)^{-1} G_0^{'} W$, $P(\theta_0) = F(\theta_0)[F(\theta_0)' F(\theta_0)]^{-1} F(\theta_0)'$, $F(\theta_0) = S^{-1/2} G_0$ and e_j is a $(p \times 1)$ vector whose j^{th} element is one and remaining elements are all zero. As Newey and Smith (2004) observe these four components of the bias have an interesting interpretation. To motivate this part of the discussion, it is useful to first recall the Method of Moments interpretation of GMM derived from the first order conditions, that is the two step GMM estimator is the MM estimator based on $G_0^{'} S^{-1} E[f(v_t, \theta_0)] = 0$.[23] If both G_0 and S are known, then the GMM estimator is just the value of θ that sets this linear combination of the sample moments equal to zero, that is $\bar{\theta}_T$, the solution to $G_0^{'} S^{-1} g_T(\bar{\theta}_T) = 0$. It is easily recognized that this version of the estimator converges to the same limiting distribution as the two step estimator. We thus refer to $\bar{\theta}_T$ as an infeasible optimal GMM estimator – infeasible as G_0 and S are unknown, optimal in the sense that it is a minimum variance estimator based on $E[f(v_t, \theta_0)] = 0$. With this in mind, we now consider the components of the bias in turn. B_I is the approximate asymptotic bias of the infeasible optimal GMM estimator; B_G is a bias term that arises due to the estimation of G_0; B_S is a bias term that arises due to the need to estimate S; B_W is a bias term arising from the first step estimator. Two other general features of this decomposition

[22] See (3.102) in Section 3.7.
[23] See Section 3.3.

are worth noting. First, if the parameter vector is just identified then B_G, B_S and B_W are all equal to zero. Therefore, overidentification introduces bias from a variety of sources. Secondly, it is interesting to note that these sources of the bias depend on the same features of the model that play a crucial role in the limiting distribution of the GMM estimator in misspecified models.[24]

Newey and Smith (2004) show that the corresponding bias of the continuous updating GMM estimator is given by

$$E[\tilde{\theta}_T] - \theta_0 = T^{-1}\{B_I + B_S\} + o(T^{-1}) \tag{6.23}$$

In comparison to (6.22), it can be seen that there are fewer sources of bias. Specifically, there are no longer bias terms associated with the first step estimation or the estimation of the derivative matrix. The absence of the first of these is to be expected because there is no longer a first step estimation. The second is less easy to explain from a GMM perspective. Newey and Smith (2004) show that the absence of B_G is to be expected because the continuous updating GMM estimator is a member of the class of Generalized Empirical Likelihood estimators. However, further elaboration here would constitute a major detour, and so the interested reader is refered to Newey and Smith (2004).[25]

While these general formulae provide some useful insights into the bias, the specific form of the terms is difficult to interpret. Newey and Smith (2004) specialize these formulae to three cases of interest: the IV estimator in the linear model described in Chapter 2; the Generalized IV estimators described in Section 7.2; and separable moment conditions, that is $f(v_t, \theta) = f_1(v_t) - f_2(\theta_0)$. In all cases, Newey and Smith (2004) show that the bias of the GMM estimator increases with the number of overidentifying restrictions *ceteris paribus* – however some caution is needed in making such comparisons as noted by Buse (1992) because the introduction of additional moment conditions alters other aspects of the model.[26] Imbens (2002) reports a similar calculation for a very simple example in which only one moment condition provides information and all the remaining moment conditions are redundant. He shows that the approximate bias increases linearly with the number of redundant moment conditions.

6.3 Simulation Evidence from Nonlinear Dynamic Models

As we have just seen, finite sample distribution theory provides some useful insights into what aspects of the distribution affect the quality of the asymptotic approximation in static models. Intuition suggests that these aspects of the specification are going to play a similarly important role in nonlinear dynamic models. At the same time, it would also be anticipated that the presence of nonlinearity and/or dynamics introduces additional complications. In recent

[24] See Chapter 4.
[25] There is a brief introduction to empirical likelihood estimators in Section 10.2.
[26] See discussion earlier in this section.

years, concern has grown about the adequacy of the asymptotic approximation in the sample sizes encountered in practice, and this has spawned a number of computer based simulation studies calibrated to the types of model which appear in Table 1.1.[27] An overview of these studies is provided by Table 6.2. In this section we review the main findings from this literature.

Table 6.2
Simulation studies of the finite sample properties of GMM

Economic or statistical topic	Type	
Asset pricing	NV, NP	Tauchen (1986), Kocherlakota (1990), Hansen, Heaton, and Yaron (1996), Smith (1999)
	LV, NP	Ferson and Foerster (1994)
Business cycles	NV, NP	Burnside and Eichenbaum (1996), Christiano and den Haan (1996)
Covariance structures	NV, LP	Altonji and Segal (1996)
	NV, NP	Clark (1996)
Inventories	LV, LP	Fuhrer, Moore, and Schuh (1995), West and Wilcox(1994,1996)
Stochastic volatility	NV, NP	Andersen and Sørensen (1996)

Note: *Type* indicates the functional form of the model with NV (LV) denoting nonlinear (linear) in variables and NP (LP) denoting nonlinear (linear) in the parameters.

As in the previous section, there are two main questions of interest here – does asymptotic theory provide a good approximation in the sample sizes encountered in practice? – and, what aspects of the specification affect the quality of this approximation? The answer to the first question is going to be model specific, but the answer to the second is likely to be generic on some level and so is our main focus here. In spite of this, it is pedagogically more convenient to organize the discussion around four specific studies. We begin with the studies by Tauchen (1986) and Kocherlakota (1990) which are calibrated to the consumption based asset pricing model used in our empirical example. We then briefly summarize the results reported in Hansen, Heaton, and Yaron (1996) for a slightly more sophisticated version of this model. Finally, we consider the study by Andersen and Sørensen (1996) based on the stochastic volatility model described in Section 1.3.5. Together these four studies provide a good overview of the qualitative findings from this literature.

Asymptotic theory has been used to justify the GMM estimation and also to develop a vast array of inference procedures based on the estimator. In our discussion here, we focus on how well this theory approximates finite sample

[27] As an illustration of the level of this interest, the July 1996 issue *Journal of Business and Economic Statistics* has a special section devoted to seven papers on this topic.

behaviour of the two most important components of this framework: the estimator $\hat{\theta}_T$ and the overidentifying restrictions test J_T. Specifically, we consider the following five questions:

1. *Is the GMM estimator approximately unbiased?*
2. *How reliable are confidence intervals based on asymptotic theory?*
3. *Is the finite sample distribution of the overidentifying restrictions test well approximated by a* χ^2_{q-p}*?*
4. *How does iteration affect the answers to 1.–3.?*
5. *How does the use of the continuous updating estimator affect the answers to 1.–3.?*

Apropos the fourth question, it can be recalled from Section 3.6 that iteration beyond the second step has no effect on the asymptotic distribution and was proposed purely because of potential gains in finite samples. At that stage in our discussion we could only anticipate some advantage, now we can learn whether these gains are realized in practice. With these five questions in mind, we now turn to the simulation evidence.

Tauchen (1986) examines the behaviour of GMM in Hansen and Singleton's (1982) version of the consumption based asset pricing model.[28] His design assumes there is only one asset, and estimation is based on the population moment condition,

$$E[z_t u_t(\theta_0)] \;=\; 0 \tag{6.24}$$

where

$$\begin{aligned} u_t(\theta) &= \delta(c_{t+1}/c_t)^{\gamma-1}(r_{t+1}/p_t) \\ z_t &= (1, c_t/c_{t-1}, \ldots, c_{t-L}/c_{t-L-1}, r_t/p_{t-1}, \ldots, r_{t-L}/p_{t-L-1})' \end{aligned}$$

The degree of overidentification is controlled by L, and Tauchen considers the cases $L = 1, 2, 3, 4$. Notice that $L = 2$ gives the instrument vector used in our estimation of the model earlier in the text. Two sample sizes are considered: $T = 50, 75$. A large part of Tauchen's (1986) contribution is to have developed a method for generating artificial data consistent with the underlying model. However, we only comment very briefly on this aspect of his study. To this end, note that the asset return, r_{t+1}, is given by $r_{t+1} = p_{t+1} + d_{t+1}$ where d_{t+1} denotes the dividends paid out during the period. Therefore, the model can be viewed as depending on three stocastic variables: c_t, d_t, and p_t. Tauchen generates data on the first two of these variables from a $VAR(1)$ model for $[ln(c_{t+1}/c_t), ln(d_{t+1}/d_t)]$. Given this data and the Euler equation for $t = 1, 2, \ldots T$, it is possible to solve for $\{p_t\}$. Tauchen reports results for various choices of parameters in the VAR; he sets $\gamma = 0.3, 1.30$ and $\delta = 0.97$. The secondly step weighting matrix is $\hat{S}_{SU}^{-1}$ defined in (3.40). It should be noted that

[28] See Section 1.3.1.

Tauchen only considers the two step estimator because that was the conventional practice at his time of writing. So his results cannot help us with questions 4 or 5 above. However, in terms of the other three questions, his study reveals the following answers in order.

1. *Bias:* For $L = 1$ (*i.e.* $q - p = 1$) the estimator is approximately unbiased, but there is a tendency for the bias to increase as L, and hence $q - p$, increases. At the same time, increases in L reduce the variance and so $\hat{\theta}_T$ becomes concentrated at a value away from the truth. Interestingly, this mirrors Anderson and Sawa's (1979) finding for the IV estimator in the linear model discussed in the previous section.

2. *C.I.'s*: For $L = 1, 2$ (*i.e.* $q-p = 1, 3$) the empirical coverage of the asymptotic confidence intervals is approximately equal to the nominal value.[29] However, for $L = 3, 4$ (i.e. $q - p = 5, 7$) the empirical coverage tends to be less than the nominal value.

3. J_T: the empirical size of the overidentifying restrictions test tends to be close to its nominal value in all cases considered.[30] If anything, the test rejects slightly less frequently than would be anticipated from asymptotic theory.

Based on this evidence, Tauchen recommends that $q - p$ be kept less than or equal to three in this model with these sample sizes. However, there is one aspect of Tauchen's (1986) study which should be borne in mind when considering this recommendation. The degree of overidentification is controlled by L, and so an expansion of the instrument vector involves the inclusion of lagged values of consumption growth and the asset return from further back in time. Now $u_t(\theta)$ depends on $(c_{t+1}/c_t, r_{t+1}/p_t)$, and within his design the autocorrelations of these variables decays as the lag length increases. Therefore, as L increases z_t becomes augmented with variables whose association with $u_t(\theta)$ is decaying. In other words, every increase in L introduces instruments whose quality is worse than those already included. This is not a criticism of Tauchen's (1986) design because this strategy is commonly used for instrument selection in Euler equation models in practice. However, it is probably more appropriate to view Tauchen's recommendation within the context of this instrument selection strategy than as a more general comment about the desirable degree of overidentification *per se*.

Kocherlakota (1990) uses Tauchen's (1986) simulation method to investigate the behaviour of GMM in Hansen and Singleton's (1982) model with multiple assets. In this case, estimation is based on the population moment condition

$$E[z_t \otimes u_t(\theta_0)] \;=\; 0 \tag{6.25}$$

[29] "Empirical coverage" is the term used for the proportion of the replications in which the calculated confidence interval contains the true parameter value. So for a 95% confidence interval, say, to be perfectly accurate, its empirical coverage must equal its nominal value, which is 95%.

[30] "Empirical size" is term given to the proportion of replications in which the test is significant.

where $u_t(\theta)$ is $(s \times 1)$ vector whose i^{th} element is $\delta(c_{t+1}/c_t)^{\gamma-1}(r_{i,t+1}/p_{i,t})$ and z_t is $(k \times 1)$ vector of instruments. Kocherlakota considers models with up to three assets, i.e. $s = 1, 2, 3$, and $k = 1, 3, 4$. However, of the seven particular combinations chosen, six involve $q - p = 1$ and one involves $q - p = 6$. Since the design involves multiple assets, Kocherlakota is able to confine attention to instrument vectors whose elements come from the set $\{1, c_t/c_{t-1}, r_{i,t}/p_{i,t-1}\}$ – in other words, $L = 1$ in terms of the notation used to describe Tauchen's (1986) study. Unlike Tauchen (1986), Kocherlakota (1990) evaluates the performance of both the two step and iterated estimator – the latter with $I_{max} = 70$ – with $\hat{S}_T(i) = \hat{S}_{SU}$ for $i > 1$. Two other differences between the two studies are also worth noting: Kocherlakota (1990) sets $T = 90$ for the most part but also reports results for $T = 200, 500, 2000$; he also sets $\gamma = 13.7$ and $\delta = 1.139$.[31] We begin our discussion of his results with the case in which $T = 90$ because these most closely parallel Tauchen's settings. As a whole, Kocherlakota's (1990) evidence suggests that the iteration beyond the second step considerably improves the quality of the asymptotic theory as an approximation to finite sample behaviour. So strong is the evidence that he focuses entirely on the iterated estimator in the published version of his paper – and so our discussion of his results must do the same. In terms of the other three questions, his findings are as follows.

1. *Bias:* There is evidence of bias in some cases, and not others. This bias does not appear to be linked to the degree of overidentification *per se*, that is to the limited extent this can be assessed within this design.

2. *C.I.'s:* The empirical coverage of the asymptotic confidence intervals is too low in nearly every case and in some cases the strikingly so – e.g. $\approx 60\%$ instead of the nominal value of 95%.

3. J_T : The quality of the asymptotic approximation is good in some cases but not in others. In the latter, the empirical size of the test tends to be around 20% when the nominal size is 5%.

Buried within this summary is an interesting pattern to the results. Although the choices $s = 1, k = 3$ and $s = 3, k = 1$ both imply $q - p = 1$ the estimator behaves very differently in the two cases. If there are multiple assets and one instrument $(s = 3, k = 1)$ then the finite sample behaviour is well approximated by the asymptotic theory, but if there is one asset and multiple instruments $(s = 1, k = 3)$ then the estimator is biased, the asymptotic confidence intervals are unreliable and the overidentifying restriction test rejects too frequently. Therefore, low values of $q - p$ are no guarantee that the asymptotic approximation is good.

One attractive feature of Kocherlakota's (1990) study is that he also considers what happens as T increases. As T moves from 90 to 200, 500 and finally

[31] The parameter values are calibrated to replicate certain features of annual data for the U.S spanning 1889–1978. In contrast, Tauchen's (1986) parameter settings were chosen to be "reasonable" from an economic theoretic standpoint. It should be noted that Kocherlakota (1990) also reports a limited number of simulation results using data generated with other parameter values including $\gamma = 0.3$, $\delta = 0.97$ which were used by Tauchen.

2000, the quality of the asymptotic approximation improves. However, in the worst cases, it is only at the largest sample size that asymptotic theory accurately predicts the empirical coverage of the asymptotic confidence intervals and the empirical size of the overidentifying restrictions test. While this is not an encouraging conclusion, Kocherlakota finds that the situation is worse with the two step estimator. He finds that after only two steps, the overidentifying restrictions test converges very slowly to its asymptotic distribution.

Clearly, some aspect of the asymptotic theory is not providing a good approximation to finite sample behaviour. It would clearly be useful to diagnose where the problem lays, and Kocherlakota (1990) provides some useful guidance in this direction for the overidentifying restrictions test. To describe what he did, we must remind ourselves of the structure of the estimated sample moment again. Equation (3.35) shows that

$$W_T^{1/2}T^{1/2}g_T(\hat{\theta}_T) \;=\; N_T(\hat{\theta}_T)W_T^{1/2}T^{1/2}g_T(\theta_0) \;=\; \tilde{N}_T T^{1/2}g_T(\theta_0), \text{ say} \quad (6.26)$$

It can be recalled from Section 3.4.3 that the asymptotic normality of $W_T^{1/2}$ $T^{1/2}g_T(\hat{\theta}_T)$ rested on the convergence in probability of $\tilde{N}_T$ to a matrix of constants and the application of the Central Limit Theorem to $T^{1/2}g_T(\theta_0)$. Interestingly, Kocherlakota (1990) finds that $T = 90$ is large enough for $T^{1/2}g_T(\theta_0)$ to be approximately normally distributed in *all* the cases he considers. The problem stems from $\tilde{N}_T$. Kocherlakota (1990) finds that all the cases in which the χ^2 approximation is poor are exactly the cases in which $\tilde{N}_T$ is still exhibiting considerable variability. Since $\tilde{N}_T$ is the product of matrices, Kocherlakota's (1990) evidence points to two possible culprits: $\hat{S}_T^{-1}$ and $G_T(\hat{\theta}_T)$. Interestingly, this evidence highlights two of the sources of bias in the Nagar type expansion for the GMM estimator described in the previous sub-section; see equation (6.22). The involvement of $G_T(\hat{\theta}_T)$ here also creates an interesting tie in with our discussion of the finite sample distribution of IV estimator in the static linear model. It can be recalled from the previous section that the convergence of the IV estimator to its asymptotic distribution depends on the concentration parameter, $T\tilde{\mu}^2$, and that this convergence is likely to be slow if θ_0 is "weakly" identified. Now the matrix $G_T(\hat{\theta}_T)$ has a similar link to identification because it is the sample analog of G_0.[32] This suggests that weak identification may be one source of the problems noted in Kocherlakota's (1990) study – an explanation which would certainly accord with our empirical experience of the model in Chapter 3.[33]

Before we move on to discuss the other two studies mentioned above, it is worth reflecting what we have learnt from Tauchen's (1986) and Kocherlakota's (1990) results about the interpretation of our empirical results. It can be recalled that choice of z_t is a special case of Tauchen's (1986) design with $L = 2$, and one in which he found asymptotic theory provided a reasonable approximation even in his much smaller sample sizes. However, Kocherlakota's (1990) study reveals that the quality of the approximation can be sensitive to θ_0 as well as

[32] Recall that the condition for local identification is $rank(G_0) = p$; see Assumption 3.6 in Section 3.1.

[33] In particular see the discussion in Section 3.6.

other aspects of the data generation process. In particular, it seems reasonable to be concerned about the quality of the identification and how it has affected finite sample behaviour. So it may be premature to draw a line under the results obtained so far, and we return to this example in the next two chapters as we explore various methods for improving inference based on GMM estimation.

Hansen, Heaton, and Yaron (1996) also examine the behaviour of GMM and its associated statistics in a consumption based asset pricing model. However, their study builds from those described above in two important ways. First, they allow for time non-separability in the utility function of the representative agent. Second, they simulate the behaviour of continuous updating estimator as well as the two step and iterated estimators.

Hansen, Heaton, and Yaron (1996) consider the case in which the representative agent's utility function takes the form,

$$U(c_t) \;=\; \frac{(c_t \;+\; \eta_0 c_{t-1})^{1-\gamma_0} \;-\; 1}{1 \;-\; \gamma_0}$$

Notice if $\eta_0 = 0$ then this utility function reduces to the CRRA utility function used by Tauchen and Kocherlakota.[34] The agent is assumed to invest in two assets: a bond, whose payoff is denoted $R_{1,t}$, and a stock, whose payoff is denoted $R_{2,t}$.[35] Hansen, Heaton, and Yaron (1996) simulate artificial data from this model for a number of scenarios of empirical relevance.[36] For brevity, we focus on two scenarios here: in the first, the data generation process is calibrated to annual US data and the sample size is set to $T = 100$; and in the second, the data generation process is calibrated to monthly US data and the the sample size is set to $T = 400$. In both cases, they consider the case in which estimation is based on the population moment condition

$$E[z_t \otimes e_{t+2}(\theta_0)] \;=\; 0 \tag{6.27}$$

where $z_t \in \Omega_t$, $\theta = (\gamma, \delta, \eta)'$, δ is the discount factor[37], and $e_{t+2}(\theta)$ is the (2×1) vector with i^{th} element given by[38]

$$\begin{aligned} e_{i,t+2}(\theta) \;=\; 1 \quad &+ \quad \delta\eta \left\{ \frac{c_{t+1} + \eta c_t}{c_t + \eta c_{t-1}} \right\}^{-\gamma} \;-\; \delta R_{i,t+1} \left\{ \frac{c_{t+1} + \eta c_t}{c_t + \eta c_{t-1}} \right\}^{-\gamma} \\ &- \quad \delta^2\eta \left\{ \frac{c_{t+2} + \eta c_{t+1}}{c_t + \eta c_{t-1}} \right\}^{-\gamma} \end{aligned}$$

Two choices of instrument are used: $z_{1,t} = (1, c_t/c_{t-1})'$ and $z_{2,t} = (z_{1,t}, R_{1,t}, R_{2,t})'$ for which $q - p$ equals 1 and 5 respectively.

[34] If $\eta_0 = 0$ then utility is time separable in the sense that utility in period t depends on consumption in period t; otherwise, utility is said to be time non-separable because utility in period t depends on both contemporaneous and lagged consumption.

[35] In terms of the notation above, $R_{2,t} = (p_t + d_t)/p_{t-1}$.

[36] Hansen, Heaton, and Yaron (1996) use a variation on Tauchen's (1986) method to simulate the data.

[37] See Section 1.3.1.

[38] It should be noted that $e_{t+2}(\theta)$ is a transformed version of the Euler equation associated with this model.

Hansen, Heaton, and Yaron's (1996) evidence for the two step and iterated estimators tends to corroborate the findings from the previous studies. Therefore, we do not report specific details save to note that they find asymptotic theory tends not to be a good guide in samples $T = 100$; whereas, it is reasonably accurate for the iterated estimator with $T = 400$ for the model with $q - p = 1$, but not in the model with $q - p = 5$. Instead, we focus our discussion on how their results illuminate the relative properties of the iterated and continuous updating estimators. The most striking feature of this comparison is that the continuous updating estimator converges to very extreme values in a small but significant number of the replications whereas the iterated estimator does not.[39] This behaviour is the source of two key differences between the simulated distributions of the estimators. First, the simulated distribution of the continuous updating estimator exhibits far longer tails than those of the iterated estimator. Secondly, these extreme values are not evenly distributed between the left and right tails and so cause an asymmetry in the simulated distribution of the continuous updating estimator which does not appear to be present for the iterated estimator. Both these features manifest themselves in the moments of the simulated distribution, and so impact on the comparison of the estimators. For example, if bias is measured as the difference between the true value and the median of the simulated distribution then in most cases – but not all – the continuous updating estimator exhibits less bias than the iterated estimator. However, if the median is replaced by the mean in the previous calculation, then in most cases the ranking is reversed. This tail behaviour leads Hansen, Heaton, and Yaron (1996) "from the standpoint of obtaining estimates, we see no particular advantage to using continuous updating when minimizing GMM criterion functions"[p.278]. However, they also note that the use of the continuous updating estimator may be advantageous for inference. Specifically, they find that the overidentifying restrictions test based on the continuous updating estimator tends to exhibit empirical size closer to its nominal value than its counterpart based on the iterated estimator. It is worth noting that this conclusion regarding the relative merits of the iterated GMM and the continuous updating estimator appears to be in conflict with that based on their Nagar type expansions; see (6.22)–(6.23) in the previous sub-section. These differing conclusions may reflect the different contexts: the Nagar expansions are for static models and the simulation results are for a dynamic model. Further work is needed to reconcile the results from these two approaches.

The simulation studies described above shed little light on what factors effect the behaviour of $\hat{S}_T^{-1}$. To gain some insight into this question, it is useful to recall both the form of the long run variance and the estimators. It can be recalled from Section 3.5 that

$$S \;=\; \Gamma_0 \;+\; \sum_{i=1}^{\infty}(\Gamma_i \;+\; \Gamma_i^{'})$$

[39] Also see Section 3.7.

and our basic strategy for estimating this matrix is to use a weighted sum of the sample autocovariance matrices.[40] So there are two natural questions: – what factors affect the convergence of the sample autocovariances to their population counterparts? – and what factors affect the convergence of our weighted sum of autocovariances to S? The answer to the first depends on the nature of the nonlinearity in $f(v_t, \theta)$. The answer to the second depends in part on the weighted sum involved. We have reviewed the extensive literature on covariance matrix estimation already, and below we discuss some further simulation evidence on this issue. However, before that, it is useful to expand a little on the answer to the first question.

For the purposes of this discussion, we can confine attention to polynomial powers of a scalar random variable v_t. For simplicity, assume that $\{v_t; t = 1, 2, \ldots T\}$ is an independent sequence and $v_t \sim N(0, 1)$. As we have seen, the GMM estimation strategy exploits the convergence in probability of sample to population moments, that is

$$T^{-1} \sum_{t=1}^{T} v_t^k \xrightarrow{p} E[v_t^k] = \mu_k, \; say. \tag{6.28}$$

While this result holds for any k, the variability of the sample moment depends on k in a rather simple – but striking – fashion. It is straightforward to show that

$$Var[T^{-1} \sum_{t=1}^{T} v_t^k] = \sigma_k^2 / T$$

where $\sigma_k^2 = Var[v_t^k]$, and under our assumptions it follow that $\sigma_1^2 = 1$, $\sigma_2^2 = 2$, $\sigma_3^2 = 15$, $\sigma_4^2 = 96$, $\sigma_5^2 = 945$ and so on.[41] So, for example, $T^{-1} \sum_{t=1}^{T} v_t^4$ exhibits 96 times as much variability as the sample mean in any sample size! Or put another way, the variance of the sample mean is 0.1 when $T = 10$, but it takes a sample of size 960 to achieve the same precision for $T^{-1} \sum_{t=1}^{T} v_t^4$. These simple calculations indicate that the convergence of sample moments is very sensitive to the form of the nonlinearity. This example is not without practical relevance either. Polynomial powers naturally occur in the population moment conditions used in many studies, and these calculations provide a simple intuition behind the findings in a number of the simulation studies listed in Table 6.2.

We now turn our attention to a simulation study of GMM in stochastic volatility models which involves moment conditions of the type in (6.28) and also HAC estimators. Andersen and Sørensen (1996) consider the following simplified version of the model in Section 1.3.5

$$\begin{aligned} y_t &= \sqrt{x_t} e_t \\ ln(x_t) &= \theta_1 + \theta_2 ln(x_{t-1}) + \theta_3 u_t \end{aligned}$$

[40] For the purposes of this discussion, we exclude $\hat{S}_{VARMA}$ which is not considered in any of the simulation studies listed in Table 6.2.

[41] For the standard normal distribution, $E[v_t^k] = (k-1)(k-3)\ldots 3.1.$; see Johnson and Kotz (1970) [p.47].

where $(e_t, u_t) \sim i.i.d.N(0, I_2)$.[42] Note that this version of the model is used both to generate the data and is also the assumed specification for the estimation. Therefore, there are only three parameters to be estimated: $\theta = (\theta_1, \theta_2, \theta_3)'$. It can be recalled from Section 1.3.5 that the normality assumptions yields an infinite number of possible population moment conditions. Andersen and Sørensen (1996) consider estimation based on various permutations of the moment conditions but for our purposes here it is sufficient to concentrate on just four choices. Below we just list the moments of y_t involved; the exact form of the associated population moment condition can then be deduced from (1.48).[43] To this end, we define $mi = |y_t|^i$, for $i = 1, 2, 3, 4$; $mi = E[|y_t y_{t-i}|]$, for $i = 4 + j$, $j = 1, 2 \ldots 10$; $mi = E[y_t^2 y_{t-i}^2]$, for $i = 14 + j$, $j = 1, 2, \ldots 10$. The four sets of population moment condition are then given by:

$$
\begin{aligned}
M5 &: \; m1,\, m2,\, m4,\, m6,\, m15 \\
M9 &: \; m1 - m5,\, m7,\, m9,\, m16,\, m18 \\
M14 &: \; m1 - m4,\, m6,\, m8,\, m10,\, m12,\, m14,\, m15,\, m17,\, m19,\, m21,\, m23 \\
M24 &: \; m1 - m24
\end{aligned}
$$

Andersen and Sørensen (1996) report results for the two- and three-step estimators.[44] They consider different choices of θ for the parameter generation, and sample sizes of $T = 500, 1000, 2000, 4000$ and $10,000$. While the latter may seem large numbers, they are not uncommon in the the high frequency data to which these models are applied. In spite of these sizes, Andersen and Sørensen (1996) report that their numerical algorithm experienced non-convergence problems in the smaller sample sizes; further details of the source of these problems and how they were addressed can be found in their paper.

Andersen and Sørensen (1996) report results for various choices of kernel and bandwidth in HAC estimator. We begin, as they do, with the case in which a Bartlett kernel is used with $b_T = 10$. In terms of our four questions above, the results suggest the following:

1. *Bias:* There are quite substantial biases at $T = 500$ but tend to disappear quickly as the sample size increases. The bias tends to be smallest with $M9$ at $T = 2000$ and with $M14$ for the larger samples.

2. *C.I.'s:* For $T > 1000$, the empirical coverage is reasonably close to the nominal value if $M9$ or $M14$ are used. However, if $M24$ is used then the studentized coefficient – that is $(\hat{\theta}_{T,i} - \theta_{0,i})/s.e.(\hat{\theta}_{T,i})$ – exhibits a marked leftward skewness even at $T = 10,000$.

3. J_T : The results reveal an interesting pattern. As the number of moment conditions increase the distribution of J_T shifts to the right, and, for a given set of moment conditions, the distribution shifts to the left as T

[42] This model can be obtained using the following restrictions in (1.45)–(1.47): $y(\tau_t) = y_t$, $x(\tau_t) = x_t$, $d_t = 1$, $\eta = \theta_2 - 1$, $\alpha = 0$, $\beta = -1$, $\gamma = 0$, $\delta = \theta_1$, $\zeta = \theta_3$ and $\rho = 0$.

[43] Note that in this simple model $w_t(\theta) = y_t$.

[44] The "three-step" estimator is the iterated estimator with $I_{max} = 3$.

increases. However, there is little evidence that the statistic is converging to its asymptotic distribution even at these sample sizes. The empirical size is closest to its nominal value at $T = 1000$, $T = 2000$ and $T = 4000$ for respectively $M9$, $M14$ and $M24$. If fewer moment conditions are used than this prescription at a given sample size then the test rejects too frequently; if too many are used then the test rejects too infrequently.

4. *Iteration:* There is no systematic difference between the two- and three-step estimators.

In qualitative terms, these results are broadly similar for various types of HAC estimators. However, Andersen and Sørensen (1996) report that the quality of the asymptotic approximation is improved by the use of the prewhitening and recoloring advocated by Andrews and Monahan (1992), and also the data based bandwidth selection method proposed by Newey and West (1994). The evidence also suggests the Bartlett kernel is to be prefered over the quadratic spectral in this model, which is counter to asymptotic theory.[45] At the same time, it should be noted that the evidence suggests the choice between these HAC estimators is of second order of importance. All the HAC estimators converge fairly slowly to their limit in this class of models. A similar finding is reported by Burnside and Eichenbaum (1996), Christiano and den Haan (1996) and West and Wilcox (1996) albeit to differing degrees depending on the setting in question.

Two features of Andersen and Sørensen's (1996) results stand out. First, a large sample size is needed for asymptotic theory to approximate finite sample behaviour in these models, and secondly the quality of the approximation depends on the choice of moments. The culprit in the first case is $\hat{S}_T$. Andersen and Sørensen (1996) compare $\hat{S}_T$ with its simulated population long run variance, and find that the former is clearly exhibiting considerable bias and variation even at $T = 10,000$. Given that the moment conditions involve polynomial powers, such behaviour would be anticipated for the reason given above. However, there may be another facet to this explanation. Altonji and Segal (1996) argue that in models of covariance structure this slow convergence means that $\hat{S}_T^{-1}$ and $T^{1/2} g_T(\hat{\theta}_T)$ exhibit a correlation in sample sizes which has a negative impact on the quality of the asymptotic approximation.[46] Since stochastic volatility models involve variances Altonji and Segal's (1996) arguments may well apply here as well. Andersen and Sørensen (1996) also uncover an interesting explanation for the sensitivity of the quality of the asymptotic approximation to the choice of moment conditions. They calculate the asymptotic variances of the GMM estimator implied by the four choices. These figures reveal a dramatic drop in variance with the move from $M5$ to $M9$, a smaller, but still marked, drop in variance with the move from $M9$ to $M14$, but only a slight drop in

[45] In contrast, the studies by Burnside and Eichenbaum (1996) and Christiano and den Haan (1996) find no clear ranking between the two kernels is possible. See Section 3.5.3 for further discussion of their relative merits.

[46] Recall they are statistically independent in the limit because the former converges in probability to S^{-1}, a matrix of constants.

variance with the move from $M14$ to $M24$. It can be recognized that these calculations are a good predictor of the finite sample behaviour described above, that is the properties of the estimator tended to improve as q increased – at least in the larger samples – until we reached $M14$, but then deteriorated with move from $M14$ to $M24$. These calculations indicate that whether or not it is beneficial to expand the population moment condition in finite samples from $E[f_1(v_t, \theta_0)] = 0$ to $E[f_1(v_t, \theta_0)] = 0, E[f_2(v_t, \theta_0)] = 0$ depends on both the precise definitions of $E[f_1(v_t, \theta_0)] = 0$ and $E[f_2(v_t, \theta_0)] = 0$ and also their interrelationship. So in these terms, it can be recognized that this conclusion echoes the one drawn from the calculations based on the Nagar type approximation to the bias of the linear IV estimator reported in the previous section.

Since the move from $M14$ to $M24$ has only a marginal impact on the asymptotic variance of the estimator, this expansion of the population moment condition appears to introduce what might be viewed as nearly redundant moment conditions. Therefore, this last set of results appears to suggest that the inclusion of redundant or nearly redundant moment conditions can lead to a deterioration in the finite sample properties of the estimator. Such an explanation would certainly accord with the intuition gained from the Nagar type approximation calculated by Imbens (2002) discussed in Section 6.2.2. However, this example gives no sense of whether the inclusion of redundant moment conditions can have such dramatic effects on the quality of the asymptotic approximation. While Andersen and Sørensen (1996) did not explicitly pursue this issue further, Hall and Peixe (2003) provide simulation evidence which does corroborate this conclusion albeit in a different setting. They consider the following linear model

$$y_t = x_t\theta_0 + u_t \tag{6.29}$$

$$x_t = \Pi_0^{'} z_t + e_t \tag{6.30}$$

where x_t is a scalar and z_t is a 12×1 vector. Putting $v_t^{'} = [u_t, e_t, z_t^{'}]$, artifical data are generated using $v_t \sim IN(0, \Sigma_v)$ where the main diagonal of Σ_v are all set to unity, and the only non-zero off diagonal elements are $\Sigma_v(1,2) = \Sigma_v(2,1) = cov(u_t, e_t) = 0.5$. The parameters are set to $\theta_0 = 0$ and $\Pi_0^{'} = [0.5, 0.5, 0, \ldots 0]$. Notice that within this design, $[z_{t,3}, z_{t,4}, \ldots z_{t,12}]$ are redundant given $(z_{t,1}, z_{t,2})$. Hall and Peixe (2003) consider the behaviour of the set of IV estimators $\{\hat{\theta}_T(i);\ i = 1, 2, \ldots 12\}$ where $\hat{\theta}_T(i)$ is given by (2.8) evaluated at $W_T = (T^{-1}Z'Z)^{-1}$, $z_t = z_t(i)$ and $z_t(i)$ is the $i \times 1$ vector $(z_{t,1}, z_{t,2}, \ldots z_{t,i-1}, z_{t,i})'$.[47] This definition implies that, for $i > 2$, $z_t(i)$ contains $i-2$ redundant instruments. Table 6.3 contains the simulated bias and root mean square error of $\hat{\theta}_T(i)$ along with the mean and empirical rejection frequency of the t-statistic for the hypothesis $H_0 : \theta_0 = 0$ based on $10,000$ replications in the case where $T = 100$. It is evident from these results that the quality of the asymptotic approximation deteriorates as the number of redundant instruments increases. For example, if there are up to three redundant instruments then the empirical rejection frequency of the t-statistic is close to the nominal value of 10%; however, if there

[47] Notice that, for this design, $\hat{\theta}_T(i)$ is the optimal two step estimator; see Section 2.4.

are nine or ten redundant instruments then the empirical rejection frequency is twice the nominal size.

Table 6.3
Consequences of the inclusion of redundant instruments

i	*bias*	*rmse*	*tstat*	*size*
1	−0.030	0.371	0.084	0.075
2	−0.000	0.146	0.140	0.095
3	0.010	0.143	0.211	0.099
4	0.021	0.142	0.282	0.106
5	0.030	0.141	0.351	0.114
6	0.039	0.141	0.418	0.126
7	0.048	0.142	0.484	0.137
8	0.057	0.143	0.550	0.148
9	0.065	0.144	0.617	0.161
10	0.073	0.147	0.682	0.177
11	0.080	0.149	0.746	0.195
12	0.088	0.152	0.810	0.210

Source: Hall and Peixe (2003). Copyright Marcel Dekker; reprinted with permission.
Notes: *bias* and *rmse* are the simulated bias and rmse of $\hat{\theta}_T(i)$. *tstat* denotes the simulated mean of t-statistic for $H_0 : \theta_0 = 0$. *size* denotes the empirical size of the t-test with nominal size 0.1.

While we have reviewed only four studies in detail, their results are representative of this literature. In terms of our five questions, the overall findings for the first four are as follows.

1. *Bias: the estimator is approximately unbiased in some settings and not in others. The bias tends to increase with $q - p$, the degree of overidentification, and particularly with the inclusion of a number of redundant moment conditions. However, a low value for $q - p$ is not a guarantee of the absence of bias because the bias is also sensitive to other aspects of the model such as the functional form of moment condition, the time series properties of the data and the choice of long run covariance matrix estimator.*

2. *C.I.'s: the empirical coverage of the asymptotic confidence intervals is sometimes close to the nominal value but more often tends to be less than the nominal value. This means the asymptotic confidence intervals tend to overstate the precision of the estimation in finite samples. The empirical coverage tends to deteriorate with the inclusion of a number of redundant moment conditions or in the presence of weak identification. The reliability of the asymptotic approximation is also sensitive to the time series properties of the data and the functional form of the moment condition; the approximation can be extremely unreliable in circumstances where these two features of the model interact to cause the long run covariance matrix estimator to be ill-behaved.*

3. *J_T: in some cases it is well approximated by a χ^2_{q-p}, but in others this approximation can be poor. In the latter cases, the test may either reject or fail to reject too frequently depending on the model in question. While there does not appear to be a systematic pattern to the relationship between the empirical and nominal size, the discrepancy between them appears to be larger in the presence of weak identification. The reliability of the asymptotic approximation is also sensitive to the time series properties of the data and the functional form of the moment condition; the approximation can be extremely unreliable in circumstances where these two features of the model interact to cause the long run covariance matrix estimator to be ill behaved.*

4. *Iteration: the quality of asymptotic approximation tends to be greatly improved by iteration.*

Since only one study has examined the continuous updating estimator to date, our conclusions about the fifth question are more tentative. Nevertheless, for completeness, we summarize them here.

5. *Continuous updating: this version of the estimator tends to exhibit fat tails which may have undesirable consequences for parameter estimation, but the associated overidentifying restrictions test may be closer to its asymptotic distribution than its counterpart based on the iterated estimator.*

6.4 Summary and Link to Following Chapters

In this chapter, we have investigated how well asymptotic theory approximates behaviour in samples of the size encountered in practice. It seems fair to say that the evidence is mixed. In some models of interest the approximation can be good at samples of size 100, and in others it is bad even at samples a hundred times larger. Furthermore, for a given functional form, the adequacy of the approximation may be very sensitive to the parameter values used to generate the artificial data. Perhaps only one thing can be said for certain, that is finite sample behaviour is far more complex than would be predicted by asymptotic theory.

In spite of this complexity, the following factors appear to play an important role in determining the quality of the asymptotic approximation:

- the functional form of $f(v_t, \theta_0)$;
- the degree of overidentification, $q - p$;
- the interrelationship between the elements of $f(v_t, \theta_0)$;
- the quality of the identification;
- the estimator of the long run variance.

All these factors collectively point to the following conclusion: the exact choice of population moment condition is crucial to the performance of the method. This observation motivates the material covered in the next chapter. Two specific questions are addressed: – is there a feasible optimal choice of population moment condition? – how can we select the right set of population conditions for the problem in hand? Progress has been made with both questions, but at the end of the day, there is still a need to explore methods for improving the properties of inference techniques in finite samples. In Chapter 8, we examine three methods for achieving this goal. The first is the use of the bootstrap to provide more accurate critical points, the second is an asymptotic theory which has been developed for the case in which some or all of the parameters are weakly identified, and the third is an asymptotic theory in which the HAC estimator converges to a random matrix.

7

Moment Selection in Theory and in Practice

A researcher is typically faced with a large set of alternatives from which to choose the q elements of the population moment condition. This choice can be made in an ad hoc fashion, but it is clearly preferable to base selection of $f(.)$ upon statistical criteria which reflect the ultimate purpose of the analysis. Throughout this chapter, we focus almost exclusively on the common case in which the objective is to make inferences about θ_0 based on the asymptotic distribution theory developed in previous chapters. From this perspective, the optimal choice of moment condition is the score vector because the resulting GMM estimator is the MLE and the latter is known to be asymptotically efficient in the class of consistent uniformly asymptotically normal estimators.[1] Unfortunately, as argued in Section 1.1, ML is infeasible in the types of model listed in Table 1.1. Therefore, if any useful guidance is to be provided for these settings then optimality must be judged relative to the class of moment conditions employed in practice for the model under consideration. There have been two distinct phases to the literature on moment selection within the GMM framework. From the mid-1980s until the mid-1990s, attention focused on the use of theoretical arguments to characterize the optimal choice of moment condition within the class of GMM estimators known as *Generalized Instrumental Variables*. More recently, attention has focused on data based methods for moment selection using information criteria. Both phases of the literature are reviewed in this chapter.

To begin this discussion, it is useful to consider what properties it is desirable for the selected moment condition to possess. Using the material from the previous chapters, it is argued in Section 7.1 that the selected moment condition should satisfy three conditions: the *orthogonality condition*, the *efficiency condition* and the *non-redundancy condition*. The latter two are most naturally considered together and their combination is refered to as the *relevance*

[1] See Section 3.8.

condition. In addition, this section describes both the ways in which moment selection complicates the concept of identification, and also how the use of the data in moment selection has the potential to contaminate subsequent inferences about θ_0.

Section 7.2 reviews the available results on the efficient choice of moment condition within Generalized Instrumental Variables (GIV) estimation. Within this class of problems, the optimal choice of moment condition is found by characterizing the optimal choice of instrument vector. The choice of instrument vector involves two decisions: which elements of the information set should be used? and, which functions of these elements (or variables) should be used as instruments? Since the first question depends on the particular model under consideration, the answer varies from case to case. Therefore, with few exceptions, the literature on optimal instruments has focused exclusively on the second question. It turns out that the optimal functional form is relatively straightforward to characterize in static models, but far more problematic in dynamic models. The relative simplicity of the solution in static models makes it far easier to develop an intuition for the form of the optimal instrument in this context. It is therefore instructive to start by considering the static case, and then to use these results as a stepping stone to the dynamic models which are the focus of this book. Accordingly, we split our discussion into two parts with Section 7.2.1 covering the static case and Section 7.2.2 considering the extension to the dynamic case. In either case, the optimal functional form depends on aspects of the data generation process which are typically unknown in practice. One possible way forward is to estimate these unknown features of the data generation process from the sample, and then substitute these estimates into the formula for the optimal instrument. However, these auxilliary estimations encounter a number of practical problems which are also described in Section 7.2. In fact, these problems tend to be of sufficient magnitude that the "optimal instrument" is rarely used in applications. Therefore, this literature is best viewed as providing an efficiency bound for GIV estimators rather than a practical method for instrument selection. This bound can be used to compare the efficiency of GIV with other estimators, and Section 7.2.3 provides a brief review of the available results of this type.

In contrast to the setting just described, a researcher must decide which moments to choose without knowledge of the underlying data generation process. In such circumstances, moment selection must perforce be based upon the data, and this is a key feature of all the methods recently proposed in the literature. These methods are reviewed in Section 7.3. Section 7.3.1 deals with selection based on the orthogonality condition and Section 7.3.2 deals with selection based on the relevance condition. Section 7.3.3 discusses their sequential use to provide a practical method for moment selection and illustrates it using Hansen and Singleton's (1982) consumption based asset pricing model. The methods reviewed in Section 7.3.1–7.3.3 can be applied to any GMM estimator that satisfies the types of regularity condition in Chapter 3. There has also been some related work within the more restrictive setting of GIV estimation. These methods are briefly reviewed in Section 7.3.4.

7.1 Preliminaries

To consider the problem of moment selection, it is necessary to introduce some additional notation. It is assumed that the candidate set of scalar functions which can form the basis for the population moment condition is finite. It is convenient to stack these scalar functions into a single vector $f_{max}(.)$ whose dimension is denoted by q_{max}. Following Andrews (1999), we use a $(q_{max} \times 1)$ selection vector c to denote which elements of the candidate set are included in a particular moment condition. We therefore now index $f(.)$ by c; $c_j = 1$ implies the j^{th} element of $f_{max}(.)$ is included in $f(.;c)$, and $c_j = 0$ implies this element is excluded. Note that $|c| = c'c$ equals the number of elements in $f(.;c)$. The set of all possible selection vectors is denoted by C, that is

$$\begin{aligned} C \quad = \quad & \{\, c \in \Re^{q_{max}};\ c_j \,=\, 0, 1,\ \text{for } j = 1, 2, \ldots q_{max}, \\ & \qquad \text{and } c = (c_1, c_2, \ldots c_{q_{max}}),\ |c| \geq p \,\} \end{aligned}$$

Below we use c_{sel} to denote the element of c that indexes the "selected" moment condition. For the present, we need not concern ourselves with how this element is selected.

To assess what properties are desirable for the selected moment condition, it is necessary to consider the objective of the estimation. Throughout this chapter, we follow the empirical literature and assume that this objective is to make inferences about θ_0 based on the two step (or iterated) estimator using the GMM asymptotic distribution theory developed in previous chapters.[2] For purposes of discussion, it is useful to restate the appropriate version of Theorem 3.2 in terms of the notation used here. Accordingly, define $\hat{\theta}_T(c)$ to be the GMM estimator based on $E[f(v_t, \theta; c)] = 0$, and let $V_\theta(c)$ denote the matrix $[G_0(c)'S(c)^{-1}G_0(c)]^{-1}$ where $G_0(c) = E[\partial f(v_t, \theta_0; c)/\partial\theta']$ and $S(c) = \lim_{T\to\infty} Var[T^{-1/2}\sum_{t=1}^{T} f(v_t, \theta_0; c)]$. The distributional result in Theorem 3.2 can then be restated as

$$T^{1/2}[\hat{\theta}_T(c) \,-\, \theta_0] \ \xrightarrow{d}\ N\,(0,\, V_\theta(c)\,) \tag{7.1}$$

Our list of three desirable properties for the selected moment condition arises from a consideration of the first and second moment properties of this asymptotic distribution, and of its quality as an approximation to finite sample behaviour.

The distribution in (7.1) has a mean of zero, and so embodies the assumption that the GMM estimator is consistent for the true value θ_0. From Section 3.4, it is clear that Assumptions 3.3 plays a crucial role in the derivation of this result, and so it is desirable for this condition to be satisfied by the selected vector. This observation leads to the following condition.

[2] It should be noted that while this objective is common to many of the studies in Table 1.1, it is not shared by all. For example, in some cases the main focus of the study is a point estimate of a particular parameter and this may necessitate alternative criterion for moment selection; see Section 7.3.4 for further discussion.

Definition 7.1 Orthogonality Condition
The selected moment condition satsifies Assumption 3.3, that is $E[f(v_t, \theta_0; c_{sel})] = 0$.

If this condition is satisfied, then the asymptotic distribution can be viewed as having the desirable first moment properties.

In most cases, there is more than one element of C which yields a moment condition satisfying the orthogonality condition. It is clearly most desirable to base inference on the moment condition with the smallest variance in a matrix sense. This leads to the following efficiency condition.

Definition 7.2 Efficiency Condition
The selected moment condition is efficient, that is $V_\theta(c) - V_\theta(c_{sel})$ *is positive semi-definite for all* $c \in C$ *such that* $E[f(v_t, \theta_0; c)] = 0$.

If this condition is satisfied, then the asymptotic distribution can be viewed as having desirable second moment properties.

It can be recalled from Theorem 6.1 that asymptotic variance can never increase as q increases. Therefore, the efficiency condition can be met by basing the estimation upon the moment condition consisting of all elements of the candidate set that satisfy the orthogonality condition. However, simulation evidence indicates that the inclusion of redundant moment conditions can lead to a deterioration in the quality of the asymptotic approximation to finite sample behaviour.[3] This consideration motivates the non-redundancy condition.

Definition 7.3 Non-Redundancy Condition
No individual element of $E[f(v_t, \theta_0; c_{sel})] = 0$ *is redundant given the remaining elements.*

It should be noted that, to date, there are no theoretical results on the determinants of the quality of the asymptotic approximation in nonlinear dynamic models that might provide a basis for selecting moments so that this approximation is good. Selection based on non-redundancy is best viewed, therefore, as a way of avoiding a situation in which the quality of the approximation can be very bad.

Both the efficiency and non-redundancy conditions relate to the asymptotic variance of the estimator, and so it proves useful to treat them simultaneously on occasion in moment selection. For expositional brevity, we refer to this combination as the relevance condition.

Definition 7.4 Relevance Condition
The selected moment condition is said to be relevant for the estimation of θ_0 *if it satisfies both the efficiency and non-redundancy conditions.*

The remainder of this chapter focuses on methods for moment selection based on the conditions above. Section 7.2 reviews the literature on the characterization of choice of moment condition that satisfies the efficiency condition in a

[3] See Section 6.3

class of Generalized Instrumental Variables estimators. Sections 7.3.1 and 7.3.2 describe methods for moment selection based on the orthogonality condition and relevance condition respectively, and Section 7.3.3 considers their combined use in a sequential fashion.

It might be wondered why the identification condition (Assumption 3.4) did not enter the discussion, particularly since this assumption played a crucial role in the analysis in Chapter 3. The reason is that the issue of identification is going to become much more complex once allowance is made for moment selection. In Chapters 3 and 4, all the analysis is conditional on a *given* choice of $f(.)$. For example, the model is said to be correctly specified if there exists a value θ_0 such that $E[f(v_t, \theta_0)] = 0$, and θ_0 was said to be identified if there is no other value of θ which satisfies this moment condition. The setting here is different because, by its very nature, moment selection means we must consider different choices of $f(.)$. Now it is entirely possible for two different choices of moment condition to satisfy the orthogonality condition at different parameter values, that is both $E[f(v_t, \theta_1; c_1)] = 0$ and $E[f_2(v_t, \theta_2; c_2)] = 0$. As seen below, each of the proposed moment selection methods involves its own particular assumption about identification that must hold if the method is to have the desired properties.

We conclude this section by considering a further way in which moment selection complicates the analysis. The methods described in Section 7.3 are based on the data. However, once the data are employed in this way, a potential problem emerges. All the asymptotic theory developed in Chapters 3 and 5 is premised on the assumption that $f(.)$ is fixed a priori. If $f(.)$ is selected from the data then the choice of moment condition may be random, and hence the asymptotic properties of the resulting estimator would depend on the statistical properties of the selection method. From a practical perspective, it is simpler by far if we can proceed with our inference about θ_0 as if the selected moment condition had been fixed a priori. We refer to this requirement as the *inference condition*. This issue has been addressed in the literature by providing conditions under which the data based selection vector, $\hat{c}_T$ say, converges in probability to a constant vector because in this case the validity of the inference condition can be deduced from the following lemma due to Pötscher (1991).

Lemma 7.1 Sufficient Conditions for the Inference Condition
Let $\hat{c}_T, c_0 \in C$ and let $h_T(c)$ be any statistic based on $E[f(v_t, \theta_0; c)] = 0$. If $\hat{c}_T \xrightarrow{p} c_0$ then $h_T(\hat{c}_T) - h_T(c_0) = o_p(1)$.

If $h_T(\hat{c}_T) - h_T(c_0) = o_p(1)$ then $h_T(\hat{c}_T)$ has the same asymptotic properties as $h_T(c_0)$. This lemma provides a theoretical justification for proceeding with inference as if c has been set equal to c_0 a priori but, as Pötscher (1991) observes, this result must be interpreted with some caution. The convergence is only pointwise, and Pötscher (1991) shows that the convergence may not be uniform in some cases of interest. As a result, the asymptotic distribution of $h_T(c_0)$ may provide a very poor approximation to the distribution of $h_T(\hat{c}_T)$ even in large samples. Pötscher (1991) demonstrates this lack of uniform convergence in an example where the dimension of the parameter vector is related to the dimension

of the model. To date, it is unclear whether these arguments translate to the setting here in which the dimension of the parameter vector is independent of the moment selection. In the absence of a theoretical resolution, the only guidance available is from simulation studies and these studies are reviewed on a case by case basis below. Finally, it is useful to introduce an item of terminology. It is customary in the model selection literature to say that $\hat{c}_T$ is consistent for c_0 to describe the situation in which $\hat{c}_T \xrightarrow{p} c_0$, and we follow this practice below.

7.2 The Optimal Instrument

In this section, we restrict attention to a class of GMM estimators known as Generalized Instrumental Variables (GIV) for which the efficient choice of moment condition is characterized by finding the efficient choice of instrument. It is customary to refer to this efficient choice as the "optimal instrument" for reasons that become apparent, and we follow this practice. Although our running empirical illustration is actually an example of GIV, we have not yet discussed this particular class of GMM estimators in its general form.[4] Since the structure of the associated population moment condition is crucial for our analysis here, we begin by providing a formal definition of the GIV estimator.

Within the GIV framework, the population moment condition is based on the statistical orthogonality of two vectors. These two vectors are denoted here by $u_t(\theta_0)$ and z_{t-m}. The vector $u_t(\theta_0)$ consists of functions of the data and the unknown parameter vector, and satisfies the conditional moment restriction

$$E[u_t(\theta_0)|\Omega_{t-m}] = 0 \tag{7.2}$$

where Ω_{t-m} is the infomation set at time $t-m$ for some non-negative integer m. The exact definitions of Ω_{t-m} and m depend on the assumptions about the dynamic structure, and so are provided below on a case by case basis. In applications, (7.2) represents the information derived from the underlying economic/statistical model. The instrument vector z_{t-m} consists of a vector of functions of elements of the information set, and so satisfies

$$z_{t-m} \in \Omega_{t-m} \tag{7.3}$$

Using an iterated expectations argument,[5] equations (7.2) and (7.3) can be combined to deduce the population moment condition,

$$E[z_{t-m} \otimes u_t(\theta_0)] = 0 \tag{7.4}$$

Hansen and Singleton (1982) refer to GMM estimation based on (7.4) as *Generalized Instrumental Variables* estimation. In view of the genesis of (7.4), the researcher needs only to decide which z_{t-m} to use in order to implement GIV.

[4] GIV is also used to estimate the conditional capital asset pricing model (Section 1.3.3) and the inventory holdings model (Section 1.3.4).

[5] See Section 1.3.1.

Therefore, within this framework, the problem of moment selection reduces to one of instrument selection.

In the literature on optimal instruments, it is customary to work with a slightly modified version of the population moment condition.[6] Instead of (7.4), the population moment condition takes the form

$$E[f(v_t, \theta_0)] = E[Z_{t-m} u_t(\theta_0)] \; = \; 0 \tag{7.5}$$

where $u_t(\theta_0)$ is a $(s \times 1)$ vector of functions which satisfies (7.2), Z_{t-m} is a $(q \times s)$ matrix and $Z_{t-m} \in \Omega_{t-m}$. If GMM estimation is based on the population moment condition in (7.5) with the optimal choice of weighting matrix then it follows from Theorems 3.2 and 3.4 that

$$T^{1/2}(\hat{\theta}_T \; - \; \theta_0) \; \xrightarrow{d} \; N\,(0,\, V(Z)) \tag{7.6}$$

where

$$V(Z) = \{E[\left(\frac{\partial u_t(\theta_0)}{\partial \theta'}\right)' Z'_{t-m}]\; S_Z^{-1}\; E[Z_{t-m} \frac{\partial u_t(\theta_0)}{\partial \theta'}]\}^{-1} \tag{7.7}$$

for $S_Z = \lim_{T\to\infty} Var[T^{-1/2} \sum_{t=1}^{T} Z_{t-m} u_t(\theta_0)]$. Since this distribution is centred on zero by construction, the optimal choice of Z_t is the one which minimizes $V(Z)$ in a matrix sense.

Below we use the notation Z^0_{t-m} to denote the optimal instrument. Since this optimality is relative to the class of instruments which lead to an asymptotic distribution of the form (7.6)–(7.7), it is necessary that the optimal instrument satisfies the regularity conditions for Theorem 3.2. It is most convenient to impose these regularity conditions up front. Since our focus here is on the functional form of the optimal instrument, we adopt the following high level assumption.[7]

Assumption 7.1 Regularity Conditions for the Optimal Instrument
$f(v_t, \theta_0) = Z^0_{t-m} u_t(\theta_0)$ *satisfies the regularity conditions for Theorem 3.2.*

7.2.1 Static Models

For this part of our discussion, ergodicity is replaced by the following more restrictive assumption.

Assumption 7.2 Independence
$\{v_t;\; t = 1, 2, \ldots T\}$ *forms an independent sequence.*

Notice that Assumptions 3.1 and 7.2 together imply $\{v_t\}$ forms an independent and identically distributed process.

To proceed further, it is necessary to put some structure on the information set which appears in (7.2). Throughout this book, $\{v_t\}$ is taken to be a time

[6] This difference facilitates the analysis but makes no difference to the ultimate result.

[7] More primitive conditions can be found in either Gallant (1987), Newey (1993) (for the iid case) or Wooldridge (1994).

series. Assumption 7.2 implies that v_t is independent of the history of the process, $V_{t-1} = (v_{t-1}, v_{t-2}, \ldots)$. Furthermore, by construction V_{t-1} is observable at time t and so must lie in the information set. However, the information set must contain more than this if GMM estimation is to work here. To see why, consider the static linear model in Chapter 2 and suppose that x_{t-1} is used as an instrument for x_t. In this case, it follows from Assumption 2.3 that the condition for identification is $rank\{E[x_{t-1}x_t^{'}]\} = p$. However, if v_t, and hence x_t, is i.i.d. then

$$E[x_{t-1}x_t^{'}] \;=\; E[x_{t-1}]E[x_t]^{'} \;=\; \mu_x\mu_x^{'}, \text{ say}$$

which is rank one by construction. To avoid this problem, it is necessary for the information set to contain some contemporaneous variables. Therefore, we partition v_t into $v_t = (v_{1,t}, v_{2,t})'$ and define the information set to be $\Omega_t = \{v_{2,t}, V_{t-1}\}$. This structure also means that expectations conditional on Ω_t are identical to those conditional on $v_{2,t}$, and so we use the notation $E[\,.\,|\,v_{2,t}]$ for $E[\,.\,|\,\Omega_t]$ below.

The optimal choice of Z_t is given by the following theorem.

Theorem 7.1 The Optimal Choice of Instrument in Static Models
If (i) v_t satisfies Assumptions 3.1 and 7.2; (ii) Assumption 7.1 holds with $m = 0$; then the optimal choice of Z_t in (7.5) is given by

$$Z_t^0 \;=\; K\,E[\partial u_t(\theta_0)/\partial\theta' \mid v_{2,t}]'\Sigma_{u|v_2}^{-1}$$

where K is any $(p \times p)$ nonsingular matrix of finite constants and $\Sigma_{u|v_2} = E[u_t(\theta_0)u_t(\theta_0)'|v_{2,t}]$. This optimal choice leads to a GMM estimator with asymptotic covariance matrix

$$V(Z^0) \;=\; \left\{E\left[\,E[\partial u_t(\theta_0)/\partial\theta' \mid v_{2,t}]'\Sigma_{u|v_2}^{-1}E[\partial u_t(\theta_0)/\partial\theta' \mid v_{2,t}]\right]\right\}^{-1}$$

Proof:
Let $\hat{\theta}_T(Z)$ denote the GIV estimator based on (7.5) with the optimal weighting matrix, and $\hat{\theta}_T(Z^0)$ denote the GIV estimator based on (7.5) with $Z_t = Z_t^0$. Notice that Z_t^0 is $(p \times s)$ and so the choice of weighting matrix is immaterial in this case.

The proof rests on using

$$\hat{\theta}_T(Z) \;=\; \hat{\theta}_T(Z^0) \;+\; [\hat{\theta}_T(Z) \;-\; \hat{\theta}_T(Z^0)] \tag{7.8}$$

to derive an explicit formula for $V(Z) - V(Z^0) = D(Z)$. It is then shown that $D(Z)$ is positive semi-definite for any choice of Z, which establishes the desired result.

The matrix $D(Z)$ depends on certain asymptotic variances and covariances. It is most convenient to define these terms prior to the derivation. These definitions rest on the random vectors which determine the asymptotic distributions of $\hat{\theta}_T(Z)$ and $\hat{\theta}_T(Z^0)$. From (3.26) it follows that

$$T^{1/2}[\hat{\theta}_T(Z) \;-\; \theta_0] \;=\; T^{-1/2}\sum_{t=1}^{T} m_t(Z) \;+\; o_p(1) \tag{7.9}$$

where

$$m_t(Z) \;=\; -[F_Z(\theta_0)'F_Z(\theta_0)]^{-1}F_Z(\theta_0)'S_Z^{-1/2}Z_t u_t(\theta_0) \tag{7.10}$$

and $F_Z(\theta_0) = S_Z^{-1/2}E[Z_t\partial u_t(\theta_0)/\partial\theta']$. Similarly, the corresponding expression for the optimal GIV estimator is given by

$$T^{1/2}[\hat{\theta}_T(Z^0) - \theta_0] \;=\; T^{-1/2}\sum_{t=1}^{T} m_t(Z^0) \;+\; o_p(1) \tag{7.11}$$

where

$$m_t(Z^0) \;=\; -\{E[Z_t^0\partial u_t(\theta_0)/\partial\theta']\}^{-1}Z_t^0 u_t(\theta_0) \tag{7.12}$$

and we have set $K = I_p$ without loss of generality.[8] The derivation of $D(Z)$ is most readily understood if we adopt a notation which explicitly reflects the variance/covariance nature of the terms. Accordingly, we introduce the following definitions

$$\begin{aligned}
Avar[\hat{\theta}_T(Z)] &= \lim_{T\to\infty} Var[T^{-1/2}\sum_{t=1}^{T} m_t(Z)] \\
Avar[\hat{\theta}_T(Z^0)] &= \lim_{T\to\infty} Var[T^{-1/2}\sum_{t=1}^{T} m_t(Z^0)] \\
Acov[\hat{\theta}_T(Z),\hat{\theta}_T(Z^0)] &= \lim_{T\to\infty} E[T^{-1}\sum_{t=1}^{T} m_t(Z)\{\sum_{t=1}^{T} m_t(Z^0)\}'] \\
Avar[\hat{\theta}_T(Z) - \hat{\theta}_T(Z^0)] &= \lim_{T\to\infty} Var[T^{-1/2}\sum_{t=1}^{T} d_t(Z)] \\
Acov[\hat{\theta}_T(Z^0),\hat{\theta}_T(Z) - \hat{\theta}_T(Z^0)] &= \lim_{T\to\infty} E[T^{-1}\sum_{t=1}^{T} m_t(Z^0)\{\sum_{t=1}^{T} d_t(Z)\}'] \\
&= C, \text{ say}
\end{aligned}$$

where $d_t(Z) = m_t(Z) - m_t(Z^0)$ and the "A" prefix stands for asymptotic. Notice that $Avar[\hat{\theta}_T(Z)]$ and $Avar[\hat{\theta}_T(Z^0))]$ are just the matrices $V(Z)$ and $V(Z^0)$ given in (7.7) and Theorem 7.1.

We are now in a position to derive $D(Z)$. From (7.8), it follows that

$$T^{1/2}[\hat{\theta}_T(Z) - \theta_0] \;=\; T^{1/2}[\hat{\theta}_T(Z^0) - \theta_0] + T^{1/2}[\hat{\theta}_T(Z) - \hat{\theta}_T(Z^0)]$$

and so

$$\begin{aligned}
Avar[\hat{\theta}_T(Z)] \;=\;& Avar[\hat{\theta}_T(Z^0)] + Avar[\hat{\theta}_T(Z) - \hat{\theta}_T(Z^0)] \\
&+ C \;+\; C'
\end{aligned} \tag{7.13}$$

[8] Note that $V(Z^0)$ is invariant to K.

Equation (7.13) can be rearranged to show that

$$D(Z) \;=\; Avar[\hat{\theta}_T(Z) \,-\, \hat{\theta}_T(Z^0)] \,+\, C \,+\, C' \tag{7.14}$$

Since $Avar[\hat{\theta}_T(Z) - \hat{\theta}_T(Z^0)]$ is positive semi-definite by construction, it is sufficient to establish that $C = 0$ in order for $D(Z)$ to be positive semi-definite. So we now focus on the matrix C.

From the definition of a covariance, it follows that

$$\begin{aligned} C &= Acov[\hat{\theta}_T(Z), \hat{\theta}_T(Z^0)] \,-\, Avar[\hat{\theta}_T(Z^0)] \\ &= Acov[\hat{\theta}_T(Z), \hat{\theta}_T(Z^0)] \,-\, V(Z^0) \end{aligned}$$

and so

$$C \,=\, 0 \iff Acov[\hat{\theta}_T(Z), \hat{\theta}_T(Z^0)] \,=\, V(Z^0)$$

It is at this stage that the static nature of the model is exploited because Assumption 7.2 implies $Acov[\hat{\theta}_T(Z), \hat{\theta}_T(Z^0)] \,=\, E[m_t(Z)m_t(Z^0)']$. Using an iterated conditional expectations argument, it follows that

$$\begin{aligned} E[m_t(Z)m_t(Z^0)'] &= [F_Z(\theta_0)'F_Z(\theta_0)]^{-1}F_Z(\theta_0)'S_Z^{-1/2}E[Z_t\, E[u_t(\theta_0)u_t(\theta_0)' \\ &\qquad \times|v_{2,t}]\; Z_t^{0\prime}]\{E[(\partial u_t(\theta_0)/\partial\theta')'(Z_t^0)']\}^{-1} \\ &= [F_Z(\theta_0)'F_Z(\theta_0)]^{-1}F_Z(\theta_0)'S_Z^{-1/2}E[Z_t\Sigma_{u|v_2}Z_t^{0\prime}] \\ &\qquad \times\{E[(\partial u_t(\theta_0)/\partial\theta')'(Z_t^0)']\}^{-1} \end{aligned} \tag{7.15}$$

Using the definition of Z_t^0, it follows that

$$E[Z_t\Sigma_{u|v_2}Z_t^{0\prime}] \;=\; E\,[Z_t\, E[\partial u_t(\theta_0)/\partial\theta' \mid v_{2,t}]] \tag{7.16}$$

Now,

$$\partial u_t(\theta_0)/\partial\theta' \;=\; E[\partial u_t(\theta_0)/\partial\theta' \,|\, \Omega_t] \;+\; A_t \tag{7.17}$$

where $E[A_t \,|\, \Omega_t] = 0$. Since $Z_t \in \Omega_t$, it follows from (7.16)-(7.17) that

$$\begin{aligned} E[Z_t\Sigma_{u|v_2}Z_t^{0\prime}] &= E\,[Z_t\, E[\partial u_t(\theta_0)/\partial\theta' \mid v_{2,t}]] \\ &= E\,[Z_t\, E[\partial u_t(\theta_0)/\partial\theta' \mid \Omega_t]] \\ &= E[Z_t\partial u_t(\theta_0)/\partial\theta'] \end{aligned} \tag{7.18}$$

Substsituting (7.18) into (7.15), we obtain

$$\begin{aligned} E[m_t(Z)m_t(Z^0)'] &= [F_Z(\theta_0)'F_Z(\theta_0)]^{-1}F_Z(\theta_0)'F_Z(\theta_0)\{E[(\partial u_t(\theta_0)/\partial\theta')'Z_t^{0\prime}]\}^{-1} \\ &= V(Z^0) \end{aligned}$$

where the last identity follows from (7.17) by similar logic to (7.18). Therefore $C = 0$ and so $D(Z)$ is positive semi-definite which establishes the desired result. ◇

One aspect of the proof is worth commenting on. Notice that $C = 0$ implies that $T^{1/2}[\hat{\theta}_T(Z^0) - \theta_0]$ is asymptotically uncorrelated with $T^{1/2}[\hat{\theta}_T(Z) - \hat{\theta}_T(Z^0)]$ for any other choice of instrument Z.[9]

At first sight, it is not obvious why Z_t^0 is the optimal instrument. To help develop an intuition for this result, we consider three simple examples involving linear models. The first example shows that Theorem 7.1 leads to an IV estimator which corresponds with the estimation approach proposed in the literature on linear simultaneous equations models. The second two examples illuminate the role of $\Sigma_{u|v_2}$ in the construction of Z_t^0. After these examples, it is shown how the intuition from linear models can also be used to understand the form of the optimal instrument in nonlinear models.

Example: Linear Model with $s = 1$ and Conditional Homoscedasticity
In Chapter 2, we consider the case in which the population moment condition takes the form,

$$E[z_t u_t(\theta_0)] \;=\; 0 \tag{7.19}$$

where $u_t(\theta_0) = y_t - x_t'\theta_0$. Notice that if we set $z_t = Z_{t-m}$ then (7.19) is a special case of (7.5). For our puposes here, it is sufficient to restrict attention to the case in which $p = 1$ and so x_t is a scalar. We also now add the restriction that $u_t(\theta_0)$ is conditionally homoscedastic, and denote this variance by $\Sigma_{u|v_2} = \sigma_0^2$. Since $\partial u_t(\theta_0)/\partial\theta = -x_t$, the optimal instrument is given by

$$Z_t^0 \;=\; -\frac{k}{\sigma_0^2} E[x_t \,|\, v_{2,t}]$$

for any non-zero finite constant k. However, since k can take any such value, we are free to set $k = -\sigma_0^2$ in which case the optimal instrument reduces to $Z_t^0 = E[x_t \,|\, v_{2,t}]$. In other words, the optimal instrument is just the part of x_t which can be explained by $v_{2,t}$. It can be verified that the resulting IV (GIV) estimation with $z_t = Z_t^0$ is identical to OLS estimation of θ_0 based on[10]

$$y_t \;=\; E[x_t \,|\, v_{2,t}]\theta_0 \;+\; \tilde{u}_t$$

The latter is described by Theil (1971)[p.452] as "an obvious estimation procedure" in his discussion of estimation in linear simultaneous equation models. ◇

Example: Linear Model with $s = 1$ and Conditional Heteroscedasticity
Suppose now that we modify the previous example by introducing conditional heteroscedasticity in $u_t(\theta_0)$, that is $E[u_t(\theta_0)^2 | v_{2,t}] = \sigma_t^2$, but leave all other

[9] West (2001) uses this property to characterize the optimal instrument, and considers conditions under which this holds in dynamic models.

[10] Recall that by definition of the conditional expectation, $x_t = E[x_t \,|\, v_{2,t}] + e_t$ where $E[e_t | v_{2,t}] = 0$.

aspects of the specification the same. In this case, the optimal instrument takes the form

$$Z_t^0 \;=\; -\frac{k}{\sigma_t^2} E[x_t \,|\, v_{2,t}]$$

for any non–zero finite constant k. This time, it is not possible to eliminate σ_t^2 by judicious choice of k – although we can set $k = -1$ to remove the minus sign. In this case, $\Sigma_{u|v_2}^{-1}$ scales $E[\partial u_t(\theta_0)/\partial\theta | v_{2,t}]$ to take account of the conditional heteroscedasity in $u_t(\theta_0)$. $\diamond$

Example: Linear Model with $s = 2$ and Conditional Homoscedasticity Now suppose that

$$u_t(\theta_0) \;=\; \begin{bmatrix} y_{1,t} - x_{1,t}^{'}\theta_{0,1} \\ y_{2,t} - x_{2,t}^{'}\theta_{0,2} \end{bmatrix}$$

where $\theta_0 = (\theta_{0,1}^{'}, \theta_{0,2}^{'})^{'}$, $\theta_{0,i}$ is $p_i \times 1$ and assume that $u_t(\theta_0)$ is conditionally homoscedastic with

$$E[u_t(\theta_0) u_t(\theta_0)^{'} \,|\, v_{2,t}] \;=\; \Sigma_0$$

With this specification, the components of $Z^0(.)$ are given by

$$\begin{aligned} E\left[\frac{\partial u_t(\theta_0)}{\partial \theta'} \;|\; v_{2,t} \right] &= E\left[\begin{array}{cc} -x_{1,t}^{'} & 0_{p_2}^{'} \\ 0_{p_1}^{'} & -x_{2,t} \end{array} \;\Big|\; v_{2,t} \right] \\ \Sigma_{u|v_2}^{-1} &= \Sigma_0^{-1} \end{aligned}$$

where 0_a is the $a \times 1$ null vector. In this case, $\Sigma_{u|v_2}^{-1}$ scales $E[\partial u_t(\theta_0)/\partial\theta' | v_{2,t}]$ to take account of any difference between the variances of $u_{1,t}(\theta_0)$ and $u_{2,t}(\theta_0)$, and any covariance between $u_{1,t}(\theta_0)$ and $u_{2,t}(\theta_0)$. $\diamond$

These examples provide an intuition for the structure of Z_t^0 in linear models. To develop a comparable understanding for the nonlinear model, it is necessary to explain why Z_t^0 depends on $\partial u_t(\theta_0)/\partial\theta'$. The explanation can be found by comparing the determinants of the asymptotic behaviour of $T^{1/2}(\hat{\theta}_T - \theta_0)$ in linear and nonlinear models. To simplify the exposition, we consider the case in which $s = 1$; we also introduce "L" and "N" subscripts on $\hat{\theta}_T$ to distinguish the linear and nonlinear cases. For the linear model in Chapter 2, we have[11]

$$T^{1/2}(\hat{\theta}_{L,T} - \theta_0) = \{(T^{-1}X'Z)W_T(T^{-1}Z'X)\}^{-1}(T^{-1}X'Z)W_T(T^{-1/2}Z'u) \tag{7.20}$$

For the nonlinear model, it can be shown that[12]

$$\begin{aligned} T^{1/2}(\hat{\theta}_{N,T} - \theta_0) \;=\; & -\{[\,T^{-1}D(\theta_0)'Z]W_T[T^{-1}Z^{'}D(\theta_0)\,]\}^{-1} \\ & \times [\,T^{-1}D(\theta_0)'Z\,]W_T\,T^{-1/2}Z^{'}u(\theta_0) \;+\; o_p(1) \end{aligned} \tag{7.21}$$

[11] See equation (2.23).
[12] See equations (7.9)–(7.10).

where $D(\theta_0)$ is the $T \times p$ matrix with t^{th} row $\partial u_t(\theta_0)/\partial\theta'$, Z is the $T \times q$ matrix with t^{th} row Z_t, and $u(\theta_0)$ is the $T \times 1$ vector with t^{th} element $u_t(\theta_0)$. A comparison of (7.20) and (7.21) reveals that the asymptotic behaviour of $T^{1/2}(\hat{\theta}_{N,T} - \theta_0)$ is identical to that of $T^{1/2}(\hat{\theta}_{L,T} - \theta_0)$ in a model with regressor vector, $x_t = -\partial u_t(\theta_0)/\partial\theta'$ and error, $u_t(\theta_0)$. This equivalence can be used to translate the intuition from linear models to their nonlinear counterparts.[13]

While Theorem 7.1 characterizes the optimal instrument, it does not by itself solve the problem of instrument selection. The function Z_t^0 depends on $E[\partial u_t(\theta_0)/\partial\theta' \mid v_{2,t}]$ and, in most cases, $\Sigma_{u|v_2}$ as well; neither of these functions are typically part of the specification of the underlying economic/statistical model. Therefore, Z_t^0 is an infeasible choice of instrument. One natural solution is to estimate the components of Z_t^0 from the data. In some cases, this approach may be plausible. One such case is the linear model in our first example above in which case the feasible optimal IV estimator is just the Two Stage Least Squares estimator as we now illustrate.

Example: Linear Model with $s = 1$ and Conditional Homoscedasticity (continued)

To construct a feasible optimal instrument, it is necessary to specify a functional form for $E[x_t|v_{2,t}]$. Therefore, we assume that x_t is itself generated by a linear regression model,

$$x_t \;=\; v_{2,t}'\gamma_0 \;+\; e_t \tag{7.22}$$

where $E[e_t|v_{2,t}] = 0$ and $E[e_t^2|v_{2,t}] \;=\; \tau_0^2$. With this specification, the optimal instrument is

$$Z_t^0 \;=\; v_{2,t}'\gamma_0$$

where we have set $k = -\sigma_0^2$ as discussed above. To construct a feasible counterpart to Z_t^0, it is necessary to estimate γ_0. Under the above conditions, it is natural to estimate γ_0 via Ordinary Least Squares applied to (7.22). If this is done then the resulting IV estimator of θ_0 is

$$\tilde{\theta}_T \;=\; \frac{x'V_2(V_2'V_2)^{-1}V_2'y}{x'V_2(V_2'V_2)^{-1}V_2'x}$$

in the obvious notation. This estimator can be recognized as the Two Stage Least Squares (2SLS) estimator of θ_0 which is familiar from the linear simultaneous equations model literature.[14] Using similar arguments to Section 2.3, it can be shown that $\tilde{\theta}_T$ has the same asymptotic distribution as the IV estimator of θ_0 with $z_t = Z_t^0$. Therefore, the 2SLS can be interpreted as the feasible optimal instrumental variable estimator within this model.[15] ◇

[13] Recall that a similar linearization of the moment condition lay behind the construction of the identifying restrictions; see Section 3.4.2.

[14] See Theil (1971)[p.451-454].

[15] This result also applies if $p > 1$.

In this example, the construction of Z_t^0 rests crucially on the assumption that $E[x_t|v_{2,t}]$ is linear. This specification may be very natural in some contexts – such as the linear simultaneous equations model – but may not be so appropriate in others. A comparable approach in nonlinear models would require an assumption about the conditional mean of $\partial u_t(\theta_0)/\partial\theta'$. Unfortunately, this is unlikely to be an aspect of the data generation process specified by the underlying economic model. One alternative is to use non-parametric methods to approximate this expectation. However, since our ultimate focus is dynamic models, we do not explore these methods further here. Instead we refer the interested reader to Newey (1990) or the survey in Newey (1993).[16]

7.2.2 Dynamic Models

A number of papers consider the extension of Theorem 7.1 to dynamic models. As might be imagined, the characterization of the optimal instrument depends crucially on specific assumptions about the dynamic structure of certain aspects of the data generation process. In many cases of interest, economic theory provides little guidance on these aspects, and even in cases where the economic model provides this type of information, the construction of a feasible version of the optimal instrument is intractable. Consequently, there have been few attempts to implement GIV with the optimal instrument in the types of model in Table 1.1. In view of this, there seems little value to reproducing here the very technical analysis needed to rigorously justify the functional form of the optimal instrument in dynamic nonlinear models. Instead, we focus on two relatively simple dynamic structures, and present only heuristic arguments. Our discussion rests heavily on the framework developed in Hansen (1985), and, to a lesser extent, the earlier work by Hansen and Sargent (1982) and Hayashi and Sims (1983); the interested reader is referred to these sources for the more technical details.[17]

Throughout this sub-section, the information set, Ω_{t-m}, is assumed to contain the information in the series up until time $t-m$. However, to present a more formal definition, it is necessary to place additional structure on v_t. In our notation, v_t represents the vector of random variables which appear in the population moment condition. In many cases in which GIV is applied to dynamic models, the instruments are lagged values of variables which appear in $u_t(\theta_0)$. For example, in Hansen and Singleton's (1982) consumption based asset pricing model, $u_t(\theta_0)$ depends on $x_{1,t+1} = c_{t+1}/c_t$ and $x_{2,t+1} = r_{t+1}/p_t$, and in our empirical implementation, the instrument vector contained the constant and lagged values of $x_{1,t+1}$ and $x_{2,t+1}$.[18] This approach is so commmon in practice that we lose little generality by assuming it is followed here. Accordingly, we partition $v_t = (v_{1,t}', v_{2,t}')'$ and assume that $v_{2,t}$ contains functions of lagged values of $v_{1,t}$. In this case, the information set is defined to be: $\Omega_{t-m} = \{v_{1,t-m}, v_{1,t-m-1}, \ldots\}$.

[16] Note that Newey (1993) considers this issue in the context of cross section data and so his information consists of $v_{2,t}$ alone.

[17] Also see Hansen, Heaton, and Ogaki (1988), Bates and White (1990) and West (2001).

[18] See Section 3.2.

We consider the form of the optimal instrument under two assumptions about the dynamic structure of $u_t(\theta_0)$. In the first, $u_t(\theta_0)$ is a martingale difference with respect to Ω_{t-1} and so $Z_{t-1}u_t(\theta_0)$ is an uncorrelated sequence. In the second, $u_t(\theta_0)$ is a VMA(n) process, and so $Z_{t-n-1}u_t(\theta_0)$ is a n-dependent process. We start with the simplest case.

Assumption 7.3 Martingale Difference with Respect to Ω_{t-1}
$u_t(\theta_0)$ is a martingale difference sequence with respect to Ω_{t-1}.

One consequence of this assumption is that $f(v_t, \theta_0) = Z_{t-1}u_t(\theta_0)$ is a serially uncorrelated process, and it is this property which is important here. An inspection of the proof of Theorem 7.1 reveals that the serial independence of $\{v_t\}$ is only important because it implies $\{f(v_t, \theta_0)\}$ is serially uncorrelated. Therefore, Theorem 7.1 extends directly to the martingale difference case.

Theorem 7.2 The Optimal Choice of Instrument in Dynamic Models (i): $u_t(\theta_0)$ is a Martingale Difference with Respect to Ω_{t-1}
If (i) v_t satisfies Assumptions 3.1 and 3.8; (ii) Assumptions 7.1 (with $m = 1$) and 7.3 hold then the optimal choice of Z_{t-1} in (7.5) is given by

$$Z_{t-1}^0 = K\, E[\partial u_t(\theta_0)/\partial\theta' \mid \Omega_{t-1}]'\Sigma_{t-1}^{-1}$$

where K is any $(p \times p)$ nonsingular matrix of finite constants and $\Sigma_{t-1} = E[u_t(\theta_0)u_t(\theta_0)'|\Omega_{t-1}]$. This optimal choice leads to a GMM estimator with asymptotic covariance matrix

$$V(Z^0) \;=\; \left\{E\left[\, E[\partial u_t(\theta_0)/\partial\theta' \mid \Omega_{t-1}]'\Sigma_{t-1}^{-1}E[\partial u_t(\theta_0)/\partial\theta' \mid \Omega_{t-1}]\,\right]\right\}^{-1}$$

Just as before, the optimal instrument is infeasible because it depends on unknown aspects of the data generation process. In view of the relative simplicity of the dynamic structure, it might be hoped that it is possible to construct a feasible counterpart to Z_{t-1}^0. However, this hope is misplaced in most cases of interest. The following example illustrates the problems encountered.

Example: Hansen and Singleton's (1982) Consumption Based Asset Pricing Model
It can be recalled from Section 1.3.1 that $u_t(\theta_0)$ is a martingale difference sequence in Hansen and Singleton's (1982) version of the consumption based asset pricing model. In our earlier discussion of this model, we denoted the key random variables c_{t+1}/c_t and r_{t+1}/p_t by $x_{1,t+1}$ and $x_{2,t+1}$. We continue this practice here. This means $x_{t+1} = (x_{1,t+1}, x_{2,t+1})'$ plays the role of $v_{1,t}$ in our discussion above, and so, for consistency, we denote the information set by Ω_t instead of Ω_{t-1}. With these adjustments in notation, Theorem 7.2 implies the optimal instrument depends on:

$$\begin{aligned} E[\partial u_t(\theta_0)/\partial\theta \mid \Omega_t] &= E\left[\begin{array}{c} \delta_0 log(x_{1,t+1})(x_{1,t+1})^{\gamma_0-1}x_{2,t+1} \\ (x_{1,t+1})^{\gamma_0-1}x_{2,t+1} \end{array} \Big| \,\Omega_t\right] \\ \Sigma_t &= E[(\delta_0 x_{1,t+1}^{\gamma_0-1}x_{2,t+1} - 1)^2 \mid \Omega_t] \end{aligned}$$

To calculate these two components of Z_t^0 directly involves knowledge of certain aspects of the conditional distribution of x_{t+1} given Ω_t. A review of Section 1.3.1 reveals that these aspects are not specified as part of the underlying economic model. There are two natural ways forward. First, in the spirit of 2SLS, models can be assumed for $E[\partial u_t(\theta_0)/\partial\theta' \mid \Omega_t]$ and Σ_t. Secondly, the conditional distribution of x_{t+1} can be estimated and then the relevant conditional expectations can be approximated using a numerical integration technique such as quadrature.[19] We consider these in turn.

The first approach shares both the strengths and weaknesses of 2SLS. If the assumed models for $E[\partial u_t(\theta_0)/\partial\theta' \mid \Omega_t]$ and Σ_t are correct then their estimated versions can be used to construct a feasible optimal instrument. If the assumed specifications are wrong, then clearly the resulting instrument while feasible is not optimal. Unfortunately, as mentioned above, economic theory provides little, if any, guidance on suitable specifications.

The second approach raises two problems. First, it requires precisely the type of distributional assumption which the use of GMM was supposed to avoid.[20] Secondly, the estimation is likely to be computationally very burdensome. Both these problems are sufficient by themselves to make the use of a feasible optimal instrument unattractive. A further disincentive is provided in a simulation study reported by Tauchen (1986).[21] In the controlled simulation environment, the data generation process is known and so the calculations are more straightforward, although still computationally burdensome. He finds that in many cases the numerical optimization routine failed to converge when the optimal instrument was used. Therefore, the attempted use of the feasible optimal instrument undermined the estimation completely. ◇

From an analytical perspective, the martingale difference case represents the best possible scenario because $\{f(v_t, \theta_0) = Z_{t-1}u_t(\theta_0)\}$ is a serially uncorrelated process and so Theorem 7.1 translates directly. Once serial correlation is introduced, the form of the optimal instrument must change. To illustrate how, we now consider the following case.

Assumption 7.4 Moving Average Case
$u_t(\theta_0)$ is generated by the following VMA(n) process

$$u_t(\theta_0) = \Lambda(L)e_t = e_t + \Lambda_1 e_{t-1} + \Lambda_2 e_{t-2} + \ldots + \Lambda_n e_{t-n}$$

where $\{e_t\}$ satisfies $E[e_t|\Omega_{t-i}] = 0$ and $Var[e_t|\Omega_{t-i}] = I_s$ for all $i > 0$, and the roots of $det[\Lambda(s^*)] = 0$ lie outside the unit circle.

Under this assumption $u_t(\theta_0)$ is a homoscedastic, invertible VMA(n) process.[22] With this specification, $E[u_t(\theta_0) \,|\, \Omega_{t-k}]$ is zero for $k > n$ but is non-zero in general for $k \leq n$. Therefore, we set $m = n + 1$ in (7.5).

[19] See Tauchen (1985*b*, 1986), Tauchen and Hussey (1991), Ghysels and Hall (1993).
[20] See Sections 1.1 and 1.3.1.
[21] See Section 6.3 for further discussion of this study.
[22] The assumptions of invertibility and conditional homoscedasticity can be relaxed; see Hansen, Heaton, and Ogaki (1988) and Heaton and Ogaki (1991).

Although Theorem 7.2 does not directly apply to this new setting, we can exploit it here as part of the following three-step strategy for deducing the form of the optimal instrument.

Step 1: Transform $u_t(\theta_0)$ into a process $\tilde{u}_t(\theta_0)$ so that $Z_{t-n-1}\tilde{u}_t(\theta_0)$ is a serially uncorrelated process.

Step 2: Use Theorem 7.2 to characterize the optimal instrument in the transformed model.

Step 3: Reverse the transformation to deduce the form of the optimal instrument in the untransformed model from the result in Step 2.

To execute this strategy, we must find the appropriate transformation. As we review possible candidates, there is one implicit consequence of the assumed specification which plays a particularly important role, and so it is useful to highlight this feature at the outset. Since $\Omega_t = \{v_{1,t}, v_{1,t-1}, \ldots\}$, $Z_t \in \Omega_t$ implies $Z_t \in \Omega_{t+j}$ for $j > 0$ but it does not imply that $Z_t \in \Omega_{t-j}$ for $j > 0$. In other words, Z_t is not strictly exogenous.[23] The key consequence of this structure is that

$$E[Z_{t-n-1}u_i(\theta_0)] = 0 \quad \text{for } i \geq t \tag{7.23}$$

$$\neq 0 \quad \text{for } i < t_0 \text{ and some } t_0 < t \tag{7.24}$$

The obvious first candidate for the transformation is $\Lambda(L)^{-1}$ because the resulting process $\Lambda(L)^{-1}u_t(\theta_0)$ equals e_t.[24] However, a closer inspection reveals this filter does not meet our requirements here. Setting $\Lambda(L)^{-1} = 1 + \sum_{i=1}^{\infty} \bar{\Lambda}_i L^i$, it follows that

$$\begin{aligned} E[Z_{t-n-1}\Lambda(L)^{-1}u_t(\theta_0)] &= E[Z_{t-n-1}\{u_t(\theta_0) + \bar{\Lambda}_1 u_{t-1}(\theta_0) + \bar{\Lambda}_2 u_{t-2}(\theta_0) + \ldots\}] \\ &= E[Z_{t-n-1}u_t(\theta_0)] + \bar{\Lambda}_1 E[Z_{t-n-1}u_{t-1}(\theta_0)] \\ &\quad + \bar{\Lambda}_2 E[Z_{t-n-1}u_{t-2}(\theta_0)] + \ldots \end{aligned} \tag{7.25}$$

Using (7.23)–(7.24) to evaluate this expectation, it is apparent that

$$E[Z_{t-n-1}\Lambda(L)^{-1}u_t(\theta_0)] \neq 0$$

Therefore, GIV estimation based on the assumption that this expectation is zero would lead to an inconsistent estimator of θ_0.

The problem here clearly stems from that backward nature of the filter $\Lambda(L)^{-1}$.[25] Fortunately, this is not the only type of filter which can be used to remove the autocovariance structure of $u_t(\theta_0)$. Hayashi and Sims (1983)

[23] See Engle, Hendry, and Richard (1983) for a discussion of various types of exogeneity.

[24] The filter $\Lambda(L)^{-1}$ is actually an infinite order polynomial in L and so would have to approximated by a finite order polynomial in practice but this can be ignored here.

[25] The filter is said to operate "backwards in time" because the filtered value of $u_t(\theta_0)$ only depends on its current and past values.

suggest using the forward filter $\Lambda(L^{-1})^{-1} = 1 + \sum_{i=1}^{\infty} \tilde{\Lambda}_i L^{-i}$.[26] This filter not only removes the autocovariance structure but also produces a sequence which is still orthogonal to Z_{t-n-1}.[27] To see this, let $\tilde{u}_t(\theta_0) = \Lambda(L^{-1})^{-1} u_t(\theta_0)$ and observe that

$$\begin{aligned} E[Z_{t-n-1}\tilde{u}_t(\theta_0)] &= E[Z_{t-n-1}\{u_t(\theta_0) + \tilde{\Lambda}_1 u_{t+1}(\theta_0) + \tilde{\Lambda}_2 u_{t+2}(\theta_0) + \ldots\}] \\ &= E[Z_{t-n-1}u_t(\theta_0)] + \tilde{\Lambda}_1 E[Z_{t-n-1}u_{t+1}(\theta_0)] \\ &\quad + \tilde{\Lambda}_2 E[Z_{t-n-1}u_{t+2}(\theta_0)] + \ldots \end{aligned} \tag{7.26}$$

Using (7.23), it is straightforward to deduce from (7.26) that $E[Z_{t-n-1}\tilde{u}_t(\theta_0)] = 0$. In view of this, it is the forward filter which is used here to remove the autocovariance structure.

The next stage in the analysis involves the characterization of the optimal instrument in the transformed model. The transformation ensures that $Z_{t-n-1}\tilde{u}_t(\theta_0)$ is a serially uncorrelated process and so we can appeal to Theorem 7.2 to deduce that the optimal choice of Z_{t-n-1} in the transformed model is given by

$$\tilde{Z}^0_{t-n-1} = E[\partial\tilde{u}_t(\theta_0)/\partial\theta' \mid \Omega_{t-n-1}]'\{E[\tilde{u}_t(\theta_0)\tilde{u}_t(\theta_0)'|\Omega_{t-n-1}]\}^{-1} \tag{7.27}$$

It only remains to reverse the transformation in order to deduce the form of the optimal instrument in the original model. At first glance, this objective would appear to be met by premultipying $\tilde{Z}^0_t$ by $\Lambda(L^{-1})$. However, this is not so because $\Lambda(L^{-1})\tilde{Z}^0_t \notin \Omega_{t-n-1}$ due to the forward nature of the filter. Instead, Hansen (1985) shows that the appropriate transformation is given by the $\Lambda(L)^{-1}$, and so the optimal instrument can be calculated via the recursion

$$Z^0_t = \Lambda_1 Z^0_{t-1} + \Lambda_2 Z^0_{t-2} + \ldots + \Lambda_n Z^0_{t-n} + \tilde{Z}^0_t \tag{7.28}$$

To construct this optimal instrument in practice, it would be necessary to truncate the infinite order filter $\Lambda(L)^{-1}$. Therefore, Hansen (1985) suggests using (7.28) with $Z^0_i = 0$ for $i = 0, -1, \ldots - n$.

For completeness, we summarize the previous discussion in the following lemma.[28]

Lemma 7.2 The Optimal Choice of Instrument in Dynamic Models (ii): Moving Average Case
If (i) v_t satisfies Assumptions 3.1 and 3.8; (ii) Assumptions 7.1 (with $m = n+1$) and 7.4 hold then the optimal choice of Z_{t-n-1} in (7.5) is given by

$$Z^0_{t-n-1} = K\,\Lambda(L)^{-1}\tilde{Z}^0_{t-n-1}$$

[26] Again we will ignore the infinite nature of the filter for the time being and concentrate on showing that the technique solves our problem. This filter is said to act "forwards in time" because the filtered value of $u_t(\theta_0)$ is a function of its current and future values.

[27] See Hayashi and Sims (1983) for further discussion of the properties of forward filters.

[28] We omit the characterization of associated asymptotic variance because the resulting expression is very complicated, and provides no additional insights; see Hansen (1985) for further details.

where K is any $(p \times p)$ nonsingular matrix of finite constants,

$$\tilde{Z}^0_{t-n-1} = E[\partial \tilde{u}_t(\theta_0)/\partial \theta' \mid \Omega_{t-n-1}]' \{E[\tilde{u}_t(\theta_0)\tilde{u}_t(\theta_0)' | \Omega_{t-n-1}]\}^{-1}$$

and $\tilde{u}_t(\theta_0) = \Lambda(L^{-1})^{-1} u_t(\theta_0)$.

To date, this result has had little, if any, impact on the empirical literature because of the complexity of the calculations involved. If n is known, then the estimation of $\Lambda(L)$ is conceptually straightforward but nevertheless computationally burdensome.[29] If n is unknown then this burden is increased by the need to estimate the order of the VMA. The estimation of $E[\partial \tilde{u}_t(\theta_0)/\partial \theta' \mid \Omega_{t-n}]$ is problematic for all the reasons described in the martingale difference case above.

While the estimation of Z^0_{t-n-1} is fraught with problems, there are grounds for anticipating that, under some circumstances, an indirect approach may yield an IV estimator which achieves the efficiency bound implied by Lemma 7.2. In order to be able to elaborate on this statement, it is useful to consider first the form of the optimal instrument in a simple example.[30]

Example: Univariate Linear Regression Model with MA(1) Errors
Suppose that $u_t(\theta_0) = y_t - x_t\theta_0$ and $p = 1$ so that x_t is a scalar. Let $u_t(\theta_0)$ satisfy Assumption 7.4 with $n = 1$. In this case, there is only one moving average parameter which is denoted by λ here for simplicity, and so $\Lambda(L) = 1 + \lambda L$. The forward filter is then

$$\Lambda(L^{-1})^{-1} \;=\; 1 \;-\; \lambda L^{-1} \;+ \lambda^2 L^{-2} \;-\; \ldots \tag{7.29}$$

Using (7.29) and $\partial u_t(\theta_0)/\partial \theta = -x_t$, it follows that

$$E[\partial \tilde{u}_t(\theta_0)/\partial \theta \,|\, \Omega_{t-2}] \;=\; -\,E[x_t \;-\; \lambda x_{t+1} \;+ \lambda^2 x_{t+2} \;-\; \ldots \,|\, \Omega_{t-2}] \tag{7.30}$$

To proceed further, it is necessary to make an assumption about the data generation process for x_t. So we now suppose that

$$\begin{aligned} x_t &= \pi w_t \;+\; e_{x,t} \\ w_t &= \psi w_{t-1} \;+\; e_{w,t} \end{aligned}$$

where $w_t \in \Omega_{t-2}$, $|\psi| < 1$, $\{e_{i,t}\}$ is i.i.d. for $i = x, w$, $E[e_{x,t}|\Omega_{t-2}] = 0$, and $E[e_{w,t}|w_{t-1}, w_{t-2}, \ldots] = 0$.[31] Note that this specification implies

$$\begin{aligned} x_{t+m} &= \pi w_{t+m} \;+\; e_{x,t+m} \\ &= \pi \left\{ \psi^m w_t \;+\; \sum_{i=0}^{m-1} \psi^i e_{w,t+m-i} \right\} \;+\; e_{x,t+m} \end{aligned}$$

[29] Recall that this burden motivated the use of a VAR approximation to a VARMA process in den Haan and Levin's (1996) covariance matrix estimator; see Section 3.5.2.

[30] This example is based on personal correspondence from Ken West, and I am very grateful for his permission to use it here.

[31] Note that for this specification to be logically consistent, w_t cannot be a lagged value of either x_t or y_t. Therefore, we must modify our definition of the information set used above to include a third variable w_t.

Therefore, it follows that

$$\begin{aligned} E[\partial\tilde{u}_t(\theta_0)/\partial\theta \,|\, \Omega_{t-2}] &= -\pi\left\{w_t - \lambda\psi w_t + (\lambda\psi)^2 w_t - \ldots\right\} \\ &= -\pi w_t/(1+\lambda\psi) \end{aligned}$$

and so the optimal instrument is

$$\begin{aligned} Z^0_{t-2} &= (1+\lambda L)^{-1}\gamma w_t \\ &= \gamma w_t - \gamma\lambda w_{t-1} + \gamma\lambda^2 w_{t-2} - \ldots \end{aligned} \tag{7.31}$$

where $\gamma = -\pi/(1+\lambda\psi)$. $\diamond$

In this example, it is necessary to estimate π, λ and ψ in order to construct a feasible version of Z^0_{t-2}. However, the structure of (7.31) suggests an alternative approach may be viable. Since Z^0_{t-2} is a linear function of $\{w_{t-j};\ j = 0, 1, 2, \ldots\}$, the "optimal" population moment condition $E[Z^0_{t-2}u_t(\theta_0)] = 0$ is implied by the set of population moment conditions $\{E[w_{t-j}u_t(\theta_0)] = 0;\ j = 0, 1, \ldots\}$. Therefore, Hayashi and Sims (1983) suggest bypassing Z^0_{t-2} and estimating θ_0 from the population moment condition

$$E[z_t(q_T)u_t(\theta_0)] = 0$$

where $z_t(q_T) = (w_t, w_{t-1}, w_{t-2}, \ldots w_{t-q_T})'$. Hayashi and Sims (1983) argue that if the optimal weighting matrix is used and $q_T \to \infty$ with T then the resulting estimator is as asymptotically efficient as the estimator based on the optimal instrument.[32] In spite of its intuitive appeal, this conclusion should be treated with some caution. Hayashi and Sims's (1983) analysis is premised on the assumption that both estimators have an asymptotic normal distribution, but their analysis does not consider the rate at which q_T must increase in order for this to be true.[33] Nevertheless, it seems plausible that the result holds under certain conditions both in the linear regression model case considered by Hayashi and Sims (1983), and also in nonlinear models as well.

7.2.3 Efficiency Comparison with Maximum Likelihood

It is remarked above that GIV estimation is only undertaken in situations in which Maximum Likelihood is infeasible. In view of this background, intuition suggests that the resulting GIV estimator is less efficient asymptotically than the Maximum Likelihood estimator. Although there have been only a few formal comparisons of the two methods in the literature, the previous statement is most likely a good guide. Nevertheless, there are a couple of exceptions which are worth noting. Both involve linear models and Maximum Likelihood under a normality assumption. The first case is the linear simultaneous equation models in which 2SLS is as asymptotically efficient as the Limited Information

[32] Hayashi and Sims (1983) analysis is confined to linear regression models but allows for $p > 1$.

[33] See Section 6.1.3.

Maximum Likelihood; see Theil (1971)[p.507]. The second case is univariate ARMA(m,n) models. Stoica, Söderström, and Friedlander (1985) show that an IV estimator of the AR parameters is asymptotically as efficient as Maximum Likelihood.[34] Linearity plays an important role in such results, and this type of equivalence is unlikely to extend to nonlinear models. To date, the only results available are in the context of nonlinear simultaneous equation models with a normality assumption on the errors. In this context, Jorgenson and Laffont (1974) and Amemiya (1977) show that IV is indeed less efficient asymptotically than Maximum Likelihood.

However, there is a sense in which GIV estimation based on the optimal instrument is the best we can do given the information available. Chamberlain (1987) shows that, in static models, the matrix $V(Z^0)$ in Theorem 7.1 represents a lower bound on the asymptotic covariance matrix of *any* consistent and asymptotically normal estimator of θ_0 in which the only substantive information used in estimation is the population moment condition in (7.5).[35]

7.3 Moment Selection in Practice

Once Maximum Likelihood is ruled out, the extant results on optimal moment selection do not provide a practical solution to the problem of moment selection. An immediate problem is that results have only been obtained for the class of moment conditions associated with GIV estimation. However, even in this case, the practical value is limited for three reasons. First, the results characterize the efficient member out of the set of moment conditions which satisfy the orthogonality condition, but provide no guidance on how to identify this reference set. Secondly, it turns out that the construction of the optimal instrument is computationally burdensome and requires the assumptions about aspects of the data generation process which are typically not specified as part of the underlying economic model. Thirdly, the optimal instrument has desirable asymptotic properties but there is no guarantee that this translate to desirable finite sample properties. Therefore, in this section, we consider methods of moment selection which are arguably of more practical relevance. Sections 7.3.1 and 7.3.2 discuss methods for moment selection based on the orthogonality condition and relevance condition respectively, and Section 7.3.3 considers a method of moment selection based on their sequential use. This section also contains an application of the methods to Hansen and Singleton's (1982) consumption based asset pricing model. Section 7.3.4 reviews some related methods which have been proposed for Instrumental Variables estimators.

[34] Also see Hansen and Singleton (1991, 1996).

[35] Chamberlain's (1987) analysis is based on a form of semiparametric Maximum Likelihood estimation known as Empirical Likelihood; see Section 10.2.

7.3.1 Selection Based on the Orthogonality Condition

To implement a data based model selection based on this criterion, it is necessary to find a statistic which can indicate whether or not the orthogonality condition is satisfied. The obvious candidate is the overidentifying restrictions test statistic, J_T, given in equation (5.2). For, although we did not use this terminology in Section 5.1, it can be recognized that the orthogonality condition is in fact the null hypothesis of this test. Andrews (1999) considers a number of ways in which this statistic can be used as a basis for moment selection, and derives their statistical properties. In this section, we concentrate on Andrews's (1999) information criterion based approach because simulation evidence suggests this method works best. However, the other methods are briefly discussed at the end of this sub–section.

Information criterion have been applied to the problem of model selection in a wide variety of settings. In the case here, the criterion is the sum of two terms: the overidentifying restrictions test and a "bonus" term which reflects the number of overidentifying restrictions. This criterion is evaluated for all possible choices of moment condition, and then the selected moment condition is the one which minimizes the criterion. To express this idea mathematically, it is necessary to index the overidentifying restrictions test statistic by c, the selection vector introduced in Section 7.1. Therefore, we define $J_T(c)$ to be equal to J_T in (5.2) evaluated at $f(.) = f(.; c)$. The *moment selection criterion* takes the form

$$MSC(c) \;=\; J_T(c) \;+\; B(T, |c|) \tag{7.32}$$

where $B(T, |c|)$ is the aforementioned bonus term. The selected moment condition is given by $\hat{c}_T$, the choice of c which minimizes the criterion, that is

$$\hat{c}_T \;=\; argmin_{c \in C}\; MSC(c) \tag{7.33}$$

Although the minimization is defined over C, Andrews (1999) observes that it may be more appropriate to consider a reduced set of possibilities in certain circumstances. For example, in our consumption based asset pricing example, all the moment conditions are derived from the same Euler condition. If one such condition is invalid, then the underlying model is wrong and it makes little sense to base the estimation upon only those moment conditions that appear valid. In this case, an argument can be made for testing the validity of the candidate set alone. In other cases, moment conditions may be naturally associated with different aspects of the underlying specification, and so it may be desired to assess the validity of different groups of moment conditions using $MSC(c)$. For example, in the stochastic volatility model in Section 1.3.5, different moment conditions are associated with different aspects of the assumed distribution of the series.

In spite of the previous remarks, we focus on the limiting properties of $\hat{c}_T$ as defined in (7.33) and what they imply about this method of moment selection.[36] In order to develop this analysis, we must: (i) make assumptions

[36] It is relatively straightforward to modify the analysis to accommodate minimization over a restricted set of possibilities, and so this is left to the interested reader.

about the limiting behaviour of $J_T(c)$; (ii) specify the properties of the bonus term, $B(T,|c|)$; (iii) impose certain identification conditions. We address these three in turn below.

In Section 5.1, it is shown that the overidentifying restrictions test statistic converges to a χ^2_{q-p} distribution if the null hypothesis is correct, but diverges to infinity if the null is invalid. It is important that both these properties hold here. However, for the specification of the bonus term below, the rate of divergence is also important. It can be recalled from Theorem 5.2 and 5.3 that rate of divergence depends on the way in which the long run variance is estimated. Following Andrews (1999), it is assumed here that this variance is estimated using a mean correction discussed in Section 4.3 so that the resulting estimator is consistent regardless of whether or not the orthogonality condition is satisfied.[37] Therefore, we impose the following high level assumption on the overidentifying restrictions test; more primitive conditions are given in Theorems 5.1 and 5.2.

Assumption 7.5 Regularity Conditions for $J_T(c)$
(i) If $E[f(v_t,\theta;c)] = 0$ for a unique $\theta \in \Theta$ then $J_T(c) \xrightarrow{d} \chi^2_{|c|-p}$; (ii) if $E[f(v_t,\theta;c)] = \mu(\theta) \neq 0$ for all $\theta \in \Theta$ then $T^{-1}J_T(c) \xrightarrow{p} a(c)$ where $a(c)$ is a finite postive constant dependent on c.

The bonus term takes the form

$$B(T,|c|) = -h(|c|)\kappa_T \tag{7.34}$$

Its constituents are assumed to satisfy the following conditions.

Assumption 7.6 Regularity Conditions for the Bonus Term
(i) $h(.)$ is strictly increasing; (ii) $\kappa_T \to \infty$ as $T \to \infty$ and $\kappa_T = o(T)$.

Notice that under these conditions, the bonus term decreases as $|c|$ increases, and so, since $MSC(c)$ is minimized, rewards selection vectors which include more elements from the candidate set. To implement the method, it is necessary to choose specific functions for $h(.)$ and κ_T. We consider two choices here, both of which are suggested by earlier work on order selection in autoregressive time series. The first involves

$$h(|c|) = |c| - p \qquad \text{and} \qquad \kappa_T = ln(T) \tag{7.35}$$

and corresponds to the BIC proposed by Schwarz (1978). The second involves

$$h(|c|) = |c| - p \qquad \text{and} \qquad \kappa_T = b\,ln[ln(T)] \tag{7.36}$$

where b is a finite constant greater than 2. Andrews (1999) recommends setting $b = 2.01$. These choices of $h(.)$ and κ_T correspond to those proposed by Hannan

[37] Hall, Inoue, and Peixe (2003) show that the consistency result in Theorem 7.3 still holds if the long run variance is estimated using an uncentred HAC with appropriate modification of Assumption 7.5. However, we assume here that a centred covariance matrix is used because this is the way in which the method is normally implemented.

and Quinn (1979), and so the implied criterion is often denoted HQIC. A third popular choice in the autoregressive time series literature is the AIC proposed by Akaike (1974), but the analogous choice of κ_T does not satisfy Assumption 7.6. Therefore we do not consider it further at this stage, but, in view of its popularity, return to it once we have established the properties of selection methods based on bonus terms which do satisfy Assumption 7.6.

As explained in Section 7.1, there are going to be two layers to the necessary identification conditions. First, there must be a unique c which minimizes the population analog to (7.33). Secondly, given this choice of c, the orthogonality condition must be satisfied at a unique value of θ – or in other words, the parameter vector must be identified by the selected moment condition. Conditions for the latter have already been presented in Section 3.1. So our focus here is on the identification of the selection vector. It is useful to derive the appropriate condition in steps. To begin, recall from above that different choices of $f(.)$ can satisfy the orthogonality condition at different values of θ. Therefore, we define $\mathcal{Z}^0$ to be the set of selection vectors for which $f(.;c)$ satisfies the orthogonality condition for some parameter value, that is

$$\mathcal{Z}^0 \;=\; \{\, c \in C \text{ such that } E[f(v_t, \theta; c)] = 0 \text{ for some } \theta \in \Theta \,\}$$

From this set, we need to distinguish those selection vectors which include the most elements of the candidate set, that is

$$\mathcal{MZ}^0 \;=\; \left\{\, c \in \mathcal{Z}^0 \text{ such that } |c| \geq |c^*| \text{ for all } c^* \in \mathcal{Z}^0 \,\right\}$$

For the population analog to (7.33) to have a unique minimum, this set must contain only one vector which we denote by c_o below. Perforce, this condition implies $|c_o| > p$.[38] We now impose this condition along with the requisite condition for parameter identification.

Assumption 7.7 Identification Conditions
(i) $\mathcal{MZ}^0 = \{c_o\}$*; (ii)* $E[f(v_t, \theta_0; c_o)] = 0$ *and* $E[f(v_t, \theta; c_o)] \neq 0$ *for all* $\theta \in \Theta \setminus \{\theta_0\}$.

With these conditions in place, the limiting behaviour of $\hat{c}_T$ is given by the following theorem.

Theorem 7.3 Consistency of $\hat{c}_T$
If Assumptions 7.5-7.7 hold then $\hat{c}_T \xrightarrow{p} c_o$.

Before presenting the proof, we note that this theorem combined with Lemma 7.1 imply that[39]

$$T^{1/2}[\hat{\theta}_T(\hat{c}_T) \,-\, \theta_0] \;\xrightarrow{d}\; N(\,0,\, V_\theta(c_o)) \tag{7.37}$$

Proof of Theorem 7.3:
Notice that the stated result holds if it can be shown that $MSC(c_o) < MSC(c)$

[38] In general, there is a $\theta(c)$ which satisfies $E[f(v_t, \theta(c); c)] = 0$ for any c such that $|c| = p$; see the preamble to Chapter 4.

[39] See the discussion of Lemma 7.1.

for any $c \neq c_o$ with probability one in the limit as $T \to \infty$. To establish the latter, it suffices to consider just two cases: (i) $c = c_1$ where $E[f(v_t, \theta_1; c_1)] = 0$ but $c_1 \neq c_o$; (ii) $c = c_2$ where $E[f(v_t, \theta; c_2)] \neq 0$ for any $\theta \in \Theta$. Notice that these two scenarios cover all other possibilities apart from $c = c_o$. We now consider them in turn.

To simplify the notation, we define $\Delta_T(c, c_o) = MSC(c) - MSC(c_o)$. From (7.32) it follows that

$$\Delta_T(c_1, c_o) \;=\; J_T(c_1) \;+\; B(T, |c_1|) \;-\; \{\, J_T(c_o) \;+\; B(T, |c_o|)\}$$

Using (7.34) and Assumption 7.5(i), we have

$$\Delta_T(c_1, c_o) \;=\; O_p(1) \;+\; [h(|c_o|) \;-\; h(|c_1|)]\kappa_T$$

As remarked above, Assumption 7.7(i) implies that $|c_o| > |c_1|$ and so

$$\Delta_T(c_1, c_o) \;=\; O_p(1) \;+\; k\kappa_T \tag{7.38}$$

where $k > 0$. The desired result then follows from (7.38) and Assumption 7.6(ii).

Now consider $\Delta_T(c_2, c_o)$. From (7.32), it follows that

$$T^{-1}\Delta_T(c_2, c_o) \;=\; T^{-1}\,\{J_T(c_2) \;+\; B_T(T, |c_2|) \;-\; J_T(c_o) \;-\; B_T(T, |c_o|)\}$$

Using Assumptions 7.5 and 7.6, it can be seen that

$$T^{-1}\Delta_T(c_2, c_o) \;=\; a(c_2) \;+\; o_p(1) \tag{7.39}$$

Since $a(c_2) > 0$ from Assumption 7.5(ii), the desired result is established. $\diamond$

Andrews (2000) reports simulation evidence on the finite sample behaviour of these methods in the context of a static linear regression model estimated by IV; see Chapter 2. Within his design, there are five regressors and the candidate set consists of eight instruments *i.e.* $p = 5$ and $q_{max} = 8$. Various parameter settings are used in which either seven or all eight instruments satisfy the orthogonality condition. The minimization in (7.33) is performed over a restricted set to make the computations managable: three cases are considered involving respectively 8, 12 and 17 possible selection vectors. The evidence suggests the model selection procedure works well for some parameter settings but not for others. The problems appear to stem from failure in the identification conditions, and we consider the ramifications of such failure in an example below. The evidence suggests that $MSC(c)$ works marginally better with the bonus term associated with BIC given in (7.35). Another feature of the design is also pertinent: the maximum degree of overidentification is 3. Hall and Peixe (2003) reports simulation results for a similar linear regression model estimated by IV in which all instruments satisfy the orthogonality condition and the maximum degree of overidentification is 7. Within their design p equals one, all the instruments satisfy the orthogonality condition but six are redundant

given the other two. Their evidence indicates that $MSC(c)$ tends to select all the orthogonal instruments with high probability as would be expected. However, the inclusion of the redundant instruments leads to a deterioration of the finite sample properties of the estimator relative to the one based on just the two non–redundant instruments. This finding motivates moment selection based on the relevance condition which is the topic of the next sub-section.

In view of its familiarity in other contexts, it is worth considering the properties of $MSC(c)$ with the bonus term associated with Akaike's (1974) information criterion (AIC), that is

$$h(|c|) \; = \; |c| \; - \; p \qquad \text{and} \qquad \kappa_T \; = \; 2 \tag{7.40}$$

It can be seen that this choice of bonus term does not satisfy Assumption 7.6 because κ_T does not tend to infinity with T. A review of the proof to Theorem 7.3 indicates that (7.39) still holds, and so the method selects moment conditions which satisfy the orthogonality condition with probability one. However, since $\kappa_T \not\to \infty$ with T, (7.38) no longer implies that $\Delta(c_1, c_o) > 0$ with probability one. Instead, the selected vector is random in the limit – a result which parallels Shibata's (1976) finding that AIC overfits the order of autoregressive time series with non-zero probability in the limit. Andrews (2000) finds this method performs worst in his simulation study.

We conclude our discussion of $MSC(c)$ by considering the consequences of identification failure. These are best illustrated within the context of a simple example.

Example: Identification Failure in a Linear Model
Consider the linear model

$$\begin{aligned} y_t &= x_t\theta_0 \; + \; u_t \\ x_t &= w_{1,t}\pi_1 \; + \; w_{2,t}\pi_2 \; + \; e_t \end{aligned}$$

where all variables are scalars. Once again, we define $u_t(\theta) = y_t - x_t\theta$. The candidate set of instruments is constructed from the (8×1) vector w_t whose i^{th} element is $w_{i,t}$. The stochastic behaviour of the model depends on $n_t^{'} = [u_t, e_t, w_t^{'}]$, and its assumed that $n_t \sim N(0, \Sigma)$ where Σ has $i - j^{th}$ element $\sigma_{i,j}$ and lower triangular elements

$$\begin{aligned} \sigma_{i,j} &= 1 \qquad \text{for } i \; = \; j \\ &= \sigma_{ue} \neq 0 \qquad \text{for } (i,j) \; = \; (1,2) \\ &= 0 \qquad \text{else} \end{aligned}$$

The candidate set, $f_{max}(v_t, \theta)$, is assumed to consist of the (8×1) vector whose i^{th} element is $z_{i,t}(y_t - x_t\theta)$ and

$$\begin{aligned} z_{i,t} &= w_{i,t} \qquad \text{for } i \; = \; 1, 2, \ldots 6 \\ &= w_{i,t} \; + \; \delta_i u_t, \qquad \delta_i \neq 0, \text{ for } i \; = \; 7, 8 \end{aligned}$$

With this specification, it is immediately apparent that

$$\begin{aligned} E[z_{i,t}u_t(\theta_0)] &= 0 \quad \text{for } i = 1,2,\dots 6 \\ &\neq 0 \quad \text{for } i = 7,8 \end{aligned}$$

However, it is also the case that

$$\begin{aligned} E[z_{i,t}u_t(\theta_1)] &= 0 \quad \text{for } i = 3,4,\dots 8 \\ &\neq 0 \quad \text{for } i = 1,2 \end{aligned}$$

for $\theta_1 = \sigma_{ue}^{-1}$. Therefore Assumption 7.7(i) fails in this case because $\mathcal{MZ}^0 = \{c_1, c_2\}$ for $c_1 = (1,1,1,1,1,1,0,0)$ and $c_2 = (0,0,1,1,1,1,1,1)$. $\diamond$

If identification fails in this way then the consequences are dramatic. $\hat{c}_T$ converges to a random vector whose probability distribution attaches non-zero probability to both c_1 and c_2.[40] Furthermore, this non-degeneracy mainfests itself in the limiting behaviour of the estimator: $\hat{\theta}_T$ converges to a random variable $\theta_0(c)$ whose distribution takes the form: $\theta(c) = \theta_0$ with probability p_c and $\theta(c) = \sigma_{ue}^{-1}$ with probability $1 - p_c$. So in these circumstances moment selection has undermined the consistency of the estimator. One further aspect of this case is worth noting. The limiting distribution of $\hat{c}_T$ only attaches non–zero probability to selection vectors containing six non-zero elements. In this example, it can be verified that there are no instrument vectors containing seven or eight elements which would satisfy the orthogonality condition for some θ. This turns out to be a general result. Andrews (1999) shows that $|\hat{c}_T|$ converges in probability to the largest $|c|$ such that $E[f(v_t, \theta; c)] = 0$ for some θ. Since it is impossible to know a priori if the identification condition is satisfied, caution must be exercised in the use of this method of moment selection. One possible way forward is to use the method to identify $|c|$, and then examine the associated $J_T(c)$ for all permutations of c with this length. However, to date, no statistical theory is available to guide this investigation.

This concludes our discussion of MSC here, but we return to it in Section 7.3.3 where the method is illustrated using our running empirical example. We end this sub-section by briefly considering other methods of moment selection based on the overidentifying restrictions test.

In view of its hypothesis testing origins, it would seem natural to develop moment selection strategy based on the outcome of repeated applications of the overidentifying restrictions test. Andrews (1999) considers two such strategies known as "upward" and "downward" testing. As the names suggest, the only difference between them is the direction of testing. The upward sequence involves considering choices of $f(.)$ of dimension $(p + i \times 1)$ in the sequence $i = 1, 2, \dots$ until a significant overidentifying restrictions test is encountered. The downward sequence involves considering choices of $f(.)$ of dimension

[40] Using a special case of this design, Peixe (2000) finds the probabilities to be 0.564 and 0.436 for MSC based on the BIC bonus term in her simulations for sample size $T = 500$.

$(q_{max} - i \times 1)$ in the sequence $i = 0, 1, \ldots$ until an insignificant overidentifying restrictions test is encountered. In some cases, it may be necessary to consider all possible choices of a given dimension; in others, it may be possible to limit the number of permutations considered. Whichever sequence is used, this approach has the potential to uncover which elements of the candidate set satisfy the orthogonality condition. However, if a fixed significance level is used then this approach does not satisfy the inference condition. The problem is that, by construction, a 5% significance level implies the null hypothesis is falsely rejected with a probability of 0.05. This makes the outcome of either the upward or downward testing sequences random in repeated samples. One way around this problem is to employ a significance level, α_T, which decays to zero with T. Pötscher (1983) shows that a suitable rate of decay is given by $ln(\alpha_T) = o(T)$. Unfortunately, this type of rule does not indicate how to pick α_T in a given sample of size T. Andrews (2000) also reports simulation evidence for these methods using $\alpha_T = 0.276/ln(T)$ where the scaling factor is chosen to yield $\alpha_{250} = 0.05$. He finds the selection procedures work reasonably well, but are dominated by MSC with the bonus term associated with BIC. Therefore, we do not consider these methods further here. The interested reader is referred to Andrews (1999, 2000).[41]

7.3.2 Selection Based on the Relevance Condition

In this section, we describe an information criterion for moment selection based upon the relevance condition. When selection is based on the orthogonality condition, there is a natural choice of statistic to capture the sample information. With the relevance condition, it is not immediately obvious what constitutes the pertinent sample statistic. It can be recalled from Section 7.1 that the relevance condition is a combination of the the efficiency and non–redundancy conditions. Since both the latter conditions are statements about the asymptotic variance of the estimator, the sample analog of this variance is the natural basis for the sample information in an information criterion. However, this sample information must be a scalar and so it is necessary to find a suitable transformation of the variance. Hall, Inoue, Jana, and Shin (2003) show that the natural logarithm of the determinant of the variance is a natural candidate because it satisfies the following properties.

Lemma 7.3 Properties of $ln[|V_\theta(c)|]$
Let $c_i \in C$ *for* $i = 1, 2$. *If* $V_\theta(c_1) - V_\theta(c_2)$ *is positive semi-definite then* $ln[|V_\theta(c_1)|] - ln[|V_\theta(c_2)|] \geq 0$ *with the equality only holding if* $V_\theta(c_1) = V_\theta(c_2)$.

Accordingly, Hall, Inoue, Jana, and Shin (2003) propose the *relevant moment selection criterion*

$$RMSC(c) \;=\; ln[|\hat{V}_{\theta,T}(c)|] \;+\; P(T, |c|) \tag{7.41}$$

[41] Also see Hall, Inoue, and Peixe (2003).

where $\hat{V}_{\theta,T}(c) = [G_T(\hat{\theta}_T(c);c)'\hat{S}_T^{-1}(c)G_T(\hat{\theta}_T(c);c)]^{-1}$, and we have now indexed into $G_T(.)$ and $S_T(.)$ by c. Note that the covariance estimator $\hat{S}_T(c)$ must be consistent for $S(c)$ (using the obvious notation) but may depend on a preliminary estimator of θ_0. The penalty term is given by $P(T,|c|)$. The selected vector is the value which minimizes the criterion over C, that is

$$\tilde{c}_T \;=\; \text{argmin}_{c\in C} RMSC(c)$$

To analyse the asymptotic behaviour of $\tilde{c}_T$, it is necessary to make certain assumptions. As with our analysis if $\hat{c}_T$ in the previous sub-section, three types of conditions are required: (i) conditions on the sample statistic; (ii) conditions on the penalty term; (iii) identification conditions. In terms of the first of these, it is far more convenient here to adopt rather high level assumptions to streamline the discussion; more primitive conditions can be found in Chapter 3 or the references therein. With that caveat, we now present and discuss each of these three types of regularity condition in turn.

Assumption 7.8 Regularity Conditions for $\hat{V}_{\theta,T}(c)$
$\hat{V}_{\theta,T}(c) = V_\theta(c) + O_p(\tau_T^{-1})$ *where* $\tau_T \to \infty$ *as* $T \to \infty$.

Notice that the statement of this assumption makes explicit reference to the rate of convergence of the covariance matrix. This rate depends on the rate of convergence of the constituents of $\hat{V}_{\theta,T}(c)$. Under the assumptions in Chapter 3, it can be shown that $G_T(\hat{\theta}_T(c);c) = G_0 + O_p(T^{-1/2})$. However, the rate of convergence for $\hat{S}_T(c)$ depends on the form of the covariance matrix. If $\hat{S}_T(c)$ is the sum of a fixed number of autocovariances – such as $\hat{S}_{SU}$ – then it can be shown that $\hat{S}_T(c) = S + O_p(T^{-1/2})$. In this case, $\tau = T^{1/2}$. If $\hat{S}_T(c)$ is an HAC estimator, then $\hat{S}_T(c) = S + O_p((b_T/T)^{-1/2})$. In this case, $\tau_T = (T/b_T)^{1/2}$.[42] The exact rate is important because it determines the exact form of the penalty term.

Assumption 7.9 Regularity Conditions for $P(T,|c|)$
For $\bar{c} \in C$ *such that* $|\bar{c}| > |c|$, $\tau_T[P(T,|\bar{c}|) - P(T,|c|)] \to +\infty$ *as* $T \to \infty$ *and* $P(T,|c|) = o(1)$.

This assumption would be met by the choice

$$P(T,|c|) \;=\; (|c| - p)\ln(\tau_T)/\tau_T \tag{7.42}$$

which corresponds to the BIC-type criterion as discussed in the previous subsection.

As with $MSC(c)$, there are two layers to the identification condition: one involving the selection vector and one involving θ_0. The first of these identification conditions defines c_r to be the selection vector associated with the relevant subset. To formalize this definition it is necessary to introduce the following

[42] See Section 4.4 for further discussion.

sets: the set of selection vectors that are asymptotically efficient relative to the candidate set,

$$\mathcal{C} \;=\; \{c;\, V_\theta(\iota_{q_{max}}) = V_\theta(c),\, c \in C\}$$

where $\iota_{q_{max}}$ is a $q_{max} \times 1$ vector of ones; and also the subset of $\mathcal{C}$ containing the selection vectors of minimum length,

$$\mathcal{C}_{min} \;=\; \{c;\, c \in \mathcal{C},\, |c| \leq |\bar{c}| \text{ for all } \bar{c} \in \mathcal{C}\}$$

Using this notation, we impose the following identification conditions.

Assumption 7.10 Identification Condition
(i) $\mathcal{C}_{min} = \{c_r\}$*; (ii)* $E[f(v_t, \theta_0; c)] = 0$ *if and only if* $\theta = \theta_0$ *for any* $c \in C$*.*

Under these conditions, Hall, Inoue, Jana, and Shin (2003) establish the following result.

Lemma 7.4 Consistency of $\tilde{c}_T$
Under Assumptions 7.8–7.10,

$$\tilde{c}_T \xrightarrow{p} c_r. \tag{7.43}$$

The proof exploits Lemma 7.3 and follows similar lines to Theorem 7.3, and so is omitted for brevity. Note that this lemma combined with Lemma 7.1 imply that[43]

$$T^{1/2}[\hat{\theta}_T(\tilde{c}_T) \;-\; \theta_0] \;\xrightarrow{d}\; N(\,0,\, V_\theta(c_r)) \tag{7.44}$$

Hall, Inoue, Jana, and Shin (2003) report simulation evidence for RMSC in the context of IV estimation of linear regression model with a single regressor x_t and $q_{max} = 8$ so that the maximum degree of overidentification is seven.[44] Within their design, all the potential instruments satisfy the orthogonality condition but six are redundant given the other two. The evidence suggests that the performance of the method is sensitive to both the R^2 from the regression of x_t on the intruments and also the degree of endogeneity of the x_t. If the R^2 equals 0.5 then the method does a good job of identifying which moment conditions are informative about the regression parameter, and the behaviour of $\hat{\theta}_T(\tilde{c}_T)$ is well approximated by conventional asymptotic theory in samples of size $T = 100$. If the R^2 equals 0.1 then RMSC has problems identifying which moment conditions are informative about the regression parameter, and the behaviour of $\hat{\theta}_T(\tilde{c}_T)$ is not well approximated by conventional asymptotic theory in samples of size of $T = 100$.[45] However, by $T = 500$, the method performs much better and $\hat{\theta}_T(\tilde{c}_T)$ is well approximated by conventional asymptotic theory except for cases where x_t is highly endogenous.[46]

[43] See the discussion of Lemma 7.1.
[44] See Chapter 2.
[45] Also see Section 8.2.
[46] Here "highly endogenous" means the correlation between x_t and the error of the equation – u_t in the notation of Chapter 2 – is 0.9.

7.3.3 A Combined Strategy

In practice, the candidate set, $f_{max}(v_t, \theta_0)$, is most likely to contain some elements which satisfy the orthogonality condition and some which do not. Of these orthogonal instruments, only a subset may satisfy the relevance condition. Therefore, it is desirable to develop a method which selects moments on the basis of both the orthogonality and relevance conditions. Selection based on either MSC(c) or RMSC(c) alone cannot meet this objective because each is based on only one of the conditions. However, intuition suggests that a combination of the two methods should achieve the desired goal. This section explores the properties of such a selection strategy.

So we now assume that the candidate set is made up as follows.

Assumption 7.11 Candidate Set
$f_{max}(v_t, \theta) = [f(v_t, \theta; c_o)', f(v_t, \theta; c_*)']'$ where c_o is defined in Assumption 7.7, and $f(v_t, \theta; c_o) = [f(v_t, \theta; c_r)', f(v_t, \theta; c_i)']'$ where c_r is defined in Assumption 7.10.

For the sake of exposition, we assume that MSC is applied first, and then RMSC. The sequence does not affect the essence of the theoretical arguments below, but may potentially have consequences in finite samples in practice. Since $RMSC(c)$ is to be applied following $MSC(c)$, it is necessary to modify the definition of $\tilde{c}_T$ to reflect the fact that the minimization is over a candidate set delineated by the first selection criterion.[47] Accordingly, we define the set

$$\begin{aligned} \hat{C}_T \quad = \quad & \Big\{ c \in \Re^{|\hat{c}_T|};\ c_j \ = \ 0, 1, \text{ for } j = 1, 2, \ldots |\hat{c}_T| \\ & \qquad \text{and } c = (c_1, c_2, \ldots c_{|\hat{c}_T|}),\ |c| \geq p \Big\} \end{aligned}$$

and redefine $\tilde{c}_T$ as follows,

$$\tilde{c}_T^{seq} \ = \ \text{argmin}_{c \in \hat{C}_T} RMSC(c)$$

The following theorem establishes the consistency of this sequential method of moment selection.

Theorem 7.4 Consistency of $\tilde{c}_T^{seq}$
If Assumptions 7.5 – 7.11 hold then: $\tilde{c}_T^{seq} \xrightarrow{p} c_r$.

The proof follows directly from a combination of Theorem 7.3 and Lemma 7.4, and so is left to the reader. It follows directly from Theorem 7.4 and Lemma 7.1 that the asymptotic distribution of $\hat{\theta}_T(\tilde{c}_T^{seq})$ is given by[48]

$$T^{1/2}[\hat{\theta}_T(\tilde{c}_T^{seq}) \ - \ \theta_0] \ \xrightarrow{d} \ N(\,0,\, V_\theta(c_r)) \tag{7.45}$$

[47] See Section 7.3.1 for discussion of circumstances in which it is desirable to minimize $MSC(c)$ over subsets of C.

[48] This assumes $f(v_t, \theta; c_o)$ satisfies the regularity conditions of Theorem 3.2. Also see the discussion of Lemma 7.1.

To date, there have been no simulation studies exploring the finite sample behaviour of this combined method of moment selection, and this is an interesting area for future research.[49] We now illustrate both MSC and RMSC using our running example.

Example: Hansen and Singleton's (1982) Consumption Based Asset Pricing Model

Our previous empirical implementation is based on the population moment condition

$$E[z_t(\delta_0 x_{1,t+1}^{\gamma_0 - 1} x_{2,t+1} - 1)] = 0$$

where $x_{1,t+1} = c_{t+1}/c_t$, $x_{2,t+1} = r_{t+1}/p_t$ and $z_t = [1, x_{1,t}, x_{2,t}, x_{1,t-1}, x_{2,t-1}]'$. It was remarked at the outset that this choice of instrument vector is arbitrary, and we now consider the performance of the model with the population moment condition

$$E[f(v_t, \theta_0; c)] = E[z_t(c)(\delta_0 x_{1,t+1}^{\gamma_0 - 1} x_{2,t+1} - 1)] = 0$$

for $c = c_i$, $i = 1, 2, \ldots 5$ where

$$\begin{aligned}
z_t(c_1) &= [1, x_{1,t}, x_{2,t}]' \\
z_t(c_2) &= [1, x_{1,t}, x_{2,t}, x_{1,t-1}, x_{2,t-1}]' \\
z_t(c_3) &= [1, x_{1,t}, x_{2,t}, x_{1,t-1}, x_{2,t-1}, x_{1,t-2}, x_{2,t-2}]' \\
z_t(c_4) &= [1, x_{1,t}, x_{2,t}, x_{1,t-1}, x_{2,t-1}, x_{1,t-2}, x_{2,t-2}, x_{1,t-3}, x_{2,t-3}]' \\
z_t(c_5) &= [1, x_{1,t}, x_{2,t}, x_{1,t-1}, x_{2,t-1}, x_{1,t-2}, x_{2,t-2}, x_{1,t-3}, x_{2,t-3}, x_{1,t-4}, x_{2,t-4}]'
\end{aligned}$$

Notice that $c = c_2$ gives the moment condition used in our earlier empirical implementation of this model. Table 7.1 reports the values of $MSC(c)$ and $RMSC(c)$ for these five choices of c.

Table 7.1
$MSC(c)$ and $RMSC(c)$ for certain choices of instrument vector in the consumption based asset pricing model

	VWR		EWR	
i	$MSC(c_i)$	$RMSC(c_i)$	$MSC(c_i)$	$RMSC(c_i)$
1	−5.546	0.970	2.141	1.682
2	−16.671	1.235	−6.308	1.929
3	−28.806	1.186	−17.920	1.848
4	−37.438	1.542	−26.617	2.193
5	−41.294	1.778	−37.446	2.428

Notes: $MSC(c_i)$ is given by (7.32) with $\hat{S}_T = \hat{S}_{SU,\mu}$ from (4.24) and bonus term given by (7.35); $RMSC(c_i)$ is given by (7.41) with $\hat{S}_T = \hat{S}_{SU}$ from (3.40) with penalty term given by (7.42) with $\tau = T^{1/2}$.

[49] See Section 7.3.4 for discussion of a related issue.

Consider first the results for value weighted returns (VWR). The value of *MSC* falls as the number of instruments increases, and so c_5 is the prefered choice from this limited set. The overidentifying restrictions test statistic associated with this choice, $J_T(c_5)$, takes the value 13.9262 which implies a p-value of 0.1250. Therefore, this choice of moments appears valid. However, *RMSC* indicates that the choice c_1 is prefered. It is interesting to contrast the parameter estimates and their associated confidence intervals for these two choices of instrument. If $c = c_5$ – the choice selected using *MSC* – then $\hat{\gamma}_T$ and $\hat{\delta}_T$ are 0.991 and 1.627, and their respective 95% asymptotic confidence intervals are $(0.985, 0.998)$ and $(-1.496, 4.751)$. If $c = c_1$ – the choice selected using *RMSC* – then $\hat{\gamma}_T$ and $\hat{\delta}_T$ are 0.994 and 0.593, and their respective 95% asymptotic confidence intervals are $(0.987, 1.001)$ and $(-3.026, 4.212)$.[50] These results provide an illustration of the sensitivity of inferences to the choice of moment condition. We now consider what happens if $f_{max}(v_t, \theta) = z_t(c_5)(\delta_0 x_{1,t+1}^{\gamma_0 - 1} x_{2,t+1} - 1)$ is treated as the candidate set and the moment selection is performed by minimizing $RMSC(c)$ over C. One immediate problem is the computational burden associated with allowing for so many possible choices of instrument vector. For purposes of comparison, we implemented the search two ways: using the two step estimator and the iterated estimator with a maximum of twenty iterations. Interestingly, both versions led to the same selected vector, $z_t(\tilde{c}_T) = (x_{1,t}, x_{1,t-1}, x_{2,t-2})$, and a minimimized value for *RMSC* of 0.807. This agreement raises the possibility that little may be lost by limiting the number of iterations, but further work is needed to explore whether this finding extends to other settings. The resulting iterated GMM parameter estimates are $(\hat{\gamma}_T, \hat{\delta}_T) = (0.994, 0.611)$, and their respective 95% asymptotic confidence intervals are $(0.987, 1.001)$ and $(-2.683, 3.906)$.

Now consider equally weighted returns (EWR). The value for *MSC* exhibits a similar pattern to the case with *VWR*. However, this time $J_T(c_5)$, takes the value 17.7745 which implies a p-value of 0.0379, and so indicates this choice of moments is invalid. Therefore, we do not consider this case further.

7.3.4 Other Methods of Instrument Selection

Both MSC and RMSC are valid for the GMM framework. There has also been some recent work on the problem of instrument selection within the framework of the GIV estimator described in Section 7.2. In this section, we review two particular methods: an information criterion for instrument selection based on the relevance condition proposed by Hall and Peixe (2003) and a method based on minimizing an approximation to the mean square error proposed by Donald and Newey (2001). Each method is designed to address the issue of instrument selection in classes of problems encountered in practice. However, neither is applicable in the general GIV framework. In view of this lack of generality, we only provide a heuristic discussion of the methods. Before we describe these two methods, it is worth noting that the same basic question was addressed in

[50] These figures are for the iterated estimator with $\hat{S}_T = \hat{S}_{SU}$.

the context of IV estimation of linear simultaneous equation models back in the 1960's. Fisher (1965) and Mitchell and Fisher (1970) respectively introduce and refine the method of "structurally ordered instrumental variables". However, we do not review their work here because it does not extend to the types of model in Table 1.1.[51]

Hall and Peixe (2003) consider the problem of instrument selection when the moment condition takes the form

$$E[z_t(c)u_t(\theta_0)] \;=\; 0$$

where $u_t(\theta) = u(v_t, \theta)$ is a scalar function, $u_t(\theta_0)$ is a martingale difference sequence with respect to Ω_{t-1}, $E[u_t(\theta_0)^2|\Omega_{t-1}] = \sigma_0^2$ and $z_t(c) \in \Omega_{t-1}$. They are concerned with developing a method for selecting the moment condition – or instrument – which satisfies the relevance condition. All the members of the candidate set are assumed to satisfy the orthogonality condition. Their method is motivated by considering the form of the asymptotic variance of the estimator. Under their conditions, the asymptotic distribution of the GIV estimator is

$$T^{1/2}[\hat{\theta}_T(c) \;-\; \theta_0] \;\xrightarrow{d}\; N(\,0, V_\theta(c))$$

where[52]

$$V(c) \;=\; \sigma_0^2 A(c)\Lambda(c)^{-2}A(c)' \tag{7.46}$$

$\Lambda(c) = diag\,(\rho_1(c), \ldots \rho_p(c))$, $\{\rho_i(c); i = 1, 2, \ldots p\}$ are defined to be the canonical correlations between $\partial u_t(\theta_0)/\partial\theta$ and $z_t(c)$, and $A(c)$ is the $p \times p$ matrix whose i^{th} row, $a_i(c)'$, contain the weights in the linear combination of $\partial u_t(\theta_0)/\partial\theta$ associated with the i^{th} canonical correlation.[53] The form of this variance suggests that the canonical correlations may provide a basis for selection based on the relevance condition. In fact, Hall and Peixe (2003) establish that $z_t(c_2)$ is redundant given $z_t(c_1)$ if and only if

$$\rho_i(c_1 + c_2) \;=\; \rho_i(c_1) \qquad i \;=\; 1, 2, \ldots p$$

Therefore, Hall and Peixe (2003) propose an information criterion for moment selection which exploits the information in these canonical correlations. They refer to this criterion as the canonical correlation information criterion (CCIC). Since θ_0 is unobservable, CCIC is based on the sample canonical correlations between and $\partial u_t(\tilde{\theta}_T)/\partial\theta$ and $z_t(c)$ where $\tilde{\theta}_T$ is some preliminary estimator. Using $r_{i,T}(c)$ to denote the i^{th} such canonical correlation, CCIC is given by[54]

$$CCIC(c) \;=\; T\sum_{i=1}^{p} ln\left[1 - r_{i,T}(c)^2\right] \;+\; (|c| - p)ln(T) \tag{7.47}$$

[51] Also see Hall and Peixe (2000) for further discussion.

[52] This type of decomposition for the asymptotic variance was first presented by Sargan (1958) in his study of IV estimators in linear models.

[53] See *inter alia* Anderson (1984) [Chap. 12] for further discussion of canonical correlations.

[54] This version of CCIC uses the BIC version of the penalty term. Hall and Peixe (2003) also consider the behaviour of their method with the HQIC and AIC type penalty terms. However, simulation evidence suggests the method works best with BIC and so we do not consider the other versions here.

The selected instrument vector is given by $z_t(\tilde{c}_T^{ccic})$ where

$$\tilde{c}_T^{ccic} \;=\; argmin_{c\in C}\; CCIC(c) \tag{7.48}$$

Hall and Peixe (2003) report simulation results for a static linear regression model. Overall, their evidence suggests the method is successful in screening out redundant instruments, and that selection based on relevance leads to a considerable improvement in the quality of the asymptotic approximation to the behaviour of GIV estimator. They also report simulation results for the case in which $MSC(c)$ and $CCIC(c)$ are used sequentially. Interestingly, they find the ordering can make a substantial difference to the performance of the sequential method in finite samples. Within their design, it proves beneficial to use $CCIC(c)$ first and then $MSC(c)$. The interested reader is refered to Hall and Peixe (2003) for further discussion of this issue.

Donald and Newey (2001) consider the problem of instrument selection for a type of model that is encountered in cross-sectional studies in labour economics. In these studies, the focus of attention is often on the point estimate of a particular parameter, and so the finite sample precision of this estimate is a more appropriate criterion for instrument selection than the orthogonality or relevance conditions. However, since the resulting method is only applicable to i.i.d. data, we confine our discussion to a special case of their method in order to illustrate the basic approach. The interested reader is refered to Donald and Newey (2001) for a more detailed discussion including simulation evidence and an empirical example. Within their framework, it is assumed that the researcher wishes to estimate the following single equation by instrumental variables

$$y_t \;=\; Y_t^{'}\gamma_0 \;+\; x_{1,t}^{'}\beta_0 \;+\; u_t \tag{7.49}$$

where $x_{1,t}$ are a vector of exogenous variables but Y_t is a vector of endogenous variables generated by

$$Y_t \;=\; h(x_t) \;+\; e_t \tag{7.50}$$

where $x_t = (x_{1,t}^{'}, x_{2,t}^{'})^{'}$. The variables $(x_t^{'}, u_t, e_t^{'})$ are assumed to be i.i.d. The errors are assumed to satisfy: $E[u_t|x_t] = 0$, $E[e_t|x_t] = 0$, $Var[(u_t, e_t^{'})^{'}(u_t, e_t^{'})|x_t] = \Omega$ and $Cov[e_t, u_t|x_t] \neq 0$. It is convenient to stack the unknown parameters into a single vector and so we set $\theta = (\gamma^{'}, \beta^{'})'$, and also to introduce the stacked system

$$N_t \;=\; \left[\begin{array}{c} Y_t \\ x_{1,t} \end{array}\right] \;=\; d(x_t) \;+\; w_t \tag{7.51}$$

where $d(x_t) = [h(x_t)^{'}, x_{1,t}^{'}]^{'}$ and $w_t = [e_t^{'}, 0']$ The candidate set of instruments is assumed to consist of elements of the form $z_{t,i} = z_i(x_t)$. Notice that by construction all instruments satisfy the orthogonality condition. Unlike the other methods described above, this framework allows for the dimension of the candidate set to increase with T at some rate which is restricted by the theory as described below. Donald and Newey (2001) propose choosing c to minimize a Nagar type approximation to the mean square error (MSE) of the estimator.[55]

[55] See Section 6.2.2 for further discussion of Nagar approximations.

To make this approach operational, Donald and Newey (2001) assume that the researcher is interested in a linear combination of the parameters rather than the parameter vector itself, and they also propose substituting preliminary estimates for any unknown nuisance parameters which appear in the formula. In the case of Two Stage Least Squares estimator, this approach leads to the following estimated approximate MSE for the linear combination $\hat{\lambda}_T^{'}\hat{\theta}_T(c)$:[56]

$$AMSE(c) \;=\; \hat{\sigma}^2_{\lambda,u}\frac{|c|^2}{T} \;+\; \hat{\sigma}^2_u[\hat{R}_\lambda(c) - \hat{\sigma}^2_\lambda\frac{|c|}{T}]$$

where $\hat{\sigma}^2_u = \tilde{u}'\tilde{u}/T$, $\hat{\sigma}^2_\lambda = \hat{\lambda}_T^{'}\tilde{D}^{-1}\tilde{w}^{'}\tilde{w}\tilde{D}^{-1}\hat{\lambda}_T/T$, $\hat{\sigma}_{\lambda,u} = \hat{\lambda}_T^{'}\hat{H}^{-1}\tilde{w}'\tilde{u}/T$ and $\tilde{D}$ is a preliminary estimator of $T^{-1}\sum_{t=1}^{T} d(x_t)d(x_t)'$, $\tilde{w}$ is a residual vector from a preliminary estimation of (7.51), $\tilde{u}$ is a residual vector from a preliminary estimation of (7.49), and $\hat{R}_\lambda(c)$ is a measure of the goodness of fit for the estimation of (7.51) using $z_t(c)$. One possible choice for $\hat{R}_\lambda(c)$ is

$$\hat{R}_\lambda(c) \;=\; \frac{\hat{\lambda}_T^{'}\tilde{D}^{-1}\hat{w}(c)^{'}\hat{w}(c)\tilde{D}^{-1}\hat{\lambda}_T}{T} \;+\; \frac{2\hat{\sigma}^2_\lambda|c|}{T}$$

$\hat{w}(c) = \{I_T - Z(c)[Z(c)'Z(c)]^{-1}Z(c)^{'}\}N$, $Z(c)$ is the $T \times |c|$ matrix with t^{th} row $z_t(c)'$ and N is the $T \times n$ matrix with t^{th} row $N_t^{'}$. The selected instrument vector is the one which minimizes the estimated approximate MSE, that is

$$\hat{c}_T^{mse} \;=\; min_{c\in C}\; AMSE(c)$$

Donald and Newey (2001) provide conditions under which $AMSE(c)$ converges in probability to the true MSE, and these include the requirement that the candidate set expands at a rate slower than $T^{1/2}$. Under these conditions, $\hat{c}_T^{mse}$ can be considered optimal in the sense that it minimizes the MSE asymptotically with probability one. However, they do not consider the asymptotic distribution of $\hat{\theta}_T(\hat{c}_T^{mse})$. Intuition suggests that the inference condition is satisfied under certain regularity conditions, but the characterization of these regularity conditions remains a topic for future research.[57]

7.4 Summary

In this chapter, we have considered the problem of moment selection. The desirable properties for the selected moment depend upon the ultimate objective of the study in question. For the majority of this chapter, it is assumed that this objective is to perform inference about θ_0 based on the asymptotic theory derived in Chapters 3 and 5. Given this context, it is argued that it is desirable for the selected vector to satisfy:

[56] Donald and Newey (2001) also consider the Limited Information Maximum Likelihood estimator and a bias adjusted version of 2SLS. See their article for further discussion of these two cases and comparisons between all three estimators.

[57] See Section 6.1.3 for a discussion of the available asymptotic distribution theory if the number of moment conditions increases with T.

- *the orthogonality condition* – so that the estimation is based on valid information;

- *the efficiency condition* – so that inference is based on the asymptotically most precise estimates;

- *the non-redundancy condition* – so that the selected moment condition does not contain any redundant elements whose inclusion can cause a deterioration in the quality of the asymptotic approximation to finite sample behaviour.

There have been two approaches to this issue in the literature. The first approach is to characterize theoretically the optimal choice of moment condition. The second approach is to develop data based methods for moment selection. We now briefly summarize the results on each.

- *The optimal moment condition:* Given that only asymptotic distribution theory is available, the optimal choice is the one that satisfies both the orthogonality and efficiency conditions. Given this criterion, the optimal moment condition is always the score vector because the resulting GMM estimator is the MLE. Unfortunately, this choice is infeasible in the types of model listed in Table 1.1. Therefore, it is necessary to restrict the search for the optimal moment to settings encountered in practice. For the class of Generalized Instrumental Variable estimators, it is possible to characterize the functional form of the optimal instrument in terms of the information set. However, the optimal instrument is infeasible in most nonlinear dynamic models because its construction requires knowledge of aspects of the data generation process which are typically not specified as part of the economic model. However, knowledge of the form of the optimal instrument facilitates efficiency comparisons between ML and GIV. To date, such comparisons indicate that GIV can be as asymptotically efficient as ML in certain linear models with normal errors, but this equivalence does not extend to the general nonlinear model.

- *Data based methods for moment selection:* In most circumstances, a researcher must decide which moments to choose without knowledge of the underlying data generation process. In such circumstances, moment selection must perforce be based upon the data, and it is therefore important that the use of the data in this way does not contaminate the limiting distrubtion theory. This consideration yields a fourth desirable property for a moment selection procedure that is termed the inference condition. To date, this problem has mostly been approached using information criterion. The moment selection criterion (MSC) is designed to select moments on the basis of the orthogonality condition, and the relevant moment selection criterion (RMSC) is designed to select moments on the basis of a combination of the efficiency and non-redundancy conditions that is

termed the relevance condition. Under certain conditions, these methods each satisfy the inference condition. MSC and RMSC can be used individually or sequentially.

The preliminary evidence suggests that the use of MSC and RMSC can help a researcher to avoid situations in which asymptotic theory provides a very poor approximation to finite sample behaviour. However, it is also clear that their use is not a panacea for the finite sample deficiencies of the conventional asymptotic theory that are documented in Chapter 6. Therefore, in the following chapter, we explore a number of alternative asymptotic approximations to the finite sample behaviour of the GMM estimator.

8

Alternative Approximations to Finite Sample Behaviour

In Chapter 6, it is seen that the available simulation evidence indicates that the asymptotic theory developed in Chapters 3 and 5 may not provide a good approximation to the finite sample behaviour of the GMM estimator in certain circumstances of interest. The situation can be ameliorated by careful selection of moment conditions, and this motivates the methods described in the previous chapter. However, while the use of such moment selection procedures may lead to an improvement, the overall quality of the asymptotic approximation may still leave something to be desired. Therefore, in this chapter, we consider three alternative methods for approximating the finite sample behaviour of the GMM estimator and its associated statistics. These three are: (i) the bootstrap; (ii) an asymptotic theory developed for the case in which the parameter vector is weakly identified by the population moment condition; (iii) and an asymptotic theory designed to provide a better approximation when the weighting matrix is based on a heteroscedasticity autocorrelation covariance (HAC) matrix estimator.

In Section 8.1, we discuss the use of the bootstrap which is a resampling technique that has – at least in theory – the potential to improve the quality of the approximation in any model. This potential has been successfully realized in many areas of statistical inference, and so the method is a natural candidate for improving the quality of inferences based on GMM estimators. However, it turns out that the extension of the bootstrap to this setting is not so simple in terms of both implementation and also the verification that it yields an improvement. In particular, complications arise in overidentified, nonlinear dynamic models. While considerable progress has been made in circumventing these complications, the available analysis does not yet encompass the general framework employed in Chapter 3. Section 8.1.1 provides a brief review of the ideas behind the bootstrap. Section 8.1.2 describes the steps involved in the application of the bootstrap to nonlinear dynamic models.

In Section 8.2, we describe an alternative asymptotic theory that has been developed for the case in which the parameter vector is weakly identified by the population moment condition. Equivalently, this scenario can also be termed as the case in which the population moment condition is "nearly uninformative" about the parameter vector. To date, this problem has mostly been encountered in models estimated by Generalized Instrumental Variables (GIV). Therefore, we focus our discussion on this case but the qualitative conclusions extend to the GMM setting. Section 8.2.1 presents the limiting behaviour of the GIV estimator. Section 8.2.2 presents methods for performing inference within this scenario. Section 8.2.3 discusses the detection of poor identification.

In Section 8.2.3, we return to the problem of bandwidth selection for HAC matrix estimators of the long run variance. In Section 3.5.3, we review the literature on bandwidth selection when the aim is to provide a consistent estimator of the long run variance. It can be recalled that, to date, there is no definitive rule for making this selection. One way to remove this ambiguity is simply to set the bandwidth equal to the sample size. While this choice does not satisfy the conditions for consistency, it does lead to an alternative asymptotic theory upon which to base inference about the parameters. This alternative theory is briefly reviewed in Section 8.2.3.

Since all three alternative approximations are relatively new and so not yet widely applied, the discussion here is less technical than before and no formal proofs are provided. Instead, the focus is placed on the intuition behind the three approaches, and on practical matters.

8.1 The Bootstrap

8.1.1 Background and Intuition

Efron (1979) introduced the term "bootstrap" as a generic name for methods of statistical inference based on resampling techniques. By their very nature, resampling techniques can be computationally burdensome but, with advances in computer technology, it has become feasible to apply the method in increasingly complex settings. These advances have stimulated a considerable literature in statistics on the bootstrap where the method has been used both for the estimation of bias, variance, and distribution functions, and also for the reduction of errors made in the use of approximate significance levels of tests or coverage probabilities of approximate confidence intervals. However, it is only relatively recently that researchers have considered applying the method in the context of GMM. Hall and Horowitz (1996) provide the first treatment of the bootstrap based on GMM in the context of nonlinear, dynamic models, and our discussion rests heavily on their work.[1] It is beyond the scope of this book to provide a comprehensive review of the more general literature on the bootstrap in statis-

[1] An alternative approach is to base the bootstrap upon the Empirical Likelihood. However, since this method has only been developed for i.i.d. cases, we do not not discuss it here but do return to it as part of the discussion of Empirical Likelihood in Section 10.2.

tics. Instead, the interested reader is refered to Hall (1994) and the references therein.

The idea behind the bootstrap is best understood by considering a simple example in which the method is used to reduce the errors made in the use of an approximate $100\alpha\%$ significance level test. Let $\{v_t; t = 1, 2, \ldots T\}$ be a sample of independent random draws from a common distribution with mean μ_0 and variance σ_0^2. In terms of our framework, the parameter vector is $\theta_0 = [\mu_0, \sigma_0^2]'$. It is natural to estimate θ_0 from the first two moment conditions, and so, using (1.1)–(1.2), it follows that:

$$\hat{\theta}_T \;=\; \left[\begin{array}{c} \hat{\mu}_T \\ \hat{\sigma}_T^2 \end{array}\right] \;=\; \left[\begin{array}{c} \bar{v}_T \\ T^{-1}\sum_{t=1}^{T}(v_t - \bar{v}_T)^2 \end{array}\right]$$

where $\bar{v}_T = T^{-1}\sum_{t=1}^{T} v_t$. Suppose that it is desired to test the hypothesis $H_0 : \mu_0 = 0$ versus $H_1 : \mu_0 \neq 0$ based on a sample of size T. The natural test statistic is the t–ratio,

$$\tau_T \;=\; \frac{T^{1/2}\hat{\mu}_T}{\hat{\sigma}_T} \tag{8.1}$$

The decision rule of the test involves the comparison of $|\tau_T|$ with a percentile from some distribution. The key question is: what is the appropriate distribution? Before we consider how the bootstrap can be used to answer this question, it is useful for purposes of comparison to review two more familiar choices of distribution and the properties of the ensuing tests.

If the true distribution of τ_T is known then it is possible to perform an exact test. If $F_T[\tau]$ denotes the true cumulative distribution function of τ_T then the decision rule for the test is as follows:

Test based on the finite sample distribution: Reject H_0 if $|\tau_T| > c_T(\alpha/2)$ (8.2)

where $F_T[c_T(\alpha/2)] = 1 - \alpha/2$. This version is said to be an "exact" $100\alpha\%$ significance level test because the probability of a type I error is α.[2] Clearly, this exact test can only be performed if the true distribution function is known. Unfortunately, this is rarely the case. Therefore, inference is most commonly based upon the limiting distribution, which for τ_T is the standard normal distribution. If $\Phi[\tau]$ denotes the cumulative distribution function of the standard normal distribution then the decision rule based on the limiting distribution takes the form:

Test based on the limiting distribution: Reject H_0 if $|\tau_T| > c_\infty(\alpha/2)$ (8.3)

where $\Phi[c_\infty(\alpha/2)] = 1-\alpha/2$. With the decision rule in (8.3), the true probability of a type one error is $2\{1 - F_T[c_\infty(\alpha/2)]\}$, and this is only guaranteed to be α in the limit. Therefore, it is not an exact $100\alpha\%$ significance level test for finite T,

[2] Such an exact test is most often encountered in circumstances when $\{v_t\}$ are random draws from a normal distribution because then $(\sqrt{(T-1)/T})\tau_T$ has a Student's t distribution with $T - 1$ degrees of freedom.

but is instead refered to as an "approximate" $100\alpha\%$ level test in finite samples. As with all approximations, there is an error and it is the desire to reduce this error that motivates the use of the bootstrap.

The bootstrap version of the test is based on an alternative approximation to the distribution of τ_T that is obtained via resampling from the observed sample. The decision rule for this version of the test takes the form:

Test based on the bootstrap distribution: Reject H_0 if $|\tau_T| > c_T^B(\alpha/2)$ (8.4)

where $c_T^B(\alpha/2)$ is calculated as follows.

- Draw N samples of size T with replacement from the observed sample. Let the n^{th} such sample be denoted $(v_1^{(n)}, v_2^{(n)}, \dots v_T^{(n)})$.
- For each of these N samples, calculate the statistic

$$\tilde{\tau}_T^{(n)} = \frac{T^{1/2}(\bar{v}_T^{(n)} - \bar{v}_T)}{\hat{\sigma}_T^{(n)}}, \qquad n = 1, 2, \dots N$$

  where $\bar{v}_T^{(n)} = T^{-1}\sum_{t=1}^T v_t^{(n)}$ and $\hat{\sigma}_T^{(n)} = \sqrt{T^{-1}\sum_{t=1}^T (v_t^{(n)} - \bar{v}_T^{(n)})^2}$.
- $c_T^B(\alpha/2)$ is the $100(1-\alpha)^{th}$ percentile of the empirical distribution of $(|\tilde{\tau}_T^{(1)}|, |\tilde{\tau}_T^{(2)}|, \dots |\tilde{\tau}_T^{(N)}|)$.

Notice that the critical point is calculated using the empirical distribution of the *absolute* value of $\tilde{\tau}_T$. This transformation is taken because it is the absolute value of τ_T that appears in the decision rule. Notice also that the t-statistic in the bootstrap, $\tilde{\tau}_T^{(n)}$, is centred about the sample mean, and so is different from the original t-statistic. This correction is needed to ensure that $E[\tilde{\tau}_T^{(n)}] = 0$, and hence that the bootstrapped distribution mimics the first moment properties of τ_T under H_0 regardless of the true value of μ_0. The bootstrap version of the test is also only an approximate $100\alpha\%$ significance level test but intuition suggests that the involvement of the data yields a test whose size is closer to α than its counterpart based on the limiting distribution. This turns out to be the case, and a formal justification comes from consideration of Edgeworth expansions of both the true and the bootstrap cumulative distribution functions.[3]

To begin, we consider the Edgeworth expansion of the true cumulative distribution function. Under certain regularity conditions, it can be shown that

$$P[\tau_T \le c] = \Phi(c) + T^{-1/2}h_{1,T}(c) + T^{-1}h_{2,T}(c) + o(T^{-1}) \tag{8.5}$$

uniformly over c where $h_{1,T}(c)$ and $h_{2,T}(c)$ are respectively even and odd functions of c for any T. The properties of $\{h_{i,T}(.);\ i = 1, 2\}$ mean that a convenient cancellation takes places when we consider the probability that $|\tau| < c$ for any $c > 0$. Specifically, it follows from (8.5) that:

$$\begin{aligned} P[|\tau_T| \le c] &= P[\tau_T < c] - P[\tau < -c] &(8.6)\\ &= \Phi(c) - \Phi(-c) + 2T^{-1}h_{2,T}(c) + o(T^{-1}) &(8.7)\end{aligned}$$

[3] Also see Section 6.2.2.

Since $\Phi(-c) = 1 - \Phi(c)$, equation (8.7) can be simplified further to yield:

$$P[|\tau_T| \leq c] \;=\; -1 \;+\; 2\Phi(c) \;+\; 2T^{-1}h_{2,T}(c) \;+\; o(T^{-1}) \tag{8.8}$$

For our purposes, it is more convenient to focus on the expansion for $P[|\tau_T| > c]$. Using (8.8), it follows that

$$\begin{aligned} P[|\tau_T| > c] &= 1 \;-\; P[|\tau_T| \leq c] \\ &= 2\left\{1 \;-\; \Phi(c) \;-\; T^{-1}h_{2,T}(c)\right\} \;+\; o(T^{-1}) \end{aligned} \tag{8.9}$$

Equation (8.9) can be used to provide insights into the nature of the "approximation" in the approximate $100\alpha\%$ significance level test based on the limiting distribution of τ_T. Putting $c = c_\infty(\alpha/2)$, it follows from (8.9) that the true significance level of this version of the test is

$$P[|\tau_T| > c_\infty(\alpha/2)] \;=\; \alpha \;+\; O(T^{-1}) \tag{8.10}$$

Therefore, the test based on the limiting distribution has an exact significance level that deviates from $100\alpha\%$ by a term of large order T^{-1}.

A similar expansion can be developed for probabilities based on the bootstrap distribution. In practice, this distribution is a function of N, the number of replications, but the theoretical justification derives from considering the limiting bootstrap distribution obtained as $N \to \infty$. This clearly raises the issue of how N should be chosen, and this is considered in Section 8.1.2.3. A second important aspect of this distribution is that it is conditional on the observed sample and so is itself subject to sampling variation. This dependence is indicated by inserting a B superscript on $P[.]$. It should also be noted that this randomness manifests itself in the percentile $c_T^B(\alpha/2)$, and this feature becomes important at certain points in the argument. With these features in mind, we can now consider the Edgeworth expansion for $P^B[.]$. Under appropriate regularity conditions, it can be shown that

$$P^B[\tilde{\tau}_T \leq c] \;=\; \Phi(c) \;+\; T^{-1/2}h_{1,T}^B(c) \;+\; T^{-1}h_{2,T}^B(c) \;+\; o_p(T^{-1}) \tag{8.11}$$

uniformly over c where $h_{1,T}^B(c)$ and $h_{2,T}^B(c)$ are respectively even and odd functions of c. Since $h_{i,T}^B(.)$ have the same properties as their counterparts in the expansion for the true CDF, we can repeat the argument above to deduce

$$P^B[|\tilde{\tau}_T| > c] \;=\; 2\left\{1 \;-\; \Phi(c) \;-\; T^{-1}h_{2,T}^B(c)\right\} \;+\; o_p(T^{-1}) \tag{8.12}$$

A comparison of (8.9) and (8.12) indicates that the probabilities based on the true and bootstrapped distributions differ. However, it can be shown that $T^{-1}h_{2,T}^B(c)$ converges almost surely to $T^{-1}h_{2,T}(c)$ as $T \to \infty$. Using this result, (8.12) can be re-written as

$$P^B[|\tilde{\tau}_T| > c] \;=\; 2\left\{1 \;-\; \Phi(c) \;-\; T^{-1}h_{2,T}(c)\right\} \;+\; o_p(T^{-1}) \tag{8.13}$$

and so it can be recognized that the probabilities based on the true and bootstrap distributions are equal through terms of order T^{-1}. All that remains

is to show that this equivalence implies that the use of the bootstrap version of the test yields more accurate inference than the test based on the limiting distribution. This is established in two steps. First, it is shown that $c_T^B(\alpha/2) = c_T(\alpha/2) + o_p(T^{-1})$. Secondly, it is shown that the preceeding relationship between the percentiles implies that $P[|\tau_T| \geq c_T^B(\alpha/2)] = \alpha + o(T^{-1})$. The details follow; for this part of the presentation we set $c_T = c_T(\alpha/2)$ and $c_T^B = c_T^B(\alpha/2)$ to avoid excessive notation.

The first step can be established by considering

$$d_T(c_T^B) \;=\; \Phi(c_T^B) \;+\; T^{-1}h_{2,T}(c_T^B) \tag{8.14}$$

Using a Mean Value Theorem expansion of $d_T(c_T^B)$ around $d_T(c_T)$, it follows that

$$d_T(c_T^B) \;=\; d_T(c_T) \;+\; D_T(\bar{c}_T)(c_T^B \;-\; c_T) \tag{8.15}$$

where $D_T(c) = \partial d_T(c)/\partial c$ and $\bar{c}_T = \lambda_T c_T^B + (1 - \lambda_T)c_T$ for some $\lambda_T \in [0, 1]$. Simple rearrangement yields

$$d_T(c_T^B) \;-\; d_T(c_T) \;=\; D_T(\bar{c}_T)(c_T^B \;-\; c_T) \tag{8.16}$$

It turns out that $|D_T(\bar{c}_T)|$ is finite and bounded away from zero for non-zero α although we do not present the details here.[4] Therefore, the desired result follows if it can be shown that $d_T(c_T^B) - d_T(c_T) = o_p(T^{-1})$. The latter can be established by manipulating the expansions derived above. Since $P[|\tau_T| > c_T] = \alpha$ by construction, it follows from (8.9) that

$$1 \;-\; \Phi(c_T) \;-\; T^{-1}h_{2,T}(c_T) = \alpha/2 \;+\; o(T^{-1}) \tag{8.17}$$

Similarly, since $P^B[|\tilde{\tau}_T| > c_T^B] = \alpha$, it follows from (8.13) that

$$1 \;-\; \Phi(c_T^B) \;-\; T^{-1}h_{2,T}(c_T^B) \;=\; \alpha/2 \;+\; o_p(T^{-1}) \tag{8.18}$$

Taken together (8.17) and (8.18) imply that $d_T(c_T^B) - d_T(c_T) = o_p(T^{-1})$, and so it follows from (8.16) that

$$c_T^B \;=\; c_T \;+\; o_p(T^{-1}) \tag{8.19}$$

This completes the first step of the argument.

To establish the second step, it is useful to express the probabilities $P[|\tau_T| > c_T]$ and $P[|\tau_T| > c_T^B]$ in terms of indicator functions. To this end, define $\mathcal{I}(A)$ to be an indicator function that takes the value one if event A occurs and is zero otherwise. Using this notation, we have

$$P[|\tau_T| > c_T] \;=\; E[\mathcal{I}(|\tau_T| > c_T)] \tag{8.20}$$

$$P[|\tau_T| > c_T^B] \;=\; E[\mathcal{I}(|\tau_T| > c_T^B)] \tag{8.21}$$

It is therefore possible to compare the probabilities by comparing the underlying indicator functions. It can be recognized that $\mathcal{I}(|\tau_T| > c_T)$ and $\mathcal{I}(|\tau_T| > c_T^B)$

[4] See Hall (1994).

agree if either $|\tau_T| > max\{c_T^B, c_T\}$ or $min\{c_T, c_T^B\} \geq |\tau_T|$ because in these cases we have

$$\begin{aligned} |\tau_T| > max\{c_T^B, c_T\} \quad &\Rightarrow \quad \mathcal{I}(|\tau_T| > c_T) \;=\; \mathcal{I}(|\tau_T| > c_T^B) \;=\; 1 \\ min\{c_T, c_T^B\} \geq |\tau_T| \quad &\Rightarrow \quad \mathcal{I}(|\tau_T| > c_T) \;=\; \mathcal{I}(|\tau_T| > c_T^B) \;=\; 0 \end{aligned}$$

However they disagree if either $c_T^B \geq |\tau_T| > c_T$ or $c_T \geq |\tau_T| > c_T^B$ because in these cases we have

$$\begin{aligned} c_T^B \geq |\tau_T| > c_T \quad &\Rightarrow \quad \mathcal{I}(|\tau_T| > c_T) \;=\; 1, \qquad \mathcal{I}(|\tau_T| > c_T^B) \;=\; 0 \\ c_T \geq |\tau_T| > c_T^B \quad &\Rightarrow \quad \mathcal{I}(|\tau_T| > c_T) \;=\; 0, \qquad \mathcal{I}(|\tau_T| > c_T^B) \;=\; 1 \end{aligned}$$

Using these relations, it is clear that

$$\mathcal{I}(|\tau_T| > c_T^B) \;=\; \mathcal{I}(|\tau_T| > c_T) - \mathcal{I}(c_T^B \geq |\tau_T| > c_T) + \mathcal{I}(c_T \geq |\tau_T| > c_T^B) \quad (8.22)$$

The substitution of (8.22) into the right hand side of (8.21) yields

$$\begin{aligned} P[|\tau_T| > c_T^B] \quad &= \quad E[\mathcal{I}(|\tau_T| > c_T)] \;-\; E[\mathcal{I}(c_T^B \geq |\tau_T| > c_T)] \\ &\quad + E[\mathcal{I}(c_T \geq |\tau_T| > c_T^B)] \end{aligned} \quad (8.23)$$

Equation (8.20) and the definition of c_T imply that the first term on the right hand side of (8.23) is α, and (8.19) implies that the other terms on the right hand side are collectively $o(T^{-1})$. Substituting these results into (8.23) yields

$$P[|\tau_T| \;>\; c_T^B] \;=\; \alpha \;+\; o(T^{-1}) \quad (8.24)$$

A comparison of (8.10) and (8.24) reveals the potential gains from the use of the bootstrap. The use of the bootstrap involves an approximation error in the significance level of $o(T^{-1})$; whereas basing inference on the limiting distribution involves an approximation error of $O(T^{-1})$. The bootstrap is therefore said to provide "asymptotic refinements". In more general settings, the bootstrap yields such asymptotic refinements in cases where inference is based on an asymptotically pivotal statistic, that is a statistic whose limiting distribution is independent of the distribution of the data. While this asymptotic refinement motivates the use of the bootstrap, it should be noted that order statements only reveal something about the rate at which the error decreases. They do not tell us anything about the magnitude of the error for a given T, and so there is no guarantee that the bootstrap yields more reliable inference procedures in every case.[5]

The above discussion has focused on a very simple case in which the data are independently and identically distributed and the parameter vector is just identified by the population moment condition. The key question is whether the method and the theoretical arguments can be extended to the case where the data are dependent, the parameter vector is overidentified and the population condition is nonlinear in the parameters. The answer is in the affirmative but subject to some important qualifications. This is the topic of the next sub–section.

[5] For example, if the order T^{-1} term in (8.9) may be $10^{-6}T^{-1}$ then the use of the bootstrap is unlikely to yield a significant improvement over inference based on the limiting distribution.

8.1.2 Nonlinear Dynamic Models

Hall and Horowitz (1996) provide the first rigorous treatment of the bootstrap based on GMM estimation of the parameters in nonlinear dynamic models. Their analysis deals with the use of the so-called block bootstrap with non-overlapping blocks to reduce the error in the significance level of the overidentifying restrictions test statistic and the two-sided t-statistic for testing $H_{0,i} : \theta_{0,i} = \bar{\theta}_{0,i}$. Andrews (2002*b*) extends Hall and Horowitz's (1996) analysis in a number of directions that include the use of the block bootstrap with either overlapping or non-overlapping blocks and a consideration of a broader array of inference procedures. Our discussion covers both versions of the block bootstrap but, for brevity, we restrict attention to just two statistics, the overidentifying restrictions test and the two-sided confidence intervals for $\theta_{0,i}$. Therefore, this sub-section is based on a synthesis of certain results in Hall and Horowitz (1996) and Andrews (2002*b*) and relies heavily on these sources.

As discussed in the previous sub-section, the theoretical justification for the bootstrap derives from Edgeworth expansions. Such expansions are only valid under certain regularity conditions, and these conditions turn out to be far more restrictive than those used to underpin the asymptotic analysis in Chapters 3 and 5. Although we do not consider these Edgeworth expansions here, we begin this sub-section with a discussion of the necessary regularity conditions in order to highlight the key differences from the earlier analysis. After that, we describe the mechanics of applying the block bootstrap within the GMM setting.

The discussion here is premised on the assumptions that the overidentifying restrictions test has the limiting chi-squared distributions given in Theorems 5.1 and $T^{1/2}(\hat{\theta}_T - \theta_0)$ has the limiting normal distribution given in Theorem 3.2. In addition to the regularity conditions required for these results, Hall and Horowitz (1996) and Andrews (2002*b*) impose a number of other conditions. For our purposes here, it suffices to focus on the conditions that most obviously restrict the model in comparison to the framework in Chapters 3 and 5. These conditions involve the dependence structure of v_t, the autocovariance structure of $f(v_t, \theta_0)$, and the composition of v_t. The interested reader is refered to the aforementioned sources for a complete listing of the required conditions.[6]

It can be recalled that our asymptotic analysis is premised on the assumption that the data are stationary and ergodic. As discussed in Appendix A, this assumption places restrictions on the memory of the process. To implement the bootstrap, it is necessary to restrict this memory further.

Assumption 8.1 Approximation of v_t by a m–Dependent Process *There is a sequence of $s \times 1$ i.i.d. vectors $\{e_t\}_{t=-\infty}^{\infty}$ with $s \geq r$ and a $r \times 1$ function h such that the $r \times 1$ vector v_t can be written as $v_t = h(e_t, e_{t-1}, e_{t-2}, \ldots)$. There is a constant $d > 0$ such that for all $t = 1, 2, \ldots$ and all $m > d^{-1}$,*

$$\|h(e_t, e_{t-1}, e_{t-2}, \ldots) - h(e_t, e_{t-1}, e_{t-2}, \ldots e_{t-m}, 0, 0, \ldots)\| \leq d^{-1}e^{-dm}$$

[6] The remaining conditions involve restrictions on $f(v_t, \theta)$ pertaining to continuity, existence of certain moments and the existence and smoothness of derivatives up to the fourth order.

This condition implies that the dynamic behaviour of v_t can be approximated by a nonlinear moving average of $\{e_{t-i};\ i = 0, 1, \ldots m\}$ in the sense described above and so $\{e_{t-i};\ t = m+1, m+2, \ldots\}$ have a negligible effect on the dynamics of v_t. This restriction implies that v_t is α–mixing with mixing parameter $\alpha_m = e^{-m}$ which is a much faster rate than is required for the Weak Law of Large Numbers or Central Limit Theorem.[7] In spite of this limitation, Assumption 8.1 is still satisfied by a number of empirically relevant models such as infinite order moving average processes with exponentially decreasing coefficients.

The earlier asymptotic analysis places fairly mild restrictions on the autocovariance structure of $f(v_t, \theta_0)$ requiring simply that the long run variance, S, exists and is positive definite. For the bootstrap analysis here, it is necessary to limit this dependence structure as follows.

Assumption 8.2 $f(v_t, \theta_0)$ is a k-Dependent Process
$E[f(v_t, \theta_0)f(v_{t-i}, \theta_0)'] = 0$ for $i > k$ and some $k < \infty$.

This assumption implies that

$$S \ = \ \Gamma_0 + \sum_{i=1}^{k} (\Gamma_i \ + \ \Gamma_i^{'}) \tag{8.25}$$

where $\Gamma_i = E[f(v_t, \theta_0)f(v_{t-i}, \theta_0)']$. While this assumption is clearly not universally valid, it is satisfied by a number of models of interest: witness our running empirical example of the consumption based asset pricing model in which the underlying economic theory implies that $f(v_t, \theta_0)$ satisfies this restriction with $k = 0$. Parenthetically, we note that there are grounds for anticipating that this assumption can be relaxed in future work. Inoue and Shintani (2003) consider the use of the bootstrap in linear models estimated by instrumental variables under very weak restrictions on the dependence structure of $f(v_t, \theta_0)$.[8] However, to date, these results have not been extended to GMM estimators of nonlinear models and so we do not consider their framework here.

The last additional restriction highlighted here involves the composition of v_t in terms of continuous and discrete random variables. To date, no assumptions have been made regarding this aspect of the model. However, now they must.

Assumption 8.3 Composition of v_t
v_t can be partitioned into $(v_t^{(c)'}, v_t^{(d)'})'$ where $v_t^{(c)} \in \Re^c$ for some $c > 0$ and $v_t^{(d)} \in \Re^d$ for $d \geq 0$ and $c + d = r$. The distributions of $v_t^{(c)}$ and $\partial f(v_t, \theta_0)/\partial \theta'$ are absolutely continuous. The distribution of $v_t^{(d)}$ is discrete.

[7] See Appendix A for a definition of α_m.

[8] Inoue and Shintani (2003) provide a theoretical justification for the bootstrap in this case but find the potential gains are not as great as those described here. They also find that the potential gains are sensitive to the choice of kernel in the HAC estimator.

In other words, v_t must contain at least one continuous random variable; the remaining elements may include discrete random variables but this is not necessary. While noteworthy, this assumption is unlikely to be particularly restrictive in practice because most economic models involve at least one continuous random variable.

We now turn to the mechanics of the bootstrap. This discussion breaks down naturally into three parts. Section 8.1.2.1 considers appropriate designs for the bootstrap sampling scheme when the data are dependent, and this leads to a discussion of the block bootstrap. Section 8.1.2.2 describes the appropriate construction of the statistics whose bootstrap distributions are used to approximate the distribution of our statistics of interest. This sub-section also includes a brief discussion of the so–called approximate bootstrap method that has been proposed to reduce the computational burden in nonlinear models. Section 8.1.2.3 presents a rule for picking N, the number of bootstrap replications. As emerges below, the precise details require fairly lengthy explanation. Therefore, Section 8.1.2.4 summarizes the necessary calculations and also illustrates them using our running empirical example.

8.1.2.1 Generation of Bootstrap Sample When the Data are Dependent

There are two basic approaches to constructing the bootstrap sample when the data are dependent. These are known as the parametric bootstrap and the non-parametric bootstrap. In the parametric bootstrap, the resampling is based on an estimated model for v_t. As an illustration, suppose this estimated model is a VAR; in this case, the bootstrap sample for v_t is generated by resampling from the residuals with replacement, and then solving the model recursively. While relatively straightforward to implement, this approach is only guaranteed to deliver the types of gain described above if the assumed model for v_t is correct. It is this caveat that makes the approach unattractive in the types of models in Table 1.1 for which the data generation process of v_t is not completely specified. Therefore, we do not pursue the parametric bootstrap further here.[9] Instead we focus on the non-parametric bootstrap. This method essentially involves sampling blocks of adjacent observations from the observed sample and so is commonly refered to as the "block bootstrap". These blocks can be non-overlapping or overlapping.[10] To illustrate the difference, consider the following example. Suppose that v_t is scalar and we have an observed sample of four observations, (v_1, v_2, v_3, v_4). Suppose further that it is decided to draw observations in blocks of two – the question of how to choose the block size is discussed below. Since the original sample has $T = 4$, it is necessary to draw two blocks with replacement from the original sample to make up one particular bootstrap sample. If the non-overlapping scheme is used then there are only two possible blocks: (v_1, v_2) and (v_3, v_4). If the overlapping scheme is

[9] The interested reader is refered to Andrews (2002*b*) and the survey in Li and Maddala (1996).

[10] The non–overlapping scheme is proposed by Carlstein (1986), and the overlapping scheme by Künsch (1989). Each method is sometimes refered to by the name of its proponent.

used then there are three possible blocks: (v_1, v_2), (v_2, v_3) and (v_3, v_4). Both schemes seem intuitively reasonable. To date, there has been no comparison of the two in the context of GMM. However, the available evidence in other contexts suggest that the overlapping scheme is to be prefered.[11] Nevertheless, we consider both below.

While the previous illustration gives the flavour of the block bootstrap, it turns out that the construction of the base sample is actually more complicated. This complication arises because the bootstrap is justified using Edgeworth expansions and these are only valid for dependent data if the statistics of interest are functions of sample moments involving the same variables and running over the same set of observations. If we view the sample in terms of the original observations $\{v_t\}_{t=1}^T$ then this structure is not present. However, if we alter our perspective on both the sampling unit and the sample size then the desired structure can be restored. To illustrate, we consider the simple case in which $k = 1$ and so $S = \Gamma_0 + \Gamma_1 + \Gamma_1'$. In this case the overidentifying restrictions test and confidence intervals are functions of the three basic sample statistics: the sample moment condition $T^{-1}\sum_{t=1}^T f(v_t, \theta)$, the derivative matrix $T^{-1}\sum_{t=1}^T \partial f(v_t, \theta)/\partial\theta'$ and the sample analog to the long run variance

$$\begin{aligned} S_T &= T^{-1}\sum_{t=1}^T f(v_t,\theta)f(v_t,\theta)' + T^{-1}\sum_{t=2}^T f(v_t,\theta)f(v_{t-1},\theta)' \\ &\quad + T^{-1}\sum_{t=2}^T f(v_{t-1},\theta)f(v_t,\theta)' \end{aligned} \tag{8.26}$$

It can be recognized that these three statistics do not collectively have the desired structure for two reasons. First, the sample moment and its derivative depend on v_t but the variance depends on v_t and v_{t-1}. Secondly, some of the summations start at $t = 1$ and some at $t = 2$. To solve the first problem, it is necessary to view the sampling unit as

$$\tilde{V}_t = \begin{bmatrix} v_t \\ v_{t-1} \end{bmatrix}$$

To solve the second problem, the sample is restricted to the observations $t = 2, 3 \ldots T$. These two amendments together imply that the sample is now viewed as consisting of $\{\tilde{V}_t; t = 2, \ldots T\}$.

These ideas extend easily to the general case defined by Assumption 8.2. The base sample for the bootstrap is $\mathcal{V}_B = \{\tilde{V}_t;\ t = k+1, k+2, \ldots T\}$ where $\tilde{V}_t$ is the $r(k+1) \times 1$ consisting of the $r \times 1$ vectors $\{v_{t-i};\ i = 0, 1, \ldots k\}$ stacked into a vector as follows

$$\tilde{V}_t = \begin{bmatrix} v_t \\ v_{t-1} \\ . \\ . \\ v_{t-k} \end{bmatrix} \tag{8.27}$$

[11] See Lahiri (1999).

It is important to realize that if the bootstrap is to yield the gains described in the previous sub-section then the GMM estimation must also be based on the same sample. This means, for instance, that the first and second step estimators must be calculated respectively as:

$$\hat{\theta}_{k,T}(1) = argmin_{\theta\in\Theta}\ g_{k,T}[\theta]' W_T g_{k,T}[\theta] \tag{8.28}$$

$$\hat{\theta}_{k,T}(2) = argmin_{\theta\in\Theta}\ g_{k,T}[\theta]' \{\hat{S}_{k,T}[\hat{\theta}_{k,T}(1)]\}^{-1} g_{k,T}[\theta] \tag{8.29}$$

where $g_{k,T}[\theta] = (T-k)^{-1}\sum_{t=k+1}^{T} f(v_t,\theta)$,

$$\hat{S}_{k,T}[\theta] \;=\; \hat{\Gamma}_{0,(k,T)}(\theta) \;+\; \sum_{i=1}^{k}\left\{\hat{\Gamma}_{i,(k,T)}(\theta) \;+\; \hat{\Gamma}_{i,(k,T)}(\theta)'\right\} \tag{8.30}$$

and $\hat{\Gamma}_{i,(k,T)}(\theta) = (T-k)^{-1}\sum_{t=k+1}^{T} f(v_t,\theta) f(v_{t-i},\theta)'$. Notice that the required structure for the bootstrap necessitates the use of the "truncated" covariance matrix estimator.[12] In comparison to the original definitions of these estimators, it is clear that there is some loss of information associated with taking this approach because, for example, the contribution of $\{f(v_i,\theta); i = 1,2,\ldots k\}$ to $\sum_t f(v_t,\theta)$ is lost. However, this is unavoidable because their retention would lead to a statistic that deviates from the required structure by terms of $O_p(T^{-1})$ and this would negate the anticipated gain from the use of the bootstrap.

Using the base sample in (8.27), we can now present the details of how to implement the block bootstrap. Needless to say, the exact details depend on the sampling scheme.

Non-overlapping blocks: The base sample, $\mathcal{V}_B$, is divided into b blocks of a pre-specified length ℓ. Denote these blocks by $\{B_i;\ i = 1,2,\ldots b\}$ where $B_1 = (\tilde{V}_{k+1}, \tilde{V}_{k+2}, \ldots \tilde{V}_{k+\ell})$, $B_2 = (\tilde{V}_{k+\ell+1}, \tilde{V}_{k+\ell+2}, \ldots \tilde{V}_{k+2\ell})$ and so forth. Notice that this means $T-k = b\ell$; if this is not the case for the desired choice of ℓ then additional observations must be dropped from the sample to ensure conformity. The n^{th} bootstrap sample is constructed by randomly sampling with replacement b blocks from $\{B_i;\ i = 1,2,\ldots b\}$.

Overlapping blocks: Let I denote the set of observations that can begin a block of ℓ observations, that is $I = \{k+1, k+2, \ldots T-\ell+1\}$. The construction of the n^{th} bootstrap sample begins with random sampling from I with replacement b times. If this random sample from I is denoted by $\{i_j;\ j = 1,2,\ldots b\}$ then the n^{th} bootstrap sample then consists of the b blocks that begin with the observations $\{i_j;\ j = 1,2,\ldots b\}$. So the first block in the bootstrap sample is $\tilde{B}_1^{(n)} = (\tilde{V}_{i_1}, \tilde{V}_{i_1+1}, \ldots \tilde{V}_{i_1+\ell})$, the second block is $\tilde{B}_2^{(n)} = (\tilde{V}_{i_2}, \tilde{V}_{i_2+1}, \ldots \tilde{V}_{i_2+\ell})$ and so forth.

In our subsequent discussion, it is necessary to express the bootstrap sample in

[12] See Section 3.5.3 for a discussion of its properties.

terms of the sampling unit instead of the blocks. Regardless of the sampling scheme used, we write the n^{th} bootstrap sample as $\tilde{\mathcal{V}}^{(n)} = \{\tilde{V}_s^{(n)};\ s = 1, 2, \ldots \tilde{T}\}$ where $\tilde{T} = T - k$.

To implement this approach, it is clearly necessary to choose the block length ℓ. Since the dynamic structure of the data is unknown, it is natural to consider rules in which the block size increases with T, such as $\ell_T = CT^{1/a}$. To date, there has been significant progress in deducing appropriate choices for a but less with regard to the selection of C. A complicating factor is that the choice of a depends on the statistic of interest. As mentioned above, we focus below on the use of the bootstrap to reduce the error in both the size of the overidentifying restrictions test and also in the confidence level of intervals for the unknown parameters. For these uses, Andrews (2002*b*) shows the optimal choice of a is 4.[13] He further shows that this is the optimal choice if the bootstrap is used to reduce the error in the size of the Wald tests and t–tests. In contrast, if the objective is to use the bootstrap to estimate the distribution function of the absolute value of the t–statistic then Hall, Horowitz, and Jing (1995) show that the optimal choice of a is 5.[14] To date there is no guidance available on the choice of C for minimizing the error in either the size of the overidentifying restrictions test or the confidence coverage of intervals based on the GMM estimator. However, there has been some progress on this issue in other settings; see Hall, Horowitz, and Jing (1995) and Bühlmann and Künsch (1996).

8.1.2.2 Calculation of the GMM Estimator and Related Statistics in the Bootstrap Samples

It can be recalled from Section 8.1.1 that even in the simple example of inference about a mean in an i.i.d. context, the functional form of the t-statistic differed in the original and bootstrap samples. Specifically, the bootstrap version of t-statistic involves a correction to ensure that it is invariant to whether or not the null hypothesis holds. A similar modification is necessary in the GMM setting. However, there is a second problem here that necessitates the introduction of an additional correction factor. While the block bootstrap seems an intuitively reasonable method for resampling from dynamic data, it does not yield samples with identical time series properties to the original data. Fortunately, it is possible to remedy the situation by the introduction of an additional correction factor. The exact nature of these corrections is discussed below as they arise in the sequence of necessary calculations. The presentation here only considers the case in which inference is based on the second step-estimator although, in principle, all definitions can be modified to accommodate inference based on the iterated estimator.[15]

Recall that the sample is now viewed in terms of the augmented vectors

[13] Optimal in the sense that this choice minimizes the error between the nominal and actual size of the test, and the nominal and actual coverage probability for the confidence interval.

[14] Optimal in the sense that it minimizes the mean squared error.

[15] Hall and Horowitz (1996) and Andrews (2002*b*) consider inference based on either the first-step or second-step estimators.

$\{\tilde{V}_s^{(n)}\}$ but the sample moment only depends on a sub-vector of $\tilde{V}_s^{(n)}$. Therefore, we decompose $\tilde{V}_s^{(n)}$ to reflect its structure as defined by (8.27), as follows

$$\tilde{V}_s^{(n)} \;=\; \begin{bmatrix} \tilde{v}_{s,1}^{(n)} \\ v_{s,2}^{(n)} \\ . \\ . \\ v_{s,k+1}^{(n)} \end{bmatrix} \text{“}=\text{”} \begin{bmatrix} v_t \\ v_{t-1} \\ . \\ . \\ v_{t-k} \end{bmatrix} \tag{8.31}$$

where the last identity is heuristic and included to remind the reader of the structure of $\tilde{V}$.

Before we proceed any further, it is necessary to address a matter relating to the notation. As with $\tilde{V}_s^{(n)}$, all the statistics calculated in the bootstrap sample should be indexed by n. However, this makes the notation extremely cumbersome and so we suppress this dependence during this part of the discussion. As emerges below, the statistics of interest are functions of both the bootstrap and also the original sample. All statistics calculated from the bootstrap sample are indicated by a tilde accent (i.e. $\tilde{a}$); all statistics calculated from the original sample are indicated by a hat accent (i.e. $\hat{a}$). No accent indicates that the statistic in question is a function of both samples.

In the original sample, the GMM estimator is obtained by minimizing a quadratic form in the sample moment. In the bootstrap sample, this moment must be centred to ensure that it is zero at $\hat{\theta}_{k,T}$ and thus mimics the property of the population moment which is zero at θ_0. This centered version of the bootstrap sample moment is calculated as:

$$\tilde{g}_{\tilde{T}}[\theta; \hat{\theta}_{k,T}] \;=\; \tilde{T}^{-1} \sum_{s=1}^{\tilde{T}} f_c(\tilde{v}_{s,1}, \theta;\, \hat{\theta}_{k,T}) \tag{8.32}$$

where

$$f_c(v, \theta;\, \hat{\theta}_{k,T}) \;=\; f(v, \theta) \;-\; m_T(\hat{\theta}_{k,T}) \tag{8.33}$$

where the c subscript stands for “centred”, and $m_T(.)$ is calculated from the original sample but the exact formula depends on the sampling scheme. If the non-overlapping scheme is used then

$$m_T(\theta) \;=\; g_{k,T}[\theta] \tag{8.34}$$

where $g_{k,T}[\theta]$ is defined under (8.29) above. If the overlapping scheme is used then

$$m_T(\theta) \;=\; (T - k - \ell + 1)^{-1} \sum_{t=k+1}^{T} w(t) f(v_t, \theta) \tag{8.35}$$

where

$$w(t) \;=\; \begin{cases} (t-k)/\ell & \text{if } t \in [k+1, \ell+k-1] \\ 1 & \text{if } t \in [\ell+k, T-\ell+1] \\ (T-t+1)/\ell & \text{if } t \in [T-\ell+2, T] \end{cases} \tag{8.36}$$

The first step GMM estimator in the bootstrap sample is calculated as follows:

$$\tilde{\theta}_{\tilde{T}}(1) \;=\; argmin_{\theta\in\Theta} \quad \tilde{g}_{\tilde{T}}[\theta;\hat{\theta}_{k,T}(1)]'\, W_T\, \tilde{g}_{\tilde{T}}[\theta;\hat{\theta}_{k,T}(1)] \tag{8.37}$$

Notice that on this first step, the bootstrap sample moment is centred using $m_T(.)$ evaluated at the first step GMM estimator defined in (8.28). The second step GMM estimator in the bootstrap sample is calculated as:

$$\tilde{\theta}_{\tilde{T}}(2) \;=\; argmin_{\theta\in\Theta} \quad \tilde{g}_{\tilde{T}}[\theta;\hat{\theta}_{k,T}(2)]' \left\{\tilde{S}_{\tilde{T}}[\tilde{\theta}_{\tilde{T}}(1);\hat{\theta}_{k,T}(1)]\right\}^{-1} \tilde{g}_{\tilde{T}}[\theta;\hat{\theta}_{k,T}(2)] \tag{8.38}$$

where

$$\begin{aligned}
\tilde{S}_{\tilde{T}}[\theta;\bar{\theta}] &= \tilde{\Gamma}_{0,\tilde{T}}(\theta;\bar{\theta}) + \sum_{i=1}^{k}\left\{\tilde{\Gamma}_{i,\tilde{T}}(\theta;\bar{\theta}) + \tilde{\Gamma}_{i,\tilde{T}}(\theta;\bar{\theta})'\right\} \\
\tilde{\Gamma}_{i,\tilde{T}}(\theta;\bar{\theta}) &= \tilde{T}^{-1}\sum_{s=1}^{\tilde{T}} f_c(\tilde{v}_{s,1},\theta;\,\bar{\theta}) f_c(\tilde{v}_{s,i+1},\theta;\,\bar{\theta})'
\end{aligned}$$

and $f_c(.)$ is defined in (8.33). Two aspects of the second step minimand are worth noting. First, the bootstrap sample moment is centred using $m_T(.)$ evaluated at the second step estimator defined in (8.29). Secondly, the long run variance estimator is calculated using the centred sample moment.

Recall that we consider here only two statistics associated with the second step estimator: the overidentifying restrictions test and a confidence interval for $\theta_{0,i}$. The distribution of the overidentifying restrictions test is approximated using the following statistic calculated in the bootstrap samples:

$$\tilde{J}_{\tilde{T}} \;= \tilde{H}_{\tilde{T}}[\tilde{\theta}_{\tilde{T}}(2)] \tag{8.39}$$

where[16]

$$\tilde{H}_{\tilde{T}}[\theta] = \tilde{T}\tilde{g}'_{\tilde{T}}[\theta;\hat{\theta}_{k,T}(1)]\{\tilde{S}_{\tilde{T}}[\theta;\hat{\theta}_{k,T}(2)]\}^{-1/2} A^{+}_{\tilde{T}}\{\tilde{S}_{\tilde{T}}[\theta;\hat{\theta}_{k,T}(2)]\}^{-1/2}\tilde{g}_{\tilde{T}}[\theta;\hat{\theta}_{k,T}(2)] \tag{8.40}$$

In the previous equation, $A^{+}_{\tilde{T}}$ denotes the Moore–Penrose generalized inverse of the matrix $A_{\tilde{T}}$ that is calculated as follows:

$$A_{\tilde{T}} \;=\; \hat{M}_{k,T}\hat{S}^{-1/2}_{k,T} B_{\tilde{T}} \hat{S}^{-1/2}_{k,T}\hat{M}_{k,T} \tag{8.41}$$

where

$$\hat{M}_{k,T} = I_q - \hat{S}^{-1/2}_{k,T}\hat{G}_{k,T}\hat{C}_{k,T}\hat{G}'_{k,T}\hat{S}^{-1/2}_{k,T} \tag{8.42}$$

$$\hat{C}_{k,T} = [\hat{G}'_{k,T}\hat{S}^{-1}_{k,T}\hat{G}_{k,T}]^{-1} \tag{8.43}$$

$$\hat{S}_{k,T} = \hat{S}_{k,T}[\hat{\theta}_{k,T}(2)] \tag{8.44}$$

$$\hat{G}_{k,T} = \hat{G}_{k,T}[\hat{\theta}_{k,T}(2)] \tag{8.45}$$

[16] Following the practice in this literature, $Z^{-1/2}$ denotes the symmetric square root of Z^{-1} for any nonsingular symmetric real matrix Z, that is if the spectral decomposition of Z is $Z_1 Z_2 Z_1'$ then $Z^{-1/2} = Z_1 Z_2^{-1/2} Z_1'$.

for $\hat{S}_{k,T}[\theta]$ is defined in (8.30), and

$$\hat{G}_{k,T}[\theta] = \tilde{T}^{-1} \sum_{t=k+1}^{T} \partial f(v_t, \theta)/\partial\theta' \tag{8.46}$$

The last component of $A_{\tilde{T}}$ is a matrix $B_{\tilde{T}}$ whose calculation depends on the sampling scheme employed. If the non-overlapping scheme is used then:

$$B_{\tilde{T}} \;=\; \tilde{T}^{-1} \sum_{i=0}^{b-1} \sum_{j=1}^{\ell} \sum_{m=1}^{\ell} h_T(i\ell + j + k) h_T(i\ell + m + k)' \tag{8.47}$$

where

$$h(t) \;=\; f(v_t, \hat{\theta}_{k,T}(2)) - m_T(\hat{\theta}_{k,T}(2))$$

If the overlapping scheme is used then:

$$B_{\tilde{T}} \;=\; b\tilde{T}^{-1}(\tilde{T} - \ell + 1)^{-1} \sum_{i=0}^{\tilde{T}-\ell} \sum_{j=1}^{\ell} \sum_{m=1}^{\ell} h_T(i + j + k) h_T(i + m + k)' \tag{8.48}$$

The bootstrap version of the confidence interval is based on approximating the distribution of the absolute value of the t-ratio by

$$a\tilde{\tau}_{\tilde{T},i} \;=\; |\tilde{\tau}_{\tilde{T},i}| \;=\; c_i \frac{\tilde{T}^{1/2} |\tilde{\theta}_{\tilde{T},i}(2) \;-\; \hat{\theta}_{k,T,i}(2)|}{\sqrt{\{\tilde{V}_{\tilde{T}}\}_{i,i}}} \tag{8.49}$$

where $\hat{\theta}_{k,T,i}(2)$ is the i^{th} element of $\hat{\theta}_{k,T}(2)$, $\{.\}_{i,i}$ denotes the $i - i^{th}$ diagonal element of the matrix in parentheses, and $\tilde{V}_{\tilde{T}}$ is given by

$$\tilde{V}_{\tilde{T}} \;=\; \left[\tilde{G}'_{\tilde{T}} \{\tilde{S}_T[\tilde{\theta}_{\tilde{T}}(2); \hat{\theta}_{k,T}(2)]\}^{-1} \tilde{G}_{\tilde{T}}\right]^{-1} \tag{8.50}$$

and

$$\tilde{G}_{\tilde{T}} = \tilde{T}^{-1} \sum_{s=1}^{\tilde{T}} \partial f(\tilde{v}_{s,1}, \tilde{\theta}_{\tilde{T}}(2))/\partial\theta' \tag{8.51}$$

The correction factor c_i is defined as

$$c_i \;=\; \sqrt{\frac{\{\hat{C}_{k,T}\}_{i,i}}{\{D_{\tilde{T}}\}_{i,i}}} \tag{8.52}$$

where $\hat{C}_{k,T}$ is defined above in (8.43), and

$$D_{\tilde{T}} \;=\; \hat{C}_{k,T} \hat{G}'_{k,T} \hat{S}^{-1}_{k,T} B_{\tilde{T}} \hat{S}^{-1}_{k,T} \hat{G}_{k,T} \hat{C}_{k,T} \tag{8.53}$$

The bootstrap percentile is based on the absolute value because the objective here is to calculate a *symmetric* confidence interval for $\theta_{0,i}$, that is of the generic form $\hat{\theta}_{k,T,i} \pm \hat{n}_T$.

Inspection of (8.39)–(8.40) and (8.49) reveals that the bootstrap versions of the overidentifying restrictions statistics and t-ratio have two types of correction relative to their sample counterparts. First, each statistic has a "centering" correction similar in spirit to the correction required to the t-statistic in our motivating example: – the sample moment is centred in the overidentifying restrictions test; – the t-ratio is centred using the corresponding GMM estimator from the original sample. However, in this more general setting, it is also necessary to make a second correction and this leads to the presence of $A_{\tilde{T}}^{+}$ in (8.40) and c_i in (8.49). These additional corrections are needed because the block bootstrap does not adequately replicate the time series properties of the original data. The problem stems from the long run variance. In the original population, this variance only involves terms of the form $E[f(v_s, \theta_0)f(v_t, \theta_0)']$ for $|s - t| < k$. However, in the population generated from the bootstrap distribution, this long run variance involves terms of the form $\tilde{E}[f(v_s, \theta_0)f(v_t, \theta_0)']$ for all s, t within the same block where $\tilde{E}[.]$ denotes expectations relative to the bootstrap distribution.[17] Note that this second correction is only needed because the block bootstrap is used in an attempt to take account of the dynamic dependence structure in the data. If the data are independent then there is no need to use the block bootstrap and so the problem goes away. It is for this reason that this second correction is unnecessary in our motivating example in Section 8.1.1.

Bootstrap methods are inherently computationally burdensome because of the resampling. However, in our setting, the burden is potentially far greater because it is necessary to perform two numerical optimizations for each bootstrap sample. Fortunately, it is not necessary to iterate gradient methods until convergence within the numerical optimization in order to gain the asymptotic refinements associated with the bootstrap.[18] Davidson and MacKinnon (1999) present a heuristic justification for this statement and Andrews (2002*b*) subsequently provides a rigorous demonstration. For our purposes here, it suffices to concentrate on the heuristic argument. To illustrate, we focus attention on the Newton–Raphson method of optimization in which the the estimator is updated on the i^{th} step of the numerical optimization according to

$$\bar{\theta}(i) \;=\; \bar{\theta}(i-1) \;-\; \left[\frac{\partial^2 Q_T\left(\bar{\theta}(i-1)\right)}{\partial\theta\partial\theta'} \right]^{-1} \frac{\partial Q_T(\bar{\theta}(i-1))}{\partial\theta}$$

Typically, this updating is continued until convergence to give the estimator $\hat{\theta}_T$. However, for our purposes here, it is important to consider the way in which this convergence occurs. Robinson (1988) shows that if $\bar{\theta}(0) - \hat{\theta}_T = O_p(T^{-1/2})$ then $\bar{\theta}(i) - \hat{\theta}_T = O_p(T^{-2^{i-1}})$. Using this property, Davidson and MacKinnon (1999) make the important observation that it is only necessary to iterate two steps before the difference between $\bar{\theta}(i)$ and $\hat{\theta}_T$ is of smaller order than asymptotic refinements associated with the bootstrap. Therefore, it suffices to use

[17] See Hall and Horowitz (1996) for further discussion.
[18] See Section 3.2 for a discussion of gradient methods.

the Newton–Raphson method with only two iterations in the numerical optimizations within the bootstrap samples. Needless to say, the specifics depend on the exact method of numerical optimization. If the Gauss–Newton method is used then at least three steps are needed in the numerical optimization; see Davidson and MacKinnon (1999) or Andrews (2002*b*). Davidson and MacKinnon (1999) introduce the term "approximate bootstrap" to describe the generic strategy of fixing the number of iterations within the numerical optimization in the bootstrap samples. To implement the approximate bootstrap, it is necessary to have a suitable starting value for the optimization. The natural choice is the corresponding GMM estimator from the observed sample, and Andrews (2002*b*) verifies that this choice is appropriate.

Clearly the approximate bootstrap has the potential to reduce the computational burden considerably. However, two caveats need to be borne in mind. First, the results above only yield a minimum number of iterations. Intuition suggests that the results can never be worse if more iterations are used within the numerical optimization routine. To date, there is no evidence on how the number of iterations affects the accuracy of subsequent inferences in overidentified nonlinear dynamic models. Secondly, the available results only cover variants of the Newton–Raphson and Gauss–Newton routines. It is an open question whether the approximate bootstrap can be extended to other gradient methods.

8.1.2.3 Choosing the Number of Replications

It can be recalled that the motivation for the bootstrap derives from its ability to provide asymptotic refinements to inference. However, the argument is based on a consideration of what is often termed the "ideal bootstrap" in which the number of replications, N, tends to infinity. Clearly this version of the bootstrap is infeasible, and so we now consider a three-step method for the selection of N proposed by Andrews and Buchinsky (2000). The precise details of this method are application specific. Following our practice above, we limit our discussion to the two statistics of interest, the overidentifying restrictions test and a two–sided symmetric confidence interval for elements of θ_0. For the former, we focus purely on using the bootstrap to calculate the critical point for a given level of significance. Therefore, in both cases, the bootstrap is being used to calculate a pre–specified percentile of a distribution. It should be noted that our discussion is specific to this precise context. The details of the method would be different if, for example, the bootstrap is used to calculate the p-value of the test or if it is used to calculate two-sided equal-tailed confidence intervals.[19]

Since both our cases of interest involve the use of the bootstrap to calculate a particular percentile, we abstract to a generic notation to simplify the presentation. Accordingly, let Ξ_T denote the statistic calculated in the original sample and $\tilde{\Xi}_T^{(n)}$ denote the statistic calculated in the n^{th} bootstrap sample. Let p_T denote the true $100(1-\alpha)\%$ percentile of Ξ_T, $\tilde{p}_{T,\infty}$ be the corresponding

[19] Andrews and Buchinsky (2000) consider both these cases along with many others of empirical relevance.

percentile based on the ideal bootstrap and $\tilde{p}_{T,N}$ be the same statistic based on the bootstrap with N replications. It turns out to be convenient to restrict attention to choices of N that satisfy:

$$\frac{\nu}{N+1} \;=\; 1 \,-\, \alpha \tag{8.54}$$

where ν is some positive integer, because then $\tilde{p}_{T,N}$ is the ν^{th} order statistic of the bootstrap sample $\{\tilde{\Xi}_T^{(n)};\, n = 1, 2, \ldots N\}$. However, it should be noted that this is a non–trivial restriction: for example, $\alpha = 0.05$ then the set of possible values for N is $\{20h - 1;\, h = 1, 2, 3, \ldots\}$; see Andrews and Buchinsky (2000) for further details.

Since the bootstrap is motivated by the properties of the ideal bootstrap, it is natural to base the selection rule for N upon some measure of the distance between $\tilde{p}_{T,N}$ and $\tilde{p}_{T,\infty}$. In Andrews and Buchinsky's (2000) three-step method, this distance is captured by the percentage deviation of $\tilde{p}_{T,N}$ from $\tilde{p}_{T,\infty}$, that is

$$\frac{100\,|\tilde{p}_{T,N} \,-\, \tilde{p}_{T,\infty}|}{\tilde{p}_{T,\infty}}$$

It can be recalled from Section 8.1.1 that the bootstrap percentiles are random because they are conditional on the observed sample. Therefore, any statements about the distance between the percentiles must have a probability attached, and so take the form

$$P^B\left[\frac{100\,|\tilde{p}_{T,N} \,-\, \tilde{p}_{T,\infty}|}{\tilde{p}_{T,\infty}} \;\leq\; pdb\right] \;=\; (1 \,-\, \beta) \tag{8.55}$$

where, once again, $P^B[\,.\,]$ denotes probability based on the bootstrap distribution. The three-step method provides a rule for selecting N for prespecified values of the percentage deviation bound, *pdb*, and the probabilty $1 - \beta$. Although it does not serve our purposes here to explore the theoretical underpinnings of the method here, it is worth noting that it derives from the following limiting result:

$$N^{1/2}\left(\frac{\tilde{p}_{T,N} \,-\, \tilde{p}_{T,\infty}}{\tilde{p}_{T,\infty}}\right) \;\xrightarrow{d}\; N(0,\, \omega) \tag{8.56}$$

where ω is application specific. The method involves the construction of consistent estimates for ω that can be used in conjunction with the distributional result in (8.56) to deduce a value for N for which (8.55) is satisfied for a given *pdb* and $1 - \beta$. The details are as follows.

Andrews and Buchinsky's (2000) three-step method for the selection of N

Define α_1 and α_2 such that $\alpha = \alpha_1/\alpha_2$ where α_1 and α_2 are positive integers with no common divisors.[20] It is also necessary to specify (pdb, β).

Step 1: Calculate an initial value for N as follows: $N_1 = \alpha_2 h_1 - 1$ where $h_1 = int[10,000 z^2_{1-\beta/2}\omega_1/(pdb^2\alpha_2)]$, ω_1 is an initial estimator of ω given on a case by case basis in Tables 8.1–8.2, z_δ is the 100δ percentile of the standard normal distribution, and $int[\,.\,]$ denotes the integer part.

Step 2: Simulate N_1 bootstrap samples $\{\tilde{\mathcal{V}}^{(n)};\ n = 1, 2, \ldots N_1\}$ and compute the updated estimator of ω, ω_2, the formula for which is given in Tables 8.1–8.2 on a case by case basis.

Step 3: Calculate an updated estimator of N as follows: $N_2 = \alpha_2 h_2 - 1$ where $h_2 = int[10,000 z^2_{1-\beta/2}\omega_2/(pdb^2\alpha_2)]$.

The selected number of replications is $N^* = max\{N_1, N_2\}$.

Table 8.1
ω_1 and ω_2 in the calculation of the $100\alpha\%$ critical value of the overidentifying restrictions test

ω_i	Quantities used in ω_i
$\omega_1 = \dfrac{\alpha(1-\alpha)}{p^2_{1-\alpha} g^2(p_{1-\alpha})}$	$g(x) = \dfrac{x^{d-1}e^{-x/2}}{\Gamma(d)2^d} \qquad d = (q-p)/2$ $p_{1-\alpha}$ = the $100(1-\alpha)\%$ percentile of the χ^2_{q-p} distribution
$\omega_2 = \dfrac{\alpha(1-\alpha)}{\tilde{p}^2_{1-\alpha}\tilde{g}^2}$	$\tilde{g} = \left\{\dfrac{N_1}{2\tilde{m}}(\tilde{J}^*_{\nu+\tilde{m}} - \tilde{J}^*_{\nu-\tilde{m}})\right\}^{-1}$ $\tilde{J}^*_i$ = the i^{th} order statistic of $\{\tilde{J}^{(n)}_{\tilde{T}};\ n = 1, 2, \ldots N_1\}$ $\nu = (N_1 + 1)(1 - \alpha)$ $\tilde{m} = int[c_\alpha N_1^{2/3}]$ $\tilde{p}_{1-\alpha} = \tilde{J}^*_\nu$ $c_\alpha = \left\{\dfrac{1.5 z^2_{1-\alpha/2} g^4(p_{1-\alpha})}{3g'(p_{1-\alpha})^2 - g(p_{1-\alpha})g''(p_{1-\alpha})}\right\}^{1/3}$ $g'(x) = g(x)\{(d-1)x^{-1} - 0.5\}$ $g''(x) = g'(x)\{(d-1)x^{-1} - 0.5\} - g(x)(d-1)x^{-2}$

Notes: In this context, $\Gamma(\,.\,)$ denotes the "gamma function" and is not to be confused with our earlier use of this symbol for an autocovariance matrix. It is defined as follows: $\Gamma(0.5) = \sqrt{\pi}$; if a is an even integer then $\Gamma(a/2) = (a/2-1)\ldots 3.2.1$; if a is an odd integer then $\Gamma(a/2) = (a/2-1)\ldots \frac{5}{2}\cdot\frac{3}{2}\cdot\frac{1}{2}\sqrt{\pi}$; *e.g.* see Hamilton (1994) [p.355].

[20] So, for example, if $\alpha = 0.10, 0.05, 0.01$ then (α_1, α_2) are respectively $(1,10), (1,20), (1,100)$.

Table 8.2
ω_1 and ω_2 in the calculation of the $100(1-\alpha)\%$ symmetric confidence interval for $\theta_{0,i}$

ω_i	Quantities used in ω_i
$\omega_1 = \dfrac{\alpha(1-\alpha)}{z^2_{1-\alpha/2}\left[2\phi(z_{1-\alpha/2})\right]^2}$	$\phi(z) = (2\pi)^{-1/2}e^{-z^2/2}$
$\omega_2 = \dfrac{\alpha(1-\alpha)}{\tilde{p}^2_{1-\alpha}\tilde{h}^2}$	$\tilde{h} = \left\{\dfrac{N_1}{2\tilde{m}}(\tilde{a\tau}^*_{i,\nu+\tilde{m}} - \tilde{a\tau}^*_{i,\nu-\tilde{m}})\right\}^{-1}$ $\tilde{a\tau}^*_{i,j}$ = the j^{th} order statistic of $\{\tilde{a\tau}^{(n)}_{\tilde{T},i}; n = 1,2,\ldots N_1\}$ $\nu = (N_1+1)(1-\alpha)$ $\tilde{m} = int[c_\alpha N_1^{2/3}]$ $\tilde{p}_{1-\alpha} = \tilde{a\tau}^*_\nu$ $c_\alpha = \left\{\dfrac{6z^2_{1-\alpha/2}[\phi(z_{1-\alpha/2})]^2}{2z^2_{1-\alpha/2}+1}\right\}^{1/3}$

8.1.2.4 Summary of Bootstrap Calculations

Pulling together the previous discussion, it can be seen that there are four major steps to implementing the bootstrap in the context here.

Step 1: Re-estimate the model using the truncated original sample as in (8.28)–(8.29).

Step 2: Generate N_1 bootstrap samples in the way described in Section 8.1.2.1 for N_1 defined in the Andrews and Buchinsky's (2000) three-step procedure in Section 8.1.2.3.

Step 3: Calculate the bootstrap distributions for the statistic of interest as described in Section 8.1.2.2 and compute N_2, and hence N^*, defined in Andrews and Buchinsky's (2000) three-step procedure in Section 8.1.2.3.

Step 4: Generate N^* bootstrap samples as in Section 8.1.2.1 and calculate the required statistics of interest as in section 8.2.1.2.[21]

Inference is then conducted as follows. For the overidentifying restrictions test the decision rule is to reject $H_0 : E[f(v_t,\theta)] = 0$ at the $100\alpha\%$ significance level if:

$$J_T > \tilde{J}^*_\nu \tag{8.57}$$

[21] Note that this step does not involve additional calculations if $N^* = N_1$ and only the generation of an additional $N_2 - N_1$ bootstrap samples if $N^* = N_2$.

where $\tilde{J}_\nu^*$ is defined in Table 8.1. This decision rule yields an approximate $100\alpha\%$ test because the true size of the test is given by:

$$P(J_T > \tilde{J}_\nu \,|\, H_0) = \alpha + o(T^{-1})$$

The $100(1-\alpha)\%$ bootstrap confidence interval for $\theta_{0,i}$ is given by

$$\hat{\theta}_{k,T,i} \pm \tilde{a\tau}^*_{i,\nu}\sqrt{\hat{V}_{i,i}/(T-k)} \tag{8.58}$$

where $\tilde{a\tau}_{i,\nu^*}$ is defined in Table 8.2, $\hat{V}_{i,i}$ is the $i-i^{th}$ element of $\hat{V}$ where

$$\hat{V} = [\hat{G}'_{k,T}\hat{S}^{-1}_{k,T}\hat{G}_{k,T}]^{-1}$$

and the component matrices are defined in (8.44)–(8.45). Once again, the attached probability statement is only approximate as the true coverage probability is:

$$P\left(\hat{\theta}_{k,T,i} - \tilde{a\tau}^*_{i,\nu}\sqrt{\hat{V}_{i,i}/(T-k)} \leq \theta_{0,i} \leq \hat{\theta}_{k,T,i} + \tilde{a\tau}^*_{i,\nu}\sqrt{\hat{V}_{i,i}/(T-k)}\right)$$
$$= 1 - \alpha + o(T^{-1})$$

We now illustrate these calculations using our running empirical example.

Example: Hansen and Singleton's (1982) Consumption Based Asset Pricing Model

We use the bootstrap to calculate the critical value associated with performing the overidentifying restrictions test at the 5% significance level, and the percentiles needed to construct 95% symmetric confidence intervals for the parameters. Our description of the necessary calculations follows the four-step sequence described above. For this particular model *Step 1* is redundant because $k = 0$. To implement *Step 2*, it is necessary to determine the sampling unit and sample size for the bootstrap sample. It can be recalled for this model

$$f(v_t, \theta) = z_t(\delta x_{1,t+1}^{\gamma-1} x_{2,t+1} - 1)$$

where $x_{1,t+1} = c_{t+1}/c_t$, $x_{2,t+1} = r_{t+1}/p_t$ and $z_t = (1, x_{1,t}, x_{2,t}, x_{1,t-1}, x_{2,t-1})'$. Therefore the sampling unit for the bootstrap sample is

$$\tilde{V}_t = \begin{bmatrix} x_{1,t+1} \\ x_{2,t+1} \\ x_{1,t} \\ x_{2,t} \\ x_{1,t-1} \\ x_{2,t-1} \end{bmatrix}$$

Since $k = 0$, the bootstrap sample size is the same as the original sample, that is $\tilde{T} = T = 465$. We next turn to the block size, ℓ. It can be recalled that from Section 8.1.2.1 that the optimal block size depends on the statistic being calculated. Since we calculate a fixed percentile of a distribution, the optimal block size is $O(T^{1/4})$. For this example, the sample size is 465, and so $T^{1/4} = 4.6437$. Obviously, the block size must be an integer and so we fix $\ell = 5$. This is a convenient choice because it means there are exactly ninety-three non-overlapping blocks in the sample. To calculate the number of replications, we set $pdb = 0.05$ and $\beta = 0.95$. The resulting choice of N_1 is 2379 for the overidentifying restrictions test and 1379 for the confidence intervals. Therefore, for simplicity, we set $N_1 = 2379$ for both types of statistics. Parenthetically, it should be noted that in a very small number of the bootstrap samples $\tilde{S}_T[\tilde{\theta}_{\tilde{T}}(2); \hat{\theta}_{k,T}(2)]$ was singular. Therefore the number of bootstrap samples was actually set equal to $N_1 + 50$ and the calculations were based on the first N_1 bootstrap samples for which the calculation of the required statitics was succesful.

We report results using both choices of first-step weighting matrix used in our empirical example. For the case in which $W_T^{(1)}$ is the inverse of the instrument cross product matrix, the first step estimation is performed in the bootstrap sample using the corresponding matrix constructed from the bootstrap sample, denoted here by $(\tilde{T}^{-1} \sum_{t=1}^{\tilde{T}} \tilde{z}_t \tilde{z}_t')^{-1}$. Given the nonlinearity of the model, it is particularly attractive to use the approximate bootstrap to reduce the computational burden. Unfortunately, to date, there are no theoretical results to offer guidance on the number of steps needed to obtain asymptotic refinements for the particular optimization routine used in *fminu* in MATLAB. So for illustrative purposes, the method is performed using $2, 4, 6, 8, 10, 20, 30, 40, 50$ and 100 steps. For this example, it turns out that the percentiles are sensitive to the number of steps allowed. As an illustration, Table 8.3 reports the appropriate percentiles based on $N = N_1$ replications for the case in which the asset is VWR using both choices of first step weighting matrix and blocking scheme. It should be noted that, for a given blocking scheme, the calculations reported in Table 8.3 are all based on the same set of bootstrap samples and so the only difference is in the number of steps in the approximate bootstrap.

As a reminder, the corresponding percentiles from the limiting distributions are 7.815 for the overidentifying restrictions test, and 1.96 for the t-statistics. Inspection reveals that the bootstrap percentiles are for the most part close to these limiting values. However, the percentile are also clearly sensitive to the blocking scheme, the number of steps in the approximate bootstrap, and also the choice of first step weighting matrix. Of these three, the nature of the Edgeworth expansion would lead us to anticipate the sensitivity of the percentiles to $W_T^{(1)}$, but the rest are less easily explained. The sensitivity of the bootstrap percentiles to the number of steps in the numerical optimization suggests that the theoretical results outlined above do not extend to the routines

Table 8.3
Bootstrap 95^{th} percentiles for the overidentifying restrictions test and the absolute values of the t-statistics with VWR

$W_T^{(1)}$	i_{max}	Non-overlapping blocks			Overlapping blocks		
		$\tilde{J}_{\tilde{T}}$	$\tilde{a\tau}_{\tilde{T},\gamma}$	$\tilde{a\tau}_{\tilde{T},\delta}$	$\tilde{J}_{\tilde{T}}$	$\tilde{a\tau}_{\tilde{T},\gamma}$	$\tilde{a\tau}_{\tilde{T},\delta}$
$10^5 I_5$	2	8.311	0.000	1.161	8.071	0.000	1.087
	4	8.311	0.000	1.161	8.071	0.000	1.088
	6	8.208	1.586	1.593	7.664	1.518	1.521
	8	7.740	2.018	1.983	7.499	1.958	1.983
	10	7.794	1.995	1.990	7.438	1.937	1.966
	20	7.747	2.035	2.009	7.322	1.985	1.990
	30	7.706	2.030	2.018	7.263	2.002	2.006
	40	7.731	2.035	2.040	7.264	2.007	2.010
	50	7.731	2.049	2.027	7.266	2.007	2.012
	100	7.736	2.048	2.027	7.266	2.007	2.012
$(\tilde{T}^{-1}\sum_{t=1}^{\tilde{T}} \tilde{z}_t\tilde{z}_t')^{-1}$	2	8.238	0.000	1.156	7.564	0.000	1.100
	4	8.238	0.000	1.156	7.564	0.000	1.103
	6	8.114	1.496	1.529	7.546	1.573	1.621
	8	7.936	2.012	1.966	7.258	2.024	1.973
	10	7.822	1.964	1.960	7.214	2.022	1.971
	≥20	7.814	1.966	1.960	7.251	2.022	1.973

Notes: $\tilde{J}_{\tilde{T}}$ is defined in (8.39), $\tilde{a\tau}_{\tilde{T},\gamma}$ and $\tilde{a\tau}_{\tilde{T},\delta}$ are defined in (8.49) with the i subscript replaced by the symbol for the parameter in question. i_{max} denotes the maximum number of steps in approximate bootstrap.

used here – at least for $i_{max} \leq 100$. A striking feature of this sensitivity is that the percentile for $\tilde{a\tau}_{\tilde{T},\gamma}$ is zero to five decimal places with $i_{max} = 2, 4$ with either choice of weighting matrix or blocking scheme. We conjecture that this reflects a feature of the moment condition noted in Section 3.6, namely that it is nearly uninformative about γ_0. However, a deeper analysis is left to future research.

The next step is to calculate N_2. For brevity, we focus on the case where $i_{\max} = 100$. Table 8.4 reports the values of N_2 and the final bootstrap percentiles for both choices of asset. As can be seen, the bootstrap percentiles for the overidentifying restrictions test do not alter our verdict about the specifications. As with inference based on the asymptotic critical values, the model is rejected for EWR but not with VWR. Given this evidence, it is only interesting to consider the confidence intervals for the parameters based on VWR. However, since the percentiles for the t-statistics for γ and δ are so close to the corresponding values from the standard normal distribution, this is left as an exercise for the reader.

Table 8.4
N_2 and final bootstrap 95^{th} percentiles for the overidentifying restrictions test and confidence intervals for γ_0 and δ_0

		Non-overlapping blocks			Overlapping blocks		
Asset	$W_T^{(1)}$	$N_2(\tilde{J}_{\tilde{T}})$	$N_2(\tilde{a\tau}_{\tilde{T},\gamma})$	$N_2(\tilde{a\tau}_{\tilde{T},\delta})$	$N_2(\tilde{J}_{\tilde{T}})$	$N_2(\tilde{a\tau}_{\tilde{T},\gamma})$	$N_2(\tilde{a\tau}_{\tilde{T},\delta})$
VWR	A	2999	1819	1819	6659	1399	1199
	B	4939	1159	1299	2199	899	1179
EWR	A	2799	939	939	3419	959	1359
	B	4079	1999	1919	2139	1499	1099
		Non-overlapping blocks			Overlapping blocks		
Asset	$W_T^{(1)}$	$\tilde{J}_{\tilde{T}}$	$\tilde{a\tau}_{\tilde{T},\gamma}$	$\tilde{a\tau}_{\tilde{T},\delta}$	$\tilde{J}_{\tilde{T}}$	$\tilde{a\tau}_{\tilde{T},\gamma}$	$\tilde{a\tau}_{\tilde{T},\delta}$
VWR	A	7.684	2.013	2.000	7.544	1.973	1.994
	B	7.566	2.045	2.020	7.251	2.022	1.973
EWR	A	7.701	*NA*	*NA*	7.667	*NA*	*NA*
	B	7.415	*NA*	*NA*	7.573	*NA*	*NA*

Notes: $N_2(.)$ denotes the value for N_2 calculated using the formulae for associated with the statistics in the parentheses. A denotes $W_T^{(1)} = 10^5 I_5$, B denotes $W_T^{(1)} = (\tilde{T}^{-1}\sum_{t=1}^{\tilde{T}} \tilde{z}_t\tilde{z}_t')^{-1}$. NA denotes "not applicable". For other definitions see Table 8.3.

8.2 Inference in the Presence of Weak Identification

The asymptotic theory in Chapters 3 and 5 is predicated on the assumption that the parameter vector is identified by the population moment condition used in the estimation. In recent years there has been a growing awareness that this proviso may not be so trivial in situations which arise in practice. In a very influential paper, Nelson and Startz (1990) draw attention to this potential problem and provided the first evidence of the problems it causes for the inference framework we have described above. Their paper has prompted considerable interest in the behaviour of GMM in cases in which the parameter vector is weakly identified. In this section we provide a review of this literature.

To begin, it is necessary to define what is meant by the term "weakly identification". The essence of the concept is most easily understood within a simple example. Accordingly, we consider the simple linear regression model

$$y_t \;=\; x_t\theta_0 \;+\; u_t \tag{8.59}$$

in which u_t is an i.i.d. process with mean zero and variance σ_0^2. Suppose the scalar parameter θ_0 is estimated by Instrumental Variables which, as we have

seen, is just GMM estimation based on the population moment condition

$$E[z_t u_t(\theta_0)] \;=\; 0 \tag{8.60}$$

where z_t is a $q \times 1$ vector of instruments and $u_t(\theta_0) = y_t - x_t\theta_0$. From Section 2.1, it can be recalled that θ_0 is identified by (8.60) if $rank\{E[z_t x_t]\} = 1$. In this simple example, θ_0 is unidentified if $E[z_t x_t]$ is the null vector, which would occur if z_t and x_t are uncorrelated and both possess zero means. In practice, it is unlikely that $E[z_t x_t]$ is exactly zero. The contribution of Nelson and Startz's (1990) paper is to demonstrate that problems occur if $E[z_t x_t]$ is non-zero but small. It is this scenario which is refered to as "weak identification". It is also convenient to have a terminology that describes this scenario in terms of the population moment condition. Therefore, if the parameter vector is weakly identified then the population moment condition is said to be nearly uninformative about the parameter vector.

To proceed, it is necessary to develop a model which can capture the idea of nearly uninformative moment conditions. Staiger and Stock (1997) solve this problem by assuming that

$$x_t \;=\; z_t^{'}\gamma_T \;+\; e_t \tag{8.61}$$

where $\gamma_T = T^{-1/2}c$, c is a non-zero $q \times 1$ vector of constants, and e_t is the unobserved error which has both a zero mean and is uncorrelated with z_t.[22] Using similar logic to the derivation of (2.3), it follows that this specification implies

$$E_T[z_t u_t(\theta)] \;=\; \{E[z_t z_t^{'}]\}T^{-1/2}c(\theta_0 \,-\, \theta) \tag{8.62}$$

Therefore, θ_0 is identified by (8.60) for finite T but is not in the limit as $T \to \infty$.[23] So the concept of nearly uninformative moment conditions is captured by assuming that the information in the population moment condition disappears at rate $T^{-1/2}$. This rate is chosen so that the effects of the nearly uninformative moment conditions manifest themselves in the limiting behaviour of the estimator. Since $p = 1$, we have

$$\hat{\theta}_T - \theta_0 \;=\; \frac{x'Z(Z'Z)^{-1}Z'u}{x'Z(Z'Z)^{-1}Z'x} \tag{8.63}$$

in the obvious notation. As in Section 2.3, the limiting behaviour of $\hat{\theta}_T - \theta_0$ depends on the limiting behaviour of the components on the right hand side of (8.63). Using the Weak Law of Large Numbers and the Central Limit Theorem respectively, it follows that: (i) $T^{-1}Z'Z \xrightarrow{p} M_{zz}$, a positive definite matrix of constants; (ii) $T^{-1/2}Z'u \xrightarrow{d} N(0, \sigma_0^2 M_{zz})$ – assuming here for simplicity that z_t

[22] Notice this design involves exactly the same type of Pitman drift that is used to set up local alternatives to hypothesis tests; see Section 5.1.3. Equation (8.61) implies the explanatory variable is a triangular array $\{x_{t,T}; t = 1, 2, \ldots T; T = 1, 2, \ldots\}$ but we suppress the second subscript for notational brevity. This structure also implies that the distribution of x_t is indexed by T and so we index the expectation operator by T when it is applied to functions of x_t.

[23] See Section 2.1 for a discussion of identification in the linear model.

is independent of u_t. Notice that neither (i) nor (ii) involve the relationship between x_t and z_t and so would equally hold if θ_0 is properly identified. The key difference comes in the behaviour of $Z'x$. From (8.61), it follows that

$$Z'x \;=\; T^{-1/2}Z'Zc \;+\; Z'e \tag{8.64}$$

where e is the $T \times 1$ vector with t^{th} element e_t. Therefore, $T^{-1}Z'x \xrightarrow{p} 0$ and $T^{-1/2}Z'x \xrightarrow{d} N(M_{zz}c, \sigma_e^2 M_{zz})$. The nature of this limiting behaviour means that,

$$\begin{aligned} \hat{\theta}_T - \theta_0 \quad &= \quad \frac{T^{-1/2}x'Z(T^{-1}Z'Z)^{-1}T^{-1/2}Z'u}{T^{-1/2}x'Z(T^{-1}Z'Z)^{-1}T^{-1/2}Z'x} \\ &\xrightarrow{d} \quad \frac{\Psi_1' M_{zz}^{-1}\Psi_2}{\Psi_1' M_{zz}^{-1}\Psi_1} \end{aligned} \tag{8.65}$$

where $\Psi_1 \sim N(M_{zz}c, \sigma_e^2 M_{zz})$ and $\Psi_2 \sim N(0, \sigma_e^2 M_{zz})$. Therefore, $\hat{\theta}_T$ converges to a random variable when parameter vector is weakly identified in the sense of (8.61). This is in marked contrast to the case when θ_0 is identified because then $\hat{\theta}_T$ converges in probability to θ_0.[24]

This simple example provides a clear indication that the asymptotic theory derived in Chapters 3 and 5 is inappropriate for the weak identification case. As a result, three questions naturally arise: – what is the behaviour of the GMM estimator in dynamic nonlinear models when the parameter vector is weakly identified? – is it possible to perform inference about θ_0 in this setting and if so how? – is it possible to test whether θ_0 is identified by the population moment condition? These three questions are covered respectively in Sections 8.2.1 through 8.2.3.

Before we proceed to this discussion, it is worth emphasising the intended interpretation of this framework. The definition of weak identification is artificial in the sense that it is not seriously believed that real economic data are generated by processes with Pitman drift. This is simply a mathematical device that is used to generate a limiting distribution theory that provides a good approximation to finite sample behaviour in cases when – in the terms of our simple example – $E[x_t z_t]$ is small but non-zero. However, it should be noted that if $E[x_t z_t]$ is non-zero then the asymptotic theory in Chapter 2 (or more generally Chapters 3 and 5) is valid. The problem is that it may take a very large T before this asymptotic theory provides a good approximation. Hahn and Inoue (2002) provide simulation evidence that suggests that conventional asymptotics can provide a satisfactory approximation in the types of large dataset encountered in microeconometrics (i.e. $T = 10,000$) unless the number of instruments is large and the correlation between the endogenous regressor and the instruments is pathologically small.[25] However, available evidence suggests that this

[24] See Section 2.3.

[25] Hahn and Inoue (2002) compare a number of methods for constructing confidence in-

is not the case in the sample sizes encountered in macroeconometrics and that in these cases the theory discussed below provides a better approximation.

It is also worth noting that the approach described above is not the only possible way to obtain an alternative distribution theory for the IV estimator in the presence of weak identification. One alternative is to use the limiting distribution theory developed by Bekker (1994) that is based on the assumption that the number of instruments increases with the sample size.[26] Hahn and Inoue (2002) find that this distribution theory provides a good approximation in the types of sample sizes encountered in microeconometrics. However, this approach has not yet been extended to nonlinear models and so we do not pursue it further here.

8.2.1 The Limiting Behaviour of the GMM Estimator

To date, weak identification has been mostly encountered in models estimated by Generalized Instrumental Variables estimators, and so our discussion focuses on this case.[27] In this setting, the problem of weak identification is commonly refered to as the "weak instrument" problem. Staiger and Stock (1997) develop the limiting distribution theory in linear models, and Stock and Wright (2000) extend the analysis to nonlinear models. It is the latter that is our focus here. One central finding of these papers is that the usual limiting distribution theory does not apply and this motivates the presentation of alternative inference methods in the following sub-section. In view of this, our discussion concentrates on the framework for capturing nearly uninformative moment conditions and the conclusions to be drawn from the nature of the limiting distributions. The interested reader is refered to Stock and Wright (2000) for detailed derivations. Although the analysis is in terms of GIV estimation, it is worth noting that the corresponding results for GMM can be obtained by setting $z_t = 1$.

As mentioned above, we focus on the following class of moment conditions.

Assumption 8.4 GIV Estimation
Let $f(v_t, \theta) = u_t(\theta) \otimes z_t$.

In our simple example above, the parameter vector consists of a single element and so, by construction, the entire parameter vector is weakly identified. In more general settings, logic dictates that some elements of the parameter vector may be identified and others weakly identified. To accommodate this scenario, we partition the parameter vector as follows: $\theta = (\phi', \psi')'$ where ϕ is $p_\phi \times 1$ and ψ is $p_\psi \times 1$ where $p = p_\phi + p_\psi$. Similarly, we write $\Theta = \Phi \times \Psi$ in the obvious notation. Below ϕ consists of the weakly identified parameters and ψ the parameters that are identified. As before, it is assumed that $q \geq p$ and so problems with identification are not due to too few population moment

tervals in the context of a simple linear regression model. Also see Section 6.2.1 for further discussion of the connection between identification and the passage to the limiting distribution.

[26] See Section 6.1.3.

[27] See Section 7.2.

conditions *per se* but rather the poor quality of the information in these moment conditions.[28]

It is pedagogically easier to present the mathematical framework used to capture this scenario and then discuss how it achieves the desired goal. This framework is given in the following assumption.

Assumption 8.5 Weak Identification

(i) $E_T[f(v_t, \theta)] = T^{-1/2} m_{1,T}(\theta) + \mu_2(\psi)$.

(ii) $m_{1,T}(\theta) \to \mu_1(\theta)$ *uniformly in* $\theta \in \Theta$, $\mu_1(\theta_0) = 0$, $\mu_1(\theta)$ *is continuous in* θ *and is bounded on* Θ.

(iii) $\mu_2(\psi_0) = 0$, $\mu_2(\psi) \neq 0$ *for* $\psi \neq \psi_0$, $M_2(\psi) = \partial \mu_2(\psi)/\partial \psi'$ *has full rank at* $\psi = \psi_0$ *and is continuous.*

From Assumption 8.5, it can be seen that the population moment condition is satisfied at θ_0. At other parameter values, $E_T[f(v_t, \theta)]$ consists of two parts. The first part, $T^{-1/2} m_{1,T}(\theta)$, decays to zero at rate $T^{-1/2}$. Therefore, this first part is nearly uninformative about both ϕ_0 and ψ_0. In contrast, the second part, $\mu_2(\psi)$, is non-zero for any $\psi \neq \psi_0$ and so is informative about ψ_0. Therefore, within this framework, ϕ_0 is weakly identified but ψ_0 is identified.

Before proceeding further, it is worth briefly contrasting this framework with the scenario of redundant moment conditions described in Section 6.1.2. It can be recalled that $E[f_2(v_t, \theta_0)] = 0$ is redundant given $E[f_1(v_t, \theta_0)] = 0$ if the asymptotic variance of the GMM estimator is the same whether estimation is based on $E[f_1(v_t, \theta_0)] = 0$ alone or on both $E[f_1(v_t, \theta_0)] = 0$ and $E[f_2(v_t, \theta_0)] = 0$. The presumption in this earlier discussion is that θ_0 is identified by $E[f_1(v_t, \theta_0)] = 0$. Therefore the literature on redundant moment conditions addresses the problems encountered if uninformative moment conditions are included along with informative moment conditions when the parameter vector is identified.[29] In contrast, the literature on weak identification addresses the problems that arise when the population moment conditions are collectively nearly uninformative about the parameter vector. As might be imagined, the consequences of redundant and nearly uninformative moment conditions are quite different.[30]

Stock and Wright (2000) [Corollary 4] present the limiting distributions of the first step, second step and continuous updating GIV estimators. For brevity, we focus on the second step estimators, $\hat{\phi}_T(2)$ and $\hat{\psi}_T(2)$. However, it emerges that these distributions depend in part on the behaviour of the first step estimators, $\hat{\phi}_T(1)$ and $\hat{\psi}_T(1)$, and so the relevant aspects of the limiting behaviour of the first step estimators are also summarized below. The analysis rests on the

[28] See Section 3.1 for a discussion of identification.

[29] See Sections 6.1.2, 7.3.2 and 7.3.4.

[30] See Hall, Inoue, Jana, and Shin (2003) for further discussion of the connections between redundancy and weak identification.

empirical process representation for the GMM objective function.[31] Within this framework, $T^{1/2}g_T(\theta)$ is treated as a function of θ that converges to a Gaussian process.[32] The limit process is assumed to have the following properties.[33]

Assumption 8.6 Functional Central Limit Theorem
$T^{1/2}g_T(\theta) \Rightarrow \Psi(\theta)$ where $\Psi(\theta)$ is a Gaussian stochastic process on Θ with mean zero and covariance function $E[\Psi(\theta_1)\Psi(\theta_2)'] = \Omega(\theta_1, \theta_2)$.

The following lemma characterizes the limiting behaviour of the two step GIV estimators where, for simplicity, it is assumed that the first step weighting matrix is I_q.

Lemma 8.1 Limiting Behaviour of GIV Estimators
Under assumptions 8.4 – 8.6 and certain other regularity conditions,[34] the limiting behaviour of the GIV estimators is as follows: (i) $\hat{\phi}_T(1) \xrightarrow{d} \phi_\infty^{(1)}$, $\hat{\psi}_T(1) \xrightarrow{p} \psi_0$; (ii)

$$\begin{bmatrix} \hat{\phi}_T(2) \\ T^{1/2}[\hat{\psi}_T(2) - \psi_0] \end{bmatrix} \xrightarrow{d} \begin{bmatrix} \phi_\infty^{(2)} \\ \Delta_\psi \end{bmatrix}$$

where

$$\begin{aligned}
\phi_\infty^{(1)} &= argmin_{\phi\in\Phi}\, Q_*^{(1)}(\phi) \\
\phi_\infty^{(2)} &= argmin_{\phi\in\Phi}\, Q_*^{(2)}(\phi) \\
\Delta_\psi &= -[F(\phi_\infty^{(1)}, \psi_0)'F(\phi_\infty^{(1)}, \psi_0)]^{-1}F(\phi_\infty^{(1)}, \psi_0)'\Omega(1)^{-1/2}[\Psi(\phi_\infty^{(2)}, \psi_0) \\
&\quad + \mu_1(\phi_\infty^{(2)}, \psi_0)]
\end{aligned}$$

and

$$\begin{aligned}
Q_*^{(1)}(\phi) &= [\Psi(\phi, \psi_0) + \mu_1(\phi, \psi_0)]'C_1(\psi_0)[\Psi(\phi, \psi_0) + \mu_1(\phi, \psi_0)] \\
C_1(\psi_0) &= I_q - M_2(\psi_0)[M_2(\psi_0)'M_2(\psi_0)]^{-1}M_2(\psi_0)' \\
Q_*^{(2)}(\phi) &= [\Psi(\phi, \psi_0) + \mu_1(\phi, \psi_0)]'C_2(\phi_\infty^{(1)}, \psi_0)[\Psi(\phi, \psi_0) + \mu_1(\phi, \psi_0)] \\
C_2(\phi_\infty^{(1)}, \psi_0) &= \Omega(1)^{-1/2'}K(\phi_\infty^{(1)}, \psi_0)\Omega(1)^{-1/2} \\
K(\phi_\infty^{(1)}, \psi_0) &= I_q - F(\phi_\infty^{(1)}, \psi_0)[F(\phi_\infty^{(1)}, \psi_0)'F(\phi_\infty^{(1)}, \psi_0)]^{-1}F(\phi_\infty^{(1)}, \psi_0)' \\
F(\phi_\infty^{(1)}, \psi_0) &= \Omega(1)^{-1/2}M_2(\psi_0) \\
\Omega(1)^{-1/2} &= [\Omega(1)^{1/2'}]^{-1} \\
\Omega(1) &= \Omega(1)^{1/2'}\Omega(1)^{1/2} = \Omega(\theta^{(1)}, \theta^{(1)}) \\
\theta^{(1)} &= (\phi_\infty^{(1)'}, \psi_0')'
\end{aligned}$$

[31] See Andrews (1994) for a review of empirical process theory.
[32] A similar device is used to develop a limiting distribution theory for structural stability tests in Section 5.4.2.1. Note that in the context of structural stability tests, the partial sum is treated as a function of the break fraction π.
[33] See Andrews (1994) or Stock and Wright (2000) for more primitive conditions under which the Functional Central Limit Theorem holds in this setting.
[34] These include an identification condition that is omitted here to simplify the presentation.

and $M_2(\psi)$ is defined in Assumption 8.5. Lemma 8.1 has two important implications. First, the estimator of the sub-vector of the weakly identified parameters, $\hat{\phi}_T$, converges to a random variable on both steps and so is not consistent. Secondly, the estimator of the sub-vector of identified parameters, $\hat{\psi}_T$, is consistent but its limiting distribution is no longer normal.[35] In other words, inference about the identified parameters is contaminated by the presence of the weakly identified parameters. The bottom line is that none of the inference procedures in Chapter 5 are valid in the presence of weak identification. The following sub–section considers alternative approaches to inference that have been proposed to circumvent these problems.

8.2.2 Inference in the Presence of Weak Identification

Since the GMM estimators exhibit non-standard limiting behaviour, the conventional estimator based approach to inference is infeasible. To resurrect inference in this setting, an alternative approach is required. This approach involves finding a statistic whose limiting distribution at θ_0 is both standard and also unaffected by the presence of weak identification, and then inverting this statistic to construct confidence sets for θ_0. In the context of linear models estimated by IV, Staiger and Stock (1997) explore this approach based on the Anderson and Rubin (1949) statistic. In the same context, Wang and Zivot (1998) show that modified versions of the Wald, LM and D statistics are bounded by a statistic of known distribution and so can also form the basis for confidence sets. These approaches are compared in Zivot, Startz, and Nelson (1998). We do not review these papers in more detail as the results only apply to the linear setting. Instead, we focus on the method proposed by Stock and Wright (2000) that is valid in nonlinear models.

As noted above, it is necessary to find a statistic whose limiting distribution at θ_0 is both standard and also unaffected by the presence of weak identification. Fortunately, such a statistic is close at hand. Under the conditions of Lemma 3.2,[36] it follows that

$$TQ_{cont,T}(\theta_0) \;=\; Tg_T(\theta_0)'S_T(\theta_0)^{-1}g_T(\theta_0) \;\xrightarrow{d}\; \chi^2_q \tag{8.66}$$

Therefore, Stock and Wright (2000) propose inverting $TQ_{cont,T}(\theta_0)$ to obtain the approximate $100(1-\alpha)\%$ confidence set

$$\{\,\theta \,:\, TQ_{cont,T}(\theta) \,<\, c_q(\alpha)\,\} \tag{8.67}$$

where $c_q(\alpha)$ is the $100(1-\alpha)$ percentile of the χ^2_q distribution. This approach to inference is discussed in Section 3.7 where it is proposed as a way of constructing confidence sets that are invariant to reparameterization. In the context of weak identification, such sets are often refered to as "S-sets", a terminology derived

[35] The distribution of the $T^{1/2}[\hat{\psi}_T(1)-\psi_0]$ is qualitatively similar to that of $T^{1/2}[\hat{\psi}_T(2)-\psi_0]$; see Stock and Wright (2000).

[36] See Section 3.4.2.

from the notation in Stock and Wright (2000). However, since our notation is different, we eschew this name. Notice that the distributional result in (8.66) holds regardless of whether or not θ_0 is identified or weakly identified, and so this approach to inference is equally valid in both cases.[37]

The confidence sets in (8.67) have the appealing feature that they can be infinite in one or more dimension and so reveal that the elements of the parameter vector in question are unidentified. This contrasts with the conventional asymptotic confidence intervals in (3.27) which are by construction of fixed length. In fact, the intervals in (3.27) are fundamentally flawed in this setting. Dufour (1997) shows that for a confidence interval to have the stated coverage probability then it must be possible for it to have infinite length.[38]

However, this approach to inference also has its drawbacks. Kleibergen (2000) has pointed out that the confidence sets in (8.67) are not centred on $\hat{\theta}_T$. Whether this is perceived as a drawback may be a matter of taste. Kleibergen (2000) uses this feature to motivate confidence sets based on inversion of a statistic based on the first order derivative of the continuous updating GMM minimand; see Kleibergen (2000) for further details. A more serious drawback is that the inversion in (8.67) is only computationally feasible for relatively small values of p. This problem can be ameliorated if the partition between the weakly identified and identified parameters is known and it is the weakly identified parameters that are of interest. In such circumstances, valid confidence sets can be based on the minimand of the restricted GMM estimation in which the minimization is performed over the identified parameter, ψ, conditional on a value for the weakly identified parameter, ϕ. To flesh out the details, it is necessary to introduce some additional notation. Let $\hat{\psi}_T(\bar{\phi})$ denote the GMM estimator of ψ conditional on $\phi = \bar{\phi}$, that is

$$\hat{\psi}_T(\bar{\phi}) \;=\; argmin_{\psi \in \Psi}\, TQ_{cont,T}(\theta)\big|_{\phi=\bar{\phi}} \tag{8.68}$$

Stock and Wright (2000) show that

$$TQ_{cont,T}\left(\phi_0, \hat{\psi}_T(\phi_0)\right) \;\xrightarrow{d}\; \chi^2_{q-p_\psi} \tag{8.69}$$

where, as before, p_ψ is the dimension of ψ, and so propose the following asymptotically valid $100(1-\alpha)\%$ confidence set for ϕ_0,

$$\{\, \phi \,:TQ_{cont,T}\left(\phi, \hat{\psi}_T(\phi)\right) \;<\; c_{q-p_\psi}(\alpha)\,\} \tag{8.70}$$

We now illustrate this alternative approach within the context of our running empirical example.

[37] See Section 3.7 for a discussion of other reasons for using this approach to inference when the parameter vector is identified.

[38] The intervals in (3.27) may also be invalid if θ_0 is identified but there is a subset of the parameter space, Θ_{un}, in which θ is unidentified; see Dufour (1997) for further discussion.

Example: Hansen and Singleton's (1982) Consumption Based Asset Pricing Model
The confidence sets in (8.67) have already been presented in Section 3.7. As a reminder, the set for (γ_0, δ_0) was non-empty but bounded for the value weighted returns (VWR) case, and empty for equally weighted returns (EWR). The latter indicates misspecification, and it is consistent with our findings based on the overidentifying restrictions test in Section 5.1. Therefore, we concentrate on the VWR case here. The available evidence suggests that this is one case in which the partition of the parameter vector might reasonably be taken to be known with the weakly identified parameter being $\phi = \gamma$ and the identified parameter, $\psi = \delta$. Using this partition, we use (8.70) to calculate a confidence set – or interval as $p_\phi = 1$ – for γ. To make the calculation feasible in practice, it is necessary to discretize the parameter space for γ. This was done as follows. To begin, the parameter space is taken to consist of all points lying between -50 and 50 on a grid with 0.01 between each point. This leads to an interval of $[-4.88, 7.19]$. To refine this interval, the calculations are redone for the finer grid consisting of all points between -5.000 and 7.200 with 0.001 between each point. The resulting interval is $[-4.886, 7.188]$. This interval is almost twice the width of the interval reported in Table 3.8 that is based on the traditional asymptotic confidence interval given in (3.27). It is also asymmetric around $\hat{\gamma}_T$, which is 0.666 for this model.[39] Therefore, the use of these alternative asymptotics leads to different conclusions about the set of plausible values for γ_0. $\diamond$

8.2.3 The Detection of Weak Identification

It is clear from the discussion in Section 8.2.1 that the presence of weak identification renders the conventional asymptotics inappropriate. This motivates the development of the alternative approach to inference described in Section 8.2.2 based on methods that are robust to the presence of nearly uninformative moment conditions. The difference between these two inference frameworks naturally raises the question of how a researcher should perform inference if the identification of the parameters is suspect. One solution is to adopt the confidence set framework in Section 8.2.2 because it is valid regardless of whether or not θ_0 is identified. However, there are at least three reasons why it may be desirable to diagnose whether or not the parameters are identified. First, the confidence set may only be feasible in cases where p is relatively small. Secondly, there is far wider array of inference procedures available if the parameter vector is identified. Finally, if the parameter vector is identified then the point estimator is consistent and this knowledge may affect our interpretation of the estimates. Therefore, in this section, we consider methods that have been proposed for testing identification. Even more than other aspects of the literature on weak identification, this topic has been addressed within the context of linear regression models estimated by IV. In spite of this limitation, we review the available results because the qualitative conclusions likely extend to nonlinear

[39] See Table 3.7

models. However, it is left to future research to extend the methods described here to the GMM framework with nonlinear dynamic models.

To initiate the discussion, we first consider how the presence of weak identification might be detected within the simple motivating example given in at the beginning of this sub-section. It can be recalled that the population moment condition in (8.60) is nearly uninformative about θ_0 because $E_T[x_t|z_t]$ is decaying to zero at rate $T^{-1/2}$. Or put more simply, the relationship between x_t and z_t is dying out as $T \to \infty$. Intuition suggests that this state of affairs can be uncovered by running the regression of x_t on z_t and examining standard diagnostics for goodness of fit. Since it is desired to develop a test for identification, the most convenient diagnostic is the F-statistic for the hypothesis that the coefficients on z_t are all zero in this regression. Bound, Jaeger, and Baker (1995) advocate that this first stage F-statistic be rountinely reported as a "rough guide" on the strength of the identification. For our purposes here, it is useful to formalize this recommendation. To this end, we denote the regression model for x_t on z_t by[40]

$$x_t \;=\; z_t'\gamma \;+\; \text{error}$$

Let F_γ be the F-statistic for the hypothesis that $\gamma = 0$ in this model. In terms of identification, the null and alternative hypotheses are interpreted as follows:

$$H_0 : \gamma \;=\; 0 \quad \Leftrightarrow \quad \theta_0 \text{ not identified} \tag{8.71}$$

$$H_1 : \gamma \;\neq\; 0 \quad \Leftrightarrow \quad \theta_0 \text{ identified} \tag{8.72}$$

Therefore, θ_0 is deemed identified if F_γ is significant.

This generic approach can be extended to more general linear models in which x_t is a vector and includes both endogenous and exogenous regressors. Hall, Rudebusch, and Wilcox (1996) propose testing for identification based on the canonical correlations between x_t and z_t. Shea (1997) proposes a method based on the partial correlations between x_t and z_t. Cragg and Donald (1993) propose a test based on the rank of the coefficient matrix in the reduced form regression of x_t on z_t. However, we do not consider the specifics of these tests further but instead focus on two aspects of this generic approach, namely the implications of testing for identification prior to inference about θ_0 and the interpretation of the alternative hypothesis.

The ultimate focus of the analysis is θ_0, and so it is important to consider how the use of such a test for identification affects subsequent inferences about the parameter vector. The answer is that it depends both on how the test for identification is used and also on the statistic used to perform inference about θ_0. There are two ways in which the test for identification could be used and for simplicity, we discuss these in the context of the simple example above. One option is to use F_γ to select the instrument vector. Within this approach, F_γ is calculated for a sequence of possible choices of instrument and the selected instrument vector, z_t^* say, is the first in the sequence for which F_γ

[40] In this context, this regression is often refered to as "the first stage regression" as it is the first stage of a Two Stage Least Squares estimation of (8.59).

is significant. Estimation is then based on the moment condition in (8.60) with $z_t = z_t^*$ and inference is performed under the assumption that θ_0 is identified. A second option is to treat z_t as given and then use F_γ to determine which statistical theory is used to perform inference about θ_0. With this approach, an insignificant F_γ value leads to inference based on a method that is valid if θ_0 is weakly identified, and a significant F_γ leads to inference based on a method that is valid if θ_0 is identified. The evidence to date suggests that the first option is not a good strategy but the second is. Hall, Rudebusch, and Wilcox (1996) report simulation evidence on a variant of the first option above in which a test for identification is used to select the instrument vector and subsequent inference about the (scalar) parameter θ_0 is performed using the confidence interval in (3.27).[41] This evidence indicates that inferences about θ_0 can be severely distorted in the sense that the actual coverage probability of the confidence interval is much smaller than the nominal level. However, there is a caveat to this finding. The confidence interval in (3.27) can be interpreted as containing all values of $\bar{\theta}$ for which $H_0: \theta_0 = \bar{\theta}$ is not rejected using the Wald statistic at the $100\alpha\%$ level. Zivot, Startz, and Nelson (1998) show that the behaviour of the Wald statistic is severely distorted by the presence of weak identification whereas the LM and LR tests are far more robust. Zivot, Startz, and Nelson (1998) report comparable evidence for the case in which the confidence interval is calculated by inverting the LR and LM tests for $H_0: \theta_0 = \bar{\theta}$. This evidence indicates that the use of the Wald test based confidence interval does indeed account for a substantial part of the distortions reported by Hall, Rudebusch, and Wilcox (1996). However, non-trivial distortions remain even if inference is based on the LR or LM tests. Zivot, Startz, and Nelson (1998) also report simulation evidence for the second option in which the choice of instrument is taken as given and F_γ is used to determine the statistical theory employed. Within their design, θ_0 is a scalar and the confidence interval is constructed by inverting either the LM or LR test for $H_0: \theta_0 = \bar{\theta}$. The value of F_γ determines the distribution used to approximate the behaviour of these statistics. Their evidence indicates that the coverage rate is very close to the nominal level regardless of whether θ_0 is unidentified, weakly identified or identified.

For all the tests of identification, the null is that θ_0 is unidentified and the alternative is that θ_0 is identified. While it is true that failure to reject the null indicates a problem, Stock and Yogo (2001) argue that rejection of the null at conventional significance levels does not necessarily imply that "conventional asymptotics" provide a good approximation. This is certainly true for the Wald statistic considered in their paper, and likely true for other statistics as well. They further argue that the definition of what constitutes weak identification should reflect the nature of the desired inference about θ_0. In the context of IV estimation of a linear model, they consider two criteria for whether the instruments are weak: one based on a measure of the bias in $\hat{\theta}_T$ and the other

[41] The variation is that the test for identification is not F_γ but a test based on the correlation between scalar x_t and scalar z_t.

based on the size distortions exhibited by the Wald test. Both are intuitively reasonable but interestingly they yield different criterion that, furthermore, are sensitive to different aspects of the specification. While this analysis is confined to linear models, it clearly reveals that the issues of defining poor identification and testing for identification are more subtle than has been previously realized.

To date, there has been far less attention on the issue of testing for identification in nonlinear models. The simple reason is that in the general nonlinear model the key determinant of local identification is the derivative matrix G_0 which depends on θ_0. This problem can be circumvented as demonstrated by Wright (2001) who proposes a test for identification in the context of GIV estimation. However, the issues described in the previous two paragraphs remain to be addressed in this context, and so we do not explore the test further here.

8.3 Inference When the Long Run Variance is Estimated by an HAC Estimator with $b_T = T$

It can be recalled from Theorem 3.2 that the asymptotic variance of $\hat{\theta}_T$ depends on the long run variance of the sample moment, S. Given a consistent estimator of S, $\hat{S}_T$ say, it is possible to perform inference about θ_0 based on this limiting distribution theory. For example, Theorem 3.2 implies that inference about $\theta_{0,i}$ can be based on,

$$\frac{T^{1/2}(\hat{\theta}_{T,i} - \theta_{0,i})}{\sqrt{\hat{V}_{T,ii}}} \xrightarrow{d} N(0,1) \tag{8.73}$$

where $\hat{V}_{T,ii}$ is the $i - i^{th}$ element of

$$\hat{V}_T = [G_T(\hat{\theta}_T)' W_T G_T(\hat{\theta}_T)]^{-1} G_T(\hat{\theta}_T)' W_T \hat{S}_T W_T G_T(\hat{\theta}_T) [G_T(\hat{\theta}_T)' W_T G_T(\hat{\theta}_T)]^{-1} \tag{8.74}$$

Section 3.5 contains a review of a number of alternative estimators for $\hat{S}_T$ with the choice between them depending upon the requisite assumptions about the dependence structure of $\{f(v_t, \theta_0)\}$. The most general of these is the class of heteroscedasticity autocorrelation covariance (HAC) estimators. To calculate the HAC estimator, it is necessary to choose a kernel, $\omega(.)$, and a bandwidth, b_T. These components must satisfy certain restrictions if the resulting covariance matrix estimator is to be consistent. However, these conditions leave a fair amount of latitude. While the literature has provided guidance on the relative merits of popular choices of kernel, there is no data based method for bandwidth selection that does not involve some kind of arbitrary decision by the practitioner. This is particularly undesirable as simulation evidence suggests that subsequent inferences about θ_0 can be sensitive to the choice of bandwidth.[42] In view of these problems, Vogelsang (2003) proposes using a HAC estimator

[42] See Sections 3.5.3 and 6.3.

with $b_T = T$. Such a rule has the twin advantages of being simple and definitive but it violates the conditions for consistency of the covariance matrix estimator with the result that (8.73) no longer holds. However, Vogelsang (2003) shows that it is possible develop an alternative asymptotic theory that can be used as a basis for inference about θ_0 in this case. In this section we briefly review this theory.

To facilitate the discussion, it is useful to introduce the following notation. In view of the structure of the kernels in Table 3.3 and our current focus on the case in which $b_T = T$, we write $\omega(i/T)$ for $\omega_{i,T}$. Below it is necessary to consider situations in which the argument of the kernel is a difference, and to avoid excessive notation we set $k_{i,j} = \omega((i-j)/T)$. Let $\hat{S}_{b_T=T}$ denote the HAC estimator in equation (3.54) with $b_T = T$, and set $g_i(\theta) = T^{-1}\sum_{t=1}^{i} f(v_t, \theta)$. Finally, let $\hat{\theta}_T$ be the GMM estimator based on weighting matrix W_T. It is important for the arguments below that $\hat{\theta}_T$ is consistent for θ_0. One implication of the analysis below is that $\hat{S}_{b_T=T}^{-1}$ does not satisfy Assumption 3.7 and so is not a valid weighting matrix. *Therefore, $\hat{S}_{b_T=T}$ is only used to perform inference after estimation is completed.*

This approach to inference works for the general case in which it is desired to test the hypothesis that the parameters satisfy a nonlinear set of restrictions, $r(\theta_0) = 0$. However, it is more convenient to introduce this framework in the context of the simple case in which $r(\theta_0) = \theta_{0,i}$. The more general case is then covered at the end of the section. The arguments presented are heuristic and the interested reader is refered to Vogelsang (2003) for a rigorous justification.

Suppose then that it is desired to test perform inference about $\theta_{0,i}$. The natural starting point is the analagous statistic to the one appearing in (8.73), namely

$$\frac{T^{1/2}(\hat{\theta}_{T,i} - \theta_{0,i})}{\sqrt{\hat{V}_{b_T=T,ii}}} \tag{8.75}$$

where $\hat{V}_{b_T=T,ii}$ is the $i-i^{th}$ element of

$$\begin{aligned} \hat{V}_{b_T=T} &= [G_T(\hat{\theta}_T)'W_T G_T(\hat{\theta}_T)]^{-1} G_T(\hat{\theta}_T)'W_T \hat{S}_{b_T=T} W_T G_T(\hat{\theta}_T) \\ &\quad \times [G_T(\hat{\theta}_T)'W_T G_T(\hat{\theta}_T)]^{-1} \end{aligned} \tag{8.76}$$

Since $\hat{S}_{b_T=T}$ is only used for inference and not estimation, the limiting behaviour of the numerator is the same as before. However, this time it is useful to express this limiting behaviour in terms of a Brownian motion.[43] To this end, let ι_i be the $(p \times 1)$ selection vector whose i^{th} element is one and whose remaining elements are all zero. Using similar arguments to Section 3.4.2, it follows from (3.26) that

$$\begin{aligned} T^{1/2}(\hat{\theta}_{T,i} - \theta_{0,i}) &= \iota_i' T^{1/2}(\hat{\theta}_T - \theta_0) \\ &= -\iota_i'[G_0' W G_0]^{-1} G_0' W T^{1/2} g_T(\theta_0) + o_p(1) \end{aligned} \tag{8.77}$$

[43] See Section 5.4.2.1.

Equation (8.77) and the Functional Central Limit Theorem (Assumption 5.8) together imply that

$$T^{1/2}(\hat{\theta}_{T,i} - \theta_{0,i}) \Rightarrow -\iota_i'[G_0'WG_0]^{-1}G_0'WS^{1/2'}B_q(1) \quad (8.78)$$
$$= \sigma B_1(1) \quad (8.79)$$

where $B_q(r)$ is a $(q \times 1)$ Brownian motion and σ is the (positive) square root of $\iota_i'[G_0'WG_0]^{-1}G_0'WSWG_0[G_0'WG_0]^{-1}\iota_i$.

Now consider the denominator of (8.75). Clearly the denominator depends on $\hat{V}_{b_T=T}$. To develop the limiting behaviour of $\hat{V}_{b_T=T}$, we start with $\hat{S}_{b_T=T}$ and gradually add in the surrounding matrices that appear in (8.76). Vogelsang (2003) shows that

$$\hat{S}_{b_T=T} = T^{-1}\sum_{\ell=1}^{T-1}\sum_{m=1}^{T-1} d_{m,\ell}Tg_\ell(\hat{\theta}_T)Tg_m(\hat{\theta}_T)' + \sum_{i=1}^{T} f(v_i,\hat{\theta}_T)k_{i,T}g_T(\hat{\theta}_T)'$$
$$+T\sum_{\ell=1}^{T-1}(k_{T,\ell} - k_{T,\ell+1})g_T(\hat{\theta}_T)g_\ell(\hat{\theta}_T)' \quad (8.80)$$

where

$$d_{m,\ell} = (k_{m,\ell} - k_{m,\ell+1}) - (k_{m+1,\ell} - k_{m+1,\ell+1})$$

Now consider $C_T = G_T(\hat{\theta}_T)'W_T\hat{S}_{b_T=T}W_TG_T(\hat{\theta}_T)$. Since the first order conditions – equation (3.12) – imply $G_T(\hat{\theta}_T)'W_Tg_T(\hat{\theta}_T) = 0$, it follows from (8.80) that

$$C_T = T^{-1}\sum_{\ell=1}^{T-1}\sum_{m=1}^{T-1} d_{m,\ell}G_T(\hat{\theta}_T)'W_TTg_m(\hat{\theta}_T)Tg_\ell(\hat{\theta}_T)'W_TG_T(\hat{\theta}_T) \quad (8.81)$$

To characterize the limiting behaviour of C_T, it is useful to introduce the step function $D_T(r)$ defined on $r \in [0,1]$ as $D_T(r) = D(j)$ for $j/T \leq r \leq (j+1)/T$, $j = 1, 2, \ldots, T-1$ where $D(x/T) = [\omega((x+1)/T) - \omega(x/T)] - [\omega(x/T) - \omega((x-1)/T)]$. Using this step function, C_T can rewritten as

$$C_T = -T^{-1}\sum_{\ell=1}^{T-1}\sum_{m=1}^{T-1} T^2D_T((m-\ell)/T)G_T(\hat{\theta}_T)'W_Tg_m(\hat{\theta}_T)g_\ell(\hat{\theta}_T)'W_TG_T(\hat{\theta}_T)$$
$$= -\int_0^1\int_0^1 T^2D_T(r_1 - r_2)G_T(\hat{\theta}_T)'W_TT^{1/2}g_{[r_1T]}(\hat{\theta}_T)$$
$$\times T^{1/2}g_{[r_2T]}(\hat{\theta}_T)'W_TG_T(\hat{\theta}_T)dr_1dr_2 \quad (8.82)$$

where $g_{[rT]}(\theta) = T^{-1}\sum_{t=1}^{[rT]} f(v_t,\theta)$. The advantage of this representation is that $T^2D_T(r) \to \omega''(r)$ where $\omega''(.)$ denotes the second derivative of $\omega(.)$ on $(-1,1)$.

From (8.82), it is clear that the limiting behaviour of C_T depends on $G_T(\hat{\theta}_T)'W_TT^{1/2}g_{[rT]}(\hat{\theta}_T)$. Using similar arguments to the derivation of Theorems 5.9–5.10 in Section 5.4.2.1, it can be shown that

$$G_T(\hat{\theta}_T)'W_TT^{1/2}g_{[rT]}(\hat{\theta}_T) \Rightarrow G_0'WS^{1/2'}BB_q(r) \quad (8.83)$$

where $BB_q(r)$ denotes a $(q \times 1)$ Brownian Bridge.[44] It follows from (8.82)–(8.83) that

$$\iota_i' \hat{V}_{b_T=T} \iota_i \;\Rightarrow\; \sigma^2 \int_0^1 \int_0^1 -\omega''(r_1 - r_2) BB_1(r_1) BB_1(r_2) dr_1 dr_2 \tag{8.84}$$

Combining (8.75), (8.79) and (8.83), it follows from the Continuous Mapping Theorem that

$$\frac{T^{1/2}(\hat{\theta}_{T,i} - \theta_{0,i})}{\sqrt{\hat{V}_{b_T=T,ii}}} \;\Rightarrow\; \frac{B_1(1)}{\{\int_0^1 \int_0^1 -\omega''(r_1 - r_2) BB_1(r_1) BB_1(r_2) dr_1 dr_2\}^{1/2}} \tag{8.85}$$

A comparsion of this distribution with the conventional limit distributions in (8.73) indicates that the difference lies purely in the denominator.[45] Since $B_1(1)$ and $BB_1(r)$ are independent by construction, it follows that the limiting distribution in (8.85) is that of the ratio of two independent random variables. This structure further implies that the limiting distribution is a mixture of normals. Notice also that the distribution in (8.85) depends on the kernel. We return to this issue below.

The more general hypothesis $r(\theta_0) = 0$ can be tested using the Wald-type test,

$$\tilde{W}_T \;=\; T r(\hat{\theta}_T)' [R(\hat{\theta}_T) \hat{V}_{b_T=T} R(\hat{\theta}_T)']^{-1} r(\hat{\theta}_T)/s$$

where $R(\theta) = \partial r(\theta)/\theta'$ and $r(.)$ is $s \times 1$. The following lemma gives the limiting distribution of this statistic under the null hypothesis. The necessary regularity conditions pertain to the consistency of $\hat{\theta}_T$, $r(.)$ and the behaviour of the partial sums and derivatives. Since all three have been presented in Chapters 3 and 5, we do not explictly repeat them in the text here.

Lemma 8.2 Limiting Distribution of $\tilde{W}_T$ under $H_0 : r(\theta_0) = 0$
If Assumptions 3.1-3.5, 3.7-3.10, 5.3, 5.7 and 5.8 hold then:

$$\tilde{W}_T \;\Rightarrow\; B_s(1)' \left[\int_0^1 \int_0^1 -\omega''(r_1 - r_2) BB_s(r_1) BB_s(r_2)' dr_1 dr_2 \right]^{-1} B_s(1)/s$$

The limiting distribution in Lemma 8.2 depends only on the number of restrictions, s, and the kernel. Keifer and Vogelsang (2002*a*) show that the Bartlett kernel has superior local power properties in a simpler setting, and so we confine our discussion to this kernel here.[46] With the Bartlett kernel, $\omega''(x) = 0$ for all $x \neq 0$ but has the drawback that $\omega(x)$ is not differentiable at $x = 0$.[47]

[44] See Definition 5.2 in Section 5.4.2.1.
[45] Recall that $B_1(1) = N(0, 1)$.
[46] Keifer and Vogelsang (2002*a*) consider the case in which the null hypothesis involves a set of linear restrictions on the regression parameters in a linear model.
[47] Recall that the Bartlett kernel is $\omega(x) = 1 - |x|$ for $x \in (-1, 1)$ and zero elsewhere; see Section 3.5.3.

However, Keifer and Vogelsang (2002*b*) show that $\omega''(0)$ can be replaced by -2 with the result that the limit distribution in Lemma 8.2 becomes:

$$B_s(1)'[2\int_0^1 BB_s(r)BB_s(r)'dr]^{-1}B_s(1)/s \tag{8.86}$$

Critical values for this distribution are given in Table 8.5 for $p \leq 12$. Both Keifer and Vogelsang (2002*a*) and Vogelsang (2003) provide critical points for $p \leq 30$. Keifer and Vogelsang (2002*a*) also provide critical points for a variety of other kernels for the case of $p = 1$.

Vogelsang (2003) evaluates the finite sample performance of this approach via a small simulation study. The sample design involves the IV estimation of the parameters of linear regression model with errors that are generated by an $AR(p)$ process for $p = 0, 1, 2$. The null hypothesis of interest is that $\theta_{0,i} \leq 0$ versus the alternative that $\theta_{0,i} > 0$, and so the test statistic is given in (8.75) evaluated at $\theta_{0,i} = 0$. As comparison, Vogelsang (2003) also considers the performance of the conventional approach using (8.73) for the case in which $\hat{S}_T$ is calculated using an HAC with a quadratic spectral kernel and bandwidth selected by a method proposed in Andrews (1991). The evidence suggests that, of the two, the test based on the HAC with $b_T = T$ exhibits an empirical size that is closer to the nominal level and the asymptotic approximation is good at $T = 200$. Size adjusted power calculations indicate that this ranking is reversed under the alternative although the difference between the two tests is relatively small.[48]

Table 8.5
Critical points for $\tilde{W}_T$ with the Bartlett kernel

p	10%	5%	1%
1	14.28	23.14	51.05
2	17.99	26.19	48.74
3	21.13	29.08	51.04
4	24.24	32.42	52.39
5	27.81	35.97	56.92
6	30.36	38.81	60.81
7	33.39	42.08	62.27
8	36.08	45.32	67.14
9	38.94	48.14	69.67
10	41.71	50.75	72.05
11	44.56	53.70	74.74
12	47.27	56.70	78.80

Source: Reprinted from Advances in Econometrics, 17, T.J. Vogelsang, "Testing in GMM models without truncation", pp. 199–233, copyright (2003), with permission from Elsevier.
Notes: the figures represent the critical points for the tests at the 10%, 5% and 1% significance levels.

[48] Vogelsang (2003) considers the case with and without pre-whitening and recolouring.

Example: Hansen and Singleton's (1982) Consumption Based Asset Pricing Model
It can be recalled that the underlying economic theory for this model implies that $f(v_t, \theta_0)$ is a serially uncorrelated process, and so the majority of our inferences have not involved the use of an HAC matrix to estimate the long run variance. Nevertheless, for completeness, we use this approach to inference to obtain alternative confidence intervals for the parameters in the model with VWR. It follows from (8.85) and Table 8.5 that an approximate 95% confidence interval for $\theta_{0,i}$ is given by

$$\hat{\theta}_{T,i} \pm \sqrt{23.14}\sqrt{\hat{V}_{b_T=T,ii}/T}$$

Using the iterated estimator based on $W_T = \hat{S}_{SU}^{-1}$, the interval for γ_0 is $(-3.082, 4.415)$ and the interval for δ_0 is $(0.985, 1.026)$. Both are slightly wider than the corresponding intervals reported in Table 3.8.[49] $\diamond$

8.4 Summary

In this chapter, we review three alternative methods for approximating the finite sample behaviour of the GMM estimator and its associated statistics. These three are: (i) the bootstrap; (ii) an asymptotic theory developed for the case in which the parameter vector is weakly identified by the population moment condition; (iii) and an asymptotic theory designed to provide a better approximation when the weighting matrix is based on a heteroscedasticity autocorrelation covariance (HAC) matrix estimator. All three of these alternative approximations are relatively new to the GMM literature, and so the associated statistical theory is less comprehensive than that derived using the conventional theoretical framework reviewed in Chapters 3 and 5. While important progress has been made in each case, lacunae remain:

- *The bootstrap:* Since the bootstrap is based on resampling, it has the potential to provide asymptotic refinements for all the inference procedures discussed in Chapters 3 and 5. However, to date, these asymptotic refinements have only be proven to occur within a class of nonlinear dynamic models that includes some, but not all, the types of model in Table 1.1. The basis on resampling means that the method can also be computationally burdensome in nonlinear models, but this burden can be reduced by using the approximate bootstrap. In the types of model in Table 1.1, the data generation process is unknown and therefore the non-parametric bootstrap must be used. With dynamic data, the non-parametric bootstrap involves resampling blocks of data and, to date, there are no definitive guidelines on how these blocks should be chosen in the settings that arise in GMM estimation.

[49] See Section 3.6.

- *Weak identification:* There have been two main branches of this literature. The first branch has focused on showing the sensitivity of the conventional asymptotic approximation to the quality of the identification. This branch of the theory provides an important caveat to the conventional asymptotic analysis because weak identification is encountered in practice. The second main branch of this literature focuses on the development of inference techniques that are robust to weak identification. To date, such techniques are only feasible in the context of nonlinear dynamic models for settings in which the number of parameters is relatively small.

- *HAC estimation with bandwidth equal to sample size:* An attractive feature of this approach is that it provides a simple, definitive rule for bandwidth selection. Although the long run variance is not consistently estimated, it is still possible to perform asymptotically valid inference about the parameters. However, this choice of bandwidth cannot be used for estimation, and, to date, there are no comparable asymptotically valid procedures for inference based on the overidentifying restrictions.

Chapters 3 through 8 have exposited the main aspects of the statistical theory of GMM estimation and its associated inference techniques. Throughout the discussion, the various techniques have been illustrated using one of the examples from Section 1.3, namely the consumption based asset pricing model. In the following chapter, we consider the estimation of the remaining four examples from Section 1.3.

9

Empirical Examples

Throughout the preceeding chapters, the various facets of GMM estimation have been illustrated using the consumption based asset pricing model described in Section 1.3.1. In this chapter, we present empirical analyses for the other four models described in Section 1.3.

Section 9.1 implements the mutual fund evaluation measure proposed by Chen and Knez (1996). This discussion illustrates the potential sensitivity of inferences based on the overidentifying restrictions test to the choice of covariance matrix estimator. We also present results based on a modified measure of performance evaluation that involves a non-negativity constraint. As a consequence, the choice of $f(v_t, \theta)$ does not satisfy the restrictions on the derivative imposed by Assumption 3.5 because the derivative of the minimand is not defined at all values of the parameter space. This non-existence causes problems for gradient methods of optimization and so necessitates the use of an alternative algorithm.

Section 9.2 explores whether the conditional capital asset pricing model can explain the variation across international stock prices indices. The adequacy of the specification is assessed using both the overidentifying restrictions test and also tests for structural stability. The analysis indicates that inferences about structural stability can be very sensitive to whether inference is based on the Wald, LM or D tests discussed in Section 5.4. One possible explanation is that the LM and D tests use the full sample GMM estimator instead of the restricted GMM estimator. To assess the impact of this substitution, alternative versions of the LM and D test are introduced. The evidence indicates that this substitution, while asymptotically valid, has a considerable impact on the behaviour of the tests in this example.

Section 9.3 reports estimation results for Eichenbaum's (1989) model for inventory holdings, and examines whether the production smoothing or production cost smoothing hypothesis best captures aggegate behaviour in non-durable manufacturing industries. The analysis of the production smoothing model is based on both the original moment condition derived in Section 1.3.3 and also on two alternatives derived by applying curvature altering transformations to

the original. The iterated GMM estimates are seen to be sensitive to the transformation employed, and these are contrasted to the continuous updating GMM estimates that are insensitive to such transformations.

Section 9.4 explores whether a stochastic volatility model can capture the time series properties of the daily U.S. dollar - Canadian dollar exchange rate. This is another example in which the moment condition does not satisfy the conditions placed on the derivative matrix by Assumption 3.5. This time the problem is due to the presence of the absolute value function. In this context, this problem has been treated by using a polynomial approximation in the neighbourhood of the point at which the derivative is not defined, and we examine the sensitivity of the estimation results to the width of this neighbourhood. In this example, the parameter vector is heavily overidentified, and so we examine whether any of the moment conditions are redundant using the moment selection criteria described in Chapter 7.

9.1 Mutual Fund Performance Evaluation

Section 1.3.2 describes a method for the evaluation of mutual fund performance proposed by Chen and Knez (1996). It can be recalled that a fund receives a zero performance measure if

$$\lambda(r_t^m, d_t) = E[r_t^m X_t' \delta_0] - 1 = 0 \tag{9.1}$$

where r_t^m is the payoff on the mutual fund, d_t is the stochastic discount factor and X_t is a $(N \times 1)$ vector of payoffs on the traded assets included in the benchmark set. A fund receives a positive performance measure if $\lambda(r_t^m, d_t) > 0$. To distinguish this measure from another that is discussed below, we follow Chen and Knez (1996) and refer to $\lambda(r_t^m, d_t)$ as the LOP measure where the acronym stands for "Law of One Price", the theorem from which the measure is deduced. As in Section 1.3.2, we re-express this condition for a zero evaluation as

$$E[Q_t X_t' \delta_0] - 1_{N+1} = 0 \tag{9.2}$$

where $Q_t = (X_t', r_t^m)'$, δ_0 is a $(N \times 1)$ parameter vector defined in Section 1.3.2 and 1_{N+1} is a $(N+1 \times 1)$ vector of ones. It can be recognized that (9.2) constitutes a set of $N+1$ population moment conditions in N parameters and so it is possible to test the null hypothesis of a zero evaluation using the overidentifying restrictions test described in Section 5.1. Notice that the alternative for this test statistic is that $E[Q_t X_t' \delta_0] - 1_{N+1} \neq 0$ and so is broader than $\lambda(r_t^m, d_t) > 0$. Therefore, a significant statistic provides evidence against a zero performance evaluation but does not necessarily provide evidence of positive performance.

Inspection of (9.2) reveals that it is linear in the parameters. It therefore follows by similar arguments to Section 2.1 that δ_0 is globally identified if $rank\{E[Q_t X_t']\} = N$, the number of parameters in the notation here. Given the structure of Q_t, a sufficient condition for identification is therefore that $E[X_t X_t']$ is nonsingular which might reasonably be anticipated to hold in the

absence of any obvious redundancies in the definition in the benchmark set. The linear structure can also be exploited to deduce a closed form solution for the GMM estimator of δ_0. Using similar arguments to Section 2.2, it can be shown that

$$\hat{\delta}_T = (M_T' W_T M_T)^{-1} M_T' W_T 1_{N+1} \tag{9.3}$$

where $M_T = T^{-1} \sum_{t=1} Q_t X_t^{'}$.

Chen and Knez (1996) evaluate the performance of sixty eight funds both individually and in the aggregate. For the aggregate analysis, funds are grouped according to their investment objective and then the group "average" return is constructed as the return on an equally weighted portfolio of the funds in the group. Chen and Knez (1996) consider five investment objectives: growth (G), income-growth (IG), income (I), stability–growth–income (SGI), and maximum capital gain (MC). In this section, we evaluate fund performance for these group averages. For the benchmark set, Chen and Knez (1996) use the risk free rate and twelve industry based portfolios.[1] The data used here are the same as Chen and Knez's (1996) study and constitute monthly returns for the period January 1968 through December 1989.[2] This gives a total of $T = 264$ observations.

To implement the estimation, Chen and Knez (1996) estimate the long run variance using a HAC estimator with the Bartlett kernel and a bandwidth $b_T = 17$ and base inference on the two step estimator.[3] Therefore, we begin our analysis with this configuration and then consider the impact of using the iterated estimator and also using two alternative covariance matrix estimators. Since there is no theoretical reason to set $b_T = 17$, we also report results using an HAC estimator with a Bartlett kernel and the bandwidth selected via Newey and West's (1994) data-based method. Finally, we consider the sensitivity of the results to the use of "prewhitening and recolouring" by reporting results based on $\hat{S}_T = \hat{S}_{SE}$ in (3.58). In all calculations, the first step weighting matrix is set equal to the identity matrix.[4]

Table 9.1 reports the results for the group averages. The top line of the table represents the configuration used by Chen and Knez (1996) and the results are close to those reported in their Table 1. The slight differences likely reflect differences in estimation routine.[5] This evidence clearly fails to reject the null of a zero performance evaluation at the 5% level in every case although there is evidence against the null at the 10% level for stability–growth–income group. It can be seen that iteration has no qualitative impact on this conclusion – even though convergence required between ten and fifteen steps in each case.[6] However, the results are far more sensitive to the choice of covariance matrix estimator. If the bandwidth is estimated from the data then the chosen value

[1] See Chen and Knez (1996) for further details.

[2] I am extremely grateful to Peter Knez for providing me with this data.

[3] See Section 3.5.3 for a discussion of HAC estimators.

[4] Chen and Knez (1996) do not report which weighting matrix they used on the first step.

[5] Although a closed form solution exists, this is only exploited in the first step and estimates on subsequent steps are actually obtained using a numerical optimization routine. The convergence criterion is set at $\epsilon_M = 10^{-6}$; see Section 3.2.

[6] Convergence criterion is implemented with $\epsilon_\theta = 10^{-6}$ and $I_{max} = 20$; see Section 3.6.

is always zero or one. This, in turn, leads to an increase in the test statistics in every case. In two cases, the income and stability–growth–income groups, the tests now provide evidence against zero performance at the 5% level. Interestingly if $\hat{S}_{SE}$ is used then the statistics fall between those obtained by fixing and estimating the bandwidth. With this choice of covariance matrix estimator, the null of a zero performance evaluation is not rejected at the 5% level for all groups although only marginally for the income and stability–growth–income groups.

Table 9.1
LOP measures of average mutual fund performance

		Investment objective				
	Stat.	G	IG	I	SGI	MC
$\hat{S}_T = \hat{S}_{HAC}(17)$						
	$J_T^{(2)}$	1.870	1.095	2.233	2.846	2.460
	$p-value$	0.172	0.295	0.135	0.092	0.117
	$J_T^{(i)}$	1.860	1.109	2.195	2.674	2.513
	$p-value$	0.173	0.292	0.139	0.102	0.113
$\hat{S}_T = \hat{S}_{HAC}(\hat{b})$						
	$J_T^{(2)}$	2.478	1.524	4.944	4.042	2.579
	$p-value$	0.116	0.217	0.026	0.044	0.108
	$J_T^{(i)}$	2.410	1.586	4.824	4.084	2.546
	$p-value$	0.121	0.208	0.028	0.043	0.111
$\hat{S}_T = \hat{S}_{SE}$						
	$J_T^{(2)}$	2.170	1.250	3.588	3.431	2.267
	$p-value$	0.141	0.264	0.058	0.064	0.132
	$J_T^{(i)}$	2.147	1.261	3.331	3.213	2.268
	$p-value$	0.143	0.261	0.068	0.073	0.132

Notes: All HAC estimators are calculated with the Bartlett kernel. $\hat{S}_T = \hat{S}_{HAC}(17)$ denotes the case in which $\hat{S}_T = \hat{S}_{HAC}$ in (3.54) and $b_T = 17$; $\hat{S}_T = \hat{S}_{HAC}(\hat{b})$ denotes the case in which $\hat{S}_T = \hat{S}_{HAC}$ and b_T is selected using Newey and West's (1994) method; $\hat{S}_T = \hat{S}_{SE}$ is given in (3.58); $J_T^{(2)}$ and $J_T^{(i)}$ are the overidentifying restrictions test in (5.2) based on the two step and iterated estimators respectively; $p-value$ is the p-value of the overidentifying restrictions test on the line above.

Clearly the evidence is sensitive to the choice of covariance matrix estimator. Unfortunately, no guidance is available regarding the performance of these estimators in this type of setting and so it is impossible to know which version of the test is more reliable here. Even allowing for the sensitivity to the choice of covariance matrix estimator, the results do not provide compelling evidence against a zero performance evaluation. As remarked by Chen and Knez (1996),

such a conclusion might not be considered particularly surprising given the aggregate nature of the group return. However, it is also possible that the choice of measure may understate performance. While it is true that the LOP measure is zero if the mutual fund does not enlarge the investment opportunity set for uninformed investors, it is also possible for the LOP measure to be zero even though the opportunity set has been expanded in the sense that some investor prefers to hold the fund over any constant composition portfolio based on X_t. This concern motivates the introduction of the modified evaluation measure that we now consider.

This modified measure rests on the assumption that the securities market satisfies a no-arbitrage condition, that is all securities with a positive pay-off have a positive price.[7] If this condition is satisfied then the stochastic discount factor is strictly positive and X_t can be priced by $d_t^+ = (X_t^{'}\delta_0)^+$ where $(X_t^{'}\delta_0)^+ = max\{X_t^{'}\delta_0, 0\}$, so that (1.24) is replaced by

$$E[X_t d_t^+] \; = \; 1_N$$

A similar modification is made in the evaluation measure to yield

$$\lambda^+(r_t^m, d_t) = E[r_t^m (X_t'\delta_0)^+] \; - \; 1 \; = \; 0$$

which Chen and Knez (1996) refer to as the NA-measure with the acronym standing for "no-arbitrage". Chen and Knez (1996) show that the NA measure is only zero if the fund does not expand the investment opportunity set and that it is positive if there is at least one investor who would prefer to hold the fund rather than any constant composition portfolio constructed from X_t.

In principle, it is possible to test zero performance using the NA measure in the same way as before. The only difference is that the overidentifying restrictions test is now based on the population moment condition

$$E[Q_t(X_t'\delta_0)^+] - 1_{N+1} = 0 \tag{9.4}$$

However, this difference raises an important issue. The population moment condition in (9.4) involves the function $(X_t'\delta)^+$ that is not differentiable with respect to δ at $X_t^{'}\delta = 0$ and so does not satisfy Assumption 3.5, one of the regularity conditions for our asymptotic distribution theory. However, there are grounds for anticipating that Theorem 5.1 still holds in this case; see Hansen, Heaton, and Luttmer (1995). Therefore, we follow Chen and Knez (1996) and proceed under the assumption that this extension is possible.

The functional form in (9.4) is far more complicated than the LOP case due to the presence of the non-negative operator $(X_t'\delta)^+$. Experimentation with different starting values revealed that this type of nonlinearity creates problems for *fminu*, the gradient method in MATLAB. In fact, it is noted in the User's Guide to the Optimization Toolbox that *fminu* does not work well if the function is discontinuous. Although the minimand is continuous here, the derivative does not exist at $X_t'\delta$ and hence is not a continuous function.

[7] See Ingersoll (1987) [Chapter 2].

Therefore, the estimations are performed using an alternative routine *fmins* within the Optimization Toolbox that employs a simplex search method. This algorithm is less efficient than *fminu* but does not require evaluation of the gradient of the minimand; see Mathworks (2000) for further details.

Further experimentation using *fmins* indicated that the minimum on the first step estimation is located in the neighbourhood of the LOP solution given by (9.3). Therefore, this value is used as the starting value for the first step estimation. The convergence criterion for the numerical optimization is $\epsilon_M = 10^{-6}$. For the iterated estimation, the convergence criterion is once again $\epsilon_\theta = 10^{-6}$ but the maximum number of steps is increased to $I_{max} = 100$. In every case, the numerical optimization failed to converge after $100,000$ iterations on the first few steps of the iterated estimation. Experimentation indicated that increasing the number of replications made no difference either to the likelihood of convergence or the value of the minimand at the end. Nevertheless, after these initial steps convergence did occur on each step and the iterated estimation did itself converge. Therefore, all calculations are performed with a maximum of $100,000$ replications on each step. However, this pattern of behaviour means that the reported values for the overidentifying restrictions test on the second step should be regarded as upper bounds on these statistics.

The results are given in Table 9.2. Once again the first row of the table replicates the configuration reported in Chen and Knez (1996) and we start our discussion with this case. Our results are once again slightly different from those reported in Chen and Knez (1996) but qualitatively similar. In comparison to the LOP measure, the NA measure leads to larger values for the overidentifying restrictions test in every case although none of the tests are significant at the 5% level. Once again, iteration tends to reduce the value of the test statistic. However, as with the LOP measure, the results are very sensitive to the choice of covariance matrix estimator. When the bandwidth is estimated from the data it is invariably zero or one, and this leads to statistics that are significant at the 5% level for both the income and stability–income–growth groups. The use of pre-whitening and recolouring leads to statistics that are lower but nevertheless still marginally significant at the 5% level for the income and stability–income–growth groups.

In their study, Chen and Knez (1996) also report distributions of p-values from applying these tests to individual funds. Although we do not replicate this part of their analysis here, it is worth briefly noting what they found. Of the sixty eight funds considered, they report that 8% of the funds provide evidence against zero performance at the 5% significance level using the LOP measure and that this percentage increases to 13.2% when the NA measure is used. It is unclear precisely how to interpret such percentages because even if all the tests are independent and the null is true in every case then we would expect to reject the null 5 per cent of the time. However, if this evidence is taken at face value then there would appear to be evidence against zero performance using these measures. At this point, it is worth reassessing what the measure actually captures. It can be recalled that these measures compare fund performance to a benchmark set of passively held portfolios. Chen and Knez (1996) argue that

this may be too low a benchmark since financial information is reported in the media. They therefore propose a "conditional" measure in which the benchmark consists of portfolios whose weights can vary in response to publicly available information. These conditional measure can also be implemented using GMM and the type of testing stratgey employed above; see Chen and Knez (1996).

Table 9.2
NA measures of average mutual fund performance

		Investment objective				
	Stat.	G	IG	I	SGI	MC
$\hat{S}_T = \hat{S}_{HAC}(17)$						
	$J_T^{(2)}$	2.135	1.494	2.501	3.306	2.869
	$p-value$	0.144	0.222	0.114	0.069	0.090
	$J_T^{(i)}$	1.919	1.179	2.211	2.781	2.628
	$p-value$	0.166	0.278	0.137	0.095	0.105
$\hat{S}_T = \hat{S}_{HAC}(\hat{b})$						
	$J_T^{(2)}$	2.862	1.879	5.497	4.720	3.180
	$p-value$	0.091	0.171	0.019	0.030	0.075
	$J_T^{(i)}$	2.692	1.838	4.765	4.757	3.135
	$p-value$	0.101	0.175	0.029	0.029	0.077
$\hat{S}_T = \hat{S}_{SE}$						
	$J_T^{(2)}$	2.772	1.656	4.232	4.604	3.061
	$p-value$	0.096	0.198	0.040	0.032	0.080
	$J_T^{(i)}$	2.649	1.634	4.040	4.124	2.968
	$p-value$	0.104	0.201	0.043	0.042	0.085

Notes: see Table 9.1.

9.2 Conditional Capital Asset Pricing Model

Section 1.3.3 describes the conditional capital asset pricing model (CCAPM). This model has been used to investigate the pricing of a wide variety of assets. In this section, we follow Harvey (1991) and investigate whether the model can explain the variation in the returns across international stock markets.

It can be recalled from Section 1.3.3 that the model implies a set of population moment conditions involving the conditional first two moments of the asset returns. It is convenient to express these moment conditions more compactly here. To this end, we set $\theta_{i,0} = (\delta_{i,0}^{'}, \delta_{m,0}^{'})'$, $u_{i,t}(\delta_i) = r_{i,t} - z_{t-1}'\delta_i$ and $u_{m,t}(\delta_m) = r_{m,t} - z_{t-1}'\delta_m$ where (as a reminder) $r_{i,t}$ denotes the excess return on holding the market portfolio for country i, $r_{m,t}$ is the excess return from

holding the "world market" portfolio below, z_{t-1} denotes a vector of relevant economic and financial variables contained in the information set Ω_{t-1}. The population moment conditions in (1.34) and (1.36) can be written as

$$E[f_i(v_t, \theta_{i,0})] \;=\; E[a_{i,t}(\theta_{i,0}) \otimes z_{t-1}] \;=\; 0 \tag{9.5}$$

where

$$a_{i,t}(\theta_i) \;=\; \begin{bmatrix} u_{i,t}(\delta_i) \\ u_{m,t}(\delta_m) \\ u_{m,t}(\delta_m)^2 z_{t-1}^{'}\delta_i - u_{m,t}(\delta_m)u_{i,t}(\delta_i)z_{t-1}^{'}\delta_m \end{bmatrix} \tag{9.6}$$

If the model is correct then these moment conditions hold simultaneously for all countries. Harvey reports results based on (9.5) for both individual countries and groups of countries. For brevity here, we only consider GMM estimation on a country by country basis and so $\theta_{i,0}$ is estimated based on (9.5) for each country i.

Given this approach to the estimation, the condition for local identification of $\theta_{i,0}$ is that $rank\{E[\partial f_i(v_t, \theta_{i,0})/\partial\theta_i^{'}]\} = p$ where, in this case, $p = 2n_z$ and n_z denotes the dimension of z_{t-1}. For this model, the derivative matrix is as follows,

$$\frac{\partial f_i(v_t, \theta_i)}{\partial \theta_i^{'}} \;=\; A_{i,t}(\theta_i) \otimes z_{t-1}z_{t-1}^{'} \tag{9.7}$$

where

$$A_{i,t}(\theta_i) \;=\; \begin{bmatrix} -1 & 0 \\ 0 & -1 \\ A_{i,t}^{(5)} & A_{i,t}^{(6)} \end{bmatrix}$$

and

$$\begin{aligned} A_{i,t}^{(5)} &= u_{m,t}(\delta_m)^2 + u_{m,t}(\delta_m)z_{t-1}^{'}\delta_m \\ A_{i,t}^{(6)} &= -2u_{m,t}(\delta_m)z_{t-1}^{'}\delta_i - u_{m,t}(\delta_m)u_{i,t}(\delta_i) + u_{i,t}(\delta_i)z_{t-1}^{'}\delta_m \end{aligned}$$

Using (1.34), it follows from (9.7) that

$$E\left[\frac{\partial f_i(v_t, \theta_{i,0})}{\partial \theta_i^{'}}\right] \;=\; E[\tilde{A}_{i,t} \otimes z_{t-1}z_{t-1}^{'}] \tag{9.8}$$

where

$$\tilde{A}_{i,t} \;=\; \begin{bmatrix} -1 & 0 \\ 0 & -1 \\ u_{m,t}(\delta_{m,0})^2, & -u_{i,t}(\delta_{i,0})u_{m,t}(\delta_{m,0}) \end{bmatrix}$$

It can easily recognized that the matrix in (9.8) is rank p provided $E[z_{t-1}z_{t-1}^{'}]$ is nonsingular. The latter condition holds as long as there are no linear redundancies among the information variables.

As mentioned above, we present here results from estimating the model for individual countries. It should be noted that this approach does not impose all

the restrictions of the model. The underlying theory implies that (9.5) holds for all i at the same value for $\delta_{m,0}$. However, the latter restriction is ignored when the estimation is on a country by country basis. Therefore, as noted by Harvey, some caution must be exercised in interpreting the results. If the model is not rejected for individual countries then this does not necessarily mean that the model can simultaneously explain the variation in returns across these countries. On the other hand, if the model is rejected for a particular country then that provides valuable information about the failings of the underlying theory.

The data are the same as those used in Harvey's study.[8] The observations are monthly and span the period 1970:02 to 1989:05; this gives a total of 232 observations. The world market portfolio is the Morgan Stanley Capital International index (MSCI) and represents a weighted combination of the returns on a variety of world-wide investments; see Harvey (1991) for specific details. Both the world return, $r_{m,t}$ and the country i return, $r_{i,t}$, are expressed in U.S. dollars in excess of the holding period return on the T-bill that is closest to 30 days to maturity. The information variables are denoted by z_{t-1} above. The vector z_t contains: a constant; $r_{m,t}$; a dummy variable for the month of January; the U.S. term structure premia, calculated as the return for holding a 90-day U.S. T-bill for one month less the return from holding a 30 day T-bill; the U.S. default risk spread calculated as the yeld on a Moody's Baa rated bond less the yield on a Moody's Aaa rated bond; the dividend yield on the Standard and Poor's 500 stock index less the return on a 30-day U.S. T-bill. Given this choice of information variables, the model yields $q = 18$ moment conditions and $p = 12$ parameters. Harvey (1991) reports results for seventeen countries.[9] However, for brevity, we restrict attention to the G-7 countries.

With regard to the specifics of the GMM estimation, Harvey (1991) uses a first step weighting matrix proportional to the identity matrix, and estimates the long run variance by $\hat{S}_{SU}$. Notice that the latter is consistent because $a_{i,t}(\theta_{0,i})$ is a martingale difference given Ω_{t-1} – provided the model is correctly specified. Therefore, we use these weighting matrices as well but also consider the sensitivity of the results to the use of $(T^{-1}\sum_{t=1}^{T} I_3 \otimes z_t z_t^{'})^{-1}$ as the first step weighting matrix. In each case, the starting values for δ_i are the least squares estimates from the regression of $r_{i,t}$ on z_{t-1}, and those for δ_m are the corresponding estimates only with $r_{m,t}$ as the dependent variable.[10]

It can be seen from Table 9.3 that the relatively insensitive to the choice of first step weighting matrix, and that in each case iterated estimation yields identical statistics. However, the qualitative conclusion can be sensitive to iteration. For both the U.S. and Japan, the overidentifying restrictions test statistic

[8] I am grateful to Eric Ghysels for providing me with the data.

[9] These countries are: Australia, Austria, Belgium, Canada, Denmark, France, Germany, Hong Kong, Italy, Japan, The Netherlands, Norway, Spain, Sweden, Switzerland, the United Kingdom, and the United States.

[10] Notice that these estimates are the GMM estimates based on (1.34) alone.

is insignificant at the 10% level after two steps but becomes significant after iteration. It is the results based on the iterated statistics that replicate those reported in Harvey (1991). Overall, there is evidence in favour of the model for Canada, France, Germany, Italy and the U.K. and evidence against the model for Japan and the U.S.

Table 9.3
Overidentifying restrictions tests for the conditional capital asset pricing model

	Case A		Case B	
Country	Two step	Iterated	Two step	Iterated
Canada	3.669	3.156	3.145	3.156
	0.721	0.789	0.790	0.789
France	8.316	10.308	7.861	10.308
	0.216	0.112	0.248	0.112
Germany	3.183	3.476	3.229	3.476
	0.785	0.747	0.780	0.747
Italy	9.538	9.821	8.337	9.821
	0.146	0.132	0.214	0.132
Japan	8.461	14.984	10.428	14.984
	0.206	0.020	0.108	0.020
U.K.	1.083	1.104	1.094	1.104
	0.982	0.981	0.982	0.981
U.S.	7.450	10.764	7.499	10.764
	0.277	0.096	0.281	0.096

Notes: Case A denotes $W_T^{(1)} = 10^5 I_{18}$, and Case B denotes $W_T^{(1)} = (T^{-1}\sum_{t=1}^{T} I_3 \otimes z_t z_t')^{-1}$. The numbers below the test statistics are the associated p-values.

It can be recalled that the innovation of the CCAPM is to allow the investment betas to vary over time. It is therefore natural to question the assumed form of this variation is appropriate. If variation is present but the assumed model is incorrect, then it would be anticipated that this would manifest itself in structural instability. While the overidentifying restrictions test provides a general diagnostic for misspecification, it is not specifically designed to test for structural instability. Furthermore, it can be recalled from Section 5.4 that the overidentifying restrictions test can have size equal to power against certain types of structural instability. Motivated by these concerns, Ghysels (1998) argues that it is important to submit the CCAPM to formal test of structural stability. He pursued the issue in the context of CCAPM's for domestic U.S. assets and found widespread evidence of instability. Therefore, we now consider if similar evidence is present here.

Since there is no reason to associate any instability with a particular moment in time, the analysis is based on the unknown break point versions of the tests described in Section 5.4.2. It can be recalled from this earlier discussion that the construction of these tests involves the calculation of the "known break point" tests for all possible break points within an interval Π. Conventional practice is to set $\Pi = [0.15, 0.85]$ and we followed this rule here in the absence of any alternative guidance. However, it is worth considering what this rule actually implies here. For $\pi = 0.15$, the first sub-sample involves only thirty-five observations. This means that the sub-sample estimation attempts to retrieve twelve parameters from eighteen moment conditions based on just 35 observations! As can readily be imagined, under such circumstances, convergence can be very sensitive to the sample. In this example, it turns out that the numerical optimization runs into problems when $\pi = 0.15$ but not when $\pi = 0.85$ (and so the second sub-sample consists of thirty-five observations). The problems stem from the near singularity of $\hat{S}_{1,T}(\pi)$. Since this occurred for every country, this particular break point is dropped from the calculations and so the tests are calculated for break points $t_B = 36, 37, ...197$ – or, equivalently, 1973.02 through 1986.07. Convergence also proves a problem for sub-samples if the maximum allowable number of iterations, I_{max} is set too high. As a result, the estimations are performed with $I_{max} = 6$, that is a six-step estimator. All calculations use the first step weighting matrix $W_T^{(1)} = (T^{-1} \sum_{t=1}^{T} I_3 \otimes z_t z_t^{'})^{-1}$.

Table 9.4 reports the $Sup-$, $Av-$ and $Exp-$ versions of the tests for both parameter variation and also stability of the overidentifying restrictions.[11] It can be recalled from Section 5.4 that the three tests of parameter variation are asymptotically equivalent under both the null hypothesis (i.e. no parameter variation) and local alternatives. However, the three statistics exhibit very diverse behaviour here. The LM versions of the test are insignificant at the 10% level in every case, the Wald versions are significantly larger and significant at the 1% level in every case, and the D versions are orders of magnitude larger still.[12] To date there is no guidance available regarding which version of these tests is more reliable in finite samples. However, one possible explanation for this discrepancy is that the LM and D tests are calculated using the full sample GMM estimator rather than the restricted estimator. While the two estimators are asymptotically equivalent under the null and local alternatives, it may be that the sample size here is too small for this equivalence to apply. To explore this possibility further, the test statistics are also calculated using the following versions of the LM and D tests based on the restricted estimator $\tilde{\theta}_T(\pi)$,

$$LM_T^0(\pi) = \sum_{i=1}^{2} T_i d_{i,T}(\tilde{\theta}_T(\pi); \pi)' \tilde{V}_{i,T}(\pi) d_{i,T}(\tilde{\theta}_T(\pi); \pi) \tag{9.9}$$

[11] Critical points for these statistics are given in Tables 5.5 and 5.6.

[12] The extremely large values of the D statistic occur when one of the sub-samples is small, because in these cases it turns out here that the estimator of the long run variance in the small sub-sample is ill-conditioned and so close to singularity.

$$D_T^0(\pi) = T[J(\tilde{\theta}_T(\pi), \tilde{\theta}_T(\pi); \pi) - J(\hat{\theta}_{1,T}(\pi), \hat{\theta}_{2,T}(\pi); \pi)] \quad (9.10)$$

where

$$d_{i,T}(\tilde{\theta}_T(\pi); \pi) = G_{i,T}(\tilde{\theta}_T(\pi); \pi)' \tilde{S}_{i,T}(\pi)^{-1} g_{i,T}(\tilde{\theta}_T(\pi); \pi)$$

$$\tilde{V}_{i,T} = [G_{i,T}(\tilde{\theta}_T(\pi); \pi)' \tilde{S}_{i,T}(\pi)^{-1} G_{i,T}(\tilde{\theta}_T(\pi); \pi)]^{-1}$$

and $\tilde{S}_{i,T}(\pi)$ denotes a consistent estimator of S_i based on $\tilde{\theta}_T(\pi)$; see Section 5.4.1 for further definitions. Both these statistics are asymptotically equivalent under the null and local alternatives to the three tests for parameter variation discussed in Section 5.4.1. Interestingly, for the example here, it can be seen from Table 9.4 that the tests based on $LM_T^0(\pi)$ and $D_T^0(\pi)$ yield statistics that are closer to the Wald based tests than their counterparts based on the full sample estimator. Nevertheless, substantial differences remain, although in nearly every case the tests based on $W_T(\pi)$, $LM_T^0(\pi)$ and $D_T^0(\pi)$ yield qualitatively the same conclusion.

The overidentifying restrictions based tests tend to be insignificant. Therefore, if these results are taken at face value then collectively they suggest that the misspecification is due to parameter variation. However, this would seem to be a big "if". As mentioned above, the discrepancies in the tests raise suspicions about the adequacy of the asymptotic approximation here. Furthermore, we note that many of the *Sup−* tests yield estimated break points close to the beginning or end of the sample, and it can be recalled that this may also be an indicator that asymptotic theory is not a good approximation. Research is currently in progress to investigate the reliability of these tests in the types of setting considered here.

Due to these concerns about the adequacy of the asymptotic approximation, we do not pursue this example further. However, it is worth briefly summarising Ghysels's (1998) conclusions regarding the ability of the CCAPM to explain the prices of domestic U.S. assets. His sample consists of monthly data from 1927:01 through 1988:01 and thus contains more than three times as many observations as the sample in our example above. Inference about structural stability is based on the *Sup-LM* test based on the statistic in (5.77) using $\Pi = [0.2, 0.8]$. The evidence indicates that the model is validated by the overidentifying restrictions test in many cases but that there is substantial evidence of misspecification due to neglected parameter variation. Interestingly, Ghysels (1998) reports that in many cases the unconditional version of the capital asset pricing model provides more accurate forecasts of asset prices than its conditional counterpart. These results highlight the importance of not relying purely on the overidentifying restrictions test to assess the adequacy of the specification.

Table 9.4
Structural stability tests for the conditional capital asset pricing model

Country	Statistic	*Sup−*	*Date*	*Av−*	*Exp−*
Canada	W	53.196	1985.08	31.444	22.404
	LM	23.151	1977.11	14.636	9.817
	LM^0	43.276	1977.01	27.467	18.895
	D	4.0×10^5	1973.07	3162.427	∞
	D^0	271.293	1973.05	42.895	130.559
	O	22.219	1983.12	13.324	8.239
France	W	248.733	1973.09	36.225	119.279
	LM	25.036	1980.09	16.162	10.054
	LM^0	56.713	1973.10	24.450	23.363
	D	5.7×10^{15}	1973.10	3.5×10^{13}	∞
	D^0	2.0×10^4	1974.01	303.526	∞
	O	30.667	1983.07	17.254	11.980
Germany	W	116.670	1986.04	36.640	53.726
	LM	19.791	1982.11	12.532	7.155
	LM^0	40.085	1975.04	23.798	15.938
	D	7.8×10^7	1973.09	7.0×10^5	∞
	D^0	1.4×10^4	1973.12	294.689	∞
	O	19.821	1974.01	11.388	7.199
Italy	W	144.147	1986.02	37.277	66.986
	LM	18.466	1978.03	11.881	6.848
	LM^0	43.198	1976.05	21.689	18.239
	D	7.0×10^7	1986.07	5.2×10^5	∞
	D^0	3.9×10^5	1973.07	3775.701	∞
	O	23.598	1984.12	17.364	9.700
Japan	W	117.877	1973.02	33.620	54.157
	LM	18.715	1977.01	12.429	7.251
	LM^0	44.201	1973.05	24.508	17.215
	D	2.8×10^{10}	1986.02	1.7×10^8	∞
	D^0	3.6×10^5	1973.05	2408.333	∞
	O	30.142	1974.10	22.460	12.521
U.K.	W	64.166	1973.02	21.053	26.996
	LM	18.566	1981.12	11.594	6.704
	LM^0	37.351	1986.07	20.611	14.801
	D	6.9×10^7	1986.04	6.6×10^5	∞
	D^0	1.0×10^4	1986.04	129.555	∞
	O	21.528	1985.06	9.983	7.957

continued

Table 9.4 (*cont.*)
Structural stability tests for the conditional capital asset pricing model

Country	Statistic	*Sup–*	*Date*	*Av–*	*Exp–*
U.S.	W	121.367	1973.05	33.069	55.611
	LM	17.225	1982.11	11.185	6.556
	LM^0	32.128	1976.10	20.865	13.328
	D	1.2×10^9	1986.02	7.3×10^6	∞
	D^0	4.9×10^4	1986.02	350.100	∞
	O	27.926	1986.07	16.748	10.084

Notes: W denotes the versions of the statistics based on the Wald test in (5.75), LM denotes the versions of the statistics based on the LM test in (5.77), LM^0 denotes the versions of the statistics based on the LM test in (9.9), D denotes versions of the statistics based on the D test in (5.78), D^0 denotes versions of the statistics based on the D test in (9.10), O denotes versions of the tests based on the statistic in (5.80). ∞ denotes results too large to represent as conventional floating-point values.

9.3 Inventory Holdings by Firms

Section 1.3.4 describes Eichenbaum's (1989) model for inventory holdings by firms. The key innovation is that the model provides a framework for determining whether the production smoothing hypothesis or the production cost smoothing hypothesis better captures firm behaviour. In this section, we examine which hypothesis – if either – captures aggregate inventory holding behaviour in non-durable manufacturing industries for the U.S.

It can be recalled from Section 1.3.4 that the difference betwen the two hypotheses rests on the the presence or absence of the stochastic shock ν_t in the cost function. If $\nu_t = 0$ for all t then stochastic cost shocks are absent and so the only incentive for holding inventories is the desire to smooth production levels. However, if $\nu_t \neq 0$ then stochastic cost shocks are present and this produces an incentive to hold inventories to smooth both levels and costs. For simplicity, we use the term "production smoothing version of the model" to refer to the case in which $\nu_t = 0$, and the term "production cost smoothing version of the model" to refer to the case in which $\nu_t \neq 0$.

Eichenbaum (1989) shows that the production smoothing version of the model implies the population moment condition,

$$E[z_t h_{t+1}(\psi_0)] \;=\; 0 \tag{9.11}$$

where as a reminder

$$h_{t+1}(\psi_0) = I_{t+1} - \{\lambda_0 + (\lambda_0\beta_0)^{-1}\}I_t + \beta_0^{-1}I_{t-1} + S_{t+1} - \phi_0\beta_0^{-1}S_t \tag{9.12}$$

where $\psi_0 = (\lambda_0, \beta_0, \phi_0)'$ and $z_t \in \Omega_t$. The production cost smoothing version of the model implies the population moment condition

$$E[\{h_{t+2}(\psi_0) - \rho_0 h_{t+1}(\psi_0)\}z_t] = 0 \tag{9.13}$$

It is clear from (9.11) and (9.13) that the production smoothing and production cost smoothing hypotheses imply different restrictions on the data. To emphasize an important aspect of this difference, it is useful to compare the two sets of moment conditions in the case where $\rho_0 = 0$. From (9.11), it can be seen that the production smoothing version of the model implies that $h_{t+1}(\psi_0)$ is orthogonal to all elements of the information set Ω_t. In contrast, from (9.13) (lagged one period), it can be seen that the production cost smoothing version implies only that $h_{t+1}(\psi_0)$ is orthogonal to any element of the information set Ω_{t-1}. Furthermore, Eichenbaum (1989) shows that if the production cost smoothing version of the model is correct then $h_{t+1}(\psi_0)$ is not orthogonal to any element of the information set Ω_t. This key difference indicates a way of discriminating between the two versions of the model: if the production smoothing version is correct then (9.11) is valid, but if the production cost smoothing version is correct then (9.11) is invalid but (9.13) is valid. Therefore, the overidentifying restrictions test associated with these two sets of moment conditions can reveal which, if either, of the two competing hypotheses are correct.

The primary focus of such inventory models is to estimate of the speed of adjustment of actual inventories to the target (or desired) level of inventories. Eichenbaum (1989) shows that the speed of adjustment is $1 - \lambda_0$ within either version of the model. This means that firms adjust inventories toward their target level at $(1 - \lambda_0)100\%$ per month. This interpretation also means that λ_0 must lie between zero and one if it is to make economic sense. Economic theory also implies a restriction on ϕ_0. It can be recalled that $\phi_0 = 1 - \delta_0\gamma_0/\alpha_0$ where $(\delta_0, \gamma_0, \alpha_0)$ are parameters of the cost function. Given their roles in the cost function, all three of these parameters would be positive if the model is correctly specified. This, in turn, translates into the restriction that $\phi_0 < 1$. One final aspect of the parameterization needs to be noted. To simplify the estimation, it is customary to fix the value of the discount factor $\beta_0 = 0.995$ a priori, and we follow this practice here – as did Eichenbaum (1989).

Eichenbaum (1989) estimates both versions of the model using aggregated data for all non-durable manufacturing industries in the U.S. as well as aggregated data for six specific industries using monthly data for 1959:1 through 1984:12.[13] Here we restrict attention to the aggregate data for all non-durable manufacturing industries, and use a revised and enlarged data set spanning 1959:1 through 1998:5, yielding a sample of 473 observations. The data are compiled by the Bureau of Economic Analysis (BEA), the US Department of Commerce.[14] The series represent end of the month inventories and sales of finished goods. The data are adjusted to constant chained 1992 dollars and are seasonally adjusted.

One immediate problem is that these data on inventories and sales are not stationary because they trend over time. Eichenbaum (1989) presents results based on detrending the data via either first differencing or using a quadratic

[13] The six industries in question are: tobacco, rubber, food, petroleum, chemicals and apparel.

[14] I am grateful to David Doorn for providing me with the data.

time trend. Although there is not complete unanimity in the literature on the appropriate method, these data are most commonly detrended using a quadratic time trend and so we use that method here.[15]

We first consider estimation of the production smoothing version of the model based on (9.11). Following Eichenbaum (1989), we set

$$z_t \;=\; (1, I_t, S_t, I_{t-1}, S_{t-1})$$

and so there are $q = 5$ moment conditions in $p = 2$ unknown parameters. The first step weighting matrix is set equal to the inverse of the instrument cross product matrix, $(T^{-1}\sum_{t=1}^{T} z_t z_t^{'})^{-1}$. Within this model, $h_{t+1}(\psi_0)$ is martingale difference with respect to Ω_t and so the long run variance can be consistently estimated by $\hat{S}_{SU}$.

The condition for local identification of $\theta_0 = (\lambda_0, \phi_0)'$ is that the rank of $E[\partial f(v_t, \theta_0)/\partial\theta']$ equal two. For this model, the derivative matrix is given by

$$E[\partial f(v_t, \theta_0)/\partial\theta'] = E[z_t \tilde{x}_t'] M(\theta_0) \tag{9.14}$$

where $\tilde{x}_t = (I_{t+1}, S_t)'$ and

$$M(\theta) = \begin{bmatrix} (0.995\lambda^2)^{-1} - 1 & 0 \\ 0 & -(0.995)^{-1} \end{bmatrix}$$

As discussed in Section 3.1,

$$rank\{E[\partial f(v_t, \theta_0)/\partial\theta']\} \leq min\{rank\,(E[z_t\tilde{x}_t'])\,, rank\,(M(\theta_0))\}$$

and so a necessary condition for identification is that both $E[z_t\tilde{x}_t^{'}]$ and $M(\theta_0)$ have rank equal to two. Inspection of $M(\theta_0)$ indicates that this matrix is of rank two. However, it is impossible to say anything regarding $E[z_t\tilde{x}_t^{'}]$ *a priori*.

Table 9.5 reports results for three different starting values, $(\lambda, \phi) = (0.5, 0.5)$, $(0.9, 0.9)$, $(1.5, 1.0)$, using a convergence criterion of 10^{-6}. It can be seen that in each case the first step estimates are the same to three decimal places. However, the estimates are not identical to a higher precision and this explains why the iterated estimates are different. In spite of these differences, all the estimates for λ exceed one and all the overidentifying restrictions statistics are significant at the 1% level, both of which are indicative of misspecification.

Instead of proceeding to the production cost smoothing version of the model, we first explore some alternative approaches to estimation of the production smoothing model based on scaled versions of the moment condition. To motivate these alternative approaches, it is useful to consider a plot of the first step minimand associated with estimation based on the original moment condition in (9.11). As can be seen in Figure 9.1, this first step minimand is very flat in the area of the starting values.[16] Of most concern is the fact that the minimand is very flat in the dimension of λ, the parameter of most interest. So for this data, it would appear that the population moment condition in (9.11) is not

[15] See Doorn (2003) for further discussion.

[16] Parenthetically, we note that this flatness is exhibited by the minimand on subsequent steps and explains the sensitivity of the iterated estimates.

Table 9.5
GMM estimates of the production smoothing model based on $c(\theta_0)E[f(v_t, \theta_0)] = 0$

$c(\theta_0)$	*Statistic*	*Starting values*, (λ, ϕ) $(0.5, 0.5)$	$(0.9, 0.9)$	$(1.5, 1.0)$
1	$(\hat{\lambda}_T^{(1)}, \hat{\phi}_T^{(1)})$	(1.003, 0.947)	(1.003,0.947)	(1.003,0.947)
	$TQ_T(\hat{\theta}_T^{(1)})$	53362295.59	53362295.59	53362295.59
	$(\hat{\lambda}_T^{(2)}, \hat{\phi}_T^{(2)})$	(1.003, 0.945)	(1.003,0.945)	(1.003,0.945)
	$J_T^{(2)}$	26.151*	26.151*	26.151*
	$(\hat{\lambda}_T^{(.)}, \hat{\phi}_T^{(.)})$	(1.135, 0.937)	(1.003,0.945)	(1.003,0.945)
	$J_T^{(.)}$	25.218*	26.154*	26.154*
$(1-\lambda)^{-1}$	$(\hat{\lambda}_T^{(1)}, \hat{\phi}_T^{(1)})$	(0.699, 0.901)	*diverges*	(1.550,0.880)
	$TQ_T(\hat{\theta}_T^{(1)})$	1.0×10^9	.	4.4×10^8
	$(\hat{\lambda}_T^{(2)}, \hat{\phi}_T^{(2)})$	(0.740, 0.907)	.	(1.427,0.894)
	$J_T^{(2)}$	42.597*	.	60.879*
	$(\hat{\lambda}_T^{(.)}, \hat{\phi}_T^{(.)})$	(0.744, 0.905)	.	(1.407,0.894)
	$J_T^{(.)}$	34.229*	.	41.460*
-0.995λ	$(\hat{\lambda}_T^{(1)}, \hat{\phi}_T^{(1)})$	(0.791, 0.927)	(0.791,0.927)	(0.790,0.927)
	$TQ_T(\hat{\theta}_T^{(1)})$	37526635.32	37526635.32	37526635.32
	$(\hat{\lambda}_T^{(2)}, \hat{\phi}_T^{(2)})$	(0.815, 0.926)	(0.815,0.926)	(0.815,0.926)
	$J_T^{(2)}$	27.118*	27.118*	27.118*
	$(\hat{\lambda}_T^{(.)}, \hat{\phi}_T^{(.)})$	(0.818,0.926)	(0.818,0.926)	(0.818,0.926)
	$J_T^{(.)}$	25.993*	25.993*	25.993*

Notes: $(\hat{\lambda}_T^{(i)}, \hat{\phi}_T^{(i)})$ denote the GMM estimators on the i^{th} step, $J_T^{(i)}$ denotes the overidentifying restrictions test on the i^{th} step; $i = .$ denotes the iterated estimator, * denotes significance at the 1% level.

particularly informative about the parameters. A similar qualitative conclusion has been drawn by researchers using other aggregate inventory data, and this has motivated an interest in scaling the Euler equation residual, $h_{t+1}(\psi_0)$, in an attempt to provide a moment condition that is more informative about λ_0 over the range of economically meaningful values. A number of scalings have been used, here we consider two. Schuh (1996) bases estimation on the scaled moment condition[17]

$$(1 - \lambda_0)^{-1} E[z_t h_{t+1}(\psi_0)] \; = \; 0 \tag{9.15}$$

[17] It should be noted that Schuh (1996) uses this transformed moment condition to estimate the model using establishment level data and not the aggregate data used here.

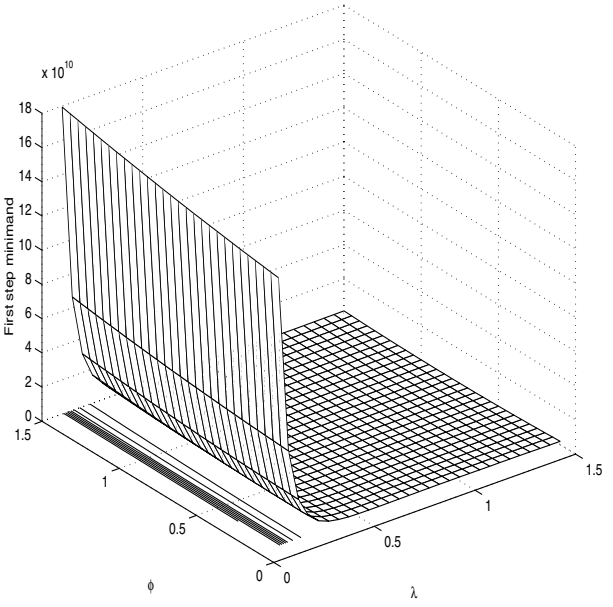

Figure 9.1: First step minimand for the production smoothing model – original version

Durlauf and Maccini (1995) base estimation on the scaled moment condition

$$-\beta_0\lambda_0 E[z_t h_{t+1}(\psi_0)] \;=\; 0 \tag{9.16}$$

Both these transformations are examples of a "curvature altering transformation of the population moment condition" that is discussed in Section 3.7. From this discussion, it can be recalled that such transformations alter the first order conditions in overidentified models and so, in turn, alter the estimates in general. However, the estimator is still consistent. We now consider the effects of each of these transformations upon our results.

First, consider the case in which the moment condition is multiplied by $(1-\lambda_0)^{-1}$. Figure 9.2 contains a plot of the first step minimand associated with estimation based on (9.15) – again using $W_T = (T^{-1}\sum_{t=1}^{T} z_t z_t^{'})^{-1}$. It can be seen that the transformation serves to create a ridge at $\lambda = 1$ so that there is now a boundary around the economically meaningful range of values for λ.[18] Nevertheless, the minimand still appears to be very flat within this region. It should also be noted that the transformation has created a function with at least two minima: one associated with λ in the economically relevant range and one associated with a value of λ greater than one. One potential numerical disadvantage of this transformation is that if $\lambda = 1$ then $(1-\lambda)^{-1}$ is infinite.

[18] Since the minimand is infinite at $\lambda_0 = 1$, the plot is truncated over the range $\lambda \in (0.95, 1.05)$.

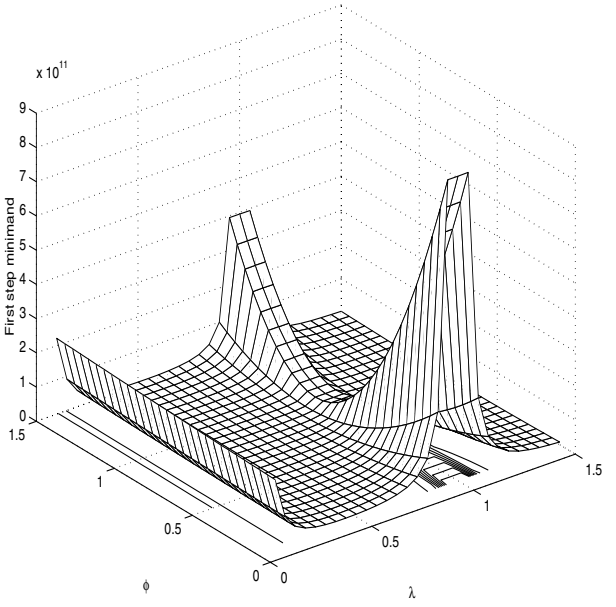

Figure 9.2: First step minimand for the production smoothing model – scaled by $(1-\lambda)^{-1}$

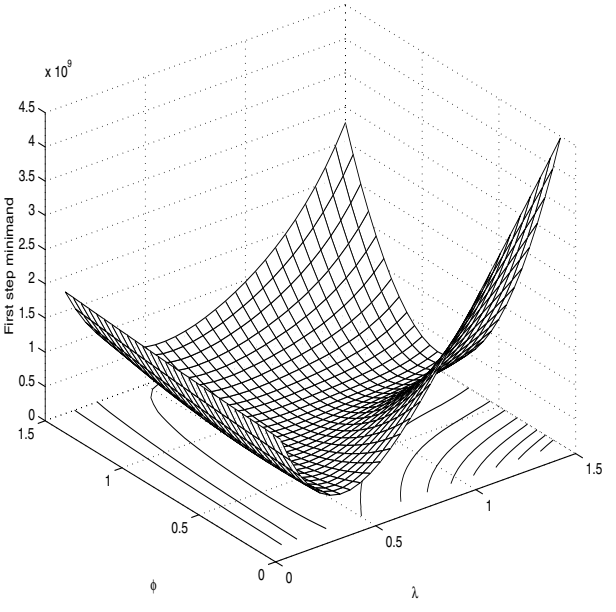

Figure 9.3: First step minimand for the production smoothing model – scaled by $-\beta\lambda$

To circumvent this problem, $(1-\lambda)^{-1}$ is computed as $(1-\lambda+eps)^{-1}$ where $eps = 2.2204 \times 10^{-16}$ and represents the floating point relative accuracy in MATLAB calculations. Estimation results are reported in Table 9.5 using the same starting values as before. It can be seen that this time the first step estimates are sensitive to the starting values. If the starting values are $(0.5, 0.5)$ then the estimates of λ are less than one. If the starting value is $(0.9, 0.9)$ – and hence close to the ridge – then the estimates diverge on the first step with the result that $\hat{S}_{SU}$ is singular. If the starting value is $(1.5, 1.0)$ then the estimate of λ is greater than one. It is the local minimum associated with $\hat{\lambda}_T > 1$ that is the smaller of the two. However, this does not necessarily mean that it is these estimates that should be chosen. The underlying economic model implies that there should be a value of θ_0 that both satisfies the population moment condition and is also economically meaningful. This does not preclude the possibility of other minima outside the economically relevant part of the parameter space. Furthermore, the asymptotic distribution theory only requires local identification. Therefore, if there is a single well-determined minimum within the economically meaningful part of the parameter space, then it is reasonable to adopt the estimates associated with that minimum. The mimimum associated with $\hat{\lambda}_T < 1$ would appear to satisfy the criteria above. However, even then, the point estimate for λ_0 implies that inventories are adjusting towards the desired level at about 30% a month which is considered to be implausibly low. The overidentifying restrictions tests also indicate misspecification. Interestingly, the values for these statistics are substantially larger than with the previous model.

Now consider the case in which the moment condition is multiplied by $-\beta_0\lambda_0$. Figure 9.3 contains a plot of the first step minimand associated with estimation based on (9.16) - again using $W_T = (T^{-1}\sum_{t=1}^{T} z_t z_t^{'})^{-1}$. It can be seen that this transformation has created more curvature in the minimand. While it is not clear exactly where the minimum lies, it certainly looks more clearly defined. This is borne out by the estimation results: the estimates are actually identical to six decimal places on each of the steps reported. As can be seen from Table 9.5, the estimates of λ_0 are all less than one, but once again the implied speed of adjustment is implausibly low. The overidentfiying restrictions tests are again significant at the 1% level.

It is evident that the estimates reported above are sensitive to the choice of transformation. Unfortunately, there is no obvious way to choose between them. It can be recalled that this sensitivity is one motivation for using the continuous updating GMM estimator. Figure 9.4 plots the minimand of the continuous updating GMM estimator. It can be seen that this function exhibits considerably more curvature. Figure 9.5 rotates this plot to reveal the valley more clearly to the eye – note, that the axis are therefore different from the previous plots. Given the potential instabilities of the continuous updating GMM minimand,[19] the estimation is implemented using all the iterated estimates reported above as starting values. The results are presented in Table 9.6. Interestingly, the different starting values lead to two different sets of estimates one with $\hat{\lambda}_T > 1$

[19] See Section 3.7.

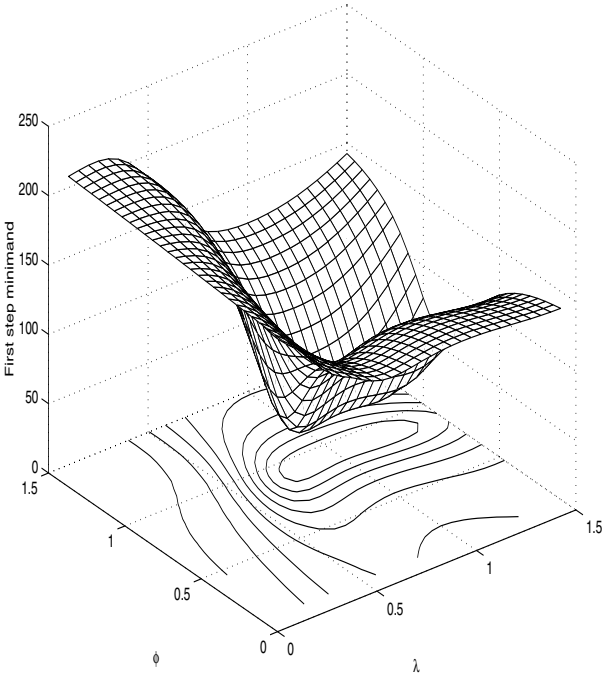

Figure 9.4: Continuous updating GMM minimand

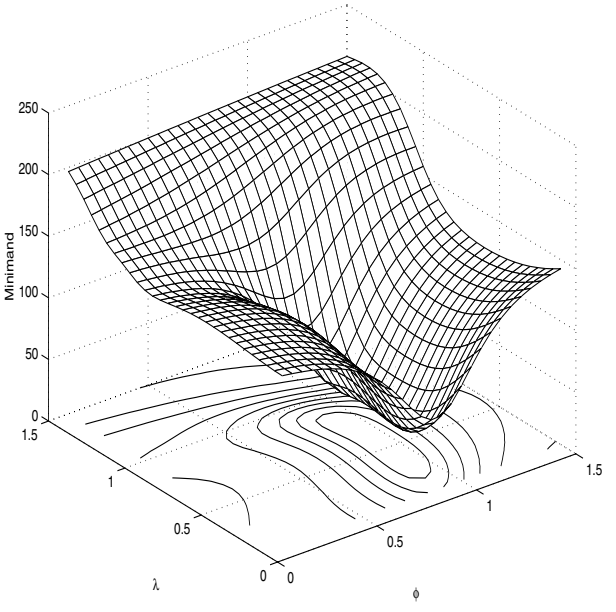

Figure 9.5: Continuous updating GMM minimand – with 90^o rotation

Table 9.6
Continuous updating GMM estimates of the production smoothing model

Starting values, (λ, ϕ)	$(\hat{\lambda}_T, \hat{\phi}_T)$	J_T
(0.818, 0.926),(1.407,0.894)	(0,865,0.935)	25.134*
(0.744,0.905), (1.003, 0.945)	(1.161,0.935)	25.134*
(1.135,0.937)		

Notes: J_T denotes the overidentifying restrictions test, * denotes significance at the 1% level.

and one with $\hat{\lambda}_T < 1$. The associated overidentifying restrictions statistics are actually identical to ten decimal places. Although the estimates differ slightly from those above, the overall conclusion is the same. Taking the minimum associated with $\hat{\lambda}_T < 1$, the implied speed of adjustment is implausibly low and the overidentifying restrictions test is significant at the 1% level.

While the estimates differ across the various estimations of the production smoothing model, the basic conclusion is the same: the model is rejected with this data. Eichenbaum (1989) reports a similar finding. We therefore now consider the production cost smoothing version of the model, and concentrate exclusively on the original version of the population moment condition in (9.13).

The condition for local identification is derived for a special case of the model in Section 3.1, and since the condition is not particularly instructive, we do not pursue it further here. The estimation is implemented using the same instrument vector as before but there are now $q = 5$ moment conditions in $p = 3$ parameters due to the introduction of ρ. The first step weighting is once again $(T^{-1} \sum_{t=1}^{T} z_t z_t')^{-1}$. Eichenbaum (1989) shows that the long run variance can be consistently estimated by $\hat{S}_{SU}$ within this model as well, and so we use this estimator here.

Experimentation with a variety of starting values indicates that the first step minimand possesses multiple minima. As above, there is one involving $\lambda < 1$ and one with $\lambda > 1$. Since there appears to be only one minima in the economically meaningful area of the parameter space, we focus attention on the results associated with this minimum. It can be seen that from Table 9.7 that these estimates provide evidence in favour of the specification. The implied speed of adjustment is approximately 70%, $\hat{\phi}_T < 1$ and the overidentifying restrictions tests are insignificant at the 10% level (although only just in one case). Table 9.7 also contains the results from a continuous updating GMM estimation using the iterated estimates as starting values. The only important difference is that the continuous updating estimator of ϕ_0 is -0.933 as opposed to the iterated GMM estimates of approximately -0.2. Therefore all the results are consistent with Eichenbaum's (1989) finding that the production cost smoothing model is not rejected using aggregate non-durables industry data.

Although our analysis is confined to aggregate non-durable industry data, it is worth noting that Eichenbaum (1989) reports a similar pattern of results

for the six industries considered in his study. Durlauf and Maccini (1995) also find the production smoothing hypothesis is rejected using aggregate data, and report evidence supportive of a variant of the production cost smoothing model. While the production smoothing hypothesis is rejected using aggregate industry data, there is evidence that this may be an artefact of aggregation. Using an establishment level data set, Schuh (1996) finds that the speeds of adjustment within the production smoothing model are an order of magnitude higher at establishment level than their counterparts estimated using industry data constructed by aggregating the establishment data.

Table 9.7
GMM estimates of the production cost smoothing model

Statistic	*First step*	*Second step*	*Iterated*	*Continuous updating*
$\hat{\lambda}_T$	0.295	0.330	0.332	0.256
$s.e.(\hat{\lambda}_T)$	0.075	0.082	0.083	0.066
$\hat{\phi}_T$	−0.205	−0.238	−0.223	−0.933
$s.e.(\hat{\phi}_T)$	0.586	0.557	0.554	0.801
$\hat{\rho}_T$	0.931	0.925	0.925	0.927
$s.e.(\hat{\rho}_T)$	0.022	0.022	0.022	0.024
J_T	.	3.671	4.136	3.629
$p-value$	.	0.160	0.126	0.163

Notes: $s.e.(.)$ denotes the standard error of the estimator calculated via (3.59), J_T denotes the overidentifying restrictions test and $p-value$ its associated p-value.

9.4 Stochastic Volatility Model of Exchange Rates

Section 1.3.5 describes the stochastic volatility model that has been used in a number studies of financial time series. In this section, we follow Melino and Turnbull (1990) and investigate whether this model can capture the time series properties of the daily U.S. dollar – Canadian dollar exchange rate.

Melino and Turnbull (1990) base their estimation on the following moment conditions:

$$
\begin{aligned}
E[w_t(\theta_0)] &= 0 \\
E[w_t^2(\theta_0)] - exp[2\mu_x + 2\sigma_x^2] &= 0 \\
E[w_t^3(\theta_0)] &= 0 \\
E[w_t^4(\theta_0)] - 3exp[4\mu_x + 8\sigma_x^2] &= 0 \\
E[|w_t(\theta_0)|] - (2/\pi)^{1/2} exp[\mu_x + 0.5\sigma_x^2] &= 0 \\
E[|w_t(\theta_0)|^3] - 2(2/\pi)^{1/2} exp[3\mu_x + 4.5\sigma_x^2] &= 0 \\
E[|w_t(\theta_0)|w_t(\theta_0)] &= 0 \\
E[w_t(\theta_0)w_{t-j}(\theta_0)] &= 0 \\
E[|w_t(\theta_0)w_{t-j}(\theta_0)|] - \ell_{1,j}(\theta_0) + \ell_{2,j}(\theta_0) &= 0 \\
E[|w_t(\theta_0)|w_{t-j}(\theta_0)] - m_j(\theta_0) &= 0 \\
E[w_t^2(\theta_0)w_{t-j}^2(\theta_0)] - n_j(\theta_0) &= 0
\end{aligned}
\tag{9.17}
$$

for $j = 1, 2, \ldots 10$ where

$$
w_t(\theta_0) = \frac{y(\tau_t) - \alpha_0 d_t - (1 + \beta_0 d_t) y(\tau_{t-1})}{[d_t \{y(\tau_{t-1})\}^{\gamma_0}]^{1/2}} \tag{9.18}
$$

and

$$
\begin{aligned}
\ell_{1,j}(\theta_0) &= (2/\pi)^{1/2} exp[2\mu_x + \sigma_x^2(1 + (1 + \eta_0 d)^j) - 0.5\rho_0^2\zeta_0^2 d(1 + \eta_0 d)^{2(j-1)}] \\
\ell_{2,j}(\theta_0) &= (2/\pi)^{1/2} \rho_0 \zeta_0 d^{1/2} (1 + \eta_0 d)^{j-1} (1 - 2\Phi(\rho_0 \zeta_0 d^{1/2} (1 + \eta_0 d)^{j-1}) \\
&\qquad \times exp[2\mu_x + \sigma_x^2(1 + (1 + \eta_0 d)^j)] \\
m_j(\theta_0) &= (2/\pi)^{1/2} \rho_0 \zeta_0 d^{1/2} (1 + \eta_0 d)^{j-1} exp[2\mu_x + \sigma_x^2(1 + (1 + \eta_0 d)^j)] \\
n_j(\theta_0) &= \{4\rho_0^2\zeta_0^2 d(1 + \eta_0 d)^{2(j-1)} + 1\} exp[4\mu_x + 4\sigma_x^2(1 + (1 + \eta_0 d)^j)] \\
\mu_x &= -\delta_0/\eta_0 \\
\sigma_x^2 &= \zeta_0^2 d/[1 - (1 + \eta_0 d)^2]
\end{aligned}
$$

and $\Phi(.)$ denotes the cumulative distribution function of a standard normal random variable. To simplify the estimations, Melino and Turnbull (1990) fix the value of γ_0 a priori and so the parameter vector is $\theta_0 = (\alpha_0, \beta_0, \delta_0, \eta_0, \zeta_0, \rho_0)$. Melino and Turnbull (1990) try three different values for γ_0, namely zero, one and two, and find the results are relatively insensitive to the choice. Throughout this section, we set $\gamma_0 = 1$ a priori.

The condition for local identification is not particularly instructive for this model, and so is omitted. However, inspection of (9.17) does reveal that not all the moment conditions have the potential to provide information about all the parameters. The following schematic summarizes the potential information content of the moment conditions with the latter being represented by the associated functions of the data:

$$
\left.
\begin{array}{c}
w_t(\theta_0) \\
w_t^3(\theta_0) \\
w_t(\theta_0)w_{t-j}(\theta_0) \\
|w_t(\theta_0)|w_t(\theta_0)
\end{array}
\right\} \longrightarrow \quad \alpha_0,\ \beta_0
$$

$$\left.\begin{array}{r} w_t^2(\theta_0) \\ w_t^4(\theta_0) \\ |w_t(\theta_0)| \\ |w_t^3(\theta_0)| \end{array}\right\} \longrightarrow \alpha_0,\ \beta_0,\ \delta_0,\ \eta_0,\ \zeta_0$$

$$\left.\begin{array}{r} |w_t(\theta_0)w_{t-j}(\theta_0)| \\ w_t^2(\theta_0)w_{t-j}^2(\theta_0) \end{array}\right\} \longrightarrow \alpha_0,\ \beta_0,\ \delta_0,\ \eta_0,\ \zeta_0,\ |\rho_0|$$

$$|w_t(\theta_0)|w_{t-j} \longrightarrow \alpha_0,\ \beta_0,\ \delta_0,\ \eta_0,\ \zeta_0,\ \rho_0$$

Clearly all of these moment conditions involve (α_0, β_0) and the majority involve $\delta_0, \eta_0, \zeta_0$. However, it is only the moment conditions involving $|w_t(\theta_0)w_{t-j}(\theta_0)|$, $w_t^2(\theta_0)w_{t-j}^2(\theta_0)$ and $|w_t(\theta_0)|w_{t-j}$, $j = 1, 2, \ldots$ that involve ρ_0, and of these, only the latter can reveal the sign of ρ_0.

Inspection of (9.17) also reveals that the set of moment conditions involves the expectations of absolute values of functions of $w_t(\theta_0)$. Such functions are non-differentiable at zero. This scenario is outside the theoretical framework developed earlier in the book beacuse the asymptotic theory is predicated on Assumption 3.5 that states $\partial f(v_t, \theta)/\partial\theta'$ exists for all $v \in \mathcal{V}$. Melino and Turnbull (1990) argue that since the necessary derivatives exist almost everywhere then it is reasonable to anticipate that the same asymptotic theory goes through with appropriate modification of the conditions. This argument is valid but its formal proof requires empirical process methods, and is not pursued here.[20] While this argument can be used to extend the asymptotic theory to cover this type of model, there still remains the secondary issue of what value to assign the derivative of $|w_t(\theta)|$, say, should $w_t(\theta) = 0$ in the computations. In fact, this is not a vacuous question since this does happen in this model with this data due to floating point accuracy of computer calculations. Melino and Turnbull (1990) report that they set this derivative to zero. One disadvantage of this approach is that the derivative can change dramatically in response to a slight perturbation of the data, and Vetzal (1992) finds that this can cause problems for the numerical optimization routine.[21] Therefore, Vetzal (1992) proposes using a sixth order polynomial to approximate the behaviour of the absolute value function in the neighbourhood of zero and then basing the derivative on the approximating polynomial. This approximation works as follows. The derivative of $|w|$ is replaced over the range $w \in [-\epsilon, \epsilon]$ by the derivative of

$$p(w) = a_0 + a_1 w + a_2 w^2 + a_3 w^3 + a_4 w^4 + a_5 w^5 + a_6 w^6 \qquad (9.19)$$

where the weights, $\{a_i\}$, are chosen to ensure that $p(w)$ mimics the behaviour of $|w|$ at $w = 0$ and the boundaries of the neighbourhood. Specifically, the constraints are: $p(0) = 0$, $p(\epsilon) = \epsilon$, $p'(\epsilon) = 1$, $p''(\epsilon) = 0$, $p(-\epsilon) = \epsilon$, $p'(-\epsilon) = -1$, $p''(-\epsilon) = 0$ – where $p'(w) = \partial p(w)/\partial w$ and $p''(w) = \partial^2 p(w)/\partial w^2$. Vetzal (1992) shows that these constraints imply $a_i = 0$ for $i = 0, 1, 3, 5$,

[20] However, see Andrews (1994) or Newey and McFadden (1994) [Section 7].

[21] For example, if $w_t(\theta)$ is -10^{-10} then the derivative of $|w_t(\theta)|$ is -1 but if $w_t(\theta)$ is 10^{-10} then the derivative of $|w_t(\theta)|$ is 1.

$a_2 = 15/(8\epsilon)$, $a_4 = -5/(4\epsilon^3)$ and $a_6 = 3/(8\epsilon^5)$. The advantage of this approach is that the derivative makes a smooth transition from -1 to 1 in the neighbourhood of zero. The disadvantage is that the derivative is actually incorrect for $w \in \{(-\epsilon, 0) \cup (0, \epsilon)\}$. This approximation is employed in the calculation of the derivatives of $|w_t(\theta)|$, $|w_t(\theta)w_{t-j}|$, and $|w_t(\theta)|w_{t-s}(\theta)$ for $s = 0, 1, \ldots 10$. Vetzal (1992) observes that $|w_t^3(\theta)|$ resembles a parabola, and is quite flat around zero. Consequently there is no need to use the approximation in this case and so the derivative of $|w_t^3(\theta)|$ is only modified by setting it to zero at $w_t(\theta) = 0$.[22]

We now turn to the estimation of the model. The data set consists of the daily spot Canada–U.S. exchange rate for the period January 2, 1975 to December 10, 1986, and is identical to the one used in Melino and Turnbull's (1990) study.[23] This gives a total of 3,010 observations on $w_t(\theta_0)$. For this data, the minimum time interval, d, is one. With regard to the specifics of the GMM estimation, all results reported in this section are calculated in the following way. An iterated GMM estimation is performed with the maximum number of iterations set at $I_{max} = 11$ and a convergence criterion set at $\epsilon_\theta = 10^6$.[24] In practice, convergence rarely occurred before this ceiling was met. Our experience was that convergence was not uniform, and some experimentation with larger values for I_{max} did not yield obvious improvement. We attribute this to the highly nonlinear nature of the moment conditions as a function of θ_0. The first step weighting matrix equal to $10^8 I_q$, and the long run variance is estimated using the prewhitened and recoloured HAC matrix with the bandwidth selected by Newey and West's (1994) data-based method and a Bartlett kernel, that is $\hat{S}_{SE}$ in (3.58). The starting values are the estimates reported by Melino and Turnbull (1990), that is $\bar{\theta}(0) = (0.042, -0.00054, -0.384, -0.091, 0.153, -0.110)$.[25]

We begin our empirical analysis of this model by considering the sensitivity of the results to the treatment of the derivative of the absolute value functions. Table 9.8 contains estimation results using four different neighbourhoods of the approximation: $[-\epsilon, \epsilon]$ with $\epsilon = 10^{-2}, 10^{-4}, 10^{-6}, 0$. In the first three cases, the derivative is based on (9.19) in the way described above. For $\epsilon = 0$, the neighbourhood collapses to the point zero and in this case the derivative of $|w_t(\theta)|$ is only modified by setting it to zero at $w_t(\theta) = 0$. It can be seen that both the estimates and standard errors exhibit some sensitivity to the approximation used in the calculation of the derivative. However, there are

[22] I am extremely grateful to Ken Vetzal for both drawing the issue discussed in this paragraph to my attention, and also for providing me with the material upon which the discussion is based.

[23] I am extremely grateful to Angelo Melino for providing me with both the data and also an unpublished appendix to their paper prepared by Ken Vetzal that derived the gradients of the moment conditions.

[24] See Section 3.6

[25] It should be noted that the specifics of our estimation differ from those in Melino and Turnbull (1990) in three respects: (i) they use an HAC estimator with $b_T = 50$; (ii) the first step weighting matrix is $\hat{S}_T^{-1}$ evaluated at a Method of Moments estimator based on a set of six undisclosed moments; (iii) the iterated estimator is iterated an unspecified number of steps.

clearly greater differences between the two step and iterated estimator for a given value of ϵ.

Table 9.8
Sensitivity of GMM estimates to derivative calculation
in stochastic volatility model

	$\epsilon = 0$		$\epsilon = 10^{-6}$		$\epsilon = 10^{-4}$		$\epsilon = 10^{-2}$	
	2-st.	iter.	2-st.	iter.	2-st.	iter.	2-st.	iter.
$\hat{\alpha}_T$	0.051	0.096	0.012	0.089	0.044	0.061	0.069	0.090
s.e.	0.090	0.060	0.100	0.067	0.097	0.083	0.072	0.061
$\hat{\beta}_T$	−0.001	−0.001	−0.000	−0.001	−0.001	−0.001	−0.000	−0.001
s.e.	0.001	0.001	0.001	0.001	0.001	0.001	0.003	0.001
$\hat{\delta}_T$	−0.334	−0.234	−0.356	−0.258	−0.368	−0.249	−0.315	−0.291
s.e.	0.050	0.106	0.043	0.107	0.042	0.061	0.062	0.138
$\hat{\eta}_T$	−0.079	−0.056	−0.085	−0.061	−0.088	−0.059	−0.075	−0.069
s.e.	0.012	0.025	0.010	0.025	0.010	0.014	0.018	0.033
$\hat{\zeta}_T$	0.163	0.114	0.172	0.121	0.176	0.123	0.156	0.127
s.e.	0.015	0.029	0.013	0.028	0.012	0.018	0.019	0.032
$\hat{\rho}_T$	−0.279	−0.179	−0.321	−0.192	−0.324	−0.310	−0.216	−0.138
s.e.	0.356	0.696	0.320	0.632	0.299	0.672	0.375	0.469
J_T	41.00	38.52	42.57	38.50	42.28	39.71	41.26	38.39
$p - value$	0.47	0.58	0.40	0.58	0.42	0.53	0.46	0.59

Notes: Estimation is based on (9.17) for $j = 1, 2, \ldots 10$. $2 - st$ and *iter.* denote the two step and iterated estimators respectively. *s.e.* denotes the standard error of the estimator on the line above. ϵ indexes the width of the neighbourhood of approximation for the absolute value function; see text.

given value of ϵ. Regardless of the permutation chosen, the results are qualitatively the same: the overidentifying restrictions test is insignificant at the 10% level; the estimated parameters of the volatility process are individually significantly different from zero at the 5% level, although the remaining estimates are insignificant. These findings are also reported by Melino and Turnbull (1990).[26]

Since the overidentifying restrictions test suggests the model is consistent with the data, we now consider the implications for the nature of the conditional variation of the exchange rate. In this context, two questions naturally arise. Is the volatility stochastic? – and if so, then for how long does the impact of shocks to the volatility process last? We now consider these issues in turn.

[26] The most striking difference between our results and those of Melino and Turnbull (1990) are in the estimate and standard error of $\hat{\rho}_T$. Vetzal (1997) estimates a stochastic volatility model for short term interest rates using different choices of covariance matrix estimator and finds that the standard errors can be very sensitive to the choice of covariance matrix estimator.

Within this model, volatility is stochastic if $\zeta_0 \neq 0$. A test of $H_0 : \zeta_0 = 0$ versus $H_1 : \zeta_0 \neq 0$ can be performed using any of the statistics described in Section 5.3. For simplicity, we use the Wald test in (5.43) which for this simple case reduces to

$$W_T \;=\; T\left[\frac{\hat{\zeta}_T}{s.e.(\hat{\zeta}_T)}\right]^2$$

Under the null, it follows from Theorem 5.7 that $W_T \xrightarrow{d} \chi^2_1$. Inspection of Table 9.8 reveals that this hypothesis is overwhelmingly rejected in every case. Therefore, volatility does indeed appear to be stochastic. The impact of the stochastic shocks on volatility is governed by the data generation process for the latent variable $x(\tau_t)$. This process is

$$ln[x(\tau_t)] \;=\; \delta_0 d \;+\; (1 \;+\; \eta_0 d) ln[x(\tau_t - d)] \;+\; \zeta_0 d^{1/2} u(\tau_t) \tag{9.20}$$

One way to assess this impact is to consider the impulse response function. Within this approach, the impact of a single shock on current and future values is calculated under the assumption that there are no future shocks, that is assuming all subsequent values of $u(\tau_t)$ are zero.[27] So for the following calculations, it is assumed that $u(\tau_{t_0}) = \bar{u}$ and $u(\tau_t) = 0$ for all $t > t_0$. Recalling that $d = 1$ for this example, it follows from (9.20) that the impact of this type of shock on $ln[x(\tau_t)]$ is given by $(1 + \eta_0)^{t-t_0}\zeta_0\bar{u}$, for $t \geq t_0$. A common summary statistic for impulse response functions is the half-life. The half-life of the shock is defined to be the interval of time it takes for the impact of the shock to be halved, that is t_{hl} such that

$$0.5\zeta_0\bar{u} \;=\; (1 \;+\; \eta_0)^{t_{hl}-t_0}\zeta_0\bar{u} \tag{9.21}$$

Without loss of generality, we can set $t_0 = 0$ and solve (9.21) to obtain

$$t_{hl} \;=\; \frac{ln[0.5]}{ln[1 \;+\; \eta_0]}$$

To illustrate the half-life, we focus attention on the case in which $\epsilon = 10^{-6}$. The associated values for the iterated estimator of $\hat{\eta}_T$ yield $t_{hl} = 11.013$ and so a half-life of approximately eleven days.

Taken at face value, the evidence appears to suggest that the stochastic volatility can capture the time series movements of this exchange rate. However, the inferences described above rest on asymptotic theory and the simulation results reported by Andersen and Sørensen (1996) cast doubt on the adequacy of this theory as an approximation to finite sample behaviour in these types of model.[28] These simulation results also indicate that the quality of the asymptotic approximation can be very sensitive to the choice of moments, and furthermore that certain moment conditions may be redundant. Therefore, we

[27] See Hamilton (1994) [Chapter 1] for further discussion.
[28] See Section 6.3.

estimate the model using different subsets of the moment conditions and compare the results using the moment selection criteria described in Section 7.3. To this end, we divided the moment conditions into five groups as follows (where once again the moments are represented by the associated functions of the data):

$$\begin{aligned} M1 &= \{w_t^i(\theta_0),\ i=1,2,3,4;\ \ |w_t^k(\theta_0)|,\ k=1,3\} \\ M2 &= \{w_t(\theta_0)w_{t-j}(\theta_0),\ j=1,2,\ldots 10\} \\ M3 &= \{|w_t(\theta_0)w_{t-j}(\theta_0)|,\ j=1,2,\ldots 10\} \\ M4 &= \{|w_t(\theta_0)|w_{t-j}(\theta_0),\ j=0,1,\ldots 10\} \\ M5 &= \{w_t^2(\theta_0)w_{t-j}^2(\theta_0)],\ j=1,2,\ldots 10 \end{aligned} \tag{9.22}$$

The model is estimated using four combinations of groups: $M1, M2, M3, M4$; $M1, M3, M4, M5$; $M1, M2, M4, M5$; $M2, M3, M4, M5$.[29] Table 9.9 reports the values for $MSC(c)$ and $RMSC(c)$ for these four subset models and the original model involving all five groups.[30] Both criteria are calculated using the penalty term associated with BIC given in (7.35). All statistics are calculated using the iterated estimators based on the derivative approximation with $\epsilon = 10^{-6}$. It can be recalled that the selected moment condition is the one that minimizes the criterion in question. Therefore, the use of $MSC(c)$ selects all the moments in (9.17), but the use of $RMSC(c)$ leads to the choice of the subset $M1, M3 - M5$. The latter suggests that the moment conditions $E[w_t(\theta_0)w_{t-j}(\theta_0)] = 0,\ j = 1, 2, \ldots 10$ are redundant given the moment conditions in $M1, M3 - M5$. Table 9.10 reports the iterated estimation results for this subset model. Interestingly, once the moments in $M2$ are omitted, $\hat{\alpha}_T$ is roughly two to three times larger than the corresponding estimate based on the full set of moments, and the estimates of both the exchange rate and volatility equations are now individually significant at the 5% level.

While it is useful to characterize the conditional variation of financial series, this is not an end in itself. Melino and Turnbull (1990) use their model to price currency options, but we do not pursue this issue here because it requires additional estimations. Parenthetically, we note that they find the stochastic volatility model performs better than a number of its competitors in this context. More generally, stochastic volatility models have been used to analyze a variety of financial time series. However, it should be noted that few of these studies employ the GMM approach described above. Following the simulation evidence reported by Andersen and Sørensen (1996), researchers have sought alternative ways to estimate these models. One such method involves moment based estimation using simulation techniques, and this is discussed in Chapter 10.

[29] Recall from our earlier discussion that $M4$ must be included to identify ρ_0.
[30] See Sections 7.3.1 and 7.3.2 respectively for discussion of MSC and RMSC.

Table 9.9
Moment selection criteria for the stochastic volatility model

Moments	q	$MSC(c)$	$RMSC(c)$
$M1 - M5$	47	−289.76	−5.41
$M2 - M5$	41	−249.16	−2.36
$M1, M3 - M5$	37	−228.39	−7.70
$M1, M2, M4, M5$	37	−214.66	−6.08
$M1 - M4$	37	−213.02	−5.49

Notes: $MSC(c)$ and $RMSC(c)$ are the moment selection criteria in (7.32) and (7.41) respectively.

Table 9.10
Iterated GMM estimates for the stochastic volatility model
based on moments $M1, M3 - M5$

$\hat{\alpha}_T$	$\hat{\beta}_T$	$\hat{\delta}_T$	$\hat{\eta}_T$	$\hat{\zeta}_T$	$\hat{\rho}_T$	J_T
0.193	−0.002	−0.266	−0.064	0.127	−0.307	19.81
0.077	0.001	0.067	0.016	0.020	0.619	0.94

Notes: The numbers below the parameter estimates are the standard errors and the number below J_T is the p-value of the test.

10

Related Methods of Estimation

In this chapter, we briefly review two other methods for exploiting moment conditions in estimation. Section 10.1 describes simulation based estimators known as Simulated Method of Moments, Indirect Inference and Efficient Method of Moments. Section 10.2 describes the method of Empirical Likelihood. The purpose of this discussion is to provide the intuition behind the methods in question and to explain their connections to GMM. References are provided for those readers interested in a rigourous analysis of the statistical properties of these estimators.

10.1 Simulation Based Estimation

Advances in computer technology have facilitated a growing interest in the use of simulation methods to estimate the parameters of economic models based on the information in population moment conditions. This approach is feasible in models where the data generation process is known apart from certain parameters, and so it is possible to generate artificial samples of data for different values of this parameter vector. The parameter estimator is then the value in the parameter space for which moments from the artificial data match the corresponding moments in the observed sample. There are two main variants of this approach with the difference depending on the choice of moments. If the moments are derived from the model of interest then this approach is known as *Simulated Method of Moments* or *Method of Simulated Moments*. If the moments are derived from some auxiliary model then the method is known as *Indirect Inference* or *Efficient Method of Moments* depending on the precise setting. In this section we provide a brief description of these methods and the asymptotic properties of the associated estimators. The focus here is on providing an intuitive introduction to the method and on relating them to GMM. The interested reader is refered to Carrasco and Florens (2002) for a recent survey

and Gourieroux and Monfort (1996) for a more comprehensive treatment.

10.1.1 Simulated Method of Moments

The basic idea behind Simulated Method of Moments (SMM) is best understood by considering a simple example. To this end, we return to a modified version of the simple example used to introduce the Method of Moments in Section 1.2. Suppose that $\{v_t\}$ is a sequence of scalar random variables that are independently and identically distributed. As in this earlier discussion, it is assumed that $E[v_t] = \mu_0$ but here it is assumed that $Var[v_t] = 1$ and that the distribution is normal. Given this specification, it is possible to generate an artificial sample for any value of the mean μ via

$$v_n(\mu) \;=\; \mu \;+\; e_n,\; n = 1, 2, \ldots N \tag{10.1}$$

where $\{e_n; n = 1, 2, \ldots N\}$ are random draws from the standard normal distribution. While the artificial sample $\{v_n(\mu);\, n = 1, 2, \ldots N\}$ can be generated for any choice of μ, there is only one such sample that comes from the same distribution as the data, namely the one for which $\mu = \mu_0$. In consequence, it follows that as both $T \to \infty$ and $N \to \infty$:

$$\begin{aligned} T^{-1}\sum_{t=1}^{T} v_t \;-\; N^{-1}\sum_{n=1}^{N} v_n(\mu_0) \;&\xrightarrow{p}\; 0 \\ T^{-1}\sum_{t=1}^{T} v_t \;-\; N^{-1}\sum_{n=1}^{N} v_n(\mu_*) \;&\xrightarrow{p}\; \mu_0 - \mu_* \neq 0,\ \text{for } \mu_0 \neq \mu_* \end{aligned} \tag{10.2}$$

SMM exploits the properties in (10.2) in the natural way: the SMM estimator of μ_0 is $\tilde{\mu}_T$, the value that satisfies $m^{(1)}_{N,T}(\tilde{\mu}_T) = 0$ where

$$m^{(1)}_{N,T}(\mu) \;=\; T^{-1}\sum_{t=1}^{T} v_t \;-\; N^{-1}\sum_{n=1}^{N} v_n(\mu) \tag{10.3}$$

In the above example, it is possible to simulate data to match the sample moment because the parameter is just identified. SMM can also be applied in overidentified models in a similar fashion to GMM. Continuing the example, suppose now that it is desired to base the estimation of μ_0 on the information in the first two moments. The resulting SMM estimator of μ_0 is the value of μ that minimizes

$$\begin{bmatrix} m^{(1)}_{N,T}(\mu) \\ m^{(2)}_{N,T}(\mu) \end{bmatrix}' W_T \begin{bmatrix} m^{(1)}_{N,T}(\mu) \\ m^{(2)}_{N,T}(\mu) \end{bmatrix} \tag{10.4}$$

where

$$m^{(2)}_{N,T}(\mu) \;=\; T^{-1}\sum_{t=1}^{T} v_t^2 \;-\; N^{-1}\sum_{n=1}^{N} v_n^2(\mu) \tag{10.5}$$

and W_T is a weighting matrix satisfying the conditions in Assumption 3.7.

Both these SMM estimators can also be considered as a special case of GMM. To elicit this connection, we return to the case where the estimator is based on the first moment alone. Some additional notation and structure is also needed. First, it is necessary to make some assumption about the relative magnitudes of N and T. We follow the common strategy in the literature of assuming that $N = kT$ for some fixed positive integer k satisfying $k > 1$. Notice that it is now possible to write the index of the generated random variable $v(\mu)$ as $n = k(t-1) + i$ where $i = 1, 2, \ldots k$ for each $t = 1, 2, \ldots T$. Secondly, using this re-indexing, we can now define

$$f_t(\mu) \;=\; v_t \;-\; k^{-1} \sum_{i=1}^{k} v_{k(t-1)+i}(\mu) \tag{10.6}$$

Finally, define $g_T(\mu) = T^{-1} \sum_{t=1}^{T} f_t(\mu)$. With these definitions, it can be seen that $m_{N,T}^{(1)}(\tilde{\mu}_T) = 0$ can be re-written as

$$g_T(\tilde{\mu}_T) = T^{-1} \sum_{t=1}^{T} f_t(\tilde{\mu}_T) \;=\; 0$$

and so the SMM estimator is the GMM estimator based on the population moment condition $E[f_t(\mu_0)] = 0$.

This interpretation means that the consistency and asymptotic normality of this SMM estimator can be deduced from Theorems 3.1 and 3.2.[1] So, if estimation is based on (10.6) then it follows from Theorem 3.2 that[2]

$$T^{1/2}(\tilde{\mu}_T \;-\; \mu_0) \;\xrightarrow{d}\; N(0, V_{SMM}) \tag{10.7}$$

where $V_{SMM} = \lim_{T\to\infty} Var[T^{-1/2} \sum_{t=1}^{T} f_t(\mu_0)]$. Assuming that the observed and generated samples are independent, it follows

$$\begin{aligned} V_{SMM} \;&=\; \lim_{T\to\infty} Var[T^{-1/2} \sum_{t=1}^{T} v_t] \\ &\quad + \lim_{T\to\infty} Var[T^{-1/2} \sum_{t=1}^{T} \{k^{-1} \sum_{i=1}^{k} v_{k(t-1)+i}(\mu_0)\}] \end{aligned} \tag{10.8}$$

Since both $\{v_t\}$ and $\{v_{k(t-1)+i}\}$ are i.i.d normal with mean μ_0 and variance 1, it follows from (10.8) that

$$V_{SMM} \;=\; (1 + k^{-1})$$

This variance bears a simple relationship to the asymptotic variance of the GMM (or MM) estimator based on $E[v_t] - \mu_0 = 0$. It follows directly from

[1] While true in this example, it is not generally true; see below.
[2] Note $p = q = 1$ and $\partial f_t(\mu)/\partial \mu = 1$.

Theorem 3.2 and our assumptions about v_t here that the asymptotic variance of this GMM estimator is $V_{GMM} = 1$ and so

$$V_{SMM} = (1 + k^{-1})V_{GMM} \tag{10.9}$$

Therefore, the SMM estimator is asymptotically less efficient than the GMM estimator that used the information in the population moment condition directly. However, note that the relative inefficiency decreases with $k = N/T$ and is zero in the limit as $k \to \infty$. The relationship in (10.9) is generic and recurs in more complicated models.

This efficiency ranking illustrates that there are typically no advantages to implementing SMM when conventional GMM is feasible. However, there are a number of circumstances in which conventional GMM is infeasible and SMM then becomes an attractive alternative. SMM was introduced into the econometric literature by McFadden (1989) in the context of discrete response models. Within this setting, estimation can be based on moment conditions involving the difference between the observed responses and the expected responses implied by the model. In some cases, the expected response may be difficult to express analytically or to compute via numerical integration, but may be readily obtained via simulation. SMM has been used to estimate a variety of microeconometric models, such as a multinomial logit model of transportation choice (McFadden and Train, 2000) and a bargaining model with asymmetric information for medical malpractice disputes (Sieg, 2000).[3] The method has also been applied to estimate a number of macroeconomic models, such as the consumption based asset pricing model (Heaton, 1995), a real business cycle model (Collard, Fève, Langot, and Perraudin, 2002) and exchange rates (Iannizzotto and Taylor, 1999).[4] Since these macroeconomic examples are closer in spirit to the types of model considered in Section 1.3 and Chapter 9, we explore one of these examples in more detail to illustrate why GMM may be infeasible but SMM can be implemented. Of the macroeconomic models listed above, the natural choice for such treatment is the consumption based asset pricing model studied by Heaton (1995) because this is a variation of the model described in Section 1.3.1 that has been used as our running empirical example.

Example: Heaton's (1995) Consumption Based Asset Pricing Model
Heaton (1995) studies a version of the consumption based asset pricing model in which the representative agent maximizes

$$E\left[\sum_{t=0}^{T} \delta_0^t \left\{\frac{s_t^{\gamma_0} - 1}{\gamma_0}\right\} \mid \Omega_t\right]$$

[3] See Gourieroux and Monfort (1996) for other examples.
[4] See Carrasco and Florens (2002) and Gourieroux and Monfort (1996) for additional references.

where δ_0 is the discount factor, Ω_t is the information set at time t,

$$s_t \;=\; \sum_{j=0}^{\infty} \alpha_j c_{t-j}$$

and c_t is consumption expenditure in period t. Notice that the functional form of the utility function is the same as in Hansen and Singleton's (1982) version of the model described in Section 1.3.1. The key difference is that the agent now derives utility in period t from a linear combination of current and past consumption expenditures, and so preferences are not time separable. This specification can be motivated by considering consumption to be a durable good and s_t as the service flow in period t from current and past consumption expenditures. It can be recalled from Section 1.3.1 that the completion of the model requires assumptions about the investment opportunities available to the agent. For simplicity here, it is assumed that the agent can invest in period t in a single asset at price p_t that matures in period $t+1$ with payoff r_{t+1}. In this case, the Euler equation is:

$$E\left[\delta_0 \frac{muc(t+1)}{muc(t)} \frac{r_{t+1}}{p_t} \,|\, \Omega_t\right] \;-\; 1 \;=\; 0 \tag{10.10}$$

where $muc(t)$ denotes the marginal expected discounted lifetime utility of consumption, that is

$$muc(t) \;=\; E\left[\sum_{j=0}^{\infty} \delta_0^j \alpha_j s_{t+j}^{\gamma_0 - 1} \,|\, \Omega_t\right]$$

Now consider the problem of how to estimate the model parameters based on the information in the Euler equation. As in Section 1.3.1, it is possible to use an iterated expectations argument to derive moment conditions based on the orthogonality of the function of the data in the Euler equations to any variable in the information set. While such moment conditions are the stepping stone to GMM estimation in the simpler version of the model in Section 1.3.1, such an estimation is infeasible here for the following two reasons:

- The Euler condition depends on an infinite sum. In practice, for GMM estimation, this sum would need to be truncated at order m, say. However, Heaton (1995) wishes to test for the existence of long run habit formation for which the $\{\alpha_j\}$ must be allowed to decay slowly, and this in turn suggests that m needs to be large. However, large values of m lead to a high order moving average structure in the error term of the Euler equation residual, and existing simulation evidence suggests that GMM estimation may be unreliable in these cases.[5]

- Heaton (1995) wishes to allow for the case in which the agent makes decisions weekly rather than monthly. However, the GMM approach in

[5] See Sections 3.5 and 6.3.

> Section 1.3.1 only works if data are observed at the frequency at which decisions are made, and aggregate consumption data are unavailable at higher frequencies than monthly.

Heaton (1995) shows that it is possible to circumvent both problems if the estimation is performed using Simulated Method of Moments. To describe his approach, it is necessary to introduce the following notation and structure. Assume that there is only one asset, as in our empirical implementation of the consumption based asset pricing model. Let t denote the week and assume that every month consists of four weeks. Define $\tilde{c}_t$, $\tilde{d}_t$ and $\tilde{p}_t$ to be weekly consumption, weekly dividend payments and the price of the asset price at the end of the week – so that the weekly asset return is $\tilde{r}_t = \tilde{p}_t + \tilde{d}_t$.

Heaton (1995) assumes that $Y_t = [ln(\tilde{c}_t/\tilde{c}_{t-1}), ln(\tilde{d}_t/\tilde{d}_{t-1})]'$ is generated by a subset VAR(12) model with a normally distributed error process. The parameters of this VAR are estimated via SMM and are chosen to ensure that the implied series for monthly consumption and dividend growth match the first moment and certain second moments of actual monthly consumption and dividend growth data. Once the model for Y_t is estimated, it is then possible to simulate values for Y_t. Given values for Y_t and the parameters, it is possible to solve (10.10) numerically for $\tilde{p}_t$. Therefore, the parameters of the Euler equation are estimated via SMM and are chosen to ensure that the implied series for monthly asset returns matches certain first and second moment properties of the actual monthly asset return data. $\diamond$

Pakes and Pollard (1989) provide an asymptotic theory for the SMM estimator in models where the data are i.i.d. but the function in the moment condition may be discontinuous, as would be the case, for instance, in discrete response models. Lee and Ingram (1991) and Duffie and Singleton (1993) provide a comparable asymptotic theory for time series models. These authors provide conditions under which the SMM estimator is consistent and asymptotically normal. These conditions are different from those employed in Chapter 3 for the corresponding of GMM estimators and their precise nature depends on both the structure of the moment condition and also on the way the data are simulated. We therefore do not pursue this theory further here but refer the reader to the aforementioned sources. Lee and Ingram (1991) also show that in overidentified models, the SMM minimand can form the basis for a model specification test along the same lines as the overidentifying restrictions test within the GMM framework.

10.1.2 Indirect Inference

Simulated Method of Moments involves simulating data from a model and choosing the parameters to match moments implied by the same model. However, in some cases, it can be desirable to simulate data from one model to match moments associated with some other model. This type of estimation is known as *Indirect Inference*, a terminology introduced by Gourieroux, Monfort, and

Renault (1993). It is common to refer to the model from which the simulations are generated as the *simulator*, and the "other" model from which the moments are constructed as the *auxilliary model*, and we adopt this terminology here. Indirect Inference is attractive in circumstances where it is possible to simulate data from the model of interest but the complexity of the model makes it impossible to estimate the parameters by conventional approaches, such as GMM. To illustrate the approach, we revisit the problem of estimating the parameters of a stochastic volatility model.

Example: Stochastic volatility model
For simplicity, consider the following special case of the stochastic volatility model described in Section 1.3.5,

$$y_t = \sqrt{x_t} e_t \tag{10.11}$$
$$ln(x_t) = \theta_{0,1} + \theta_{0,2} ln(x_{t-1}) + \theta_{0,3} u_t \tag{10.12}$$

where $(e_t, u_t)' \sim IN(0, I_2)$. It can be recalled from Section 1.3.5 that the model completely specifies the distribution, but that the structure of the model renders Maximum Likelihood estimation infeasible. However, since the data generation process is known, it is possible to simulate data from the model for a given value of θ. This opens the door to the possibility of a simulation based estimation. Let $\{y_n(\theta); n = 1, 2, \ldots N\}$ be a sample of simulated values for y from (10.11)-(10.12) given θ, and once again we set $N = kT$ for some positive integer k.

The key question is which moments to match? Gallant and Tauchen (1996) argue that the natural choice of moments are the score equations of a closely related model. Their argument applies more generally than our specific example and their reasoning is based on the following efficiency argument. First suppose that the auxilliary model encompasses the simulator; in this case the estimation is based on the true score equations and so it can be shown that the resulting estimators are asymptotically efficient provided $k \to \infty$. Now suppose that the auxilliary model does not encompass the simulator but is a good approximation in some sense; in this case then the resulting estimator can be thought of as being "nearly" asymptotically efficient. For the stochastic volatility model, a natural choice of auxilliary model is an alternative model for conditional variation such as the autoregressive conditional moving average (ARCH) model proposed by Engle (1982). The ARCH model of order d is given by,

$$y_t = h_t(\alpha_0) w_t$$

where

$$h_t^2(\alpha) = \alpha_1 + \sum_{i=1}^{d} \alpha_{i+1} w_{t-i}^2$$

and $\alpha = (\alpha_1, \alpha_2, \ldots \alpha_{d+1})'$. Under the assumption that $w_t \sim IN(0, 1)$, the

quasi-log likelihood function is tractable and the associated score equations are:[6]

$$\sum_{t=1}^{T} s[\hat{\alpha}_T;\, y_t] \;=\; 0 \tag{10.13}$$

where $\hat{\alpha}_T$ is the quasi maximum likelihood estimator of $\alpha_0 = (\alpha_{0,1}, \alpha_{0,2}, \ldots, \alpha_{0,d+1})'$,[7] and $s[.]$ is given by

$$s[\alpha;\, y_t] \;=\; z_t \left[\frac{y_t^2 - h_t^2(\alpha)}{2h_t^2(\alpha)} \right]$$

for $z_t = (1, y_{t-1}^2, y_{t-2}^2, \ldots y_{t-d}^2)'$.

The Indirect Inference estimator of θ_0 is defined to be

$$\tilde{\theta}_T \;=\; argmin_{\theta \in \Theta} \left\{ N^{-1} \sum_{n=1}^{N} s[\hat{\alpha}_T;\, y_n(\theta)] \right\}' W_T \left\{ N^{-1} \sum_{n=1}^{N} s[\hat{\alpha}_T;\, y_n(\theta)] \right\}$$

where W_T is a weighting matrix satisfying Assumption 3.7. For this estimation strategy to work, the auxilliary model must provide at least as many moment conditions as parameters to be estimated in the simulator. In this example, this restriction translates into the constraint that $d+1 \geq 3$. Notice that if $d+1 > 3$ then θ_0 is overidentified by the moments in the auxilliary score, and this is why the minimand is a quadratic form. $\diamond$

Gourieroux, Monfort, and Renault (1993) establish the consistency and asymptotic normality of the Indirect Inference estimator. They also show that there are a number of alternative ways of setting up the Indirect Inference minimand based on choosing the parameters of the simulator so that estimates from the auxilliary model based on the simulated data match the estimates from the auxilliary model based on the observed sample. However, since Gourieroux, Monfort, and Renault (1993) show these estimators are asymptotically equivalent to those based on matching the first derivative of the auxilliary model minimand, we do not discuss these alternative approaches here. Gourieroux, Monfort, and Renault (1993) also provide a number of specifications tests for models estimated by Indirect Inference. In particular, they show that in overidentified models, the Indirect Inference minimand can form the basis for a model specification test along the same lines as the overidentifying restrictions test within the GMM framework.

To conclude this sub-section, we return to the issue raised in the example above of which moments to match. Given Gallant and Tauchen's (1996) efficiency argument, it is clearly desirable that the auxilliary model provides the best possible approximation to the data generation process. Therefore Gallant and Tauchen (1996) recommend that the auxilliary model involve the

[6] See *inter alia* Hamilton (1994) [p.661].

[7] See Section 3.8.1.

specification that the probability density function of the data takes some flexible functional form capable of approximating a wide variety of distributions within the class of interest. For time series data such as the stochastic volatility example above, Gallant and Tauchen (1996) propose using a member of the class semi-non parametric (SNP) densities proposed by Gallant and Nychka (1987) to generate the score in the auxilliary model. SNP densities consist of a lead term, such as the conditional density associated with the ARCH(q) model, multiplied by an expansion involving Hermite polynomials. Gallant and Nychka (1987) show that such a density can recover a wide class of distributions as the order of the expansion tends to infinity. Therefore, this type of Indirect Inference estimator is often refered to as *Efficient Method of Moments (EMM)*. Andersen, Chung, and Sørensen (1999) provide simulation evidence on the use of EMM estimators in the stochastic volatility model, and find the method performs far better than the type of GMM estimator implemented in Section 9.4. EMM has been used to estimate the parameters of a variety of models; see Carrasco and Florens (2002) for a summary. Examples of its use to estimate stochastic volatility models are reported in Andersen and Lund (1997) and Gallant, Hsieh, and Tauchen (1997) where the applications are to interest rates and stock price indices respectively.

10.2 Empirical Likelihood

The method of Empirical Likelihood was introduced into the statistical literature by Owen (1988). In this and two subsequent articles in 1990 and 1991, Owen demonstrated that the method can be used to perform inference about aspects of a distribution or the parameters of a linear regression model. Empirical Likelihood has subsequently been extended to cases where it is desired to impose the restriction that the distribution satisfies a nonlinear moment condition indexed by an unknown parameter vector. As a result, this type of estimator is attracting an increasing amount of interest in econometrics. In this sub-section, we provide a brief introduction to the method and pay particular attention to highlighting its connections to GMM. More comprehensive treatments can be found in the recent survey article by Imbens (2002) or the recent monograph by Owen (2001).

To introduce the Empirical Likelihood approach to estimation, consider the situation in which the researcher observes a random sample of T observations on an i.i.d. random variable, v, and wishes to estimate its distribution. In the absence of any information about the form of the distribution function, the natural estimator is the empirical distribution function, that is the estimated probability of each sample value is $1/T$. This approach to estimation can be expressed as the outcome of an optimization as follows. Let $\tilde{v}_t$ denote the t^{th} outcome in the sample and π_t denote the probability that $v = \tilde{v}_t$. To be valid probabilities, it must follow that $0 \leq \pi_t \leq 1$ and $\sum_{t=1}^{T} \pi_t = 1$. The joint probability distribution function of the sample is then given by $\prod_{t=1}^{T} \pi_t$. If some parametric model had been assumed for π_t then this joint probability distribution function could

be treated as the likelihood for the data, and used as a basis for estimating the unknown parameters of the distribution. Suppose now that the same step is taken here even though there is no assumption about the form of the underlying distribution function. In this case, the "likelihood", $\prod_{t=1}^{T} \pi_t$, is treated as a function of the unknown probabilities $\{\pi_t\}$. This likelihood interpretation leads to the following method for estimating the probabilities:

$$\hat{\pi} \;=\; max_{\pi \in \Pi} \prod_{t=1}^{T} \pi_t \qquad \text{subject to } \sum_{t=1}^{T} \pi_t \;=\; 1 \tag{10.14}$$

where $\pi = (\pi_1, \pi_2, \ldots \pi_T)$ and $\Pi = [0,1]^T$. The resulting estimators can be shown to be $\pi_t = 1/T$ for all t – in other words, the distribution function is estimated by the empirical distribution function. The function $\prod_{t=1}^{T} \pi_t$ is known as the *empirical likelihood*.

This approach can be extended to incorporate information about the moments of the unknown distribution. Continuing the example, suppose now that it is known that the distribution satisfies $E[v] = 0$. The empirical distribution estimator for the probabilities does not ensure this restriction in the sample because in general

$$\sum_{t=1}^{T} \hat{\pi}_t \tilde{v}_t \;=\; T^{-1} \sum_{t=1}^{T} \tilde{v}_t \;\neq\; 0$$

However, it is possible to modify the constraint set so that this first moment condition is imposed. Specifically, suppose now that the empirical log likelihood is maximized subject to the twin constraints that the probabilities sum to one and the sample first moment is zero, that is

$$\bar{\pi} \;=\; max_{\pi \in \Pi} \prod_{t=1}^{T} ln[\pi_t] \qquad \text{subject to } \sum_{t=1}^{T} \pi_t \;=\; 1 \text{ and } \sum_{t=1}^{T} \pi_t \tilde{v}_t \;=\; 0$$

It can be shown that estimates of the probabilities are now $\bar{\pi}_t = (1 + \lambda \tilde{v}_t)^{-1}$ where λ is the Lagrange Multiplier associated with the constraint that $\sum_{t=1}^{T} \pi_t \tilde{v}_t = 0$.

This approach can also be extended to the types of population moment condition considered in the preceding chapters. Qin and Lawless (1994) derive the Empirical Likelihood estimators for the case in which it is desired to impose the moment restriction that the $(q \times 1)$ population moment condition $E[f(v, \theta_0)] = 0$ holds for some unknown $(p \times 1)$ parameter vector, θ_0. In this case, it is necessary to estimate both the unknown probabilities and also θ_0. The Empirical Likelihood estimators for this case are defined to be:

$$(\tilde{\pi}, \tilde{\theta}) \;=\; max_{\pi \in \Pi, \theta \in \Theta} \prod_{t=1}^{T} ln[\pi_t] \quad \text{subject to } \sum_{t=1}^{T} \pi_t \;=\; 1 \text{ and } \sum_{t=1}^{T} \pi_t f(\tilde{v}_t, \theta) = 0 \tag{10.15}$$

Qin and Lawless (1994) show that if $q = p$ then the resulting estimator of π_t is $\tilde{\pi}_t = 1/T$ for all t, and $\tilde{\theta}$ is the Method of Moments estimator based on

$E[f(v,\theta_0)] = 0$ – and so $\tilde{\theta}$ is also the GMM estimator in this case. However, if $q > p$ then the Empirical Likelihood and GMM estimators are different in general.

In comparison to GMM, it can be seen that Empirical Likelihood offers a very different way to exploit the information in population moment conditions. However, it turns out that the two estimators have the same asymptotic properties. Specifically, Qin and Lawless (1994) show that the Empirical Likelihood estimator is consistent for θ_0 and converges to the same limiting distribution as the GMM estimator calculated using the optimal weighting matrix.[8]

Just as within the GMM framework, one hypothesis of particular interest is whether the data are compatible with population moment condition. Imbens, Spady, and Johnson (1998) show that this hypothesis can be tested within the Empirical Likelihood framework using statistics derived from the Likelihood Ratio, Wald and Lagrange Multiplier testing principles. For brevity, we consider only the Likelihood Ratio type test here. This test compares the value of the empirical likelihood function at the restricted estimates, $\tilde{\pi}$, with the value at the unrestricted estimates, $\hat{\pi}$. The statistic is

$$LR - EL \;=\; 2\{ELLF_T(\bar{\pi}) \;-\; ELLF_T(\tilde{\pi})\}$$

where $ELLF_T(\pi) = \sum_{t=1}^{T} ln[\pi_t]$. Under the null hypothesis that $E[f(v,\theta_0)] = 0$, Imbens, Spady, and Johnson (1998) show that $LR - EL$ converges to a χ_{q-p} distribution. Notice this limiting distribution is exactly the same as the limiting distribution of the overidentifying restrictions test under the same null.[9]

In terms of these asymptotic properties, there is nothing to choose between Empirical Likelihood and GMM estimation. However, there is some recent evidence that the Empirical Likelihood estimators may exhibit better finite sample performance. Newey and Smith (2004) develop Nagar type approximations for the approximate bias of the Empirical Likelihood estimator along with those for the two step GMM and continuous updating estimators that are discussed in Section 6.2.2. They show that the approximate bias of the Empirical Likelihood estimator is equal to $T^{-1}B_I$, using the notation defined in Section 6.2.2. Therefore the Empirical Likelihood has fewer sources of bias than either the two step GMM or continuous updating GMM estimators. Given the existence of these types of bias, it is natural to consider using the bootstrap to provide more accurate finite sample inference. Since Empirical Likelihood approach generates probabilities for the data outcomes that are consistent with the moment condition, it provides a very computationally convenient way of generating articifial data consistent with the estimated model. Brown and Newey (2002) present an empirical likelihood based method for the bootstrapping with i.i.d. data and show that it is at least as efficient as the methods described in Section 8.1.

All the studies mentioned above address the behaviour of the Empirical Likelihood estimator in the context of i.i.d. data. Kitamura (1997) shows that

[8] See Section 3.4 and 3.6. Also see Chamberlain (1987) and Section 7.2.3.
[9] See Theorem 5.1 in Section 5.1.

if the data are generated by a stationary dynamic process the Empirical Likelihood estimator in (10.15) is no longer as asymptotically efficient as the two step GMM estimator. However, asymptotic equivalence can be restored if the sampling unit is taken to be blocks of observations along similar lines to the blocking schemes discussed in Section 8.1.2.1. The probabilities in the empirical likelihood are then interpreted as the probability that a particular block is sampled. Kitamura (1997) provides conditions under which the resulting Empirical Likelihood estimators have the same asymptotic distribution as the two step GMM estimator.

There have been a number of variations of the Empirical Likelihood estimator proposed in the literature. These variations involve replacing the Empirical Likelihood by some other function of the probabilities. It is not our purpose here to provide a review of this literature and instead the interested reader is refered to Imbens (2002). However, one particular extension is worth noting. Smith (1997) introduces the class of Generalized Empirical Likelihood estimators, and Newey and Smith (2004) show that this class includes the continuous updating GMM estimator but not the two step GMM estimator. This difference is the source of the estimators contrasting approximate bias properties discussed in Section 6.2.2.

Appendix A

Mixing Processes and Nonstationarity

This appendix provides a heuristic introduction to mixing processes followed by a brief summary of existing results on GMM in a nonstationary environment.

A.1 Mixing processes

As mentioned in the Section 3.4, if v_t is a mixing process then the dependence between v_t and v_{t-m} disappears as $m \to \infty$. To make this definition operational, it is necessary to make precise the notion of "dependence". Several approaches have been taken and each yields a different type of mixing process. In this appendix, we focus on so called *strong* or α–mixing processes. The discussion in this appendix relies heavily on Davidson (1994) [Chapters 13 and 14] to which the reader is refered for a rigorous treatment of this material and definitions of other types of mixing processes.

Although more accessible than ergodicity, the definition of an α-mixing process involves some sophisticated mathematical concepts. Below we build up to a formal definition in three steps. First, we introduce the measure of dependence in the context of two specific sets of elementary events. Secondly, this measure is extended to cover the dependence between two collection of sets. Finally, we show how this measure can be used to capture the dependence structure of a stochastic process.

To begin, it is useful to recall from elementary probability theory that if two events G and H are independent then $P(G \cap H) = P(G)P(H)$. If G and H are dependent then the converse is true, namely $P(G \cap H) - P(G)P(H) \neq 0$. These two basic properties suggest that

$$a'(G, H) = P(G \cap H) - P(G)P(H)$$

provides a reasonable starting place in our search for a measure of dependence between G and H. However, as it stands, $a'(G, H)$ has one unattractive feature.

The measure $a'(G,H)$ makes a distinction between cases $a'(G_1,H_1) = c$ and $a'(G_2,H_2) = -c$ whereas intuition suggests these two cases are both the same "distance" from independence. In other words, it is preferable to capture the dependence between G and H using

$$a(G,H) \;=\; |\,P(G\cap H) \;-\; P(G)P(H)\,| \tag{A.1}$$

We now turn to the extension of this measure of dependence to collection of sets. For conformity with what follows below, it is useful to give these collections of sets certain properties.

Definition A.1 σ-field
Let $\mathcal{F}$ be a collection of subsets of the set Ω. Then $\mathcal{F}$ is a σ-field (or σ-algebra) if it satisfies the following three conditions: (i) $\Omega \in \mathcal{F}$; (ii) if $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$; (iii) if $\{A_n,\ n = 1,2,\ldots\}$ is a sequence of sets in $\mathcal{F}$ then $\cup_{n=1}^{\infty} A_n \in \mathcal{F}$.

Now define $\mathcal{G}$ and $\mathcal{H}$ to be two σ-fields, and let $G \in \mathcal{G}$ and $H \in \mathcal{H}$. As we have seen G and H are independent if $a(G,H) = 0$. For $\mathcal{G}$ and $\mathcal{H}$ to be independent, it must be the case that $a(G,H) = 0$ for all $G \in \mathcal{G}$ and all $H \in \mathcal{H}$ or, more compactly, $\alpha(\mathcal{G},\mathcal{H}) = 0$ where

$$\alpha(\mathcal{G},\mathcal{H}) \;=\; sup_{G\in\mathcal{G},H\in\mathcal{H}}\, a(G,H) \tag{A.2}$$

Notice that if $\mathcal{G}$ and $\mathcal{H}$ are dependent then there must be some $G \in \mathcal{G}$ and $H \in \mathcal{H}$ for which $a(G,H) \neq 0$ and so $\alpha(\mathcal{G},\mathcal{H}) \neq 0$. As the notation suggests $\alpha(\mathcal{G},\mathcal{H})$ forms the basis for the measure of dependence in the definition of an α–mixing process.

The last step towards the formal definition of an α-mixing process involves the adaptation of the measure of dependence to time series. At this stage, it is necessary to introduce certain concepts relating to stochastic processes. These concepts are stated, but not explained, because such an explanation is beyond the scope of this book. It is hoped that the previous discussion is sufficient to convey the intuition behind the definition of a mixing process. We refer the interested reader to Davidson (1994) [Chapters 12–14] for a rigorous treatment of stochastic processes, dependence and mixing processes.

Consider the stochastic process $\{v_t(\omega)\}$ defined on the probability space $(\Omega,\mathcal{F},P)$. Let $\mathcal{F}_s^t$ be the smallest σ-field on which $(v_s, v_{s+1}, \ldots v_t)$ is measurable. Two particular σ fields are of interest here: $\mathcal{F}_{-\infty}^t$, which can be thought of as the "the information contained in the sequence up to date t"; and $\mathcal{F}_{t+m}^{\infty}$, which can be thought of as "the information contained in the sequence from $t+m$ onwards". From our previous discussion, we can capture the dependence between $\mathcal{F}_{-\infty}^t$ and $\mathcal{F}_{t+m}^{\infty}$ by

$$\alpha_m \;=\; \alpha(\mathcal{F}_{-\infty}^t, \mathcal{F}_{t+m}^{\infty})$$

With this background, we can finally present the definition towards which we have been working.

Definition A.2 α–mixing process
The sequence $\{v_t\}_{t=-\infty}^{\infty}$ is said to be α–mixing if $\lim_{m\to\infty} \alpha_m = 0$.

Therefore, an α-mixing process is one in which the dependence between two observations, α_m, decays to zero as $m \to \infty$. One implication of this definition is that the autocovariances of v_t exhibit a similar decay, that is[1]

$$Cov(v_t, v_{t-m}) \to 0 \qquad \text{as } m \to \infty \tag{A.3}$$

This particular property provides a useful glimpse into the difference between mixing and ergodicity, because the latter implies[2]

$$M^{-1} \sum_{m=1}^{M} Cov(v_t, v_{t-m}) \to 0 \qquad \text{as } M \to \infty \tag{A.4}$$

Therefore, mixing implies the autocovariances decay to zero as $m \to \infty$ but ergodicity implies the average of the first M autocovariances tends to zero as $M \to \infty$. The former implies the latter but not vice versa. From a comparison of (A.3) and (A.4), it can be seen that some generality is lost by moving from ergodic to mixing processes. However, since (A.3) is plausible for many economic series, it may be argued that not much is lost.

For the asymptotic analysis in Chapter 3 it is insufficient for the dependence simply to decay to zero, but it must do so at a particular rate.[3] The rate of decay has been captured in the literature using the concept of *size* of a mixing process.

Definition A.3 Size of a mixing process
v_t is an α–mixing process of size $-c_0$ if $\alpha_m = O(m^{-c})$ for some $c > c_0$.

This definition implies that the larger the size then the greater the dependence allowed in the series. Obviously it is desirable to allow for as much dependence as possible. However, the issue is not that simple, because dependence is just one feature of the series which must be restricted to permit conventional asymptotic analysis. It is also necessary to place restrictions on the existence of certain moments, and hence implicitly on the tail behaviour of v_t. As an illustration, consider the conditions imposed by Andrews (1991) to underpin his analysis of the asymptotic properties of covariance matrix estimators which are discussed in Section 3.5.3. He imposes the following two conditions: (i) v_t is α-mixing of size $-3\nu/(\nu-1)$; (ii) $E[\|v_t\|^{4\nu}] < \infty$. Inspection reveals that as ν increases the size increases and so the degree of dependence allowed increases. However, at the same time, an increase in ν increases the order up to which the moments of v_t must exist. The latter is implicitly a restriction on the tail behaviour of the

[1] See Davidson (1994) [p.203].
[2] See Davidson (1994) [p.201]. Note that this condition is sufficient but not necessary for ergodicity unless $\{v_t\}$ is a Gaussian process in which case it is a necessary condition as well.
[3] This includes the Weak Law of Large Numbers, Central Limit Theorem and also the consistency of the covariance matrix estimators discussed in Section 3.5.

distribution of v_t.[4] So there is a tension between these two aspects of v_t in the context of asymptotic analysis.

In the course of the analysis in Chapter 3, it is necessary to apply limit theorems to various functions of v_t. One particularly attractive feature of α-mixing processes is that certain functions of them are also α-mixing. Specifically, if v_t is an α-mixing of size $-c_0$ then $Y_t = g(v_t, v_{t-1}, \ldots v_{t-\tau})$ is also an be α-mixing of size $-c_0$ *provided* τ is finite. The reason for this proviso is readily understood. If τ is infinite, then both Y_t and Y_{t-m} are functions of $\{v_{t-n},\ n > m\}$ regardless of how large m becomes, and so Y_t cannot be an α-mixing process in general. However, it turns out that this situation can be circumvented by restricting $g(.)$ to be *near epoch dependent*. Essentially, this condition restricts $g(.)$ so that the dependence in Y_t decays sufficiently fast to allow the derivation of Laws of Large Numbers and Central Limit Theorems. We do not pursue this topic here but instead refer the interested reader to Davidson (1994) [Chapter 17].

A.2 Nonstationarity

If stationarity is relaxed then it becomes necessary to make an explicit assumption about the nature of the nonstationarity. Three approaches have been taken in the literature: (i) nonstationary mixing processes; (ii) deterministic trends; (iii) unit root processes. We now provide a brief summary of the available results in each case.

- *Mixing processes:* The concept of a mixing process can be extended to nonstationary processes by setting $\alpha_m = \sup_t \alpha(\mathcal{F}^t_{-\infty}, \mathcal{F}^\infty_{t+m})$. Subject to certain restrictions, Weak Laws of Large Numbers and Central Limit Theorems can be developed for nonstationary mixing processes; see Gallant and White (1988) or Pötscher and Prucha (1997). It is then possible to establish the consistency and asymptotic normality of the estimator within this environment; see Gallant (1987), Gallant and White (1988) and Pötscher and Prucha (1997).

- *Deterministic trends*: Andrews and McDermott (1995) consider the case in which the data are generated by $v_t = v(d_t, w_t)$ where d_t is a deterministic trend and w_t is a stationary process. They provide conditions under which the GMM estimator is consistent and asymptotically normal within this set–up.

- *Unit root processes:* If v_t contains unit root processes then the limiting distribution theory is non-standard. To date, progress has only been made

[4] For example, consider the case in which v_t is scalar and possesses a Student t distribution with δ degrees of freedom. As the parameter δ decreases the tails of the distribution become "thicker" and this has implications for the moments. Specifically, the moments of the t distribution only exist up to order δ. Therefore the restriction $E[\|v_t\|^{4\nu}] = E[v_t^{4\nu}] < \infty$ implies $\delta \geq 4\nu$, and thereby implicitly places a restriction on the thickness of the tails of the distribution. See Johnson and Kotz (1970) [Chapter 27] for a discussion of the properties of the t distribution.

for the particular case of IV estimation of linear models. Hall (1987*b*) and Pantula and Hall (1991) analyze the behaviour of unit root tests based on IV estimators. Phillips and Hansen (1990) and Kitamura and Phillips (1997) analyze the limiting behaviour of IV estimators in a multivariate setting.

Bibliography

Ahn, S. C. (1995). 'Model specification testing based on root-T consistent estimators', Discussion paper, Department of Economics, Arizona State University, Tempe, AZ, U.S.A.

——— , Good, D. H., and Sickles, R. C. (2000). 'Estimation of long run inefficiency levels: a dynamic frontier approach', *Econometric Reviews*, 19: 461–92.

Akaike, H. (1973). 'Information theory and an extension of the maximum likelihood principle', in B. N. Petrov and F. Csaki (eds.), *Second International Symposium on Information Theory*, pp. 267–81. Akademia Kiado, Budapest, Hungary.

——— (1974). 'A new look at statistical model identification', *IEEE Transactions on Automatic Control*, AC-19(6): 716–23.

Aldrich, J. (1993). 'Reiersøl, Geary and the idea of instrumental variables', *The Economic and Social Review*, 24: 247–73.

Altonji, J. G., and Segal, L. M. (1996). 'Small sample bias in GMM estimation of covariance structures', *Journal of Business and Economic Statistics*, 14: 353–66.

Amemiya, T. (1974). 'The nonlinear two-stage least squares estimator', *Journal of Econometrics*, 2: 105–10.

——— (1977). 'The maximum likelihood and nonlinear three stage least squares estimator in the general nonlinear simultaneous equations model', *Econometrica*, 45: 955–68.

Andersen, T. G., Chung, H.-J., and Sørensen, B. E. (1999). 'Efficient method of moments estimation of a stochastic volatility model: a Monte Carlo study', *Journal of Econometrics*, 91: 61–87.

——— and Lund, J. (1997). 'Estimating continuous time stochastic volatility models of the short term interest rate', *Journal of Econometrics*, 77: 343–77.

——— , and Sørensen, B. E. (1996). 'GMM estimation of a stochastic volatility model: a Monte Carlo study', *Journal of Business and Economic Statistics*, 14: 328–52.

Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis*. Wiley, NY, U.S.A., 2nd edn.

——— (1994). *The Statistical Analysis of Time Series*. Wiley, New York, NY, U.S.A.

Anderson, T. W., and Rubin, H. (1949). 'Estimation of the parameters of a single equation in a complete system of stochastic equations', *Annals of Mathematical Statistics*, 20: 46–63.

——— and Sawa, T. (1973). 'Distributions of estimates of coefficients of a single equation in a simultaneous system and their asymptotic expansions', *Econometrica*, 41: 683–714.

——— and ——— (1979). 'Evaluation of the distribution of the two–stage least squares estimate', *Econometrica*, 47: 163–82.

Andrews, D. W. K. (1995). 'Admissibility of the Likelihood Ratio test when a nuisance parameter is present only under the alternative', *Annals of Statistics*, 23: 1609–29.

——— (1987). 'Asymptotic results for Generalized Wald Tests', *Econometric Theory*, 3: 348–58.

——— (1991). 'Heteroscedasticity and autocorrelation consistent covariance matrix estimation', *Econometrica*, 59: 817–58.

——— (1993). 'Tests for parameter instability and structural change with unknown change point', *Econometrica*, 61: 821–56.

——— (1994). 'Empirical Process Methods in Econometrics', in R. F. Engle and D. L. McFadden (eds.), *Handbook of Econometrics*, vol. 4, pp.2247–94. Elsevier Science Publishers, Amsterdam, The Netherlands.

——— (1997). 'A stopping rule for the computation of Generalized Method of Moments Estimators', *Econometrica*, 65: 913–32.

——— (1999). 'Consistent moment selection procedures for Generalized Method of Moments estimation', *Econometrica*, 67: 543–64.

——— (2000). 'Consistent moment selection procedures for GMM estimation: strong consistency and simulation results', Discussion paper, Cowles Foundation for Economics, Yale University, New Haven, CT, U.S.A.

——— (2002*a*). 'Generalized method of moments estimation when a parameter is on a boundary', *Journal of Business and Economic Statistics*, 20: 530–44.

——— (2002*b*). 'Higher order improvements of a computationally attractive k-step bootstrap for extremum estimators', *Econometrica*, 70: 119–62.

——— (2003). 'Tests for parameter instability and structural change with unknown change point: a corrigendum', *Econometrica*, 71: 395–98.

——— and Buchinsky, M. (2000). 'A three step method for choosing the number of bootstrap replications', *Econometrica*, 68: 23–52.

——— and Fair, R. (1988). 'Inference in econometric models with structural change', *Review of Economic Studies*, 55: 615–40.

——— and McDermott, C. J. (1995). 'Nonlinear econometric models with deterministically trending variables', *Review of Economic Studies*, 62: 343–60.

——— and Monahan, J. C. (1992). 'An improved heteroscedasticity and autocorrelation consistent covariance matrix', *Econometrica*, 60: 953–66.

——— and Ploberger, W. (1994). 'Optimal tests when a nuisance parameter is present only under the alternative', *Econometrica*, 62: 1383–414.

Angrist, J. D. (2001). 'Estimation of limited dependent variable models with dummy endogenous regressors: simple strategies for empirical practice', *Journal of Business and Economic Statistics*, 19: 2–16.

——— and Krueger, A. B. (1992). 'The effect of age at school entry on educational attainment: an application of instrumental variables with moments from two samples', *Journal of the American Statistical Association*, 87: 328–36.

Apostol, T. (1974). *Mathematical Analysis*. Addison-Wesley, Reading, MA, U.S.A., 2nd. edn.

Arellano, M. (2002). 'Sargan's instrumental variables estimation and the generalized method of moments', *Journal of Business and Economic Statistics*, 20: 450–9.

——— and Bond, S. (1991). 'Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations', *Review of Economic Studies*, 58: 277–97.

Atkinson, S. E., Cornwell, C., and Honerkamp, O. (2003). 'Measuring and decomposing productivity change: stochastic distance function estimation versus data envelopment analysis', *Journal of Business and Economic Statistics*, 21: 284–94.

Attanasio, O., and Browning, M. (1995). 'Consumption over the life cycle and over the business cycle', *American Economic Review*, 85: 1118–37.

——— and Weber, G. (1995). 'Is consumption growth consistent with intertemporal optimization? Evidence from the Consumer Expenditure Survey', *Journal of Political Economy*, 103: 1121–57.

Backus, D., Gregory, A. W., and Telmer, C. (1993). 'Accounting for forward rates in markets for foreign currency', *Journal of Finance*, 48: 1887–908.

Bai, J., and Perron, P. (1998). 'Estimating and testing linear models with multiple structural changes', *Econometrica*, 66: 47–78.

Baltagi, B. H. (2001). *Econometric Analysis of Panel Data*. John Wiley and Sons, Chichester, U.K.

Bansal, R., Hsieh, D. A., and Viswanathan, S. (1993). 'A new approach to international arbitrage pricing', *Journal of Finance*, 48: 1719–48.

——— and Viswanathan, S. (1993). 'No arbitrage and arbitrage pricing: a new approach', *Journal of Finance*, 48: 1231–62.

Barankin, E., and Gurland, J. (1951). 'On asymptotically normal efficient estimators: I', *University of California Publications in Statistics*, 1: 86–130.

Basmann, R. L. (1961). 'A note on the exact finite sample frequency functions of generalized classical linear estimators in two leading overidentified cases', *Journal of the American Statistical Association*, 56: 619–36.

——— (1963). 'A note on the exact finite sample frequency functions of generalized classical linear estimators in a leading three equation case', *Journal of the American Statistical Association*, 58: 161–71.

Bates, C. E., and White, H. (1990). 'Efficient instrumental variables estimation of systems of implicit heterogeneous nonlinear dynamic equations with non-spherical errors', in W. Barnett, E. Berndt, and H. White (eds.), *Dynamic Econometric Modelling*, pp. 3–25. Cambridge University Press, New York, NY, U.S.A.

Bekaert, G., and Hodrick, R. J. (2001). 'Expectations hypotheses tests', *Journal of Finance*, 56: 1357–93.

——— and ——— (1992). 'Characterizing predictable components in excess returns on equity and foreign exchange markets', *Journal of Finance*, 47: 467–509.

——— and Urias, M. S. (1996). 'Diversification, integration and emerging market closed end funds', *Journal of Finance*, 51: 835–69.

Bekker, P. A. (1994). 'Alternative approximations to the distributions of instrumental variables estimators', *Econometrica*, 62: 657–81.

Bera, A., and Bilias, Y. (2002). 'MM, ME, EL, EF and GMM approaches to estimation: a synthesis', *Journal of Econometrics*, 107: 51–86.

Bernstein, J. I. (1994). 'Exports, imports and productivity growth: with an application to the Canadian softwood lumber industry', *Review of Economics and Statistics*, 76: 291–301.

Berry, S., Levinsohn, J., and Pakes, A. (1995). 'Automobile prices in market equilibrium', *Econometrica*, 63: 841–90.

Bessembinder, H., and Chan, K. (1992). 'Time-varying risk premia and forecastable returns in futures markets', *Journal of Financial Economics*, 32: 169–93.

——— , ——— , and Seguin, P. J. (1996). 'An empirical examination of information, differences of opinion and trading activity', *Journal of Financial Economics*, 40: 105–34.

Biasis, B., Hillion, P., and Spatt, C. (1999). 'Price discovery and learning during the preopening period in the Paris bourse', *Journal of Political Economy*, 107: 1218–48.

Bils, M., and Kahn, J. A. (2000). 'What inventory behaviour tells us about business cycles', *American Economic Review*, 90: 458–81.

Bjornson, B., and Carter, C. A. (1997). 'New evidence on agricultural commodity return performance under time-varying risk', *American Journal of Agricultural Economics*, 79: 918–30.

Blinder, A. S. (1986). 'More on the speed of adjustment in inventory models', *Journal of Money, Credit and Banking*, 18: 355–65.

——— and Maccini, L. (1991). 'The resurgence of inventory research: what have we learned?', *Journal of Economic Surveys*, 5: 291–328.

Blundell, R., and Bond, S. (2000). 'GMM estimation with persistent panel data: an application to production functions', *Econometric Reviews*, 19: 321–40.

——— , Browning, M., and Meghir, C. (1994). 'Consumer demand and the life cycle allocation of household expenditures', *Review of Economic Studies*, 61: 57–80.

——— , Griffith, R., and Vanreenen, J. (1995). 'Dynamic count data models of technological innovation', *Economic Journal*, 105: 333–44.

——— , Pashardes, P., and Weber, G. (1993). 'What do we learn about consumer demand patterns from micro data', *American Economic Review*, 83: 570–97.

Bodurtha, J. N., and Mark, N. C. (1991). 'Testing the CAPM with time-varying risks and returns', *Journal of Finance*, 46: 1485–505.

Boldrin, M., Christiano, L. J., and Fisher, J. D. M. (2001). 'Habit persistence, asset returns and the business cycle', *American Economic Review*, 91: 149–66.

Bollerslev, T., Chou, R. Y., and Kroner, K. F. (1992). 'ARCH modelling in finance: theory and empirical evidence', *Journal of Econometrics*, 52: 5–59.

——— , Engle, R. F., and Nelson, D. B. (1994). 'ARCH Models', in R. F. Engle and D. L. McFadden (eds.), *Handbook of Econometrics*, vol. 4, pp. 2959–3038. Elsevier Science Publishers, Amsterdam, The Netherlands.

Bond, S., and Meghir, C. (1994). 'Dynamic investment models and the firm's financial policy', *Review of Economic Studies*, 61: 197–222.

Bonham, C., and Cohen, R. (1995). 'Testing the rationality of forecasts: comment', *American Economic Review*, 85: 284–9.

——— and ——— (2001). 'To aggregate, pool, or neither: testing the rational expectations hypothesis using survey data', *Journal of Business and Economic Statistics*, 19: 278–91.

Bound, J., Jaeger, D. A., and Baker, R. (1995). 'Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak', *Journal of the American Statistical Association*, 90: 443–50.

Bourgeon, J. M., and Le Roux, Y. (2001). 'Traders' bidding strategies on European grain export refunds: an analysis with affiliated signals', *American Journal of Agricultural Economics*, 83: 563–75.

Bowden, R. J., and Turkington, D. A. (1984). *Instrumental Variables*. Cambridge University Press, Cambridge, U.K.

Bowman, K. O., and Shenton, L. R. (1975). 'Omnibus contours for departures from normality based on b_1 and b_2', *Biometrika*, 62: 243–250.

Box, G. E. P., and Jenkins, G. M. (1976). *Time Series Analysis: Forecasting and Control*. Prentice Hall, Englewood Cliffs, NJ, U.S.A.

Braun, R. A. (1994). 'Tax disturbances and real economic activity in the postwar United States', *Journal of Monetary Economics*, 33: 441–62.

Breusch, T., Qian, H., Schmidt, P., and Wyhowski, D. (1999). 'Redundancy of moment conditions', *Journal of Econometrics*, 91: 89–111.

Brown, B. W., and Newey, W. K. (2002). 'Generalized method of moments, efficient bootstrapping and improved inference', *Journal of Business and Economic Statistics*, 20: 507–17.

Brown, R. (1828). 'A brief account of the microscopical observations made in the months of June, July and August, 1827, on the paricles contained in the pollen of plants; and on the general existence of active molecules in organic and inorganic bodies', *Philosophical Magazine (2nd. series)*, 4: 161–73.

Bühlmann, P., and Künsch, H. R. (1996). 'Block selection in the bootstrap for time series', Discussion paper, unpublished mimeo.

Burguette, J. F., Gallant, A. R., and Souza, G. (1982). 'On unification of the asymptotic theory of nonlinear econometric models', *Econometric Reviews*, 1: 151–90.

Burnside, C., and Eichenbaum, M. (1996). 'Small sample properties of GMM based Wald tests', *Journal of Business and Economic Statistics*, 14: 294–308.

——— , ——— , and Rebelo, S. (1993). 'Labor hoarding and the business cycle', *Journal of Political Economy*, 101: 245–73.

Buse, A. (1992). 'The bias of instrumental variable estimators', *Econometrica*, 60: 173–80.

Campbell, J. Y. (1996). 'Understanding risk and return', *Journal of Political Economy*, 104: 298–345.

——— and Mankiw, N. G. (1990). 'Permanent income, current income and consumption', *Journal of Business and Economic Statistics*, 8: 265–80.

Carlstein, E. (1986). 'The use of subseries methods for estimating the variance of a general statistic from a stationary time series', *Annals of Statistics*, 14: 1171–9.

Carrasco, M., and Florens, J. P. (2000). 'Generalization of GMM to a continuum of moments conditions', *Econometric Theory*, 16: 797–834.

——— (2002). 'Simulation based method of moments and efficiency', *Journal of Business and Economic Statistics*, 20: 482–92.

Caselli, F., Esquivel, G., and Lefort, F. (1996). 'Reopening the convergence debate: a new look at cross-country growth empirics', *Journal of Economic Growth*, 1: 363–89.

Cecchetti, S. G., Lam, P., and Mark, N. C. (1993). 'The equity premium and the risk free rate', *Journal of Monetary Economics*, 31: 21–45.

Chamberlain, G. (1987). 'Asymptotic efficiency in estimation with conditional moment restrictions', *Journal of Econometrics*, 34: 305–34.

——— and Rothschild, M. (1983). 'Arbitrage, factor structure and mean variance analysis on large asset markets', *Econometrica*, 51: 1281–304.

Chan, K. C., Karolyi, G. A., Longstaff, F. A., and Sanders, A. B. (1992). 'An empirical comparison of alternative models of the short term interest rate', *Journal of Finance*, 47: 1209–27.

Chavas, J. P., and Thomas, A. (1999). 'A dynamic analysis of land prices', *American Journal of Agricultural Economics*, 81: 772–84.

Chen, Z., and Knez, P. (1996). 'Portfolio performance measurement: theory and applications', *Review of Financial Studies*, 9: 511–55.

Chesher, A. (1984). 'Testing for neglected heterogeneity', *Econometrica*, 52: 865–72.

Chirinko, R. S., and Schaller, H. (1996). 'Bubbles, fundamentals and investment: a multiple equation testing strategy', *Journal of Monetary Economics*, 38: 47–76.

——— and ——— (2001). 'Business fixed investment and "bubbles": the Japanese case', *American Economic Review*, 91: 663–80.

Christiano, L. J., and den Haan, W. J. (1996). 'Small sample properties of GMM for business cycle analysis', *Journal of Business and Economic Statistics*, 14: 309–27.

——— and Eichenbaum, M. (1992). 'Current real business cycle theories and aggregate labor market fluctuations', *American Economic Review*, 82: 430–50.

Clarida, R., Gali, J., and Gertler, M. (2000). 'Monetary policy rules and macroeconomic stability: evidence and some theory', *Quarterly Journal of Economics*, pp. 147–80.

Clark, T. E. (1996). 'Small sample properties of estimators of nonlinear models of covariance structure', *Journal of Business and Economic Statistics*, 14: 367–73.

Cochrane, J. H. (1996). 'A cross-sectional test of an investment asset pricing model', *Journal of Political Economy*, 104: 572–621.

Collard, F., Fève, P., Langot, F., and Perraudin, C. (2002). 'A structural model of US job flows', *Journal of Applied Econometrics*, 17: 197–223.

Considine, T. J., and Heo, E. (2000). 'Price and inventory dynamics in petroleum product markets', *Energy Economics*, 22: 527–48.

Cox, D. R., and Hinckley, D. V. (1974). *Theoretical Statistics*. Chapman and Hall, London, U.K.

Cragg, J. G., and Donald, S. G. (1993). 'Testing identifiability and specification in instrumental variables', *Econometric Theory*, 9: 222–40.

Critchley, F., Marriott, P., and Salmon, M. (1996). 'On the differential geometry of the Wald test with nonlinear restriction', *Econometrica*, 64: 1213–22.

Cumby, R. E., and Huizinga, J. (1992). 'Investigating the correlation of unobserved expectations', *Journal of Monetary Economics*, 30: 217–53.

Cushing, M. J., and Ackert, L. F. (1994). 'Interest innovations and the volatility of long term bond yields', *Journal of Money, Credit and Banking*, 26: 203–17.

Davidson, J. (1994). *Stochastic Limit Theory*. Oxford University Press, Oxford, U.K.

Davidson, R., and MacKinnon, J. G. (1993). *Estimation and Inference in Econometrics*. Oxford University Press, Oxford, U.K.

——— and ——— (1999). 'Bootstrap testing in nonlinear models', *International Economic Review*, 40: 487–508.

Deaton, A., and Laroque, G. (1992). 'On the behaviour of commodity prices', *Review of Economic Studies*, 59: 1–23.

den Haan, W. J., and Levin, A. (1996). 'Inferences from parametric and non–parametric covariance matrix estimation procedures', Discussion paper, International Finance Division, Board of Governors of the Federal Reserve System, Washington, DC, U.S.A.

——— and ——— (1997). 'A practioner's guide to robust covariance matrix estimation', in G. Maddala and C. Rao (eds.), *Handbook of Statistics, volume 15*, pp. 309–27. Elsevier, Amsterdam, The Netherlands.

de la Croix, D., and Urbain, J.-P. (1998). 'Intertemporal substitution in import demand and habit formation', *Journal of the Applied Econometrics*, 13: 589–612.

Dhrymes, P. J. (1984). *Mathematics for Econometrics*. Springer Verlag, New York, NY, U.S.A., 2nd edn.

Diba, B. T., and Oh, S. (1991). 'Money, output and the expected real interest rate', *Review of Economics and Statistics*, 73: 10–17.

Donald, S. G., and Newey, W. K. (2001). 'Choosing the number of instruments', *Econometrica*, 69: 1161–92.

Doorn, D. (2003). 'Three essays on trend analysis and misspecification in structural econometric models', Ph.D. thesis, Department of Economics, North Carolina State University, Raleigh, NC, U.S.A.

Duffie, D., and Singleton, K. J. (1993). 'Simulated moments estimation of Markov models of asset prices', *Econometrica*, 61: 929–52.

Dufour, J.-M. (1997). 'Some impossibility theorems in econometrics with applications to structural and dynamic models', *Econometrica*, 65: 1365–87.

——— , Ghysels, E., and Hall, A. R. (1994). 'Generalized predictive tests and structural change analysis in econometrics', *International Economic Review*, 35: 199–229.

Dumas, B., and Solnik, B. (1995). 'The world price of foreign exchange risk', *Journal of Finance*, 50: 445–80.

Dunn, K., and Singleton, K. J. (1986). 'Modelling the term structure of interest rates under nonseparable utility and durable goods', *Journal of Financial Economics*, 17: 27–55.

Durbin, J. (1954). 'Errors in variables', *Review of Institute of International Statistics*, 22: 23–31.

Durlauf, S. N., and Maccini, L. J. (1995). 'Measuring noise in inventory models', *Journal of Monetary Economics*, 36: 65–89.

Dutkowsky, D. H. (1993). 'Dynamic implicit cost and discount window borrowing', *Journal of Monetary Economics*, 32: 105–20.

Dynan, K. E. (2000). 'Habit formation in consumer preferences: evidence from panel data', *American Economic Review*, 90: 391–406.

Eckstein, Z., and Leiderman, L. (1992). 'Seignorage and the welfare cost of inflation', *Journal of Monetary Economics*, 29: 389–410.

Efron, B. (1979). 'Bootstrap methods: another look at the jackknife', *Annals of Statistics*, 7: 1–26.

Eichenbaum, M. (1989). 'Some empirical evidence on the production level and production cost smoothing models of inventory investment', *American Economic Review*, 79: 853–64.

——— , Hansen, L. P., and Singleton, K. J. (1988). 'A time series analysis of representative agent models of consumption and leisure choice under uncertainty', *Quarterly Journal of Economics*, 103: 51–78.

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

——— , Hendry, D. F., and Richard, J.-F. (1983). 'Exogeneity', *Econometrica*, 51: 277–304.

English, W., Miron, J. A., and Wilcox, D. W. (1989). 'Seasonal fluctuations and the life cycle–permanent income model of consumption: a correction', *Journal of Political Economy*, 97: 988–91.

Epstein, L. G., and Zin, S. E. (1991). 'Substitution, risk aversion, and the temporal behaviour of consumption and asset returns: an empirical analysis', *Journal of Political Economy*, 99: 263–86.

Fama, E. (1976). *Foundations of Finance*. Basic Books, New York, NY, U.S.A.

Ferguson, T. S. (1958). 'A method of generating best asymptotically normal estimates with application to the estimation of bacterial densities', *Annals of Mathematical Statistics*, 29: 1046–62.

Ferson, W. E. (1990). 'Are the latent variables in time–varying expected returns compensation for consumption risk?', *Journal of Finance*, 45: 397–430.

——— and Constantinides, G. M. (1991). 'Habit persistence and durability in aggregate consumption', *Journal of Financial Economics*, 29: 199–240.

——— and Foerster, S. R. (1994). 'Finite sample properties of the Generalized Method of Moments in tests of conditional asset pricing models', *Journal of Financial Economics*, 36: 29–55.

——— , ——— , and Keim, D. B. (1993). 'General tests of latent variable models and mean–variance spanning', *Journal of Finance*, 48: 131–56.

——— and Harvey, C. R. (1992). 'Seasonality and consumption based asset pricing', *Journal of Finance*, 47: 511–52.

Finn, M. G., Hoffman, D. L., and Schlagenhauf, D. E. (1990). 'Intertemporal asset pricing relationships in barter and monetary economies: an empirical analysis', *Journal of Monetary Economics*, 25: 431–51.

Fisher, F. M. (1965). 'The choice of instrumental variables in the estimation of economy–wide econometric models', *International Economic Review*, 6: 245–74.

Fisher, R. A. (1912). 'On an absolute criterion for fitting frequency curves', *Messenger of Mathematics*, 41: 155–160.

——— (1922). 'On the mathematical foundations of theoretical statistics', *Philosophical Transactions of the Royal Society, A*, 222: 309–68.

——— (1925). 'Theory of statistical estimation', *Proceedings of the Cambridge Philosophical Society*, 22: 700–25.

Fisher, S. J. (1994). 'Asset trading, transaction costs and the equity premium', *Journal of the Applied Econometrics*, 9, Suppl. S: S71–S94.

Foster, F. D., and Viswanathan, S. (1993). 'Variations in trading volume, return volatility and trading costs: evidence on recent formation models', *Journal of Finance*, 48: 187–211.

Fuhrer, J. C. (2000). 'Habit formation in consumption and its implications for monetary-policy rules', *American Economic Review*, 90: 367–90.

——— , Moore, G. R., and Schuh, S. D. (1995). 'Estimating the linear–quadratic inventory model: maximum likelihood versus Generalized Method of Moments', *Journal of Monetary Economics*, 35: 115–57.

Fuller, W. A. (1976). *Introduction to Statistical Time Series*. Wiley, New York, NY, U.S.A.

Gallant, A. R. (1987). *Nonlinear Statistical Models*. Wiley, New York, NY, U.S.A.

——— , Hsieh, D. A., and Tauchen, G. (1997). 'Estimation of stochastic volatility models with diagnostics?', *Journal of Econometrics*, 81: 159–92.

——— and Nychka, D. W. (1987). 'Semi-nonparametric maximum likelihood estimation', *Econometrica*, 55: 363–90.

Gallant, A. R., and Tauchen, G. (1989). 'Semi–nonparametric estimation of conditionally constrained heterogeneous processes: asset pricing applications', *Econometrica*, 57: 1091–120.

——— and ——— (1996). 'Which moments to match?', *Econometric Theory*, 12: 65–81.

——— and White, H. (1988). *A Unified Theory of Estimation and Inference in Nonlinear Models*. Basil Blackwell, Oxford, U.K.

Garcia, R., and Bonomo, M. (2001). 'Tests of conditional asset pricing models in the Brazillian stock market', *Journal of International Money and Finance*, 20: 71–90.

Geary, R. C. (1942). 'Inherent relations between random variables', *Proceedings of the Royal Irish Academy, Section A*, 47: 63–76.

——— (1943). 'Relations between statistics: the general and the sampling problem when the samples are large', *Proceedings of the Royal Irish Academy, Section A*, 49: 177–96.

——— (1949). 'Determination of linear relationships between systematic parts of variables with errors of observation the variances of which are unknown', *Econometrica*, 17: 30–58.

Ghysels, E. (1998). 'On stable factor structures in the pricing of risk: do time-varying betas help or hurt?', *Journal of Finance*, 53: 549–73.

Ghysels, E., Guay, A., and Hall, A. R. (1997). 'Predictive test for structural change with unknown breakpoint', *Journal of Econometrics*, 82: 209–33.

Ghysels, E., and Hall, A. R. (1990*a*). 'Are consumption based intertemporal asset pricing models structural?', *Journal of Econometrics*, 45: 121–39.

——— and ——— (1990*b*). 'Testing nonnested Euler conditions with quadrature–based methods of approximation', *Journal of Econometrics*, 46: 273–308.

——— and ——— (1990*c*). 'A test for structural stability of Euler condition parameters estimated via the Generalized Method of Moments', *International Economic Review*, 31: 355–64.

——— and ——— (1993). 'An extension of quadrature–based methods for solving Euler equation models', in D. Brillinger, P. Caines, J. Geweke, E. Parzen, M. Rosenblatt, and M. Taqqu (eds.), *New Directions in Time Series Analysis: Part II*, pp.147–51. Springer–Verlag, New York, NY, U.S.A.

——— , Harvey, A., and Renault, E. (1996). 'Stochastic volatility', in G. S. Maddala and C. R. Rao (eds.), *Handbook of Statistics*, vol. 14, pp.119–92. Elsevier Science Publishers, Amsterdam, The Netherlands.

Gilchrist, S., and Himmelberg, C. P. (1995). 'Evidence on the role of cash flow for investment', *Journal of Monetary Economics*, 36: 541–72.

Goldberger, A. S. (1970). 'Structural equation methods in social sciences', *Econometrica*, 40: 979–1001.

Gordon, S. (1992). 'Costs of adjustment, the aggregation problem and investment', *Review of Economics and Statistics*, 74: 422–9.

Gourieroux, C., and Monfort, A. (1994). 'Testing non–nested hypotheses', in R. F. Engle and D. L. McFadden (eds.), *Handbook of Econometrics*, vol. 4, pp.2583–637. Elsevier Science Publishers, Amsterdam, The Netherlands.

——— and ——— (1996). *Simulation-Based Econometric Methods*. Oxford University Press, Oxford, U.K.

——— , ——— , and Renault, E. (1993). 'Indirect Inference', *Journal of Applied Econometrics*, 8: S85–S118.

——— , ——— , and Trognon, A. (1984). 'Pseudo maximum likelihood methods: theory', *Econometrica*, 52: 681–700.

Grammig, J., and Wellner, M. (2002). 'Modelling the interdependence of volatility and intertransaction duration processes', *Journal of Econometrics*, 106: 369–400.

Green, R. C., and Odegaard, B. A. (1997). 'Are there tax effects in the relative pricing of US government bonds?', *Journal of Finance*, 52: 609–33.

Green, S. L., and Mork, K. A. (1991). 'Toward efficiency in the crude-oil market', *Journal of the Applied Econometrics*, 6: 45–66.

Greene, W. H. (2003). *Econometric Analysis*. Prentice Hall, Upper Saddle River, NJ, U.S.A., 5th edn.

Groen, J. J., and Kleibergen, F. (2003). 'Likelihood based cointegration analysis in panels of vector error correction models', *Journal of Business and Economic Statistics*, 21: 295–318.

Hagiwara, M., and Herce, M. A. (1997). 'Risk aversion and stock price sensitivity to dividends', *American Economic Review*, 87: 738–45.

Hahn, J., and Inoue, A. (2002). 'A Monte Carlo comparison of various asymptotic approximations to the distribution of instrumental variables estimators', *Econometric Reviews*, 21: 309–36.

Haile, P. A. (2001). 'Auctions with resale markets: an application to U.S. forest service timber sales', *American Economic Review*, 91: 399–427.

Hall, A. R. (1987*a*). 'The information matrix test for the linear model', *Review of Economic Studies*, 54: 257–63.

——— (1987*b*). 'Testing for a unit root in the prescence of moving average errors', *Biometrika*, 76: 49–56.

——— (1999). 'Hypothesis testing in models estimated by Generalized Method of Moments', in L. Mátyás (ed.), *Generalized Method of Moments Estimation*, pp. 75–101. Cambridge University Press, Cambridge, U.K.

——— (2000). 'Covariance matrix estimation and the power of the overidentifying restrictions test', *Econometrica*, 68: 1517–27.

——— and Inoue, A. (2003). 'The large sample behaviour of the Generalized Method of Moments estimator in misspecified models', *Journal of Econometrics*, 114: 361–94.

——— , ——— , Jana, K., and Shin, C. (2003). 'Information in Generalized Method of Moments estimation and entropy based moment selection', Discussion paper, Department of Economics, North Carolina State University, Raleigh, NC, U.S.A.

——— , ——— , and Peixe, F. P. M. (2003). 'Covariance estimation and the limiting behaviour of the overidentifying restrictions test in the presence of neglected structural instability', *Econometric Theory*, 19: 962–83.

——— and Peixe, F. P. M. (2000). 'Data mining and the selection of instruments', *Journal of Economic Methodology*, 7: 265–78.

——— and ——— (2003). 'A consistent method for the selection of relevant instruments', *Econometric Reviews*, 22: 269–88.

——— and Rossana, R. J. (1991). 'Estimating the speed of adjustment in partial adjustment models', *Journal of Business and Economic Statistics*, 9: 441–53.

——— , Rudebusch, G., and Wilcox, D. (1996). 'Judging instrument relevance in instrumental variables estimation', *International Economic Review*, 37: 283–98.

——— and Sen, A. (1999). 'Structural stability testing in models estimated by Generalized Method of Moments', *Journal of Business and Economic Statistics*, 17: 335–48.

Hall, P. (1994). 'Methodology and Theory for the Bootstrap', in R. F. Engle and D. L. McFadden (eds.), *Handbook of Econometrics*, vol. 4, pp.2342–81. Elsevier Science Publishers, Amsterdam, The Netherlands.

Hall, P. and Horowitz, J. L. (1996). 'Bootstrap critical values for tests based on Generalized Method of Moments', *Econometrica*, 64: 891–917.

——— , ——— , and Jing, B.-Y. (1995). 'On blocking rules for the bootstrap with dependent data', *Biometrika*, 82: 561–74.

Hamilton, J. D. (1994). *Time Series Analysis*. Princeton University Press, Princeton, NJ, U.S.A.

Hannan, E. J., and Quinn, B. G. (1979). 'The determination of order of an autoregression', *Journal of the Royal Statistical Society, Series B*, 41(2): 190–5.

Hansen, B. E. (1990). 'Lagrange multiplier tests for parameter instability in non-linear models', Discussion paper, Department of Economics, University of Rochester, Rochester, NY, U.S.A.

——— (1992). 'Consistent covariance matrix estimation for dependent heterogeneous processes', *Econometrica*, 60: 967–72.

——— (1997). 'Approximate asymptotic p–values for structural change tests', *Journal of Business and Economic Statistics*, 15: 60–7.

Hansen, H., and Tarp, F. (2001). 'Aid and growth regressions', *Journal of Development Economics*, 64: 547–70.

Hansen, L. P. (1978). 'Econometric modelling strategies for exhaustible resource markets with applications to nonferrous metals', Ph.D. thesis, Department of Economics, University of Minnesota, Minneapolis, MN, U.S.A.

——— (1982). 'Large sample properties of Generalized Method of Moments estimators', *Econometrica*, 50: 1029–54.

——— (1985). 'A method of calculating bounds on the asymptotic covariance matrices of generalized method of moments estimators', *Journal of Econometrics*, 30: 203–38.

——— , Heaton, J., and Luttmer, E. G. J. (1995). 'Econometric Evaluation of Asset Pricing Models', *Review of Financial Studies*, 8: 237–74.

——— , ——— , and Ogaki, M. (1988). 'Efficiency bounds implied by multiperiod conditional moment restrictions', *Journal of the American Statistical Association*, 83: 863–71.

Hansen, L. P., Heaton, J., and Yaron, A. (1996). 'Finite sample properties of some alternative GMM estimators obtained from financial market data', *Journal of Business and Economic Statistics*, 14: 262–80.

——— and Hodrick, R. J. (1980). 'Forward exchange rates as optimal predictors of future spot rates', *Journal of Political Economy*, 887: 829–53.

——— and Jaganathan, R. (1997). 'Assessing specification errors in stochastic discount factor models', *Journal of Finance*, 52: 557–90.

——— and Sargent, T. (1982). 'Instrumental variables procedures for estimating linear rational expectations models', *Journal of Monetary Economics*, 9: 263–96.

——— and Singleton, K. J. (1982). 'Generalized instrumental variables estimation of nonlinear rational expectations models', *Econometrica*, 50: 1269–86.

——— and ——— (1983). 'Stochastic consumption, risk aversion and the temporal behavior of asset returns', *Journal of Political Economy*, 91: 249–65.

——— and ——— (1984). 'Errata', *Econometrica*, 52: 267–8.

——— and ——— (1991). 'Computing semi-parametric efficiency bounds for linear time series models', in W. Barnett, J. Powell, and G. Tauchen (eds.), *Nonparametric and Seminonparametric Methods in Econometrics and Statistics*, pp. 387–412. Cambridge University Press, Cambridge, U.K.

——— and ——— (1996). 'Efficient estimation of linear asset pricing models with moving average errors', *Journal of Business and Economic Statistics*, 14: 53–68.

Hartmann, P. (1999). 'Trading volumes and transaction costs in the foreign exchange market: evidence from daily dollar yen spot data', *Journal of Banking and Finance*, 23: 801–24.

Harvey, C. (1991). 'World price of covariance risk', *Journal of Finance*, 46: 111–57.

Hausman, J. (1978). 'Specification tests in econometrics', *Econometrica*, 46: 1251–71.

Hayashi, F., and Sims, C. (1983). 'Nearly efficient estimation of time series models with predetermined, but not exogenous instruments', *Econometrica*, 51: 783–98.

Heaton, J. (1995). 'An empirical investigation of asset pricing with temporally dependent preference specifications', *Econometrica*, 63: 681–717.

——— and Ogaki, M. (1991). 'Efficiency bound calculations for a time series model with conditional heteroscedasticity', *Economics Letters*, 35: 167–71.

He, J., Kan, R., Ng, L., and Zhang, C. (1996). 'Tests of the relations among marketwide factors, firm specfic variables and stock returns using a conditional asset pricing model', *Journal of Finance*, 51: 1891–908.

Hillier, G., Kinal, T. W., and Srivastava, V. K. (1984). 'On the moments of ordinary least squares and instrumental variables estimators in a general structural equation', *Econometrica*, 52: 185–202.

Himmelberg, C. P., & Petersen, B. C. (1994). 'R & D internal finance: a panel study of small firms in high-tech industries', *Review of Economics and Statistics*, 76: 38–51.

Holman, J. A. (1998). 'GMM estimation of a money in the utility function model: the implications of functional forms', *Journal of Money, Credit and Banking*, 30: 679–98.

Ho, M. S., Perraudin, R. M., and Sørensen, B. E. (1996). 'A continuous time arbitrage pricing model with stochastic volatility and jumps', *Journal of Business and Economic Statistics*, 14: 31–44.

Huang, R. D., and Stoll, H. R. (1997). 'The components of the bid–ask spread: a general approach', *Review of Financial Studies*, 10: 995–1034.

Hubbard, R. G., and Kayshap, A. K. (1992). 'Internal net worth and the investment process: an application to U. S. agriculture', *Journal of Political Economy*, 100: 506–34.

Iannizzotto, M., and Taylor, M. (1999). 'The target zone model, nonlinearity and mean reversion: is the honeymoon really over?', *Economic Journal*, 109: C96–C110.

Ilmanen, A. (1992). 'Time-varying expected returns in the international bond markets', *Journal of Finance*, 100: 481–506.

Imbens, G. (2002). 'Generalized method of moments and empirical likelihood', *Journal of Business and Economic Statistics*, 20: 493–506.

——— , Spady, R. H., and Johnson, P. (1998). 'Information theoretic approaches to inference in moment condition models', *Econometrica*, 66: 333–57.

Imrohoroglus, S. (1994). 'GMM estimates of currency substitution between the Canadian dollar and the United States Dollar', *Journal of Money, Credit and Banking*, 26: 792–807.

Ingersoll, J. E. (1987). *Theory of Financial Decision Making*. Rowman and Littlefield, Savage, MD, U.S.A.

Inoue, A., and Shintani, M. (2003). 'Bootstrapping GMM estimators for time series', *Journal of Econometrics*, forthcoming.

Intrilligator, M. D. (1971). *Mathematical Optimization and Economic Theory*. Prentice Hall, Englewood Cliffs, NJ, U.S.A.

Jalan, J., and Ravallion, M. (1999). 'Are the poor less well insured? Evidence on vulnerability to income risk in rural China', *Journal of Development Economics*, 58: 61–81.

Jiang, G. L., and Knight, J. L. (2002). 'Estimation of continuous time processes via the empirical characteristic function', *Journal of Business and Economic Statistics*, 20: 198–212.

Johnson, N. L., and Kotz, S. (1970). *Distributions in Statistics: Continuous Univariate Distributions–2*. Wiley, New York, NY, U.S.A.

Jorgenson, D. W., and Laffont, J.-J. (1974). 'Efficient estimation of nonlinear simultaneous equations with additive disturbances', *Annals of Economic and Social Measurement*, 3/4: 615–40.

Judge, G. G., Griffiths, W. E., Hill, R. C., Lutkepohl, H., and Lee, T. C. (1985). *The Theory and Practice of Econometrics*. Wiley, New York, NY, U.S.A., 2nd edn.

Kahn, S., and Lang, K. (1991). 'The effect of hours constraints on labor supply estimates', *Review of Economics and Statistics*, 73: 605–11.

Kayshap, A., and Wilcox, D. (1993). 'Production and inventory control at the General Motors Corporation during the 1920s and 1930s', *American Economic Review*, 83: 383–401.

Keane, M., and Runkle, D. E. (1990). 'Testing the rationality of price forecasts: new evidence from panel data', *American Economic Review*, 80: 714–35.

Keifer, N. M., and Vogelsang, T. J. (2002*a*). 'Heteroscedasticity–autocorrelation robust testing using bandwidth equal to sample size', *Econometric Theory*, 18: 1350–66.

——— and ——— (2002*b*). 'Heteroscedasticity-autocorrelation robust standard errors using the Bartlett kernel without truncation', *Econometrica*, 70: 2093–5.

Kirman, A. (1992). 'Whom or what does the representative agent represent?', *Journal of Economic Perspectives*, 6: 117–36.

Kitamura, Y. (1997). 'Empirical likelihood methods with weakly dependent processes', *Annals of Statistics*, 25: 2084–102.

——— and Phillips, P. C. B. (1997). 'Fully modified IV, GIVE and GMM estimation with possibley non–stationary regressors and instruments', *Journal of Econometrics*, 80: 85–123.

Kleibergen, F. (2000). 'Testing parameters in GMM without assuming that they are identified', Discussion paper, Discussion paper # TI 2001-067/4, Tindbergen Institute, Amsterdam, The Netherlands.

Knight, J. L. (1986). 'The moments of OLS and 2SLS when the disturbances are non-normal', *Journal of Econometrics*, 27: 39–60.

Kocherlakota, N. R. (1990). 'On tests of representative consumer asset pricing models', *Journal of Monetary Economics*, 26: 285–304.

——— (1996). 'The equity premium: it's still a puzzle', *Journal of Economic Literature*, 34: 42–71.

Koenker, R., and Machado, J. A. F. (1999). 'GMM inference when the number of moment conditions is large', *Journal of Econometrics*, 93: 327–44.

Kopp, R. J., and Mullahy, J. (1990). 'Moment based estimation and testing of stochastic frontier models', *Journal of Econometrics*, 46: 165–83.

Künsch, H. R. (1989). 'The jackknife and the bootstrap for general stationary observations', *Annals of Statistics*, 17: 1217–41.

Lahiri, S. N. (1999). 'Theoretical comparisons of block bootstrap methods', *Annals of Statistics*, 27: 386–404.

Lee, B. S. (1991). 'Government deficits and the term structure of interest rates', *Journal of Monetary Economics*, 27: 425–43.

——— and Ingram, B. F. (1991). 'Simulation estimation of time series models', *Journal of Econometrics*, 47: 197–205.

Lehmann, E. L. (1959). *Testing Statistical Hypotheses*. Wiley, New York, NY, U.S.A.

Li, H., and Maddala, G. S. (1996). 'Bootstrapping time series models', *Econometric Reviews*, 15: 115–58.

Longstaff, F. A., and Schwartz, E. S. (1991). 'Interest rate volatility and the term structure: a two factor general equilibrium', *Journal of Finance*, 27: 1259–82.

Lucas, R. E. (1978). 'Asset prices in an exchange economy', *Econometrica*, 46: 1429–46.

Maasoumi, E., and Phillips, P. C. B. (1982). 'On the behaviour of inconsistent instrumental variable estimators', *Journal of Econometrics*, 19: 183–201.

McFadden, D. (1989). 'A method of simulated moments for estimation of discrete response models without numerical integration', *Econometrica*, 47: 995–1026.

——— and Train, K. (2000). 'Mixed MNL models for discrete response', *Journal of Applied Econometrics*, 15: 447–70.

MacKinlay, A. C., and Richardson, M. P. (1991). 'Using Generalized Method of Moments to test mean–variance efficiency', *Journal of Finance*, 46: 511–27.

Madhavan, A., Richardson, M., and Roomans, M. (1997). 'Why do security prices change? A transaction–level analysis of NYSE stocks', *Review of Financial Studies*, 10: 1035–64.

——— and Smidt, S. (1993). 'An analysis of changes in specialist inventories and quotations', *Journal of Finance*, 48: 1595–628.

Magnus, J. R., and Neudecker, H. (1991). *Matrix Differential Calculus with Applications in Statistics and Econometrics*. Wiley, New York, NY, U.S.A.

Malkiel, B. G. (1987). *A Random Walk Down Wall Street*. W. W. Norton and Co., New York, NY, U.S.A.

Mankiw, N. G., Rotemberg, J., and Summers, L. H. (1985). 'Intertemporal substitution in macroeconomics', *Quarterly Journal of Economics*, 100: 225–52.

——— and Zeldes, S. P. (1991). 'The consumption of stockholders and non-stockholders', *Journal of Financial Economics*, 29: 97–112.

Mark, N. (1985). 'On time varying risk premia in the foreign exchange market', *Journal of Monetary Economics*, 16: 3–18.

Marshall, D. A. (1992). 'Inflation and asset retruns in a monetary economy', *Journal of Finance*, 47: 1315–42.

Mathworks (2000). *MATLAB*. The Mathworks Inc., Natick, MA, U.S.A.

Meghir, C., and Weber, G. (1996). 'Intertemporal nonseparability or borrowing restrictions? A diaggregate analysis using a U.S. consumption panel', *Econometrica*, 64: 1151–81.

Melino, A., and Turnbull, S. M. (1990). 'Pricing foreign currency options with stochastic volatility', *Journal of Econometrics*, 45: 239–66.

Miron, J. A. (1986). 'Seasonal fluctuations and the life cycle-permanent income model of consumption', *Journal of Political Economy*, 94: 1258–79.

——— and Zeldes, S. P. (1988). 'Seasonality, cost shocks and the production smoothing model of inventories', *Econometrica*, 56: 877–908.

Mishkin, F. S. (1995). *Financial Markets, Institutions, and Money*. Harper Collins, New York, NY, U.S.A.

Mitchell, B. M., and Fisher, F. M. (1970). 'The choice of instrumental variables in the estimation of economy–wide econometric models: some further thoughts', *International Economic Review*, 11: 226–34.

Mizon, G. E., and Richard, J. F. (1986). 'The encompassing principle and its application to testing non–nested hypotheses', *Econometrica*, 54: 657–78.

Modjtahedi, B. (1991). 'Multiple maturities and time–varying risk premia in forward exchange markets', *Journal of International Economics*, 30: 69–86.

Morgan, M. (1990). *The History of Econometric Ideas*. Cambridge University Press, New York, NY, U.S.A.

Morimune, K. (1983). 'Asymptotic distributions of k-class estimators when the degree of overidentifiability is large compared with the sample size', *Econometrica*, 51: 821–42.

Morrison, D. F. (1976). *Multivariate Statistical Methods*. McGraw–Hill, Tokyo, Japan, 2nd edn.

Nagar, A. L. (1959). 'The bias and moment matrix of the general k-class estimators of the parameters in simultaneous equations', *Econometrica*, 27: 575–95.

Nakamura, A., and Nakamura, M. (1998). 'Model specification and endogeneity', *Journal of Econometrics*, 83: 213–37.

Nelson, C. R., and Startz, R. (1990). 'The distribution of the instrumental variables estimator and its t ratio when the instrument is a poor one', *Journal of Business*, 63: S125–S140.

Nevo, A. (2003). 'Using weights to adjust for sample selection when auxilliary information is available', *Journal of Business and Economic Statistics*, 21: 43–52.

Newey, W. K. (1984). 'A method of moments interpretation of sequential estimators', *Economics Letters*, 14: 201–6.

——— (1985*a*). 'Generalized Method of Moments specification testing', *Journal of Econometrics*, 29: 229–56.

——— (1985*b*). 'Maximum likelihood specification testing and conditional moment tests', *Econometrica*, 53: 1047–70.

——— (1990). 'Efficient instrumental variables estimation of nonlinear models', *Econometrica*, 58: 809–38.

——— (1993). 'Efficient estimation of models with conditional moment restrictions', in G. S. Maddala, C. R. Rao, and H. D. Vinod (eds.), *Handbook of Statistics*, vol. 11, pp.419–54. Elsevier Science Publishers, Amsterdam, The Netherlands.

——— and McFadden, D. L. (1994). 'Large sample estimation and hypothesis testing', in R. Engle and D. L. McFadden (eds.), *Handbook of Econometrics*, vol. 4, pp.2113–247. Elsevier Science Publishers, Amsterdam, The Netherlands.

——— and Smith, R. J. (2004). 'Higher order properties of GMM and generalized empirical likelihood estimators', *Econometrica*, 72, 219–56.

——— and West, K. D. (1987*a*). 'A simple positive semi–definite heteroscedasticity and autocorrelation consistent covariance matrix', *Econometrica*, 55: 703–8.

——— and ——— (1987*b*). 'Hypothesis testing with efficient method of moments testing', *International Economic Review*, 28: 777–87.

——— and ——— (1994). 'Automatic lag selection in covariance matrix estimation', *Review of Economic Studies*, 61: 631–53.

Neyman, J. (1949). 'Contribution to the theory of the χ^2 test', in *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability*, pp. 239–73. University of California Press, Berkeley, CA, U.S.A.

——— and Pearson, E. S. (1928). 'On the use and interpretation of certain test criteria for purposes of statistical inference: part II', *Biometrika*, 20A: 263–94.

Ni, S. (1995). 'An empirical analysis on the substitutability between private consumption and government purchases', *Journal of Monetary Economics*, 36: 593–605.

Nunes, L. C., Kuan, C.-M., and Newbold, P. (1995). 'Spurious break', *Econometric Theory*, 11: 736–49.

Nyblom, J. (1989). 'Testing for the constancy of parameters over time', *Journal of the American Statistical Association*, 84: 223–30.

Ogaki, M., and Zhang, Q. (2001). 'Decreasing relative risk aversion and tests of risk sharing', *Econometrica*, 69: 515–26.

Ogawa, K., and Suzuki, K. (1998). 'Land values and corporate investment: evidence from Japanese panel data', *Journal of the Japanese and International Economies*, 12: 232–49.

Oliner, S. D., Rudebusch, G. D., and Sichel, D. (1996). 'The Lucas critique revisited: assessing the stability of empirical Euler equations for investment', *Journal of Econometrics*, 70: 291–316.

Owen, A. B. (1988). 'Empirical likelihood ratio confidence intervals for a single functional', *Biometrika*, 75: 237–49.

——— (1990). 'Empirical likelihood confidence regions', *Annals of Statistics*, 18: 90–120.

——— (1991). 'Empirical likelihood for linear models', *Annals of Statistics*, 19: 1725–47.

——— (2001). *Empirical Likelihood*. Chapman & Hall, London, U.K.

Pagan, A. R. (1984). 'Econometric issues in the analysis of regressions with generated regressors', *International Economic Review*, 25: 221–47.

Pakes, A., and Pollard, D. (1989). 'Simulation and the asymptotics of optimization estimators', *Econometrica*, 47: 1027–57.

Palacios-Huerta, I. (2003). 'An empirical analysis of the risk properties of human capital returns', *American Economic Review*, 93: 948–64.

Pantula, S., and Hall, A. R. (1991). 'Testing for unit roots in autoregressive moving average models', *Journal of Econometrics*, 48: 325–54.

Pearson, K. (1893). 'Asymmetrical frequency curves', *Nature*, 48: 615–6.

——— (1894). 'Contributions to the mathematical theory of evolution', *Philosophical Transactions of the Royal Society of London (A)*, 185: 71–110.

——— (1895). 'Contributions to the mathematical theory of evolution, II: skew variation', *Philosophical Transactions of the Royal Society of London (A)*, 186: 343–414.

——— (1900). 'On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling', *Philosophical Magazine*, 5th series, 50: 157–75.

Peixe, F. P. M. (2000). 'Instrument selection in econometric models: consequences and methods', Ph.D. thesis, Department of Economics, University of Birmingham, Birmingham, U.K.

——— and Hall, A. R. (2000). 'The mean squared error of the instrumental variables estimator when the disturbance has an elliptical distribution', Discussion paper, Department of Economics, North Carolina State University, Raleigh, NC, U.S.A.

Pesaran, M. H. (1987). 'Global and partial non-nested hypotheses and asymptotic local power', *Econometric Theory*, 3: 69–97.

Pfann, G. A., and Palm, F. C. (1993). 'Asymmetric adjustment costs in non-linear labour demand models for the Netherlands and U.K. manufacturing sectors', *Review of Economic Studies*, 60: 397–412.

Phillips, P. C. B. (1980). 'The exact distribution of instrumental variables estimators in an equation containing $n+1$ endogenous variables', *Econometrica*, 52: 861–78.

——— (1982). 'On the consistency of nonlinear FIML', *Econometrica*, 50: 1307–24.

——— (1983). 'Exact small sample theory', in Z. Griliches and M. D. Intrilligator (eds.), *Handbook of Econometrics*, vol. 1, pp.449–516. Elsevier Science Publishers, Amsterdam, The Netherlands.

——— and Hansen, B. E. (1990). 'Statistical inference in instrumental variables regression with I(1) processes', *Review of Economic Studies*, 57: 99–125.

Pindyck, R., and Rotemberg, J. (1983). 'Dynamic factor demands and the effects of energy price shocks', *American Economic Review*, 73: 1066–79.

Pitman, E. J. G. (1949). *Notes on Non-Parametric Statistical Inference*. Columbia University, New York, NY, U.S.A.

Popp, D. C. (2001). 'The effect of new technology on energy consumption', *Resource and Energy Economics*, 23: 215–39.

Pötscher, B. M. (1983). 'Order estimation in ARMA models by lagrangian multiplier tests', *Annals of Statistics*, 11: 872–85.

——— (1991). 'Effects of model selection on inference', *Econometric Theory*, 7: 163–85.

——— and Prucha, I. R. (1997). *Dynamic Nonlinear Econometric Models*. Springer–Verlag, Berlin, Germany.

Press, H., and Tukey, J. W. (1956). 'Power spectral methods of analysis and their applications to problems in airplane dynamics', *Flight Test Manual, NATO, Advisory Group for Aeronautical Research and Development*, IV–C: 1–41.

Priestley, M. B. (1981). *Spectral Analysis and Time Series*. Academic Press, New York, NY, U.S.A.

Qin, J., and Lawless, J. (1994). 'Empirical likelihood and generalized estimating equations', *Annals of Statistics*, 22: 300–25.

Quandt, R. E. (1983). 'Computational problems and methods', in Z. Grilliches and M. D. Intrilligator (eds.), *Handbook of Econometrics*, vol. 1, pp.699–764. Elsevier Science Publishers, Amsterdam, The Netherlands.

Rao, C. R. (1973). *Linear Statistical Inference and its Applications*. Wiley, New York, NY USA, 2nd edn.

Reiersøl, O. (1941). 'Confluence analysis by means of lag moments and other methods of confluence analysis', *Econometrica*, 9: 1–24.

——— (1945). 'Confluence analysis by means of instrumental sets of variables', *Arkiv foer Mathematik, Astronomi och Fysik*, 32: 1–119.

Reinsel, G. C. (1993). *Elements of Multivariate Time Series Analysis*. Springer Verlag, New York, NY, U.S.A.

Richardson, D. H. (1963). 'The exact distribution of a structural coefficient estimator', *Journal of the American Statistical Association*, 63: 1214–26.

Richardson, M., and Smith, T. (1993). 'A test for multivariate normality of stock returns', *Journal of Business*, 66: 295–321.

Robinson, P. M. (1988). 'The stochastic difference between econometric estimators', *Econometrica*, 56: 531–48.

Rudin, W. (1976). *Principles of Mathematical Analysis*. McGraw Hill, New York, NY, U.S.A, 3rd edn.

Runkle, D. E. (1991). 'Liquidity constraints and the permanant–income hypothesis', *Journal of Monetary Economics*, 27: 73–98.

Sargan, J. D. (1958). 'The estimation of economic relationships using instrumental variables', *Econometrica*, 26: 393–415.

——— (1959). 'The estimation of relationships with autocorrelated residuals by the use of instrumental variables', *Journal of the Royal Statistical Society B*, 21: 91–105.

——— (1974). 'The validity of Nagar's expansion for the moments of econometric estimators', *Econometrica*, 42: 169–76.

——— (1975). 'The Gram–Charlier approximations to t ratios of k-class estimators', *Econometrica*, 43: 327–46.

——— and Mikhail, W. M. (1971). 'A general approximation to the distribution of instrumental variables estimators', *Econometrica*, 39: 131–69.

Sawa, T. (1969). 'The exact sampling distribution of ordinary least squares and two–stage least squares estimators', *Journal of the American Statistical Association*, 64: 923–37.

Schellhorn, M. (2001). 'The effect of variable health insurance deductibles on the demand for physician visits', *Health Economics*, 10: 441–56.

Schuh, S. (1996). 'Evidence on the link between firm–level and aggregate inventory behaviour', Discussion paper, Finance and Economics Discussion Series # 1996-46, Board of Governors of the Federal Reserve System, Washington, DC, U.S.A.

Schwarz, G. (1978). 'Estimating the dimension of a model', *Annals of Statistics*, 6: 461–4.

Sen, A. (1997). 'New tests of structural stability and applications to consumption based asset pricing models', Ph.D. thesis, Department of Economics, North Carolina State University, Raleigh, NC, U.S.A.

——— and Hall, A. R. (1999). 'Two further aspects of some new tests for structural stability', *Structural Change and Economic Dynamics*, 10: 431–43.

Shea, J. (1997). 'Instrument relevance in multivariate linear models', *Review of Economics and Statistics*, 79: 348–52.

Shibata, R. (1976). 'Selection of the order of an autoregressive model by Akaike's information criterion', *Biometrika*, 63: 117–26.

Sieg, H. (2000). 'Estimating a bargaining model with asymmetric information: evidence from medical malpractice suits', *Journal of Political Economy*, 108: 1006–21.

Silva, J. M. C. S., and Windmeijer, F. A. G. (2001). 'Two–part spell models for health care demand', *Journal of Econometrics*, 104: 67–89.

Singleton, K. J. (1985). 'Testing specifications of economic agents' intertemporal optimum problems in the presence of alternative models', *Journal of Econometrics*, 30: 391–413.

——— (1988). 'Econometric issues in the analysis of equilibrium business cycle models', *Journal of Monetary Economics*, 21: 361–86.

Smith, D. C. (1999). 'Finite sample properties of tests of the Epstein–Zin asset pricing model', *Journal of Econometrics*, 93: 113–48.

Smith, K. (1916). 'On the 'best' values of the constants in frequency distributions', *Biometrika*, 11: 262–76.

Smith, R. J. (1997). 'Alternative semi-parametric likelihood approaches to generalized method of moments estimation', *Economics Journal*, 107: 503–19.

Smith, V. K., and Pattanayak, S. K. (2002). 'Is meta-analysis a Noah's ark for non-market valuation?', *Environmental and Resource Economics*, 22: 271–96.

Snow, K. N. (1991). 'Diagnosing asset pricing models using the distribution of asset returns', *Journal of Finance*, 46: 955–83.

Sowell, F. (1996). 'Optimal tests of parameter variation in the Generalized Method of Moments framework', *Econometrica*, 64: 1085–108.

Spanos, A. (1999). *Probability Theory and Statistical Inference*. Cambridge University Press, New York, NY, U.S.A.

Srinivasan, T. N. (1970). 'Approximations to finite sample moments of estimators whose exact sampling distributions are unknown', *Econometrica*, 38: 533–41.

Staiger, D., and Stock, J. H. (1997). 'Instrumental variables regression with weak instruments', *Econometrica*, 65: 557–86.

Stigler, S. M. (1986). *The History of Statistics*. Belknap Harvard, Cambridge, MA, U.S.A.

Stock, J. H., and Wright, J. H. (1995). 'Asymptotics for GMM estimators with weak instruments', Discussion paper, Kennedy School of Government, Harvard University, Cambridge, MA, U.S.A.

——— and ——— (2000). 'GMM with weak identification', *Econometrica*, 68: 1055–96.

——— and Yogo, M. (2001). 'Testing for weak instruments in linear IV regression', Discussion paper, Kennedy School of Government, Harvard University, Cambridge, MA, U.S.A.

Stoica, P., Söderström, T., and Friedlander, B. (1985). 'Optimal instrumental variable estimates of the AR parameters of an ARMA process', *IEEE Transactions on Automatic Control*, AC-30(11): 1066–74.

Strang, G. S. (1988). *Linear Algebra and its Applications*. Harcourt, Brace and Jovanovich, San Diego, CA, U.S.A., 3rd edn.

Stuart, A., and Ord, J. K. (1987). *Kendall's Advanced Theory of Statistics: volume 1.* Oxford University Press, New York, NY, U.S.A., 5th edn.

Tauchen, G. (1985*a*). 'Diagnostic testing and evaluation of maximum likelihood models', *Journal of Econometrics*, 30: 415–43.

——— (1985*b*). 'Finite state Markov chain approximations to univariate and vector autoregressions', *Economics Letters*, 20: 177–81.

——— (1986). 'Statistical properties of Generalized Method of Moments estimators of structural parameters obtained from financial market data', *Journal of Business and Economic Statistics*, 4: 397–416.

——— and Hussey, R. (1991). 'Quadrature–based methods for obtaining approximate solutions to nonlinear asset pricing models', *Econometrica*, 59: 371–96.

Theil, H. (1971). *Principles of Econometrics.* Wiley, New York, NY, U.S.A.

Thijssen, G. (1996). 'Farmers' investment behaviour: an empirical assessment of two specifications of expectations', *American Journal of Agricultural Economics*, 78: 166–74.

Timmerman, A. (2001). 'Structural breaks, incomplete information, and stock prices', *Journal of Business and Economic Statistics*, 19: 299–314.

Vetzal, K. R. (1992). 'Stochastic short rate volatility and the pricing of bonds and bond options', Ph.D. thesis, Faculty of Management, University of Toronto, Toronto, Ontario, Canada.

——— (1997). 'Stochastic volatility, movements in short term interest rates and bond options', *Journal of Banking and Finance*, 21: 169–96.

Vissing-Jørgenson, A., and Attanasio, O. P. (2003). 'Stock market participation, intertemporal substitution, and risk aversion', *American Economic Review*, 93: 383–91.

Vogelsang, T. J. (2003). 'Testing in GMM models without truncation', in T. Fomby and R. C. Hill (eds.), *Advances in Econometrics*, vol. 17, pp.199–233. Elsevier Science Publishers, Amsterdam, The Netherlands.

Wang, J., and Zivot, E. (1998). 'Inference on structural parameters in instrumental variables regression with weak instruments', *Econometrica*, 66: 1389–404.

Weber, C. E. (2000). ' "Rule of thumb" consumption, intertemporal substitution and risk aversion', *Journal of Business and Economic Statistics*, 18: 497–502.

West, K. D. (1997). 'Another heteroscedasticity and autocorrelation–consistent covariance matrix estimator', *Journal of Econometrics*, 76: 171–91.

West, K. D. (2001). 'On optimal instrumental variables estimation of stationary time series models', *International Economic Review*, 42: 29–33.

——— and Wilcox, D. W. (1994). 'Some evidence on finite sample distributions of instrumental variables estimators of the linear quadratic inventory model', in R. Fiorito (ed.), *Inventory Cycles and Monetary Policy*, pp.253–82. Springer–Verlag, Berlin, Germany.

——— and ——— (1996). 'A comparison of alternative instrumental variables estimators of a dynamic linear model', *Journal of Business and Economic Statistics*, 14: 281–93.

Whited, T. M. (1992). 'Debt, liquidity constraints and corporate investment: evidence for panel data', *Journal of Finance*, 47: 1425–60.

White, H. (1982). 'Maximum likelihood in misspecified models', *Econometrica*, 50: 1–25.

——— (1984). *Asymptotic Theory for Econometricians*. Academic Press, New York, NY, U.S.A.

——— (1994). *Estimation, Inference and Specification Analysis*. Cambridge University Press, New York, NY, U.S.A.

——— and Domowitz, I. (1984). 'Nonlinear regression with dependent observations', *Econometrica*, 52: 143–61.

Windmeijer, F. A. G., and Silva, J. M. C. S. (1997). 'Endogeneity in count data models: an application to the demand for health care', *Journal of the Applied Econometrics*, 12: 281–94.

Wooldridge, J. M. (1994). 'Estimation and inference for dependent processes', in R. Engle and D. L. McFadden (eds.), *Handbook of Econometrics*, vol. 4, pp.2641–739. Elsevier Science Publishers, Amsterdam, The Netherlands.

——— (2002). *Econometric Analysis of Cross Section and Panel Data*. MIT Press, Cambridge, MA, U.S.A.

Wright, J. (2001). 'Detecting lack of identification in GMM', Discussion paper, Board of Governors of the Federal Reserve System, Washington, DC, U.S.A.

Wright, P. G. (1928). *The Tariff on Animal and Vegetable Oils*. MacMillan, New York, NY, U.S.A.

Wright, S. (1925). 'Corn and hog correlations', Discussion paper, U. S. Department of Agriculture Bulletin No. 1300, Washington, DC, U.S.A.

Wu, D. (1973). 'Alternative tests of independence between stochastic regressors and disturbances : finite sample results', *Econometrica*, 42: 529–46.

Yashiv, E. (2000). 'The determinants of equilibrium unemployment', *American Economic Review*, 90: 1297–322.

Young, D. (1991). '2–stage modelling of resource owner behaviour – an applications to Canadian copper mining', *Resources and Energy*, 13: 263–84.

——— (1992). 'Cost specification and firm behaviour in a Hotelling model of resource extraction', *Canadian Journal of Economics*, 25: 41–59.

Yuan, M. W., and Li, W. L. (2000). 'Dynamic employment and hours effects of government spending shocks', *Journal of Economic Dynamics and Control*, 24: 1233–63.

Zhou, G. F. (1994). 'Analytical GMM tests - asset pricing with time varying risk premiums', *Review of Financial Studies*, 7: 687–709.

Zivot, E., Startz, R., and Nelson, C. R. (1998). 'Valid confidence intervals and inference in the presence of weak instruments', *International Economic Review*, 39: 1119–44.

This page intentionally left blank

Author Index

Subject Index