

©2001–2012 Masao Ogaki, Kyungho Jang, Hyoung-Seok Lim, Youngsoo Bae, and  
Yuko Imura

# STRUCTURAL MACROECONOMETRICS

MASAO OGAKI  
The Ohio State University

KYUNGHO JANG  
Inha University

HYOUNG-SEOK LIM  
Korea Institute of Finance

YOUNGSOO BAE  
Lehigh University

YUKO IMURA  
The Ohio State University

First draft: May, 2000

This version: February 5, 2012

## PREFACE

This book presents various structural econometric tools used in macroeconomics. The word “structural” has been defined in many ways. In this book, “structural” means that explicit assumptions are made in econometric methods so that estimators or test statistics can be interpreted in terms of an economic model (or models) as explained in Chapter 1.

Many applied macroeconomists link macroeconomic models with econometric methods in this sense of structural econometrics. In principle, recent advances of theoretical time series econometrics make this task easier because they often relax the very restrictive assumptions made in conventional econometrics. There are many textbooks that explain these advanced econometric methods. It is often difficult, however, for applied researchers to exploit these advances because few textbooks in time series econometrics explain how macroeconomic models are mapped into advanced econometric models.<sup>1</sup> To fill this gap, this book presents methods to apply advanced econometric procedures to structural macroeconomic models. The econometric methods covered are mainly those of time series econometrics, and include the generalized method of moments, vector autoregressions, and estimation and testing in the presence of nonstationary variables.

Since this book focuses on applications, proofs are usually omitted with references given for interested readers. When proofs are helpful to understand issues that are important for applied research, they are given in mathematical appendices. Many examples are given to illustrate concepts and methods.

---

<sup>1</sup>For example, Hamilton (1994) contains exceptional volume of explanations of applications for a time series econometrics textbook, but its main focus is on econometrics, and not on the mapping of economic models into econometric models.

This book is intended for an advanced graduate course in time series econometrics or macroeconomics. The prerequisites for this course would include an introduction to econometrics. This book is also useful to applied macroeconomic researchers interested in learning how recent advances in time-series econometrics can be used to estimate and test structural macroeconomic models.

# Contents

<b>1</b>	<b>INTRODUCTION</b>	<b>1</b>
<b>2</b>	<b>STOCHASTIC PROCESSES</b>	<b>5</b>
2.1	Review of Probability Theory . . . . .	5
2.2	Stochastic Processes . . . . .	7
2.3	Conditional Expectations . . . . .	8
2.4	Stationary Stochastic Processes . . . . .	12
2.5	Conditional Heteroskedasticity . . . . .	16
2.6	Martingales and Random Walks . . . . .	18
2.A	A Review of Measure Theory . . . . .	19
2.B	Convergence in Probability . . . . .	29
2.B.1	Convergence in Distribution . . . . .	30
2.B.2	Propositions 2.2 and 2.3 for Infinite Numbers of R.V.'s (Incomplete) . . . . .	31
<b>3</b>	<b>FORECASTING</b>	<b>33</b>
3.1	Projections . . . . .	33
3.1.1	Definitions and Properties of Projections . . . . .	33
3.1.2	Linear Projections and Conditional Expectations . . . . .	35
3.2	Some Applications of Conditional Expectations and Projections . . . . .	39
3.2.1	Volatility Tests . . . . .	39
3.2.2	Parameterizing Expectations . . . . .	41
3.2.3	Noise Ratio . . . . .	42
3.A	Introduction to Hilbert Space . . . . .	43
3.A.1	Vector Spaces . . . . .	44
3.A.2	Hilbert Space . . . . .	46
<b>4</b>	<b>ARMA AND VECTOR AUTOREGRESSION REPRESENTATIONS</b>	<b>53</b>
4.1	Autocorrelation . . . . .	53
4.2	The Lag Operator . . . . .	54
4.3	Moving Average Representation . . . . .	55
4.4	The Wold Representation . . . . .	57

4.5	Autoregression Representation . . . . .	61
4.5.1	Autoregression of Order One . . . . .	61
4.5.2	The $p$ -th Order Autoregression . . . . .	63
4.6	ARMA . . . . .	64
4.7	Fundamental Innovations . . . . .	65
4.8	The Spectral Density . . . . .	67
<b>5</b>	<b>STOCHASTIC REGRESSORS IN LINEAR MODELS</b>	<b>70</b>
5.1	The Conditional Gauss Markov Theorem . . . . .	72
5.2	Unconditional Distributions of Test Statistics . . . . .	77
5.3	The Law of Large Numbers . . . . .	79
5.4	Convergence in Distribution and Central Limit Theorem . . . . .	80
5.5	Consistency and Asymptotic Distributions of OLS Estimators . . . . .	85
5.6	Consistency and Asymptotic Distributions of IV Estimators . . . . .	86
5.7	Nonlinear Functions of Estimators . . . . .	87
5.8	Remarks on Asymptotic Theory . . . . .	88
5.9	Monte Carlo Methods . . . . .	89
5.9.1	Random Number Generators . . . . .	89
5.9.2	Estimators . . . . .	91
5.9.3	Tests . . . . .	92
5.10	Bootstrap . . . . .	93
5.A	Weakly dependence process . . . . .	99
5.A.1	Independent Process . . . . .	100
5.A.2	Mixing Process . . . . .	100
5.A.3	Martingale Difference Process . . . . .	102
5.A.4	Mixingale Process . . . . .	103
5.A.5	Near-Epoch Dependent (NED) Process . . . . .	104
5.B	Functional Central Limit Theorem . . . . .	104
5.B.1	Central Limit Theorem . . . . .	106
5.B.2	Functional Central Limit Theorem . . . . .	108
5.C	Consistency of Bootstrap . . . . .	109
5.D	Hansen's (1999) Grid Bootstrap . . . . .	110
5.E	Monte Carlo Methods with GAUSS . . . . .	111
5.E.1	Random Number Generators . . . . .	111
5.E.2	Estimators . . . . .	113
5.E.3	A Pitfall in Monte Carlo Simulations . . . . .	113
5.E.4	An Example Program . . . . .	115
<b>6</b>	<b>ESTIMATION OF THE LONG-RUN COVARIANCE MATRIX</b>	<b>124</b>
6.1	Serially Uncorrelated Variables . . . . .	125
6.2	Serially Correlated Variables . . . . .	126
6.2.1	Unknown Order of Serial Correlation . . . . .	126

6.2.2	Known Order of Serial Correlation . . . . .	131
<b>7</b>	<b>TESTING LINEAR FORECASTING MODELS</b>	<b>135</b>
7.1	Forward Exchange Rates . . . . .	135
7.2	The Euler Equation . . . . .	139
7.3	The Martingale Model of Consumption . . . . .	141
7.4	The Linearized Euler Equation . . . . .	142
7.5	Optimal Taxation . . . . .	144
7.6	Tests of Forecast Accuracy . . . . .	145
7.6.1	The Monetary Model of Exchange Rates . . . . .	146
7.7	The Taylor Rule Model of Exchange Rates . . . . .	147
7.7.1	? . . . . .	150
7.7.2	? and ? . . . . .	151
<b>8</b>	<b>VECTOR AUTOREGRESSION TECHNIQUES</b>	<b>156</b>
8.1	OLS Estimation . . . . .	157
8.2	Granger Causality . . . . .	159
8.3	The Impulse Response Function . . . . .	162
8.4	Forecast error decomposition . . . . .	165
8.5	Structural VAR Models . . . . .	166
8.6	Identification . . . . .	169
8.6.1	Short-Run Restrictions for Structural VAR . . . . .	169
8.6.2	Identification of block recursive systems . . . . .	171
8.6.3	Two-step ML estimation . . . . .	172
8.A	Asymptotic Interval Method . . . . .	174
8.B	Bias-Corrected Bootstrap Method . . . . .	176
8.C	Monte Carlo Integration . . . . .	178
<b>9</b>	<b>GENERALIZED METHOD OF MOMENTS</b>	<b>182</b>
9.1	Asymptotic Properties of GMM Estimators . . . . .	182
9.1.1	Moment Restriction and GMM Estimators . . . . .	182
9.1.2	Asymptotic Distributions of GMM Estimators . . . . .	184
9.1.3	Optimal Choice of the Distance Matrix . . . . .	185
9.1.4	A Chi-Square Test for the Overidentifying Restrictions . . . . .	186
9.2	Special Cases . . . . .	186
9.2.1	Ordinary Least Squares . . . . .	187
9.2.2	Linear Instrumental Variables Regressions . . . . .	187
9.2.3	Linear GMM estimator . . . . .	188
9.2.4	Nonlinear Instrumental Variables Estimation . . . . .	189
9.3	Important Assumptions . . . . .	190
9.3.1	Stationarity . . . . .	191
9.3.2	Identification . . . . .	192

9.4	Extensions . . . . .	192
9.4.1	Sequential Estimation . . . . .	192
9.4.2	GMM with Deterministic Trends . . . . .	194
9.4.3	Other GMM Estimators . . . . .	194
9.5	Hypothesis Testing and Specification Tests . . . . .	195
9.6	Numerical Optimization . . . . .	197
9.7	The Optimal Choice of Instrumental Variables . . . . .	198
9.8	Small Sample Properties . . . . .	199
9.9	Weak Identification . . . . .	201
9.10	Identification Robust Methods . . . . .	202
9.A	Asymptotic Theory for GMM . . . . .	205
9.A.1	Asymptotic Properties of Extremum Estimators . . . . .	206
9.A.2	Consistency of GMM Estimators . . . . .	208
9.A.3	A Sufficient Condition for the Almost Sure Uniform Convergence	209
9.A.4	Asymptotic Distributions of GMM Estimators . . . . .	214
9.B	The Conditional Likelihood Ratio Statistic . . . . .	218
9.C	A Procedure for Hansen's $J$ Test (GMM.EXP) . . . . .	220
<b>10</b>	<b>EMPIRICAL APPLICATIONS OF GMM</b>	<b>229</b>
10.1	Euler Equation Approach . . . . .	229
10.2	Habit Formation and Durability . . . . .	232
10.3	State-Nonseparable Preferences . . . . .	234
10.4	Time Aggregation . . . . .	235
10.5	Multiple-Goods Models . . . . .	236
10.6	Seasonality . . . . .	239
10.7	Monetary Models . . . . .	240
10.8	Calculating Standard Errors for Estimates of Standard Deviation, Correlation, and Autocorrelation . . . . .	241
10.9	Dynamic Stochastic General Equilibrium Models and GMM Estimation	243
10.10	GMM and an ARCH Process . . . . .	248
10.11	Estimation and Testing of Linear Rational Expectations Models . . .	251
10.11.1	The Nonlinear Restrictions . . . . .	252
10.11.2	Econometric Methods . . . . .	255
10.12	GMM for Consumption Euler Equations with Measurement Error . .	257
<b>11</b>	<b>EXTREMUM ESTIMATORS</b>	<b>270</b>
11.1	Asymptotic Properties of Extremum Estimators . . . . .	270
11.1.1	Convergence . . . . .	271
11.1.2	Identification . . . . .	271
11.2	Two Classes of Extremum Estimators . . . . .	271
11.2.1	Minimum Distance Estimators . . . . .	271
11.2.2	M-Estimators . . . . .	272

11.3	Examples of Minimum Distance Estimators . . . . .	272
11.3.1	Two-Step Minimum Distance Estimators . . . . .	272
11.3.2	Two-Step Minimum Distance Estimation with Impulse Responses	273
11.3.3	Minimum Distance to Estimate Data Statistics . . . . .	276
11.4	The Kalman Filter . . . . .	277
11.4.1	Evaluation of the Likelihood Function using the Kalman Filter	281
11.A	Examples of State-Space Representations . . . . .	283
<b>12</b>	<b>INTRODUCTION TO BAYESIAN APPROACH</b>	<b>286</b>
12.1	Bayes Theorem . . . . .	287
12.2	Parameter Estimates . . . . .	288
12.3	Bayesian Intervals and Regions . . . . .	288
12.4	Posterior Odds Ratio and Hypothesis Testing . . . . .	289
12.A	Numerical Approximation Methods . . . . .	292
12.A.1	Importance Sampling . . . . .	292
12.A.2	Markov Chain Monte Carlo . . . . .	293
12.B	Application of the MCMC methods . . . . .	296
<b>13</b>	<b>UNIT ROOT NONSTATIONARY PROCESSES</b>	<b>301</b>
13.1	Definitions . . . . .	302
13.2	Decompositions . . . . .	303
13.3	Tests for the Null of Difference Stationarity . . . . .	305
13.3.1	Dickey-Fuller Tests . . . . .	306
13.3.2	Said-Dickey Test . . . . .	307
13.3.3	Phillips-Perron Tests . . . . .	309
13.3.4	Park's $J$ Tests . . . . .	310
13.4	Testing the Null of Stationarity . . . . .	311
13.5	Near Observational Equivalence . . . . .	312
13.6	Asymptotics for Unit Root Processes . . . . .	313
13.6.1	Continuous Mapping Theorem . . . . .	313
13.6.2	Dickey-Fuller test with serially uncorrelated disturbances . . .	314
13.6.3	Said-Dickey test with serially correlated disturbances . . . . .	318
13.6.4	Phillips-Perron test . . . . .	324
13.A	Asymptotic Theory . . . . .	331
13.A.1	Functional Central Limit Theorem . . . . .	331
13.B	Procedures for Unit Root Tests . . . . .	331
13.B.1	Said-Dickey Test (ADF.EXP) . . . . .	331
13.B.2	Park's $J$ Test (JPQ.EXP) . . . . .	332
13.B.3	Park's $G$ Test (GPQ.EXP) . . . . .	333

<b>14 COINTEGRATING AND SPURIOUS REGRESSIONS</b>	<b>338</b>
14.1 Definitions . . . . .	339
14.2 Exact Finite Sample Properties of Regression Estimators . . . . .	342
14.2.1 Spurious Regressions . . . . .	342
14.2.2 Cointegrating Regressions . . . . .	346
14.3 Large Sample Properties . . . . .	347
14.3.1 Canonical Cointegrating Regression . . . . .	348
14.3.2 Estimation of Long-Run Covariance Parameters . . . . .	351
14.4 Tests for the Null Hypothesis of No Cointegration . . . . .	352
14.5 Tests for the Null Hypothesis of Cointegration . . . . .	354
14.6 Generalized Method of Moments and Unit Roots . . . . .	355
14.A Procedures for Cointegration Tests . . . . .	356
14.A.1 Park's CCR and $H$ Test (CCR.EXP) . . . . .	356
14.A.2 Park's $I$ Test (IPQ.EXP) . . . . .	358
14.B Weak Convergence to Stochastic Integral . . . . .	359
<b>15 ECONOMIC MODELS AND COINTEGRATING REGRESSIONS</b>	<b>363</b>
15.1 The Permanent Income Hypothesis of Consumption . . . . .	364
15.2 Present Value Models of Asset Prices . . . . .	367
15.3 Applications to Money Demand Functions . . . . .	369
15.4 The Cointegration Approach to Estimating Preference Parameters . . . . .	369
15.4.1 The Time Separable Addilog Utility Function . . . . .	371
15.4.2 The Time Nonseparable Addilog Utility Function . . . . .	375
15.4.3 Engel's Law and Cointegration . . . . .	380
15.5 The Cointegration-Euler Equation Approach . . . . .	383
15.5.1 The Economy . . . . .	386
15.5.2 The 2-Step Estimation Method . . . . .	390
15.5.3 Measuring Intertemporal Substitution: The Role of Durable Goods . . . . .	392
15.6 Purchasing Power Parity . . . . .	392
<b>16 VECTOR AUTOREGRESSIONS WITH UNIT ROOT NONSTA- TIONARY PROCESSES</b>	<b>400</b>
16.1 Identification on Structural VAR Models . . . . .	401
16.1.1 Long-Run Restrictions for Structural VAR Models . . . . .	401
16.1.2 Short-run and Long-Run Restrictions for Structural VAR Models	403
16.2 Representations for the Cointegrated System . . . . .	406
16.2.1 Vector Moving Average Representation . . . . .	406
16.2.2 Phillips' Triangular Representation . . . . .	408
16.2.3 Vector Error Correction Model Representation . . . . .	410
16.2.4 Common Trend Representation . . . . .	411
16.3 Long-Run Restrictions on Phillips' Triangular Representation . . . . .	412

16.3.1	Long-run Restrictions and VECM . . . . .	415
16.3.2	Identification of Permanent Shocks . . . . .	416
16.3.3	Impulse Response Functions . . . . .	418
16.3.4	Forecast-Error Variance Decomposition . . . . .	420
16.3.5	Summary . . . . .	421
16.4	Structural Vector Error Correction Models . . . . .	422
16.5	An Exchange Rate Model with Sticky Prices . . . . .	424
16.6	The System Method . . . . .	429
16.7	Tests for the Number of Cointegrating Vectors . . . . .	430
16.8	How Should an Estimation Method be Chosen? . . . . .	432
16.8.1	Are Short-Run Dynamics of Interest? . . . . .	432
16.8.2	The Number of the Cointegrating Vectors . . . . .	433
16.8.3	Small Sample Properties . . . . .	434
16.A	Estimation of the Model with Long-Run Restrictions . . . . .	435
16.B	Monte Carlo Integration . . . . .	440
16.C	Johansen's Maximum Likelihood Estimation and Cointegration Rank Tests . . . . .	442
<b>17</b>	<b>PANEL AND CROSS-SECTIONAL DATA</b>	<b>451</b>
17.1	Generalized Method of Moments . . . . .	451
17.2	Tests of Risk Sharing . . . . .	453
17.3	Decreasing Relative Risk Aversion and Risk Sharing . . . . .	455
17.4	Euler Equation Approach . . . . .	457
17.5	Panel Unit Root Tests . . . . .	458
17.6	Cointegration and Panel Data . . . . .	460
<b>A</b>	<b>INTRODUCTION TO GAUSS</b>	<b>465</b>
A.1	Starting and Exiting GAUSS . . . . .	465
A.1.1	The Windows Version . . . . .	465
A.1.2	The DOS Version . . . . .	465
A.2	Running a Program Stored in a File from the COMMAND Mode . . . . .	466
A.3	Editing a File . . . . .	466
A.4	Rules of Syntax . . . . .	466
A.4.1	Statements . . . . .	466
A.4.2	Case . . . . .	466
A.4.3	Comments . . . . .	467
A.4.4	Symbol Names . . . . .	467
A.5	Reading and Storing Data . . . . .	467
A.6	Operators . . . . .	467
A.6.1	Operators for Matrix Manipulations . . . . .	467
A.6.2	Numeric Operators . . . . .	469
A.7	Commands . . . . .	470

A.7.1	Functions . . . . .	470
A.7.2	Printing . . . . .	471
A.7.3	Preparing an Output File . . . . .	472
A.8	Procedure . . . . .	472
A.9	Examples . . . . .	472
<b>B</b>	<b>COMPLEX VARIABLES, THE SPECTRUM, AND LAG OPERATOR</b>	<b>473</b>
B.1	Complex Variables . . . . .	474
B.1.1	Complex Numbers . . . . .	474
B.1.2	Analytic Functions . . . . .	475
B.2	Hilbert Spaces on $\mathbb{C}$ . . . . .	480
B.3	Spectrum . . . . .	481
B.4	Lag Operators . . . . .	484
<b>C</b>	<b>ANSWERS TO SELECTED QUESTIONS</b>	<b>487</b>

# List of Tables

5.1	Dependence between $X_t$ and $X_{t+m}$ . . . . .	103
13.1	Critical Values of Park's $J(p, q)$ Tests for the Null of Difference Stationarity . . . . .	310
13.2	Probability of smaller values . . . . .	334
14.1	Critical Values of Park's $I(p, q)$ Tests for Null of No Cointegration . . . . .	354
C.1	GMM Results . . . . .	498
C.2	Data moments and model moments . . . . .	498
C.3	GPQ tests . . . . .	503
C.4	ADF tests . . . . .	503
C.5	CCR estimation and H(p,q) tests . . . . .	504

# Chapter 1

## INTRODUCTION

The word “structural” has various different meanings in econometrics. In this book, “structural” means that explicit assumptions are made in econometric methods so that estimators or test statistics can be interpreted in terms of an economic model (or models). In some cases, some properties of the estimators and test statistics are known when they are applied to data generated from an economic model. We then use the economic model to interpret empirical results obtained by applying the econometric tools to real data. This is important because an economic model is used to analyze causal relationships between economic variables, and understanding causal relationships is essential for policy evaluations and forecasting.

As a very simple example, consider a model of demand for a good:

$$(1.1) \quad Q_t^d = a - bP_t + e_t,$$

where  $P_t$  is the price and  $Q_t^d$  is the market quantity demanded. In this model  $a$  and  $b$  are constants and  $e_t$  is the demand shock. The model assumes that the observed quantity,  $Q_t$ , is equated with  $Q_t^d$ ,  $P_t$  is nonstochastic,  $e_t$  has mean zero,  $E(e_t^2) = \sigma^2$ , and  $E(e_t e_s) = 0$  if  $t \neq s$ . With these assumptions the Gauss-Markov Theorem can be applied to this model. If the Ordinary Least Squares (OLS) slope coefficient estimator

is applied to data of  $Q_t$  and  $P_t$  for  $t = 1, \dots, T$  in this model, then the estimator is the Best Linear Unbiased Estimator (BLUE) for the demand slope coefficient,  $b$ .

One benefit of having this structural model is that we know exactly what the limitations are when we interpret OLS results applied to real data in terms of the model. This knowledge is helpful because we can then study how to improve our econometric methods for better interpretation of data.

For example, consider the assumption made in the model that  $P_t$  is nonstochastic. This assumption is sometimes motivated by saying that the price is taken as given by the individual market participants. It is easy to see that this motivation is problematic by considering the supply side of the market. Consider a model of supply of the good:

$$(1.2) \quad Q_t^s = c + dP_t + u_t,$$

where  $Q_t^s$  the market quantity supplied and  $u_t$  is the supply shock. In equilibrium, the observed quantity,  $Q_t$ , is equal to  $Q_t^d$  and  $Q_t^s$ . Equating the right hand sides of (1.1) and (1.2), and solving for  $P_t$ , we obtain

$$(1.3) \quad P_t = \frac{1}{d+b}(a - c + e_t - u_t).$$

Hence  $P_t$  is stochastic. Moreover, (1.3) makes it clear that  $P_t$  is correlated with  $e_t$  and  $u_t$ . This means that the OLS slope coefficient estimator is not even a consistent estimator for  $b$  or  $d$  as discussed in Chapter 5. This leads us to consider an improved econometric method, an instrumental variable method, for example.

The structural demand model tells us under what assumptions we can interpret the OLS slope estimator as an unbiased estimator for  $b$ . By studying the assumptions, we can see what will happen when they are violated. This process leads to better

econometric methods.

Another consideration is the trend observed in most aggregate data. The demand model with trends leads to cointegrating regressions as discussed in Chapter 15.

Instead of starting with a demand function, one can start with a utility function as in the Euler Equation Approach discussed in Chapter 10. When data contain trends, cointegrating regressions can be used to estimate preference parameters, and this Cointegration Approach can be combined with the Euler Equation Approach as described in Chapter 15.

We do not claim that structural econometrics as defined here is better than non-structural econometrics. They are tools that serve different purposes. Just as it does not make sense to argue whether a hammer is better than a screwdriver, we cannot compare structural and non-structural econometrics without specifying the purposes. For the purpose of summarizing data properties and finding stylized facts, non-structural econometrics is better. This purpose is obviously very important in economics. Using a structural econometric model that enforces a certain economic interpretation is not good for this purpose. On the other hand, after finding stylized facts with non-structural econometrics, one may wish to understand causal relationships that explain stylized facts and make policy recommendations based on causal relationships. For that purpose, structural econometrics is better than non-structural econometrics.

Similarly, we do not claim that the definition of “structural” in this book is better than other definitions. For example, Hendry (1993) and Ericsson (1995) define a structural model as an econometric model that is invariant over extensions of

the information set in time, interventions or variables. Their definition is useful for their purpose of finding invariant relationships between economic variables in data, but cannot be used for our purpose of interpreting empirical results in terms of an economic model.

### References

- ERICSSON, N. R. (1995): “Conditional and Structural Error Correction Models,” *Journal of Econometrics*, 69, 159–171.
- HENDRY, D. F. (1993): “The Roles of Economic Theory and Econometrics in Time Series Economics,” Invited paper presented at the Econometric Society European Meeting, Uppsala, Sweden.

# Chapter 2

## STOCHASTIC PROCESSES

In most macroeconomic models, expectations conditional on information sets are used to model the forecasting conducted by economic agents. Economic agents typically observe stochastic processes of random variables (collections of random variables indexed by time) to form their information sets. This chapter defines the concepts of conditional expectations and information sets for the case of a finite number of elements in the probability space.<sup>1</sup>

### 2.1 Review of Probability Theory

Since the probability statements made in asymptotic theory involve infinitely many random variables instead of just one random variable, it is important to understand basic concepts in probability theory. Thus, we first review those basic concepts.

Imagine that we are interested in making probability statements about a set of the states of the world (or a probability space), which we denote by  $S$ . For the purpose of understanding concepts, nothing is lost by assuming that there is a finite number of states of the world. Hence we adopt the simplifying assumption that  $S$

---

<sup>1</sup>For the general probability space, these concepts are defined with measure theory (see Appendix 2.A). For our purpose, it is not necessary for the reader to understand measure theory.

consists of  $N$  possible states:  $S = \{s_1, \dots, s_N\}$ . We assign a probability  $\pi_i = Pr(s_i)$  to  $s_i$ , depending on how likely  $s_i$  is to occur. It is assumed that  $\sum_{i=1}^N \pi_i = 1$  and  $0 \leq \pi_i \leq 1$  for all  $i$ . Note that we can now assign a probability to all subsets of  $S$ . For example, let  $\Lambda$  be  $\{s_1, s_2\}$ . Then the probability that the true  $s$  is in  $\Lambda$  is denoted by  $Pr(s \in \Lambda)$ , where  $Pr(s \in \Lambda) = \pi_1 + \pi_2$ .

**Example 2.1** The state of the world consists of  $s_1$ : it rains tomorrow, and  $s_2$ : it does not rain tomorrow. According to a weather forecast,  $\pi_1 = 0.8$  and  $\pi_2 = 0.2$ . ■

A *random variable* assigns a real value to each element  $s$  in  $S$  (that is, it is a real valued function on  $S$ ). Let  $X(s)$  be a random variable (we will often omit the arguments  $s$ ). For a real value  $x$ , the *distribution function*,  $F(x)$ , of the random variable is defined by  $F(x) = Pr\{s : X(s) \leq x\}$ . A random variable is assigned an expected value or mean value

$$(2.1) \quad E(X) = \sum_{i=1}^N X(s_i)\pi_i.$$

**Example 2.2** Continuing Example 2.1, let  $X(s)$  be the profit of an umbrella seller in terms of dollars with  $X(s_1) = 100$  and  $X(s_2) = 10$ . Then  $E(X) = 100 \times 0.8 + 10 \times 0.2 = 82$ . The distribution function  $F(x)$  is given by  $F(x) = 0$  for  $x < 10$ ,  $F(x) = 0.2$  for  $10 \leq x < 100$ , and  $F(x) = 1$  for  $x \geq 100$ . ■

A random vector is a vector of random variables defined on the set of states. For a  $k$ -dimensional random vector  $\mathbf{X}(s) = (X_1(s), \dots, X_k(s))'$ , the joint distribution function  $F$  is defined by

$$(2.2) \quad F(x_1, \dots, x_k) = Pr[X_1 \leq x_1, \dots, X_k \leq x_k].$$

## 2.2 Stochastic Processes

A collection of random variables indexed by time is called a *stochastic process* or a *time series*. Let  $X_t(s)$  be a random variable, then a collection  $\{X_t : X_0(s), X_1(s), X_2(s), \dots\}$  is a univariate stochastic process. It is sometimes more convenient to consider a stochastic process that starts from the infinite past,  $\{\dots, X_{-2}(s), X_{-1}(s), X_0(s), X_1(s), X_2(s), \dots\}$ . In general,  $\{X_t(s) : t \in A\}$  for any set  $A$  is a stochastic process. If  $A$  is a set of integers, then time is *discrete*. It is also possible to consider a *continuous* time stochastic process for which the time index takes any real value. For example,  $\{X_t(s) : t \text{ is a nonnegative real number}\}$ . Here, if we take  $X_t$  as a random vector rather than a random variable, then it is a vector stochastic process. When we observe a sample of size  $T$  of a random variable  $X$  or a random vector  $\mathbf{X} : \{X_1, \dots, X_T\}$ , it is considered a particular realization of a part of the stochastic process.

Note that once  $s$  is determined, the complete history of the stochastic process becomes known. For asymptotic theory, it is usually easier to think about the stochastic nature of economic variables this way rather than the alternative, which is to consider a probability space for each period based on independent disturbances.

In a sense, the stochastic process modeled in this manner is deterministic because everything is determined at the beginning of the world when  $s$  is determined. However, this does not mean that there is no uncertainty to economic agents because they do not learn  $s$  until the end of the world. In order to illustrate this, let us consider the following example:

**Example 2.3** Imagine an economy with three periods and six states of the world. The world begins in period 0. We observe two variables, aggregate output ( $Y_t$ ) and

the interest rate ( $i_t$ ), in period 1 and period 2. The world ends in period 2. In each period,  $Y_t$  can take two values, 150 and 300, and  $i_t$  can take two values, 5 and 10. We assume that  $i_2$  is equal to  $i_1$  in all states of the world, and that the  $i_1 = 5$  in all states in which  $Y_1 = 150$ . The six states of the world can be described by the triplet,  $[Y_1, i_1, Y_2]$ .

The six states of the world are,  $s_1 = [300, 10, 300]$ ,  $s_2 = [300, 10, 150]$ ,  $s_3 = [300, 5, 300]$ ,  $s_4 = [300, 5, 150]$ ,  $s_5 = [150, 5, 300]$ , and  $s_6 = [150, 5, 150]$ . To illustrate,  $s_1$  means the economy is in a boom (higher output level) with a high interest rate in period 1, and is in a boom in period 2. In period 0, the economic agents assign a probability to each state:  $\pi_1 = 0.20$ ,  $\pi_2 = 0.10$ ,  $\pi_3 = 0.15$ ,  $\pi_4 = 0.05$ ,  $\pi_5 = 0.15$ , and  $\pi_6 = 0.35$ . Unconditional expected values are taken with these probabilities. ■

In this example, let  $\mathbf{X}_t(s) = [Y_t(s), i_t(s)]$ . Then  $[\mathbf{X}_1(s), \mathbf{X}_2(s)]$  is a stochastic process. The whole history of the process is determined at the beginning of the world when  $s$  is chosen, and the agents learn which state of the world they are in at the end of the world in period 2. In period 1, however, the agents only have partial information as to which state of the world is true. For example, if  $Y_1 = 300$  and  $i_1 = 5$ , the agents learn that they are in either  $s_3$  or  $s_4$ , but cannot tell which one they are in until they observe  $Y_2$  in period 2.

## 2.3 Conditional Expectations

Economic agents use available information to learn the true state of the world and make forecasts of future economic variables. This forecasting process can be modeled using conditional expectations.

Information can be modeled as a partition of  $S$  into mutually exclusive subsets:

$\mathcal{F} = \{\Lambda_1, \dots, \Lambda_M\}$  where  $\Lambda_1 \cup \dots \cup \Lambda_M = S$ , and  $\Lambda_j \cap \Lambda_k = \emptyset$  if  $j \neq k$ . For example, information  $\mathcal{F}$  consists of two subsets:  $\mathcal{F} = \{\Lambda_1, \Lambda_2\}$ . Here  $\Lambda_1 = \{s_1, \dots, s_M\}$ , and  $\Lambda_2 = \{s_{M+1}, \dots, s_N\}$ . The information represented by  $\mathcal{F}$  tells us which  $\Lambda$  contains the true  $s$ , but no further information is given by  $\mathcal{F}$ .

In this situation, once agents obtain the information represented by  $\mathcal{F}$ , then the agents know which subset contains the true  $s$ , and they can assign a probability of zero to all elements in the other subset. There is no reason to change the ratios of probabilities assigned to the elements in the subset containing the true  $s$ . Nonetheless, the absolute level of each probability should be increased, so that the probabilities add up to one. The probability conditional on the information that the true  $s$  is in  $\Lambda_j$  is denoted by  $Pr\{s_i | s \in \Lambda_j\}$ . The considerations given above lead to the following definition of conditional probability:

$$(2.3) \quad Pr\{s_i | s \in \Lambda_j\} = \frac{Pr\{s_i\}}{Pr\{s \in \Lambda_j\}},$$

when  $s_i$  is in  $\Lambda_j$ . Here each probability is scaled by the probability of the subset containing the true  $s$ , so that the probabilities add up to one.

We use conditional probability to define the *conditional expectation*. The expectation of a random variable  $Y$  conditional on the information that the true  $s$  is in  $\Lambda_j$  is

$$(2.4) \quad E(Y | s \in \Lambda_j) = \sum_{s \in \Lambda_j} Y(s) \frac{Pr\{s_i\}}{Pr\{s \in \Lambda_j\}},$$

where the summation is taken over all  $s$  in  $\Lambda_j$ .

It is convenient to view the conditional expectation as a random variable. For this purpose, the conditional expectation needs to be defined over all  $s$  in  $S$ , not just for  $s$  in a particular  $\Lambda_j$ . Given each  $s$ , we first find out which  $\Lambda$  contains  $s$ .

When  $\Lambda_j$  contains  $s$ , the expected value of  $Y$  conditional on  $\mathcal{F}$  for  $s$  is given by  $E(Y|\mathcal{F})(s) = E(Y|s \in \Lambda_j)$ .

Instead of a partition, we can use a random variable or a random vector to describe information. Consider information represented by a partition  $\mathcal{F} = \{\Lambda_1, \dots, \Lambda_M\}$ . Consider the set  $I$ , which consists of all random variables that take the same value for all elements in each  $\Lambda_j$ :  $I = \{X(s) : X(s_i) = X(s_k) \text{ if } s_i \in \Lambda_j \text{ and } s_k \in \Lambda_j \text{ for all } i, j, k\}$ . Then the information set  $I$  represents the same information as  $\mathcal{F}$  does. A random variable  $X$  is said to be in this information set, when  $X(s_i) = X(s_k)$  if both  $s_i$  and  $s_k$  are in the same  $\Lambda_j$ .<sup>2</sup> A random vector  $\mathbf{X}$  is said to be in this information set when each element of  $\mathbf{X}$  is in the information set.

If  $X$  is in the information set  $I$ , and if  $X$  takes on different values for all different  $\Lambda$  ( $X(s_i) \neq X(s_k)$  when  $s_i$  and  $s_k$  are not in the same  $\Lambda$ ), then we say that the random variable  $X$  generates the information set  $I$ . If a random vector  $\mathbf{X}$  is in  $I$ , and if at least one element of  $\mathbf{X}$  takes on different values for different  $\Lambda$ , then the random vector  $\mathbf{X}$  is said to generate the information set  $I$ . When a random variable  $X$  or a random vector  $\mathbf{X}$  generates the information set  $I$ , which represents the same information as a partition  $\mathcal{F}$ , we define  $E(Y|I)$  as  $E(Y|\mathcal{F})$ . If  $I$  is generated by  $X$ , we define  $E(Y|X) = E(Y|I)$ ; and if  $I$  is generated by a random vector  $\mathbf{X}$ , we define  $E(Y|\mathbf{X}) = E(Y|I)$ . It should be noted that  $E(Y|I)$  is in the information set  $I$ .

**Example 2.4** Continuing Example 2.3, let  $I$  be the information set generated by  $X_1 = (Y_1, i_1)$ , and let  $\mathcal{F}$  be the partition that represents the same information as  $I$ . Then  $\mathcal{F} = \{\Lambda_1, \Lambda_2, \Lambda_3\}$ , where  $\Lambda_1 = \{s_1, s_2\}$ ,  $\Lambda_2 = \{s_3, s_4\}$ , and  $\Lambda_3 = \{s_5, s_6\}$ .

<sup>2</sup>In the terminology of probability theory, we consider a set of all possible unions of  $\Lambda$ 's in  $\mathcal{F}$  plus the null set. This set of subsets of  $S$  is called a  $\sigma$ -field, and used to describe information. When a random variable  $X$  is in the information set  $I$ , we say that the random variable is measurable with respect to this  $\sigma$ -field.

Using (2.3),  $Pr(s_1|s \in \Lambda_1) = \frac{0.20}{0.20+0.10} = \frac{2}{3}$  and  $Pr(s_2|s \in \Lambda_1) = \frac{0.10}{0.20+0.10} = \frac{1}{3}$ . Hence  $E(Y_2|s \in \Lambda_1) = 300 \times \frac{2}{3} + 150 \times \frac{1}{3} = 250$ . Similarly,  $Pr(s_3|s \in \Lambda_2) = \frac{3}{4}$ ,  $Pr(s_4|s \in \Lambda_2) = \frac{1}{4}$ ,  $Pr(s_5|s \in \Lambda_3) = \frac{3}{10}$ ,  $Pr(s_6|s \in \Lambda_3) = \frac{7}{10}$ ,  $E(Y_2|s \in \Lambda_2) = 262.5$ , and  $E(Y_2|s \in \Lambda_3) = 195$ . Hence the random variable  $E(Y_2|I)$  is given by

$$(2.5) \quad E(Y_2|I)(s) = \begin{cases} 250 & \text{if } s \in \Lambda_1 \\ 262.5 & \text{if } s \in \Lambda_2 \\ 195 & \text{if } s \in \Lambda_3 \end{cases} .$$

■

**Example 2.5** Continuing Example 2.4, consider the information set  $J$  which is generated by  $Y_1$ . Then  $J$  is a smaller information set than  $I$  in the sense that  $J \subset I$ . Similar computations as those in Example 2.4 yield

$$(2.6) \quad E(Y_2|J)(s) = \begin{cases} 255 & \text{if } s \in \{s_1, s_2, s_3, s_4\} \\ 195 & \text{if } s \in \{s_5, s_6\} \end{cases} .$$

■

Two properties of conditional expectations are very important in macroeconomics.

**Proposition 2.1 (Properties of Conditional Expectations)**

- (a) If a random variable  $Z$  is in the information set  $I$ , then

$$(2.7) \quad E(ZY|I) = ZE(Y|I)$$

for any random variables  $Y$  with finite  $E(|Y|)$ , assuming that  $E(|ZY|)$  is finite.

- (b) *The Law of Iterated Expectations:* If the information set  $J$  is smaller than the information set  $I$  ( $J \subset I$ ), then

$$(2.8) \quad E(Y|J) = E[E(Y|I)|J]$$

for any random variable  $Y$  with finite  $E(|Y|)$ .

■

Expectation can be viewed as a special case of conditional expectation in which the information set consists of constants. Since a constant is a random variable which takes the same value for all states of the world, any information set includes all constants. Therefore, the Law of Iterated Expectations implies

$$(2.9) \quad E(Y) = E[E(Y|I)].$$

When we wish to emphasize the difference between expectations and conditional expectations, expectations are called *unconditional expectations*. Relation (2.9) states that an unconditional expected value of a random variable  $Y$  can be computed as an unconditional expected value of the expectation of the random variable conditional on any information set. For a proof of Proposition 2.1 in the general case, see, e.g., Billingsley (1986, Theorem 34.3 and Theorem 34.4).

## 2.4 Stationary Stochastic Processes

A stochastic process  $\{\dots, \mathbf{X}_{-1}, \mathbf{X}_0, \mathbf{X}_1, \dots\}$  is *strictly stationary* if the joint distribution function of  $(\mathbf{X}_t, \mathbf{X}_{t+1}, \dots, \mathbf{X}_{t+h})$  is the same for all  $t = 0, \pm 1, \pm 2, \dots$  and all  $h = 0, 1, 2, \dots$ . A stochastic process  $\{\dots, \mathbf{X}_{-1}, \mathbf{X}_0, \mathbf{X}_1, \dots\}$  is *covariance stationary* (or *weakly stationary*) if  $\mathbf{X}_t$  has finite second moments ( $E(\mathbf{X}_t \mathbf{X}_t') < \infty$ ) and if  $E(\mathbf{X}_t)$  and  $E(\mathbf{X}_t \mathbf{X}_{t-h}') do not depend on the date  $t$  for all  $t = 0, \pm 1, \pm 2, \dots$  and all  $h = 0, 1, 2, \dots$ .$

Because all moments are computed from distribution functions, if  $\mathbf{X}_t$  is strictly stationary and has finite second moments, then it is also covariance stationary. If  $\mathbf{X}_t$  is covariance stationary, then its mean  $E(\mathbf{X}_t)$  and its  $h$ -th *autocovariance*  $\Phi(h) = E[(\mathbf{X}_t - E(\mathbf{X}_t))(\mathbf{X}_{t-h} - E(\mathbf{X}_{t-h}))'] = E(\mathbf{X}_t \mathbf{X}_{t-h}') - E(\mathbf{X}_t)E(\mathbf{X}_{t-h}') does not depend on date  $t$ .$

**Proposition 2.2** If a  $k$ -dimensional vector stochastic process  $\mathbf{X}_t$  is strictly stationary, and if a continuous function  $f(\cdot) : \mathbb{R}^k \mapsto \mathbb{R}^p$  does not depend on date  $t$ , then  $f(\mathbf{X}_t)$  is also strictly stationary.<sup>3</sup> ■

This follows from the fact that the distribution function of  $f(\mathbf{X}_t), f(\mathbf{X}_{t+1}), \dots, f(\mathbf{X}_{t+h})$  is determined by  $f$  and the joint distributions of  $\mathbf{X}_t, \mathbf{X}_{t+1}, \dots, \mathbf{X}_{t+h}$  (see Appendix 2.A). Proposition 2.2 will be used frequently to derive the cointegrating properties of economic variables from economic models in Chapter 15.

The next proposition is for covariance stationary processes.

**Proposition 2.3** If a  $k$ -dimensional vector stochastic process  $\mathbf{X}_t$  is covariance stationary, and if a linear function  $f(\cdot) : \mathbb{R}^k \mapsto \mathbb{R}^p$  does not depend on date  $t$ , then  $f(\mathbf{X}_t)$  is also covariance stationary. ■

This proposition is true because  $f(\mathbf{X}_t)$  has finite second moments, and the first and second moments of  $f(\mathbf{X}_t)$  do not depend on date  $t$ . However, unlike Proposition 2.2 for strictly stationary processes, a nonlinear function of a covariance stationary process may not be covariance stationary. For example, suppose that  $X_t$  is covariance stationary. Imagine that  $X_t$ 's variance is finite but  $E(|X_t|^4) = \infty$ . Consider  $Z_t = f(X_t) = (X_t)^2$ . Then  $Z_t$ 's variance is not finite, and hence  $Z_t$  is not covariance stationary.

In order to model strictly stationary and covariance stationary processes, it is convenient to consider white noise processes. A univariate stochastic process  $\{e_t : t =$

---

<sup>3</sup>This proposition holds for any measurable function  $f(\cdot) : \mathbb{R}^k \mapsto \mathbb{R}^p$  (see Appendix 2.A). The term “measurable” is avoided because this book does not require knowledge of measure theory. All continuous functions are measurable but not vice versa. Thus the continuity condition in Proposition 2.2 is more stringent than necessary. This is not a problem for the purpose of this book because continuous functions are used in all applications of this proposition.

$\dots, -1, 0, 1, \dots$  } is *white noise* if  $E(e_t) = 0$ , and

$$(2.10) \quad E(e_t e_j) = \begin{cases} \sigma^2 & \text{if } t = j \\ 0 & \text{if } t \neq j \end{cases},$$

where  $\sigma$  is a constant. For a vector white noise, we require

$$(2.11) \quad E(\mathbf{e}_t \mathbf{e}'_j) = \begin{cases} \Sigma & \text{if } t = j \\ 0 & \text{if } t \neq j \end{cases},$$

where  $\Sigma$  is a matrix of constants. A white noise process is covariance stationary.

If a process is independent and identically distributed (i.i.d.), then it is strictly stationary. The simplest example of an i.i.d. process is an i.i.d. white noise. A *Gaussian white noise process*  $\{e_t : -\infty < t < \infty\}$  is an i.i.d. white noise process for which  $e_t$  is normally distributed with zero mean. In these definitions,  $e_t$  can be a vector white noise process.

All linear functions of white noise random variables are covariance stationary because of Proposition 2.3. In addition, by Proposition 2.2, all functions of i.i.d. white noise random variables are strictly stationary. A simple example of this case is:

**Example 2.6** Let  $X_t = \delta + e_t$ , where  $e_t$  is a white noise process, and  $\delta$  is a constant. Then  $E(X_t) = \delta$ , and  $X_t$  is covariance stationary. If  $e_t$  is an i.i.d. white noise process, then  $X_t$  is strictly stationary. ■

If  $X_t$  is strictly stationary with finite second moments,  $X_t$  is covariance stationary. Therefore,  $X_t$ 's first and second moments cannot depend on date  $t$ . In empirical work, the easiest case to see that an observed variable is *not* strictly stationary is when a variable's mean shifts upward or downward over time. A simple example of this case is:

**Example 2.7** Let  $X_t = \delta + \theta t + e_t$ , where  $e_t$  is an i.i.d. white noise random variable and  $\delta$  and  $\theta \neq 0$  are constants. Then  $X_t$  is *not* stationary because  $E(X_t) = \delta + \theta t$  depends on time.<sup>4</sup> ■

Strictly stationary and covariance stationary processes can be serially correlated, that is, their  $h$ -th order autocovariances can be nonzero for  $h \neq 0$  as in the next two examples.

**Example 2.8** (*The first order Moving Average Process*) Let  $X_t = \delta + e_t + Be_{t-1}$ , where  $e_t$  is a white noise which satisfies (2.10), and  $\delta$  and  $B$  are constant. This is a moving average process of order 1 (see Chapter 4). Then  $X_t$  is covariance stationary for any  $B$  because of Proposition 2.3.<sup>5</sup>  $E(X_t) = \delta$ , and its  $h$ -th autocovariance is

$$(2.12) \quad \phi_h = E[(X_t - \delta)(X_{t-h} - \delta)] = \begin{cases} \sigma^2(1 + B^2) & \text{if } h = 0 \\ \sigma^2 & \text{if } |h| = 1 \\ 0 & \text{if } |h| > 1 \end{cases} .$$

In this example, if  $e_t$  is an i.i.d. white noise, then  $X_t$  is strictly stationary. ■

**Example 2.9** (*The first order Autoregressive Process*) Consider a process  $X_t$  which is generated from an initial random variable  $X_0$ , where

$$(2.13) \quad X_t = AX_{t-1} + e_t \quad \text{for } t \geq 1,$$

where  $e_t$  is a Gaussian white noise random variable, and  $A$  is a constant. This is an autoregressive process of order 1 (see Chapter 4). If  $|A| < 1$  and  $X_0$  is a normally distributed random variable with mean zero and variance of  $\frac{\text{Var}(e_t)}{1-A^2}$ , then  $X_t$  is strictly

<sup>4</sup>Because  $X_t$  is stationary after removing a deterministic trend in this example, we say that  $X_t$  is trend stationary as we will discuss in Chapter 13. Trend stationarity is a way to model nonstationarity.

<sup>5</sup>Even though  $X_t$  is stationary for any  $B$ , it is often convenient to impose a restriction  $|B| \leq 1$  as explained in Chapter 4.

stationary (see Exercise 2.3). The methods explained in Chapter 4 can be used to show that  $X_t$  is not strictly stationary when  $X_0$ 's distribution is different from the one given above. ■

## 2.5 Conditional Heteroskedasticity

Using conditional expectations, we can define variance and covariance conditional on an information set just as we use unconditional expectations to define (unconditional) variance and covariance. The variance of  $Y$  conditional on an information set  $I$  is

$$(2.14) \quad \text{Var}(Y|I) = E[(Y - E(Y|I))^2|I],$$

and the covariance of  $X$  and  $Y$  conditional on an information set  $I$  is

$$(2.15) \quad \text{Cov}(X, Y|I) = E[(X - E(X|I))(Y - E(Y|I))|I].$$

Consider a stochastic process  $[Y_t : t \geq 1]$ . If the unconditional variance of  $Y_t$ ,  $\text{Var}(Y_t)$ , depends on date  $t$ , then the  $Y_t$  is said to be *heteroskedastic*; if not, it is *homoskedastic*. If  $Y_t$ 's variance conditional on an information set  $I_t$ ,  $\text{Var}(Y_t|I_t)$ , is constant and does not depend on the information set, then  $Y_t$  is said to be *conditionally homoskedastic*; if not, it is *conditionally heteroskedastic*.

**Example 2.10** Let  $Y_t = \delta + h_t e_t$ , where  $e_t$  is an i.i.d. white noise with unit variance ( $E(e_t^2) = 1$ ), and  $\{h_t : -\infty < t < \infty\}$  is a sequence of real numbers. Then the (unconditional) variance of  $Y_t$  is  $h_t$ , and  $Y_t$  is heteroskedastic as long as  $h_t \neq h_j$  for some  $t$  and  $j$ . ■

A heteroskedastic process is not strictly stationary because its variance depends on date  $t$ . It should be noted, however, that a strictly stationary random variable can

be conditionally heteroskedastic. This fact is important because many of the financial time series have been found to be conditionally heteroskedastic. For example, the growth rates of asset prices and foreign exchange rates can be reasonably modeled as strictly stationary processes. However, the volatility of such a growth rate at a point in time tends to be high if it has been high in the recent past. Therefore, such a growth rate is often modeled as a conditionally heteroskedastic process. A popular method to model conditional heteroskedasticity, introduced by Engle (1982), is an *autoregressive conditional heteroskedastic* (ARCH) process. The following is a simple example of an ARCH process.

**Example 2.11** (*An ARCH Process*) Let  $I_t$  be an information set, and  $e_t$  be a univariate stochastic process such that  $e_t$  is in  $I_t$ , and  $E(e_t|I_{t-1}) = 0$ . Assume that

$$(2.16) \quad e_t^2 = \eta + \alpha e_{t-1}^2 + w_t,$$

where  $\eta > 0$ ,  $w_t$  is another white noise process in  $I_t$  with  $E(w_t|I_{t-1}) = 0$  and

$$(2.17) \quad E(w_k w_j | I_t) = \begin{cases} \lambda^2 & \text{if } k = j \\ 0 & \text{if } k \neq j \end{cases},$$

where  $\lambda$  is a constant. Relation (2.16) implies that  $e_t$ 's conditional variance depends on  $I_t$ :

$$(2.18) \quad E(e_t^2 | I_{t-1}) = \eta + \alpha e_{t-1}^2,$$

and thus  $e_t$  is conditionally heteroskedastic.

In order to see whether or not  $e_t$ 's unconditional variance is constant over time, take expectations of both sides of (2.18) to obtain

$$(2.19) \quad E(e_t^2) = \eta + \alpha E(e_{t-1}^2).$$

Hence if the variance of  $e_t$  is a constant  $\sigma^2$ , then  $\sigma^2 = \eta + \alpha\sigma^2$ , and  $\sigma^2 = \frac{\eta}{1-\alpha}$ . Because  $\sigma^2$  is positive, this equation implies that  $\alpha < 1$ . When  $\alpha < 1$ , an ARCH process can be covariance stationary and strictly stationary. ■

## 2.6 Martingales and Random Walks

Consider a stochastic process  $[Y_t : -\infty < t < \infty]$ , and a sequence of information sets  $[\mathbf{I}_t : -\infty < t < \infty]$  that is increasing ( $\mathbf{I}_t \subset \mathbf{I}_{t+1}$ ). If  $Y_t$  is in  $\mathbf{I}_t$  and if

$$(2.20) \quad E(Y_{t+1}|\mathbf{I}_t) = Y_t,$$

then  $Y_t$  is a *martingale* adapted to  $\mathbf{I}_t$ . Rational expectations often imply that an economic variable is a martingale (see Section 3.2). If  $Y_t$  is a martingale adapted to  $\mathbf{I}_t$  and if its conditional variance,  $E((Y_{t+1} - Y_t)^2|\mathbf{I}_t)$ , is constant (that is,  $Y_t$  is conditionally homoskedastic), then  $Y_t$  is a *random walk*.

As we will discuss later in this book, most of the rational expectations models imply that certain variables are martingales. The models typically do not imply that the variables are conditionally homoskedastic, and hence do not imply that they are random walks. However, if the data for the variable does not show signs of conditional heteroskedasticity, then we may test whether or not a variable is a random walk. It is often easier to test whether or not the variable is a random walk than to test whether or not it is a martingale.

Consider a stochastic process  $[e_t : -\infty < t < \infty]$ , and a sequence of information sets  $[\mathbf{I}_t : -\infty < t < \infty]$  which is increasing ( $\mathbf{I}_t \subset \mathbf{I}_{t+1}$ ). If  $e_t$  is in  $\mathbf{I}_t$  and if

$$(2.21) \quad E(e_{t+1}|\mathbf{I}_t) = 0,$$

then  $e_t$  is a *martingale difference sequence* adapted to  $\mathbf{I}_t$ . If  $Y_t$  is a martingale adapted

to  $I_t$ , then  $e_t = Y_t - Y_{t-1}$  is a martingale difference sequence (see Exercise 2.4). A covariance stationary martingale difference sequence is a white noise process (see Exercise 2.5). However, a white noise process may not be a martingale difference sequence for any sequence of information sets. An i.i.d. white noise process is a martingale difference sequence (see Exercise 2.6).

In these definitions, a martingale or a martingale difference sequence can be a vector stochastic process.

## Appendix

### 2.A A Review of Measure Theory

Let  $S$  be an arbitrary nonempty set of points  $s$ . An event is a subset of  $S$ . A set of subsets is called a class. A class  $\mathcal{F}$  of subsets of  $S$  is called a *field* if

- (i)  $S \in \mathcal{F}$ ;
- (ii)  $A \in \mathcal{F}$  implies  $A^c \in \mathcal{F}$ , where  $A^c$  is the complement of  $A$ ;
- (iii)  $A, B \in \mathcal{F}$  implies  $A \cup B \in \mathcal{F}$ .

A class  $\mathcal{F}$  is a  $\sigma$ -field if it is a field and if

- (iv)  $A_1, A_2, \dots \in \mathcal{F}$  implies  $A_1 \cup A_2 \cup \dots \in \mathcal{F}$ .

A *set function* is a real-valued function defined on some class of subsets of  $S$ . A set function  $Pr$  on a field  $\mathcal{F}$  is a *probability measure* if it satisfies these conditions:

- (i)  $0 \leq Pr(A) \leq 1$  for  $A \in \mathcal{F}$ ;
- (ii)  $Pr(\emptyset) = 0, Pr(S) = 1$ ;

(iii) if  $A_1, A_2, \dots$  is a disjoint sequence of  $\mathcal{F}$ -sets and if  $\bigcup_{k=1}^{\infty} A_k \in \mathcal{F}$ , then

$$Pr(\bigcup_{k=1}^{\infty} A_k) = \sum_{k=1}^{\infty} Pr(A_k).$$

If  $\mathcal{F}$  is a  $\sigma$ -field in  $S$  and  $Pr$  is a probability measure on  $\mathcal{F}$ , the triple  $(S, \mathcal{F}, Pr)$  is called a *probability space*. Given a class  $\mathcal{A}$ , consider the class which is the intersection of all  $\sigma$ -fields containing  $\mathcal{A}$ . This class is the smallest  $\sigma$ -field which contains  $\mathcal{A}$ , and is called the  *$\sigma$ -field generated by  $\mathcal{A}$*  and is denoted by  $\sigma(\mathcal{A})$ .

**Proposition 2.A.1** A probability measure on a field has a unique extension to the generated  $\sigma$ -field. ■

In Euclidean  $k$ -space  $\mathbb{R}^k$ , consider the class of the bounded rectangles

$$[\mathbf{x} = (x_1, \dots, x_k) : a_i \leq x_i \leq b_i, i = 1, \dots, k].$$

The  $\sigma$ -field generated from this class is called the  *$k$ -dimensional Borel sets*, and denoted by  $\mathcal{R}^k$ .

Let  $\mathcal{F}$  be a  $\sigma$ -field of subsets of  $S$  and  $\mathcal{F}'$  be a  $\sigma$ -field of subsets of  $S'$ . For a mapping  $T : S \rightarrow S'$ , consider the inverse images  $T^{-1}(A') = [s \in S : T(s) \in A']$ . The mapping  $T$  is measurable  $\mathcal{F}/\mathcal{F}'$  if  $T^{-1}(A') \in \mathcal{F}$  for each  $A' \in \mathcal{F}'$ .

For a real-valued function  $f$ , the image space  $S'$  is the line  $\mathbb{R}^1$ , and in this case  $\mathcal{R}^1$  is always tacitly understood to play the role of  $\mathcal{F}'$ . A real-valued function on  $S$  is measurable  $\mathcal{F}$  (or simply measurable when it is clear from the context what  $\mathcal{F}$  is involved) if it is measurable  $\mathcal{F}/\mathcal{R}^1$ . If  $(S, \mathcal{F}, Pr)$  is a probability space, then a real-valued measurable function is called a *random variable*. For a random variable  $X$ , we can assign a probability to the event that  $X(s)$  belongs to a Borel set  $\mathcal{B}$  by  $Pr(X^{-1}(\mathcal{B}))$ .

For a mapping  $f : S \mapsto \mathbb{R}^k$ ,  $\mathcal{R}^k$  is always understood to be the  $\sigma$ -field in the image space. If  $(S, \mathcal{F}, Pr)$  is a probability space, then a measurable mapping  $X : S \mapsto \mathbb{R}^k$  is called a *random vector*. It is known that  $X$  is a random vector if and only if each component of  $X$  is a random variable.

A mapping  $f : \mathbb{R}^i \mapsto \mathbb{R}^k$  is defined to be measurable if it is measurable  $\mathcal{R}^i/\mathcal{R}^k$ . Such functions are called *Borel functions*.

**Proposition 2.A.2** If  $f : \mathbb{R}^i \mapsto \mathbb{R}^k$  is continuous, then it is measurable. ■

If  $X$  is a  $j$ -dimensional random vector, and  $g : \mathbb{R}^j \mapsto \mathbb{R}^i$  is measurable, then  $g(X)$  is an  $i$ -dimensional random vector. If the distribution of  $X$  is  $\mu$ , the distribution of  $g(X)$  is  $\mu g^{-1}$ . Proposition 2.2 can be proven by taking  $X = [Y'_t, \dots, Y'_{t+k}]'$ .

We now introduce two definitions of conditional expectation. One definition is standard in measure theory. The other definition is given because it is convenient for the purpose of stating a version of the conditional Gauss-Markov theorem used in this book. Intuitively, the conditional Gauss-Markov theorem is obtained by stating all assumptions and results of the Gauss-Markov theorem conditional on the stochastic regressors. Formally, it is necessary to make sure that the conditional expectations of the relevant variables are well defined.

Let  $S$  be a probability space,  $\mathcal{F}$  be a  $\sigma$ -field of  $S$ , and  $Pr$  be a probability measure defined on  $\mathcal{F}$ . The random variables we will consider in this section are defined on this probability space. Let  $\mathbf{X} = (X_1, X_2, \dots, X_T)'$  be a  $T \times K$  matrix of random variables, which will be the regressor matrix of the regression to be considered. Let  $\mathbf{y} = (y_1, y_2, \dots, y_T)$  and  $\mathbf{e} = (e_1, e_2, \dots, e_T)$  be  $T \times 1$  vectors of random variables. We are concerned with a linear model of the form:  $\mathbf{y} = \mathbf{X}\mathbf{b}_0 + \mathbf{e}$ , where  $\mathbf{b}_0$  is a  $K \times 1$  vector of real numbers.

For  $s$  such that  $\mathbf{X}(s)'\mathbf{X}(s)$  is nonsingular, the OLS estimator is

$$(2.A.1) \quad \mathbf{b}_T = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

In order to apply a conditional version of the Gauss-Markov Theorem, it is necessary to define the expectation and variance of  $\mathbf{b}_T$  conditional on  $\mathbf{X}$ .

Let  $Z$  be an integrable random variable (namely,  $E(|Z|) < \infty$ ), and  $\sigma(\mathbf{X})$  be the smallest  $\sigma$ -field with respect to which of the random variables in  $\mathbf{X}$  are measurable. The standard definition of the expectation of  $Z$  given  $\mathbf{X}$  is obtained by applying the Radon-Nikodym theorem (see, e.g., Billingsley, 1986). Throughout this paper, we use the notation  $E[Z|\sigma(\mathbf{X})]$  to denote the usual conditional expectation of  $Z$  conditional on  $\mathbf{X}$  as defined by Billingsley (1986) for a random variable  $Z$ .<sup>6</sup>  $E[Z|\sigma(\mathbf{X})]$  is a random variable, and  $E[Z|\sigma(\mathbf{X})]_s$  denotes the value of the random variable at  $s$  in  $S$ . It satisfies the following two properties:

- (i)  $E(Z|\sigma(\mathbf{X}))$  is measurable and integrable given  $\sigma(\mathbf{X})$ .
- (ii)  $E(Z|\sigma(\mathbf{X}))$  satisfies the functional equation:

$$(2.A.2) \quad \int_G E(Z|\sigma(\mathbf{X}))dPr = \int_G ZdPr, \quad G \in \sigma(\mathbf{X}).$$

There will in general be many such random variables which satisfy these two properties; any one of them is called a version of  $E(Z|\sigma(\mathbf{X}))$ . Any two versions are equal with probability 1.

It should be noted that this definition is given under the condition that  $Z$  is integrable, namely  $E(|Z|) < \infty$ . This condition is too restrictive when we define

---

<sup>6</sup>If  $\mathbf{z}$  is a vector, the conditional expectation is taken for each element in  $\mathbf{z}$ .

the conditional expectation and variance of the OLS estimator in many applications<sup>7</sup> because the moments of  $(\mathbf{X}'\mathbf{X})^{-1}$  may not be finite even when  $\mathbf{X}$  has many finite moments. For this reason, it is difficult to confirm that  $E(\mathbf{b}_T|\sigma(\mathbf{X}))$  can be defined in each application even if  $\mathbf{X}$  is normally distributed. Thus, Judge et al. (1985) conclude that the Gauss-Markov theorem based on  $E(\cdot|\sigma(\mathbf{X}))$  is not very useful.

We avoid this problem by adopting a different definition of conditional expectation based on conditional distribution. For this purpose, we first define conditional probabilities following Billingsley (1986). Given  $A$  in  $\mathcal{F}$ , define a finite measure  $v$  on  $\sigma(\mathbf{X})$  by  $v(G) = \Pr(A \cap G)$  for  $G$  in  $\sigma(\mathbf{X})$ . Then  $\Pr(G) = 0$  implies that  $v(G) = 0$ . The Radon-Nikodym theorem can be applied to the measures  $v$  and  $Pr$ , and there exists a random variable  $f$  that is measurable and integrable with respect to  $Pr$ , such that  $\Pr(A \cap G) = \int_G f dPr$  for all  $G$  in  $\sigma(\mathbf{X})$ . Denote this random variable by  $\Pr(A|\sigma(\mathbf{X}))$ . This random variable satisfies these two properties:

- (i)  $\Pr(A|\sigma(\mathbf{X}))$  is measurable and integrable given  $\sigma(\mathbf{X})$ .
- (ii)  $\Pr(A|\sigma(\mathbf{X}))$  satisfies the functional equation

$$(2.A.3) \quad \int_G \Pr(A|\sigma(\mathbf{X})) dPr = \Pr(A \cap G), \quad G \in \sigma(\mathbf{X}).$$

There will in general be many such random variables, but any two of them are equal with probability 1. A specific such random variable is called a version of the conditional probability.

Given a random variable  $Z$ , which may not be integrable, we define a conditional distribution  $\mu(\cdot, s)$  given  $\mathbf{X}$  for each  $s$  in  $S$ . Let  $\mathcal{R}^1$  be the  $\sigma$ -field of the Borel sets

---

<sup>7</sup>Loeve (1978) slightly relaxes this restriction by defining the conditional expectation for any random variable whose expectation exists (but may not be finite) with an extension of the Radon-Nikodym theorem. This definition can be used for  $E(\cdot|\sigma(\mathbf{X}))$ , but this slight relaxation does not solve our problem.

in  $\mathcal{R}^1$ . By Theorem 33.3 in Billingsley (1986, p.460), there exists a function  $\mu(H, s)$ , defined for  $H$  in  $\mathcal{R}^1$  and  $s$  in  $S$ , with these two properties:

- (i) For each  $s$  in  $S$ ,  $\mu(H, s)$  is, as a function of  $H$ , a probability measure on  $\mathcal{R}^1$ .
- (ii) For each  $H$  in  $\mathcal{R}^1$ ,  $\mu(H, s)$  is, as a function of  $s$ , a version of  $Pr(Z \in H | \sigma(\mathbf{X}))_s$ .

For each  $s$  in  $S$ , we define  $E(Z|\mathbf{X})_s$  to be  $\int_{\mathcal{R}^1} z\mu(dz, s)$ . It should be noted that  $E(Z|\mathbf{X})_s$  does not necessarily satisfy the usual properties of conditional expectation such as the law of iterated expectations. In general,  $E(Z|\mathbf{X})_s$  may not even exist for some  $s$ . If  $\int_{\mathcal{R}^1} |z|\mu(dz, s)$  is finite, then,  $E(Z|\mathbf{X})_s$  is said to exist and be finite.

Given a  $T \times K$  matrix of real numbers  $x$ ,  $E(Z|\mathbf{X})_s$  is identical for all  $s$  in  $\mathbf{X}^{-1}(x)$ . Therefore, we define  $E(Z|\mathbf{X} = x)$  as  $E(Z|\mathbf{X})_s$  for  $s$  in  $\mathbf{X}^{-1}(x)$ . This is the definition of the conditional expectation of  $Z$  given  $\mathbf{X} = x$  in this paper.

We are concerned with a linear model of the form:

**Assumption 2.A.1**  $\mathbf{y} = \mathbf{X}\mathbf{b}_0 + \mathbf{e}$

where  $\mathbf{b}_0$  is a  $K \times 1$  vector of real numbers. Given a  $T \times K$  matrix of real numbers  $x$ , we assume that the conditional expectation of  $\mathbf{e}$  given  $\mathbf{X} = x$  is zero:

**Assumption 2.A.2**  $E[\mathbf{e}|\mathbf{X} = x] = 0$ .

Next, we assume that  $\mathbf{e}$  is homoskedastic and  $e_t$  is not serially correlated given  $\mathbf{X} = x$ :

**Assumption 2.A.3**  $E[\mathbf{e}\mathbf{e}'|\mathbf{X} = x] = \sigma^2\mathbf{I}_T$ .

The OLS estimator can be expressed by (2.A.1) for all  $s$  in  $\mathbf{X}^{-1}(x)$  when the next assumption is satisfied:

**Assumption 2.A.4**  $x'x$  is nonsingular.

Under Assumptions 2.A.1–2.A.4,  $E[\mathbf{b}_T | \mathbf{X} = x] = \mathbf{b}_0$  and  $E[(\mathbf{b}_T - \mathbf{b}_0)'(\mathbf{b}_T - \mathbf{b}_0) | \mathbf{X} = x] = \sigma^2(x'x)^{-1}$ . The conditional version of the Best Linear Unbiased Estimator (BLUE) given  $\mathbf{X} = x$  can be defined as follows: An estimator  $\mathbf{b}_T$  for  $\mathbf{b}_0$  is BLUE conditional on  $\mathbf{X} = x$  if (1)  $\mathbf{b}_T$  is linear conditional on  $\mathbf{X} = x$ , namely,  $\mathbf{b}_T$  can be written as  $\mathbf{b}_T = \mathbf{A}\mathbf{y}$  for all  $s$  in  $\mathbf{X}^{-1}(x)$  where  $\mathbf{A}$  is a  $K \times T$  matrix of real numbers; (2)  $\mathbf{b}_T$  is unbiased conditional on  $\mathbf{X} = x$ , namely,  $E(\mathbf{b}_T | \mathbf{X} = x) = \mathbf{b}_0$ ; (3) for any linear unbiased estimator  $\mathbf{b}^*$  conditional on  $\mathbf{X} = x$ ,  $E[(\mathbf{b}_T - \mathbf{b}_0)(\mathbf{b}_T - \mathbf{b}_0)' | \mathbf{X} = x] \leq E[(\mathbf{b}^* - \mathbf{b}_0)(\mathbf{b}^* - \mathbf{b}_0)' | \mathbf{X} = x]$ , namely,  $E[(\mathbf{b}^* - \mathbf{b}_0)(\mathbf{b}^* - \mathbf{b}_0)' | \mathbf{X}(s) = x] - E[(\mathbf{b}_T - \mathbf{b}_0)(\mathbf{b}_T - \mathbf{b}_0)' | \mathbf{X}(s) = x]$  is a positive semidefinite matrix.

With these preparations, the following theorem can be stated:

**Theorem 2.A.1** (The Conditional Gauss-Markov Theorem) Under Assumptions 2.A.1–2.A.4, the OLS estimator is BLUE conditional on  $\mathbf{X} = x$ . ■

Applying any of the standard proofs of the (unconditional) Gauss-Markov theorem can prove this theorem by replacing the unconditional expectation with  $E(\cdot | \mathbf{X} = x)$ .

Modifying some assumptions and adding another yields the textbook version of the conditional Gauss-Markov theorem based on  $E(\cdot | \sigma(\mathbf{X}))$ .

**Assumption 2.A.2'**  $E[\mathbf{e} | \sigma(\mathbf{X})] = 0$ .

Since  $E[\mathbf{e} | \sigma(\mathbf{X})]$  is defined only when each element of  $\mathbf{e}$  is integrable, Assumption 2.A.2' implicitly assumes that  $E(\mathbf{e})$  exists and is finite. It also implies  $E(\mathbf{e}) = 0$  because of the law of iterated expectations. Given  $E(\mathbf{e}) = 0$ , a sufficient condition for Assumption 2.A.2' is that  $\mathbf{X}$  is statistically independent of  $\mathbf{e}$ . Since Assumption 2.A.2' does not imply that  $\mathbf{X}$  is statistically independent of  $\mathbf{e}$ , Assumption 2.A.2'

is weaker than the assumption of independent stochastic regressors. With the next assumption, we assume that  $\mathbf{e}$  is conditionally homoskedastic and  $e_t$  is not serially correlated:

**Assumption 2.A.3'**  $E[\mathbf{e}\mathbf{e}'|\sigma(\mathbf{X})] = \sigma^2\mathbf{I}_T$ .

The next assumption replaces Assumption 2.A.4.

**Assumption 2.A.4'**  $\mathbf{X}'\mathbf{X}$  is nonsingular with probability one.

From Assumption 2.A.1,  $\mathbf{b}_T = \mathbf{b}_0 + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}$ . Hence we can prove a version of the conditional Gauss-Markov theorem based on  $E(\cdot|\sigma(\mathbf{X}))$  when the expectations of  $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}$  and  $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}\mathbf{e}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$  exist and are finite. For this purpose, we consider the following assumption:

**Assumption 2.A.5**  $E[\text{trace}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}\mathbf{e}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1})]$  exists and is finite.

The problem with Assumption 2.A.5 is that it is not easy to verify the assumption for many distributions of  $\mathbf{X}$  and  $\mathbf{e}$  that are often used in applications and Monte Carlo studies. However, a sufficient condition for Assumption 2.A.5 is that the distributions of  $\mathbf{X}$  and  $\mathbf{e}$  have finite supports.

Under Assumptions 2.A.1, 2.A.2'–2.A.4', and 2.A.5,

$$E(\mathbf{b}_T|\sigma(\mathbf{X})) = \mathbf{b}_0 + E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}|\sigma(\mathbf{X})] = \mathbf{b}_0.$$

Moreover,  $E[(\mathbf{b}_T - \mathbf{b}_0)'(\mathbf{b}_T - \mathbf{b}_0)|\sigma(\mathbf{X})]$  can be defined, and  $E[(\mathbf{b}_T - \mathbf{b}_0)'(\mathbf{b}_T - \mathbf{b}_0)|\sigma(\mathbf{X})] = E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}\mathbf{e}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}|\sigma(\mathbf{X})] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E[\mathbf{e}\mathbf{e}'|\sigma(\mathbf{X})]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ .

We now consider a different definition of the conditional version of the Best Linear Unbiased Estimator (BLUE). The *Best Linear Unbiased Estimator (BLUE)*

conditional on  $\sigma(\mathbf{X})$  is defined as follows. An estimator  $\mathbf{b}_T$  for  $\mathbf{b}_0$  is BLUE conditional on  $\sigma(\mathbf{X})$  in  $\mathbf{H}$  if (1)  $\mathbf{b}_T$  is linear conditional on  $\sigma(\mathbf{X})$ , namely,  $\mathbf{b}_T$  can be written as  $\mathbf{b}_T = \mathbf{A}\mathbf{y}$  where  $\mathbf{A}$  is a  $K \times T$  matrix, and each element of  $\mathbf{A}$  is measurable given  $\sigma(\mathbf{X})$ ; (2)  $\mathbf{b}_T$  is unbiased conditional on  $\sigma(\mathbf{X})$  in  $\mathbf{G}$ , equivalently,  $E(\mathbf{b}_T|\sigma(\mathbf{X})) = \mathbf{b}_0$ , (3) for any linear unbiased estimator  $\mathbf{b}^*$  conditional on  $\sigma(\mathbf{X})$  for which  $E(\mathbf{b}^*\mathbf{b}^{*\prime})$  exists and is finite,  $E[(\mathbf{b}_T - \mathbf{b}_0)(\mathbf{b}_T - \mathbf{b}_0)'\sigma(\mathbf{X})] \leq E[(\mathbf{b}^* - \mathbf{b}_0)(\mathbf{b}^* - \mathbf{b}_0)'\sigma(\mathbf{X})]$  with probability 1, namely,  $E[(\mathbf{b}^* - \mathbf{b}_0)(\mathbf{b}^* - \mathbf{b}_0)'\sigma(\mathbf{X})] - E[(\mathbf{b}_T - \mathbf{b}_0)(\mathbf{b}_T - \mathbf{b}_0)'\sigma(\mathbf{X})]$  is a positive semidefinite matrix with probability 1.

**Proposition 2.A.3** Under Assumptions 2.A.1, 2.A.2'-2.A.4', and 2.A.5, the OLS estimator is BLUE conditional on  $\sigma(\mathbf{X})$ . Moreover, it is unconditionally unbiased and has the minimum unconditional covariance matrix among all linear unbiased estimators conditional on  $\sigma(\mathbf{X})$ . ■

*Proof* The proof of this proposition is given in Greene (1997, Section 6.7).

In this proposition, the covariance matrix of  $\mathbf{b}_T$  is  $\sigma^2 E[(\mathbf{X}'\mathbf{X})^{-1}]$ , which is different from  $\sigma^2 [E(\mathbf{X}'\mathbf{X})]^{-1}$ . This property may seem to contradict the standard asymptotic theory, but it does not. Asymptotically,  $(1/T)\mathbf{X}'\mathbf{X}$  converges almost surely to  $E[\mathbf{X}'_t\mathbf{X}_t]$  if  $\mathbf{X}_t$  is stationary and ergodic. Hence the limit of the covariance matrix of  $\sqrt{T}(\mathbf{b}_T - \mathbf{b}_0)$ ,  $\sigma^2 E[\{(1/T)(\mathbf{X}'\mathbf{X})\}^{-1}]$ , is equal to the asymptotic covariance matrix,  $\sigma^2 [E(\mathbf{X}'_t\mathbf{X}_t)]^{-1}$ .

In order to study the distributions of the  $t$  ratios and  $F$  test statistics we need an additional assumption:

**Assumption 2.A.6** Conditional on  $\mathbf{X}$ ,  $\mathbf{e}$  follows a multivariate normal distribution.

Given a  $1 \times K$  vector of real numbers  $R$ , consider a random variable

$$(2.A.4) \quad N_R = \frac{R(\mathbf{b}_T - \mathbf{b}_0)}{\sigma[R(\mathbf{X}'\mathbf{X})^{-1}R]^{1/2}}$$

and the usual  $t$  ratio for  $R\mathbf{b}_0$

$$(2.A.5) \quad t_R = \frac{R(\mathbf{b}_T - \mathbf{b}_0)}{\hat{\sigma}[R(\mathbf{X}'\mathbf{X})^{-1}R]^{1/2}}.$$

Here  $\hat{\sigma}$  is the positive square root of  $\hat{\sigma}^2 = (\mathbf{y} - \mathbf{X}\mathbf{b}_T)'(\mathbf{y} - \mathbf{X}\mathbf{b}_T)/(T - K)$ . With the standard argument,  $N_R$  and  $t_R$  can be shown to follow the standard normal distribution and Student's  $t$  distribution with  $T - K$  degrees of freedom with appropriate conditioning, respectively, under either Assumptions 2.A.1–2.A.6 or Assumptions 2.A.1, 2.A.2', 2.A.3', and 2.A.5–2.A.6. The following proposition is useful in order to derive the unconditional distributions of these statistics.

**Proposition 2.A.4** If the probability density function of a random variable  $Z$  conditional on a random vector  $\mathbf{Q}$  does not depend on the values of  $\mathbf{Q}$ , then the marginal probability density function of  $Z$  is equal to the probability density function of  $Z$  conditional on  $\mathbf{Q}$ . ■

This proposition is obtained by integrating the probability density function conditional on  $\mathbf{Q}$  over all possible values of the random variables in  $\mathbf{Q}$ . Since  $N_R$  and  $t_R$  follow a standard normal distribution and a  $t$  distribution conditional on  $\mathbf{X}$ , respectively, Proposition 2.A.4 implies the following proposition:

**Proposition 2.A.5** Suppose that Assumptions 2.A.1, 2.A.5, and 2.A.6 are satisfied and that Assumptions 2.A.2 and 2.A.3 are satisfied for all  $x$  in a set  $H$  such that  $Pr(\mathbf{X}^{-1}(H)) = 1$ . Then  $N_R$  is a standard normal random variable and  $t_R$  is a  $t$  random variable with  $T - K$  degrees of freedom. ■

Alternatively, the assumptions for Proposition 2.A.3 with Assumption 2.A.6 can be used to obtain a similar result:

**Proposition 2.A.5'** Suppose that Assumptions 2.A.1, 2.A.2'–2.A.3', 2.A.5, and 2.A.6 are satisfied for  $s$  and that Assumptions 2.A.2 and 2.A.3 are satisfied for all  $x$  in a set  $H$  such that  $Pr(\mathbf{X}^{-1}(H)) = 1$ . Then  $N_R$  is a standard normal random variable and  $t_R$  is a  $t$  random variable with  $T - K$  degrees of freedom. ■

Similarly, the usual  $F$  test statistics also follow (unconditional)  $F$  distributions. These results are sometimes not well understood by econometricians. For example, a standard textbook, Judge et al. (1985, p.164), states that “our usual test statistics do not hold in finite samples” on the ground that the (unconditional) distribution of  $\mathbf{b}'_T s$  is not normal. It is true that  $\mathbf{b}_T$  is a nonlinear function of  $\mathbf{X}$  and  $\mathbf{e}$ , so it does not follow a normal distribution even if  $\mathbf{X}$  and  $\mathbf{e}$  are both normally distributed. However, the usual  $t$  and  $F$  test statistics have the usual (unconditional) distributions as a result of Proposition 2.A.4.

## 2.B Convergence in Probability

Let  $c_1, c_2, \dots, c_T, \dots$  be a sequence of real numbers and  $c$  be a real number. The sequence is said to *converge* to  $c$  if for any  $\varepsilon$ , there exists an  $N$  such that  $|c_T - c| < \varepsilon$  for all  $T \geq N$ . We write  $c_T \rightarrow c$  or  $\lim_{T \rightarrow \infty} c_T = c$ . This definition is extended to a sequence of vectors of real numbers  $\{\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_T, \dots\}$  by interpreting  $|c_T - c|$  as the Euclidean distance  $(\mathbf{c}_T - \mathbf{c})'(\mathbf{c}_T - \mathbf{c})$ .

Consider a univariate stochastic process  $[X_T : T \geq 1]$ , and a random variable  $X$ . Fix  $s$ , and then  $[X_T(s) : T \geq 1]$  is a sequence of real numbers and  $X(s)$  is a real

number. For each  $s$ , verify whether or not  $X_T(s) \rightarrow X(s)$ . Then collect  $s$  such that  $X_T(s) \rightarrow X(s)$ , and calculate the probability that  $X_T(s) \rightarrow X(s)$ . If the probability is one, we say the sequence of random variables,  $[X_T : T \geq 1]$ , converges to  $X$  *almost surely* or *with probability one*. We write  $X_T \rightarrow X$  almost surely. This definition is extended to a sequence of random vectors by using convergence for a sequence of vectors for each  $s$ . In general, if a property holds for all  $s$  except for a set of  $s$  with probability zero, we say that the property holds *almost surely* or *with probability one*.

If  $\Omega$  has finite elements, almost sure convergence is the same thing as convergence of  $X_T(s)$  to  $X(s)$  in all states of the world. In general, however, almost sure convergence does not imply convergence in all states.

The sequence of random variables  $[X_T : T \geq 1]$  *converges in probability* to the random variable  $X_T$  if, for all  $\varepsilon > 0$ ,  $\lim_{T \rightarrow \infty} \text{Prob}(|X_T - X| > \varepsilon) = 0$ . This is expressed by writing  $X_T \xrightarrow{P} c$  or  $\text{plim}_{T \rightarrow \infty} X_T = X$ . This extension to the vector case is done by using the Euclidean distance. Almost sure convergence implies convergence in probability.

*Slutsky's Theorem* is important for working with probability limits. It states that, if  $\text{plim} \mathbf{X}_T = \mathbf{X}$  and if  $f(\cdot)$  is a continuous function, then  $\text{plim}(f(\mathbf{X}_T)) = f(\text{plim}(\mathbf{X}_T))$ .

### 2.B.1 Convergence in Distribution

Consider a univariate stochastic process  $[X_T : T \geq 1]$ , and a random variable  $X$  with respective distribution functions  $F_T$  and  $F$ . If  $F_T(x) \rightarrow F(x)$  for every continuity point  $x$  of  $F$ , then  $X_T$  is said to *converge in distribution* to  $X$ ; this is expressed by writing  $X_T \xrightarrow{D} X$ . The distribution  $F$  is called the *asymptotic distribution* or the

limiting distribution of  $X_T$ .

## 2.B.2 Propositions 2.2 and 2.3 for Infinite Numbers of R.V.'s (Incomplete)

In Propositions 2.2 and 2.3, we only allow for a finite number of random variables. In many applications, we are often interested in infinite sums of covariance or strictly stationary random variables. We need the convergence concepts explained in Appendix 2.B. A sequence of real numbers  $\{a_j\}_{j=0}^{\infty}$  is *square summable* if  $\sum_{j=0}^{\infty} a_j^2$  is finite. A sufficient condition for  $\{a_j\}_{j=0}^{\infty}$  is that it is *absolutely summable*, that is,  $\sum_{j=0}^{\infty} |a_j|$  is finite. In the following propositions, the infinite sum  $\sum_{j=0}^{\infty} a_j X_{t-j}$  means the convergence in mean square of  $\sum_{j=0}^T a_j X_{t-j}$  as  $T$  goes to infinity.

**Proposition 2.B.1** If  $X_t$  is a scalar covariance stationary process, and if  $\{a_j\}_{j=0}^{\infty}$  is square summable, then  $X = \sum_{j=0}^{\infty} a_j X_{t-j}$  is covariance stationary. ■

The vector version of this proposition is:

**Proposition 2.B.2** If  $\mathbf{X}_t$  is a  $k$ -dimensional vector covariance stationary process, and if the absolute value of the  $i$ -th row of a sequence of a  $k \times k$  matrix of real numbers  $\{\mathbf{A}_j\}_{j=0}^{\infty}$  is square summable for  $i = 1, \dots, k$ , then  $\mathbf{X}_t = \sum_{j=0}^{\infty} \mathbf{A}_j \mathbf{X}_{t-j}$  is covariance stationary. ■

## Exercises

**2.1** In Example 2.3, assume that  $\pi_1 = 0.15$ ,  $\pi_2 = 0.05$ ,  $\pi_3 = 0.20$ ,  $\pi_4 = 0.30$ ,  $\pi_5 = 0.10$ , and  $\pi_6 = 0.20$ . As in Example 2.4, compute  $E(Y_2|I)(s)$  and  $E(Y_2|J)(s)$ . Then compute  $E(E(Y_2|I)|J)(s)$ . Verify that  $E(Y_2|J)(s) = E(E(Y_2|I)|J)(s)$  for all  $s \in S$ .

**2.2** In example 2.9, assume that  $|A| < 1$ . This condition does not ensure that  $Y_t$  is strictly stationary. In order to see this, suppose that  $Y_0 = 0$ . Then compute the expected values of  $Y_1$  and  $Y_2$  and the variance of  $Y_1$  and  $Y_2$ , and show that  $Y_t$  is not strictly stationary if  $A \neq 0$ .

**2.3** In example 2.9, assume that  $|A| < 1$  and that  $Y_0$  is  $N(0, \frac{\sigma^2}{1-A^2})$ . Then compute the expected values of  $Y_1$  and  $Y_2$ , the variance of  $Y_1$  and  $Y_2$ , and the  $k$ -th autocovariance of  $Y$ . Prove that  $Y_t$  is strictly stationary in this case. (Hint: Remember that first and second moments completely determine the joint distribution of jointly normally distributed random variables.)

**2.4** Let  $Y_t$  be a martingale adapted to  $I_t$ . Then prove that  $e_t = Y_t - Y_{t-1}$  is a martingale difference sequence.

**2.5** Prove that a covariance stationary martingale difference sequence is a white noise process.

**2.6** Prove that an i.i.d. white noise process is a martingale difference sequence.

## References

- BILLINGSLEY, P. (1986): *Probability and Measure*. Wiley, New York.
- ENGLE, R. F. (1982): "Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation," *Econometrica*, 50(4), 987–1008.
- GREENE, W. H. (1997): *Econometric Analysis*. Prentice-Hall, 3rd edn.
- JUDGE, G. G., W. E. GRIFFITHS, R. C. HILL, H. LÜTKEPOHL, AND T. LEE (1985): *The Theory and Practice of Econometrics*. Wiley, New York, 2nd edn.
- LOEVE, M. (1978): *Probability Theory II*. Springer-Verlag, New York, 4th edn.

# Chapter 3

## FORECASTING

### 3.1 Projections

In macroeconomics, forecasting is important in many ways. For structural macroeconomic models, we usually need to specify the forecasting rules that economic agents are using and the information set used by them to forecast future economic variables. Taking the conditional expectation is one way to model forecasting. This method generally requires nonlinear forecasting rules which are difficult to estimate. For the purpose of testing the models and parameter estimation, it is sometimes possible for an econometrician to use a simpler forecasting rule and a smaller information set.

In this section, we study projections as a forecasting method. Projections are used to explain the Wold representation, which forms a basis for studying linear and nonlinear stochastic processes.

#### 3.1.1 Definitions and Properties of Projections

In this chapter, we consider random variables with finite second moments unless otherwise noted. We consider the problem of forecasting  $y$ , using a set  $H$  of random variables. Typically,  $y$  is a future random variable such as the growth rate of the Gross Domestic Product (GDP) or the growth rate of a stock price, and  $H$  contains

current and past economic variables that are observed by economic agents and/or econometricians. Let us denote a forecast of  $y$  based on  $H$  by  $y^f$ , so that the forecasting error is  $y - y^f$ . In most economic applications, we choose the forecast,  $y^f$ , so that  $y^f$  minimizes

$$(3.1) \quad E[(y - y^f)^2].$$

In other words,  $y^f$  is in  $H$ , and for all  $h$  in  $H$ ,

$$(3.2) \quad E[(y - y^f)^2] \leq E[(y - h)^2].$$

The expression (3.1) is called the *mean squared error* associated with the forecast,  $y^f$ .

When two random variables  $h_1$  and  $h_2$  satisfy

$$(3.3) \quad E(h_1 h_2) = 0,$$

they are said to be *orthogonal* to each other. When either  $h_1$  or  $h_2$  has mean zero, orthogonality means that they are uncorrelated. The concept of orthogonality is closely related to the problem of minimizing the mean squared error. Under certain conditions on  $H$ , the Classical Projection Theorem (see, e.g., Luenberger, 1969) states that there exists a unique random variable  $y^f$  in  $H$  that minimizes the mean squared error, and that  $y^f$  is the minimizer if and only if the forecasting error is orthogonal to all members of  $H$ :

$$(3.4) \quad E((y - y^f)h) = 0$$

for all  $h$  in  $H$ ; this is called the *orthogonality condition*. When such a forecast exists, we call the forecast,  $y^f$ , a *projection* of  $y$  onto  $H$ , and denote it by  $\hat{E}(y|H)$ . When  $\mathbf{Y}$  is a random vector with finite second moments, we apply the projection to each element of  $\mathbf{Y}$  and write  $\hat{E}(\mathbf{Y}|H)$ .

Some properties of projections are very important:

**Proposition 3.1 (Properties of Projections)**

- (a) Projections are linear:  $\hat{E}(aX + bY|H) = a\hat{E}(X|H) + b\hat{E}(Y|H)$  for any random variables,  $X$  and  $Y$ , with finite variance and constants,  $a$  and  $b$ .
- (b) If a random variable  $Z$  is in the information set  $H$ , then

$$\hat{E}(ZY|H) = Z\hat{E}(Y|H).$$

- (c) *The Law of Iterated Projections:* If the information set  $H$  is smaller than the information set  $G$  ( $H \subset G$ ), then

$$\hat{E}(Y|H) = \hat{E}[\hat{E}(Y|G)|H].$$

■

### 3.1.2 Linear Projections and Conditional Expectations

The meaning of projection depends on how the information set  $H$  used for the projection is constructed. Let  $\mathbf{X}$  be a  $p \times 1$  vector of random variables with finite second moments. Let  $H = \{h \text{ is a random variable such that } h = \mathbf{X}'\mathbf{b} \text{ for some } p\text{-dimensional vector of real numbers } \mathbf{b}\}$ . Since  $\hat{E}(y|H)$  is also a member of  $H$ , there exists  $\mathbf{b}_0$  such that

$$(3.5) \quad \hat{E}(y|H) = \mathbf{X}'\mathbf{b}_0.$$

In this sense,  $\hat{E}(y|H)$  uses a linear forecasting rule. When we use an information set such as  $H$ , which only allows for linear forecasting rules, the projection based on such an information set is called a *linear projection*. We write  $\hat{E}(y|H) = \hat{E}(y|\mathbf{X})$ .

Let  $H^N = \{h \text{ is a random variable with a finite variance such that } h = f(\mathbf{X}) \text{ for a function } f\}$ .<sup>1</sup> In this case, there exists a function  $f_0(\cdot)$  such that

$$(3.6) \quad \hat{E}(y|H^N) = f_0(\mathbf{X}).$$

In this sense,  $\hat{E}(y|H^N)$  allows for a nonlinear forecasting rule. It can be shown that

$$(3.7) \quad \hat{E}(y|H^N) = E(y|\mathbf{X}).$$

Hence the projection and conditional expectation coincide when we allow for nonlinear forecasting rules. For this reason, the projections we use in this book are linear projections unless otherwise noted.

An important special case is when  $y$  and  $\mathbf{X}$  are jointly normally distributed. In this case, the expectation of  $y$  conditional on  $\mathbf{X}$  is a linear function of  $\mathbf{X}$ . Hence the linear projection of  $y$  onto the information set generated by  $\mathbf{X}$  is equal to the expectation of  $y$  conditional on  $\mathbf{X}$ .

When it is necessary to distinguish the information set  $I$  generated by  $\mathbf{X}$  for conditional expectations introduced in Chapter 2 and the information set  $H$  generated by  $\mathbf{X}$  for linear projections,  $H$  will be called the *linear information set* generated by  $\mathbf{X}$ . (????? Unclear! from Billy)

Masao  
needs to  
check this!

Linear projections are important because it is easy to estimate them in many applications. Note that the orthogonality condition states that

$$(3.8) \quad E[(y - \mathbf{X}'\mathbf{b}_0)h] = 0$$

for any  $h$  in  $H$ . Since each element of  $\mathbf{X}$  is in  $H$ , using the  $i$ -th element  $\mathbf{X}_i$  for  $h$ , we obtain

$$(3.9) \quad E[(y - \mathbf{X}'\mathbf{b}_0)\mathbf{X}_i] = 0$$

<sup>1</sup>As in Proposition 2.2, we require that the function  $f$  is measurable.

for  $i = 1, 2, \dots, p$ , or

$$(3.10) \quad E[\mathbf{X}(y - \mathbf{X}'\mathbf{b}_0)] = 0.$$

Therefore

$$(3.11) \quad E(\mathbf{X}y) = E(\mathbf{X}\mathbf{X}')\mathbf{b}_0.$$

Assuming that  $E(\mathbf{X}\mathbf{X}')$  is nonsingular, we obtain

$$(3.12) \quad \mathbf{b}_0 = E(\mathbf{X}\mathbf{X}')^{-1}E(\mathbf{X}y)$$

and

$$(3.13) \quad \hat{E}(y|\mathbf{H}) = \mathbf{X}'\mathbf{b}_0,$$

where  $\mathbf{H}$  is the linear information set generated by  $\mathbf{X}$ . As we will discuss, if  $\mathbf{X}$  and  $y$  are strictly stationary, Ordinary Least Squares (OLS) can be used to estimate  $\mathbf{b}_0$ .

Following examples show differences between conditional expectation and linear projection.

Youngsoo  
needs to  
check this!

**Example 3.1** Let  $X$  and  $Y$  be random variables with non-zero mean. The linear projection of  $Y$  on  $X$  is

$$(3.14) \quad \hat{E}(Y|1, X) = a + bX.$$

Then, from (3.12) and  $E(Y) = a + bE(X)$  we have

$$(3.15) \quad b = \frac{E(XY)}{E(X^2)} = \frac{Cov(X, Y)}{Var(X)}$$

$$(3.16) \quad a = E(Y) - bE(X).$$

■

Note that the linear projection is a population regression; that is,  $a$  and  $b$  are defined by population moments. corresponding sample moments can be used to estimate  $\hat{a}$  and  $\hat{b}$ .

**Example 3.2** Let  $X$  be a standard Normal random variable, and  $Y = X^2$ . Note that  $Y$  is  $\chi(1)$  random variable and  $E(Y) = 1$ . The linear projection of  $Y$  on  $X$  is

$$(3.17) \quad \hat{E}(Y|1, X) = a + bX = 1.$$

This is because from Example 3.1, we have

$$(3.18) \quad b = \frac{E(XY)}{E(X^2)} = \frac{E(X^3)}{Var(X)} = 0$$

and

$$(3.19) \quad a = E(Y) = E(X^2) = 1.$$

Note that  $E(X^3) = 0$  because the distribution of  $X$  is symmetric. Whereas the conditional expectation of  $Y$  on  $X$  is<sup>2</sup>

$$(3.20) \quad E(Y|X) = X^2.$$

■

**Example 3.3** Let  $X_0$  be a standard Normal random variable, and  $\varepsilon_1$  be a Normal random variable with mean 0 and variance  $\sigma^2$ . Assume that  $X_0$  and  $\varepsilon$  are independent each other. Define  $X_1 = a + bX_0 + cX_0^2 + \varepsilon_1$ . Then, the unconditional expectation of  $X_1$ , the linear projection and the conditional expectation of  $X_1$  on  $X_0$  are, respectively,

$$(3.21) \quad E(X_1) = E(a + bX_0 + cX_0^2 + \varepsilon_1) = a + c,$$

---

<sup>2</sup>Note that since  $X^0 = 1$ , 1 is always in the information set for conditional expectation. However, 1 may not be in the linear information set.

$$\begin{aligned}
 (3.22) \quad \hat{E}(X_1|1, X_0) &= \hat{E}(a + bX_0 + cX_0^2 + \varepsilon_1|1, X_0) \\
 &= a + bX_0 + c.
 \end{aligned}$$

Note that  $\hat{E}(X_0^2|1, X_0) = 1$  by (3.17).

$$(3.23) \quad E(X_1|X_0) = a + bX_0 + cX_0^2.$$

■

## 3.2 Some Applications of Conditional Expectations and Projections

This section presents some applications of conditional expectations and projections in order to illustrate their use in macroeconomics. More explanations of some of these applications and presentations of other applications will be given in later chapters. In this chapter, all random variables are assumed to have finite second moments.

### 3.2.1 Volatility Tests

Many rational expectations models imply

$$(3.24) \quad X_t = E(Y_t|I_t)$$

for economic variables  $X_t$  and  $Y_t$ . Here  $I_t$  is the information set  $I_t$  which is available to the economic agents at date  $t$  while  $Y_t$  is not. A testable implication of (3.24) can be obtained by comparing the volatility of  $X_t$  with that of  $Y_t$ . Relation (3.24) implies

$$(3.25) \quad Y_t = X_t + \epsilon_t$$

where  $\epsilon_t = Y_t - E(Y_t|I_t)$  is the forecast error. Since  $E(\epsilon_t|I_t) = 0$ ,

$$(3.26) \quad E(\epsilon_t h_t) = 0$$

for any random variable  $h_t$  that is in  $I_t$ . We can interpret (3.26) as an orthogonality condition. The forecast error must be uncorrelated with any variable in the information set. Since  $X_t$  is in  $I_t$ , (3.26) implies  $E(\epsilon_t X_t) = 0$ . Therefore, from (3.25) we obtain

$$(3.27) \quad E(Y_t^2) = E(X_t^2) + E(\epsilon_t^2).$$

Since (3.24) implies that  $E(X_t) = E(Y_t)$ , (3.27) implies

$$(3.28) \quad \text{Var}(Y_t) = \text{Var}(X_t) + E(\epsilon_t^2).$$

Since  $E(\epsilon_t^2) \geq 0$ , we conclude

$$(3.29) \quad \text{Var}(Y_t) \geq \text{Var}(X_t).$$

Thus, if  $X_t$  forecasts  $Y_t$ ,  $X_t$  must be less volatile than  $Y_t$ . Various volatility tests have been developed to test this implication of (3.24).

LeRoy and Porter (1981) and Shiller (1981) started to apply volatility tests to the present value model of stock prices. Let  $p_t$  be the real stock price (after the dividend is paid) in period  $t$  and  $d_t$  be the real dividend paid to the owner of the stock at the beginning of period  $t$ . Then the no-arbitrage condition is

$$(3.30) \quad p_t = E[b(p_{t+1} + d_{t+1})|I_t],$$

where  $b$  is the constant real discount rate, and  $I_t$  is the information set available to economic agents in period  $t$ . Solving (3.30) forward and imposing the no bubble condition,<sup>3</sup> we obtain the present value formula:

$$(3.31) \quad p_t = E\left(\sum_{i=1}^{\infty} b^i d_{t+i} | I_t\right).$$

---

<sup>3</sup>It rules out the exploding solution of the difference equation

Applying the volatility test, we conclude that the variance of  $\sum_{i=1}^{\infty} b^i d_{t+i}$  is greater than or equal to the variance of  $p_t$ . One way to test this is to directly estimate these variances and compare them. However,  $\sum_{i=1}^{\infty} b^i d_{t+i}$  involves infinitely many data points for the dividend. When we have data for the stock price and dividend for  $t = 1, \dots, T$ , we use (3.31) to obtain

$$(3.32) \quad p_t = E\left(\sum_{i=1}^{T-t} b^i d_{t+i} + b^{T-t} p_T \mid \mathcal{I}_t\right).$$

Let  $Y_t = \sum_{i=1}^{T-t} b^i d_{t+i} + b^{T-t} p_T$ . Then we have data on  $Y_t$  from  $t = 1$  to  $t = T$  when we choose a reasonable number for the discount rate  $b$ . We can estimate the variance of  $p_t$  and the variance of  $Y_t$ , and compare them to form a test statistic.<sup>4</sup>

### 3.2.2 Parameterizing Expectations

As discussed in Section 3.1, conditional expectations allow for nonlinear forecasting rules. For example, consider  $E(Y|\mathcal{I})$  for a random variable  $Y$  and an information set  $\mathcal{I}$  generated from a random variable  $X$ . Then  $E(Y|\mathcal{I})$  can be written as a function of  $X$ :  $E(Y|\mathcal{I}) = f(X)$ . The function  $f(\cdot)$  can be nonlinear here. In most applications involving nonlinear forecasting rules, the functional form of  $f(\cdot)$  is not known. In order to simulate rational expectations models, it is often necessary to have a method to estimate  $f(\cdot)$ .

Marcet's (1989) parameterizing expectations method (also see den Haan and Marcet, 1990) is based on the fact that the conditional expectation is a projection, and thus minimizes the mean square error. We take a class of functions that approximate any function. For example, take a class of polynomial functions and let  $f_N(X) = a_0 + a_1 X + a_2 X^2 + \dots + a_N X^N$ . We choose  $a_0, \dots, a_N$  to minimize the mean square

---

<sup>4</sup>There are some problems with this procedure. One problem is nonstationarity of  $p_t$  and  $Y_t$ . For more detailed explanation of volatility tests, see Campbell, Lo, and MacKinlay (1997).

error,  $E[(Y - f_N(X))^2]$ . Intuitively,  $f_N(\cdot)$  should approximate  $f(X)$  for a large enough  $N$ . This method is used to simulate economic models with rational expectations.

### 3.2.3 Noise Ratio

In econometrics, we often test an economic model with test statistics whose probability distributions are known under the null hypothesis that the model is true. Hansen's  $J$  test, which will be discussed in Chapter 9, is an example. Given that all economic models are meant to be approximations, however, it seems desirable to measure how good a model is in approximating reality. Durlauf and Hall (1990) and Durlauf and Maccini (1995) propose such a measure called the noise ratio.<sup>5</sup>

Consider an economic model which states

$$(3.33) \quad E(g(\mathbf{Y})|\mathbf{I}) = 0$$

for an information set  $\mathbf{I}$  and a function  $g(\cdot)$  of a random vector  $\mathbf{Y}$ . For example, let  $S$  be the spot exchange rate of a currency in the next period,  $F$  be the forward exchange rate observed today for the currency to be delivered in the next period,  $g(S, F) = S - F$ , and  $\mathbf{I}$  be the information set available to the economic agents today. Then under the assumption of risk neutral investors, we obtain (3.33).

Let  $\nu = g(\mathbf{Y}) - E(g(\mathbf{Y})|\mathbf{I})$ . If the model is true, then  $g(\mathbf{Y}) = \nu$ . Since this model is an approximation, however,  $g(\mathbf{Y})$  deviates from  $\nu$ . Let  $N$  be the deviation:  $N = g(\mathbf{Y}) - \nu$ , which is called the model noise. A natural measure of how well the model approximates reality is  $Var(N)$ . Durlauf and Hall (1990) propose a method to estimate a lower bound of  $Var(N)$  using  $\eta = Var(\hat{E}(g(\mathbf{Y})|\mathbf{H}))$ , where  $\mathbf{H}$  is an information set generated from some variables in  $\mathbf{I}$ .<sup>6</sup>

<sup>5</sup>See Konuki (1999) for an application of the noise ratio to foreign exchange rate models.

<sup>6</sup>For example, in the forward exchange rate model mentioned above, some lagged values of  $S - F$

Using the law of iterated projections<sup>7</sup>, we have  $\hat{E}(\nu|\mathbf{H}) = 0$ . Thus,  $\hat{E}(g(\mathbf{Y})|\mathbf{H}) = \hat{E}(N|\mathbf{H})$ , and therefore  $\eta = \text{Var}(\hat{E}(N|\mathbf{H}))$ . Because  $N = \hat{E}(N|\mathbf{H}) + (N - \hat{E}(N|\mathbf{H}))$ , and the forecast error,  $N - \hat{E}(N|\mathbf{H})$ , is orthogonal to  $\hat{E}(N|\mathbf{H})$ ,  $E(N^2) = E[(\hat{E}(N|\mathbf{H}))^2] + E[(N - \hat{E}(N|\mathbf{H}))^2]$ . Since  $E[(N - \hat{E}(N|\mathbf{H}))^2] \geq 0$ ,  $E(N^2) \geq E[(\hat{E}(N|\mathbf{H}))^2]$ . Therefore,  $\text{Var}(N) = E(N^2) - (E(N))^2 \geq E[(\hat{E}(N|\mathbf{H}))^2] - \{E[\hat{E}(N|\mathbf{H})]\}^2 = \eta$ .<sup>8</sup> Thus  $\eta$  is a lower bound of  $\text{Var}(N)$ .

In a sense,  $\eta$  is a sharp lower bound. Since we do not know much about the model noise,  $N$ , it may or may not be in  $\mathbf{H}$ . If  $N$  happens to be in  $\mathbf{H}$ , then  $\hat{E}(N|\mathbf{H}) = N$ . Therefore, in this case  $\text{Var}(N) = \eta$ .

The noise ratio,  $NR$ , is defined by  $NR = \frac{\eta}{\text{Var}(g(\mathbf{Y}))}$ . Since  $\hat{E}(g(\mathbf{Y})|\mathbf{H})$  is orthogonal to  $g(\mathbf{Y}) - \hat{E}(g(\mathbf{Y})|\mathbf{H})$ ,

$$(3.34) \quad \text{Var}(g(\mathbf{Y})) = \eta + \text{Var}(g(\mathbf{Y}) - \hat{E}(g(\mathbf{Y})|\mathbf{H})).$$

Therefore, the  $0 \leq NR \leq 1$ .

## Appendix

### 3.A Introduction to Hilbert Space

This Appendix explains Hilbert space techniques used in this book.<sup>9</sup> Projections explained in this chapter are defined in a Hilbert space. In Appendix B, we will consider another Hilbert space, which provides the foundation for the lag operator methods and the frequency domain analysis which are useful in macroeconomics and time series econometrics.

---

and a constant can be used to generate a linear information set  $\mathbf{H}$ .

<sup>7</sup>We assume that the second moment exists and is finite. Therefore, the conditional expectation is a projection.

<sup>8</sup>Here, we assumed that the constants are included in  $\mathbf{H}$ , so that  $E(S) = E[\hat{E}(S|\mathbf{H})]$ .

<sup>9</sup>All proofs of the results can be found in Luenberger (1969) or Hansen and Sargent (1991).

A pre-Hilbert space is a vector space on which an inner product is defined. The inner product is used to define a distance. If all Cauchy sequences of a pre-Hilbert space converge, then it is said to be complete. A Hilbert space is a complete pre-Hilbert space. One reason why a Hilbert space is useful is that the notion of orthogonality can be defined with the inner product. Since a Hilbert space is complete, we can prove that the limit of a sequence exists once we prove that the sequence is Cauchy. For example, this technique can be used to prove that a projection can be defined.

Section 3.A.1 reviews definitions regarding vector spaces. Section 3.A.2 gives an introduction to Hilbert space.

### 3.A.1 Vector Spaces

Given a set of scalars  $K$  (either the real line,  $\mathbb{R}$ , or the complex plane,  $\mathbb{C}$ )<sup>10</sup>, a *vector space* (or a *linear space*)  $X$  on  $K$  is a set of elements, called vectors, together with two operations (addition and scalar multiplication) which satisfy the following conditions:

For any  $\mathbf{x}, \mathbf{y}, \mathbf{z}$  in  $X$  and for any  $\alpha, \beta$  in  $K$ , we require

$$(3.A.1) \quad \mathbf{x} + \mathbf{y} = \mathbf{y} + \mathbf{x} \quad (\text{commutative law})$$

$$(3.A.2) \quad (\mathbf{x} + \mathbf{y}) + \mathbf{z} = \mathbf{x} + (\mathbf{y} + \mathbf{z}) \quad (\text{associative law})$$

$$(3.A.3) \quad \text{There is a null vector } \mathbf{0} \text{ in } X \text{ such that } \mathbf{x} + \mathbf{0} = \mathbf{x} \text{ for all } \mathbf{x} \text{ in } X.$$

$$(3.A.4) \quad \left. \begin{array}{l} \alpha(\mathbf{x} + \mathbf{y}) = \alpha\mathbf{x} + \alpha\mathbf{y} \\ (\alpha + \beta)\mathbf{x} = \alpha\mathbf{x} + \beta\mathbf{x} \end{array} \right\} \quad (\text{distributive laws})$$

$$(3.A.5) \quad (\alpha\beta)\mathbf{x} = \alpha(\beta\mathbf{x}) \quad (\text{associative law})$$

$$(3.A.6) \quad 0\mathbf{x} = \mathbf{0}, 1\mathbf{x} = \mathbf{x}.$$

---

<sup>10</sup>In general, an additive group  $X$  for which scalar multiplication satisfies (3.A.4)-(3.A.6) for any field  $K$  is a vector space on  $K$ . In this book  $K$  is either the real line or the complex plane.

Using  $\alpha = -1$ , we define  $\mathbf{x} - \mathbf{y} = \mathbf{x} + (-1)\mathbf{y}$ . In this Appendix, we give examples of vector spaces on  $\mathbb{R}$ , but state results that are applicable when  $K = \mathbb{C}$ . Examples of vector spaces on  $\mathbb{C}$  are given in Appendix B.

A nonempty subset  $H$  of a vector space  $X$  is called a (*linear*) *subspace* of  $X$  if every vector of the form  $\alpha\mathbf{x} + \beta\mathbf{y}$  is in  $H$  whenever  $\mathbf{x}$  and  $\mathbf{y}$  are both in  $H$  and  $\alpha$  and  $\beta$  are in  $K$ . A subspace always contains the null vector  $\mathbf{0}$ , and satisfies conditions (3.A.1)-(3.A.6). Hence a subspace is itself a vector space.

If a subset  $H$  of  $X$  is not a subspace, it is often convenient to construct the smallest subspace containing  $H$ . For this purpose, we use linear combinations of vectors in  $H$ . A *linear combination* of the vectors  $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$  is a sum of the form  $\alpha_1\mathbf{x}_1 + \alpha_2\mathbf{x}_2 + \dots + \alpha_n\mathbf{x}_n$  where  $\alpha_i$  is a scalar ( $i = 1, \dots, n$ ). The set consisting of all vectors in  $X$  which are linear combinations of vectors in  $H$  is called the (*linear*) *subspace generated* by  $H$ .

A *normed vector space* is a vector space  $X$  on which a norm is defined. The norm is a real-valued function that maps each element of  $\mathbf{x}$  in  $X$  into a real number  $\|\mathbf{x}\|$ , which satisfies

$$(3.A.7) \quad \|\mathbf{x}\| \geq 0 \text{ for all } \mathbf{x} \text{ in } X \text{ and } \|\mathbf{x}\| = 0 \text{ if and only if } \mathbf{x} = \mathbf{0}.$$

$$(3.A.8) \quad \|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\| \quad (\text{The triangle inequality})$$

$$(3.A.9) \quad \|\alpha\mathbf{x}\| = |\alpha| \|\mathbf{x}\| \text{ for all } \alpha \text{ in } K \text{ and } \mathbf{x} \text{ in } X.$$

A norm can be used to define a metric  $d$  on  $X$  by  $d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|$ .

A sequence  $\{\mathbf{x}_n\}_{n=1}^{\infty}$  in a normed vector space *converges* to  $\mathbf{x}_0$  if the sequence  $\{\|\mathbf{x}_n - \mathbf{x}_0\|\}_{n=1}^{\infty}$  of real numbers converges to zero, which is denoted by  $\mathbf{x}_n \rightarrow \mathbf{x}_0$  or  $\lim \mathbf{x}_n = \mathbf{x}_0$ . A sequence  $\{\mathbf{x}_n\}_{n=1}^{\infty}$  in a normed vector space is a *Cauchy sequence* if

for any  $\epsilon > 0$ , there exists an integer  $N$  such that  $\|\mathbf{x}_n - \mathbf{x}_m\| < \epsilon$  for all  $n, m > N$ . In a normed vector space, every convergent sequence is a Cauchy sequence. A space in which every Cauchy sequence has a limit is said to be *complete*. A complete normed vector space is called a *Banach space*.

**Example 3.A.1** The real line,  $\mathbb{R}$ , is a vector space on  $K = \mathbb{R}$  with addition and scalar multiplication defined in the usual way. When the norm of a real number is defined as its absolute value,  $\mathbb{R}$  is a Banach space. ■

**Example 3.A.2** Vectors in the space consist of sequences of  $n$  real numbers,  $\mathbb{R}^n$ , which is a vector space on  $\mathbb{R}$  when  $\mathbf{x} + \mathbf{y}$  for  $\mathbf{x} = (x_1, x_2, \dots, x_n)'$  and  $\mathbf{y} = (y_1, y_2, \dots, y_n)'$  is defined by  $(x_1 + y_1, x_2 + y_2, \dots, x_n + y_n)'$  and  $\alpha \mathbf{x}$  for  $\alpha$  in  $\mathbb{R}$  is defined by  $(\alpha x_1, \alpha x_2, \dots, \alpha x_n)'$ . When we define a norm of  $\mathbf{x}$  as  $\|\mathbf{x}\| = \sqrt{\sum_{i=1}^n x_i^2}$ ,  $\mathbb{R}^n$  is a Banach space. ■

### 3.A.2 Hilbert Space

A *pre-Hilbert space* is a vector space  $X$  on  $K$  for which an inner product is defined. The inner product is a scalar-valued function that maps each element of  $(\mathbf{x}, \mathbf{y})$  in  $X \times X$  into an element  $(\mathbf{x}|\mathbf{y})$  in  $K$ , which satisfies

$$(3.A.10) \quad (\mathbf{x}|\mathbf{y}) = \overline{(\mathbf{y}|\mathbf{x})}$$

$$(3.A.11) \quad (\mathbf{x} + \mathbf{z}|\mathbf{y}) = (\mathbf{x}|\mathbf{y}) + (\mathbf{z}|\mathbf{y})$$

$$(3.A.12) \quad (\alpha \mathbf{x}|\mathbf{y}) = \alpha(\mathbf{x}|\mathbf{y})$$

$$(3.A.13) \quad (\mathbf{x}|\mathbf{x}) \geq 0 \text{ and } (\mathbf{x}|\mathbf{x}) = 0 \text{ if and only if } \mathbf{x} = \mathbf{0}.$$

for any  $\mathbf{x}, \mathbf{y}, \mathbf{z}$  in  $X$  and  $\alpha$  in  $K$ . The bar on the right side on (3.A.10) denotes complex conjugation, which can be ignored if  $K$  is  $\mathbb{R}$ . By (3.A.10),  $(\mathbf{x}|\mathbf{x})$  is real for each  $\mathbf{x}$  even when  $K$  is  $\mathbb{C}$ .

A norm can be defined from an inner product by  $\|\mathbf{x}\| = \sqrt{(\mathbf{x}|\mathbf{x})}$ . Thus a pre-Hilbert space is a normed vector space. A complete pre-Hilbert space is called a *Hilbert space*.

**Example 3.A.3** When we define  $(\mathbf{x}|\mathbf{y}) = \sum_{i=1}^n x_i y_i$ ,  $\mathbb{R}^n$  is a Hilbert space on  $\mathbb{R}$ . ■

The following Hilbert space of random variables with finite second moments is the one we used in Chapter 3.

**Example 3.A.4** Let  $(S, \mathcal{F}, Prob)$  be a probability space. Let  $L^2(Prob) = \{h : h \text{ is a (real-valued) random variable and } E(|h|^2) < \infty\}$ . Then with an inner product defined by  $(h_1|h_2) = E(h_1 h_2)$ ,  $L^2(Prob)$  is a Hilbert space on  $\mathbb{R}$ . If two different random variables  $h_1$  and  $h_2$  satisfy  $E[(h_1 - h_2)^2] = 0$ , then  $h_1$  and  $h_2$  are the same element in this space. If  $E[(h_1 - h_2)^2] = 0$ , then  $h_1 = h_2$  with probability one. Hence this definition does not cause problems for most purposes. In this space, the distance is defined by the mean square, so the convergence in this space is the convergence in mean square. ■

One reason why an inner product is useful is that we can define the notion of orthogonality. In a Hilbert space, two vectors  $\mathbf{x}$  and  $\mathbf{y}$  are said to be *orthogonal* if  $(\mathbf{x}|\mathbf{y}) = 0$ . A vector  $\mathbf{x}$  is said to be orthogonal to a set  $H$  if  $\mathbf{x}$  is orthogonal to each element  $h$  in  $H$ . Some useful results concerning the inner product are:<sup>11</sup>

**Proposition 3.A.1** (*The Cauchy-Schwarz Inequality*) For all  $\mathbf{x}, \mathbf{y}$  in a Hilbert space,  $|(\mathbf{x}|\mathbf{y})| \leq \|\mathbf{x}\| \|\mathbf{y}\|$ . Equality holds if and only if  $\mathbf{x} = \lambda \mathbf{y}$  for some  $\lambda$  in  $K$ , or  $\mathbf{y} = \mathbf{0}$ . ■

<sup>11</sup>These three propositions hold for a pre-Hilbert space. See Luenberger (1969, p.47 and p.49).

**Proposition 3.A.2** (*Continuity of the Inner Product*) Suppose that  $\mathbf{x}_n \rightarrow \mathbf{x}$  and  $\mathbf{y}_n \rightarrow \mathbf{y}$  in a Hilbert space. Then  $(\mathbf{x}_n | \mathbf{y}_n) \rightarrow (\mathbf{x} | \mathbf{y})$ . ■

**Proposition 3.A.3** If  $\mathbf{x}$  is orthogonal to  $\mathbf{y}$  in a Hilbert space, then  $\|\mathbf{x} + \mathbf{y}\|^2 = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2$ . ■

**Example 3.A.5** In  $L^2(Prob)$ , the Cauchy-Schwarz Inequality becomes  $|E(xy)| \leq \sqrt{E(x^2)}\sqrt{E(y^2)}$  for any random variables with finite second moments. Proposition 3.A.3 states that if  $x$  and  $y$  satisfy  $E(xy) = 0$ , then  $E[(x + y)^2] = E(x^2) + E(y^2)$ . ■

Projections can be defined on a Hilbert space due to the following result:

**Proposition 3.A.4** (*The Classical Projection Theorem*) Let  $X$  be a Hilbert space and  $H$  be a closed linear subspace of  $X$ . Corresponding to any vector  $\mathbf{x}$  in  $X$ , there is a unique vector  $\mathbf{h}_0$  in  $H$  such that  $\|\mathbf{x} - \mathbf{h}_0\| \leq \|\mathbf{x} - \mathbf{h}\|$ . Furthermore, a necessary and sufficient condition that  $\mathbf{h}_0$  in  $H$  be the unique minimizing vector is that  $\mathbf{x} - \mathbf{h}_0$  be orthogonal to  $H$ . ■

Given a closed linear space  $H$ , we define a function  $\hat{E}(\cdot | H)$  on  $X$  by  $\hat{E}(\mathbf{x} | H) = \mathbf{h}_0$  where  $\mathbf{h}_0$  is an element in  $H$  such that  $\mathbf{x} - \mathbf{h}_0$  is orthogonal to  $H$ .  $\hat{E}(\mathbf{x} | H)$  is the projection of  $\mathbf{x}$  onto  $H$ . The projection defined in Section 3.1 in  $L^2(Prob)$  is one example.

If a sequence  $\{\mathbf{e}_t\}_{t=1}^{\infty}$  in a Hilbert space satisfies  $\|\mathbf{e}_t\| = 1$  for all  $t$  and  $(\mathbf{e}_t | \mathbf{e}_s) = 0$  for all  $t \neq s$ , then it is said to be an *orthonormal sequence*. We are concerned with an infinite series of the form  $\sum_{t=1}^{\infty} \alpha_t \mathbf{e}_t$ . An infinite series of the form  $\sum_{t=1}^{\infty} \mathbf{x}_t$  is said to *converge* to the element  $\mathbf{x}$  in a Hilbert space if the sequence of partial sums  $s_T = \sum_{t=1}^T \mathbf{x}_t$  converges to  $\mathbf{x}$ . In that case we write  $\mathbf{x} = \sum_{t=1}^{\infty} \mathbf{x}_t$ . A necessary

and sufficient condition for an infinite series of orthonormal sequence to converge in Hilbert space is known (see Luenberger, 1969, p.59):

**Proposition 3.A.5** Let  $\{\mathbf{e}_j\}_{j=1}^{\infty}$  be an orthonormal sequence in a Hilbert space  $X$ . A series of the form  $\sum_{j=1}^{\infty} \alpha_j \mathbf{e}_j$  converges to an element  $\mathbf{x}$  in  $X$  if and only if  $\sum_{j=1}^{\infty} |\alpha_j|^2 < \infty$ , and in that case we have  $\alpha_j = (\mathbf{x}|\mathbf{e}_j)$ . ■

**Example 3.A.6** Applying the above proposition in  $L^2(Prob)$ , we obtain a necessary and sufficient condition for an MA( $\infty$ ) representation  $\sum_{j=0}^{\infty} b_j v_{t-j}$  to converge for a white noise process  $\{v_{t-j}\}_{j=0}^{\infty}$  with  $E(v_t^2) = \sigma_v^2 > 0$ . Define  $e_t = \frac{v_t}{\sigma_v}$ , and  $\alpha_j = b_j \sigma_v$ , so that  $\{e_{t-j}\}_{j=0}^{\infty}$  is orthonormal because  $E(e_t^2) = 1$  and  $E(e_t e_s) = 0$  for  $t \neq s$ . From the above proposition,  $\sum_{j=1}^{\infty} b_j v_j = \sum_{j=1}^{\infty} \alpha_j e_j$  converges in  $L^2(Prob)$ , if and only if  $\sum_{j=1}^{\infty} |\alpha_j|^2 < \infty$ . Since  $\sum_{j=1}^{\infty} |\alpha_j|^2 < \infty$  if and only if  $\sum_{j=1}^{\infty} |b_j|^2 < \infty$ ,  $\sum_{j=1}^{\infty} b_j v_j$  converges in mean square if and only if  $\{b_j\}_{j=1}^{\infty}$  is square summable. ■

Given an orthonormal sequence  $\{\mathbf{e}_j\}_{j=1}^{\infty}$ , we started from a square summable sequence  $\{\alpha_j\}$  and constructed  $\mathbf{x} = \sum_{j=1}^{\infty} \alpha_j \mathbf{e}_j$  in  $X$  in the above proposition. We now start with a given  $\mathbf{x}$  in  $X$  and consider a series

$$(3.A.14) \quad \sum_{j=1}^{\infty} (\mathbf{x}|\mathbf{e}_j) \mathbf{e}_j.$$

The series is called the *Fourier series* of  $\mathbf{x}$  relative to  $\{\mathbf{e}_j\}_{j=1}^{\infty}$ , and  $(\mathbf{x}|\mathbf{e}_j)$  is called the *Fourier coefficient* of  $\mathbf{x}$  with respect to  $\mathbf{e}_j$ .

In general,  $\mathbf{x}$  is not equal to its Fourier series. Given a subset  $H$  of a Hilbert space, the *closed subspace generated by*  $H$  is the closure of the linear subspace generated by  $H$ . Let  $M$  be the closed subspace generated by  $\{\mathbf{e}_j\}_{j=1}^{\infty}$ . If  $\mathbf{x}$  is in  $M$ , then  $\mathbf{x}$  is equal to its Fourier series as implied by the next proposition:

**Proposition 3.A.6** Let  $\mathbf{x}$  be an element in a Hilbert space  $X$  and  $\{\mathbf{e}_j\}_{j=1}^{\infty}$  be an orthonormal sequence in  $H$ . Then the Fourier series  $\sum_{j=1}^{\infty}(\mathbf{x}|\mathbf{e}_j)\mathbf{e}_j$  converges to an element  $\hat{\mathbf{x}}$  in the closed subspace  $M$  generated by  $\{\mathbf{e}_j\}_{j=1}^{\infty}$ . The difference vector  $\mathbf{x} - \hat{\mathbf{x}}$  is orthogonal to  $M$ . ■

This proposition shows that the Fourier series of  $\mathbf{x}$  is the projection of  $\mathbf{x}$  onto  $M$ :

$$\hat{E}(\mathbf{x}|M) = \sum_{j=1}^{\infty}(\mathbf{x}|\mathbf{e}_j)\mathbf{e}_j.^{12}$$

## Exercises

**3.1** Let  $S_t$  be a spot exchange rate at time  $t$  and  $F_t$  be a forward exchange rate observed at time  $t$  for delivery of one unit of a currency at  $t + 1$ . Assume that  $F_t = E(S_{t+1}|I_t)$  where  $I_t$  is the information set available for the economic agents at  $t$ . Prove that  $Var(F_t) \leq Var(S_{t+1})$ .

**3.2** Let  $i_{n,t}$  be the  $n$  year interest rate observed at time  $t$ . The expectations hypothesis of the term structure of interest rates states that  $i_{n,t} = E(A_t|I_t)$  where

$$(3.E.1) \quad A_t = \frac{1}{n} \sum_{\tau=0}^{n-1} i_{1,t+\tau},$$

where  $I_t$  is the information available at time  $t$ . Imagine that data on interest rates clearly indicate that  $Var(i_{n,t}) \leq Var(A_t)$ . Does the data support the expectations theory? Explain your answer.

**3.3** Let  $p_t$  be the real stock price,  $d_t$  be the real dividend, and  $b$  be the constant ex ante discount rate. Assume that  $p_t$  and  $d_t$  are stationary with zero mean and finite

---

<sup>12</sup>See Luenberger (1969, p.60).

second moments. Let

$$(3.E.2) \quad p_t^e = \sum_{\tau=1}^{\infty} b^\tau E(d_{t+\tau} | I_t),$$

where  $I_t$  is the information set available in period  $t$  that includes the present and past values of  $p_t$  and  $d_t$ . Let  $\hat{E}(\cdot | H_t)$  be the linear projection onto an information set  $H_t$ .

Define the model noise  $N_t$  by

$$(3.E.3) \quad N_t = p_t - p_t^e.$$

Let  $\eta = \text{Var}(\hat{E}(N_t | H_t))$ .

- (a) Assume that  $H_t$  is generated by  $\{d_t\}$ . Show that  $\eta \leq \text{Var}(N_t)$  for any noise  $N_t$ .
- (b) Assume that  $H_t$  is generated by  $\{d_t, d_{t-1}, d_{t-2}\}$ . Show that  $\eta \leq \text{Var}(N_t)$  for any noise  $N_t$ .

**3.4** Derive (3.34) in the text.

## References

- CAMPBELL, J. Y., A. W. LO, AND A. C. MACKINLAY (1997): *The Econometrics of Financial Markets*. Princeton University Press, Princeton, New Jersey.
- DEN HAAN, W. J., AND A. MARCET (1990): "Solving the Stochastic Growth Model by Parameterizing Expectations," *Journal of Business and Economic Statistics*, 8, 31–34.
- DURLAUF, S. N., AND R. E. HALL (1990): "Bounds on the Variances of Specification Errors in Models with Expectations," Manuscript.
- DURLAUF, S. N., AND L. J. MACCINI (1995): "Measuring Noise in Inventory Models," *Journal of Monetary Economics*, 36, 65–89.
- HANSEN, L. P., AND T. J. SARGENT (1991): *Rational Expectations Econometrics*. Westview, London.
- KONUJI, T. (1999): "Measuring Noise in Exchange Rate Models," *Journal of International Economics*, 48(2), 255–270.
- LEROY, S. F., AND R. D. PORTER (1981): "The Present-Value Relation: Tests Based on Implied Variance Bounds," *Econometrica*, 49(3), 555–574.

LUENBERGER, D. G. (1969): *Optimization by Vector Space Methods*. Wiley, New York.

MARCEY, A. (1989): "Solving Non-Linear Stochastic Models by Parameterizing Expectations," Manuscript.

SHILLER, R. J. (1981): "Do Stock Prices Move Too Much to be Justified by Subsequent Changes in Dividends?," *American Economic Review*, 71, 421–436.

# Chapter 4

## ARMA AND VECTOR AUTOREGRESSION REPRESENTATIONS

### 4.1 Autocorrelation

The Wold representation of a univariate process  $\{X_t : -\infty < t < \infty\}$  provides us with a description of how future values of  $X_t$  depend on its current and past values (in the sense of linear projections). A useful description of this dependence is *autocorrelation*. The  $j$ -th autocorrelation of a process (denoted by  $\rho_j$ ) is defined as the correlation between  $X_t$  and  $X_{t-j}$ :

$$\text{Corr}(X_t, X_{t-j}) = \frac{\text{Cov}(X_t, X_{t-j})}{\sqrt{\text{Var}(X_t)}\sqrt{\text{Var}(X_{t-j})}}.$$

In general,  $\rho_j$  depends on  $t$ . If the process is covariance stationary,  $\rho_j$  does not depend on  $t$ , and is equal to its  $j$ -th autocovariance divided by its variance:

$$(4.1) \quad \rho_j = \frac{\gamma_j}{\gamma_0},$$

where  $\gamma_j = \text{Cov}(X_t, X_{t-j})$  is the  $j$ -th autocovariance, and  $\gamma_0 = \text{Var}(X_t)$ . For covariance stationary processes,  $\gamma_j = \gamma_{-j}$ , hence  $\rho_j = \rho_{-j}$ . When we view  $\rho_j$  as a function of  $j$ , it is called the autocorrelation function. Note that  $\rho_0 = 1$  for any process by

definition. For a white noise process,  $\rho_j = 0$  for  $j \neq 0$ . The autocorrelation function is a population concept, and can be estimated by its sample counterpart as explained in Chapter 5.

## 4.2 The Lag Operator

In order to study ARMA representations, it is convenient to use the *lag operator*, denoted by the symbol  $L$ . When the operator is applied to a sequence  $\{X_t : -\infty < t < \infty\}$  of real numbers, it results in a new sequence  $\{Y_t : -\infty < t < \infty\}$ , where the value of  $Y$  at date  $t$  is equal to the value  $X$  at date  $t - 1$ :

$$Y_t = X_{t-1},$$

and we write

$$(4.2) \quad LX_t = X_{t-1}.$$

When we apply the lag operator to a univariate stochastic process  $\{X_t : -\infty < t < \infty\}$ , the lag operator is applied to all sequences of real numbers  $\{X_t(\omega) : -\infty < t < \infty\}$  given by fixing the state of the world  $\omega$  to generate a new stochastic process  $\{X_t : -\infty < t < \infty\}$  that satisfies  $X_{t-1}(\omega) = LX_t(\omega)$  for each  $\omega$ .

When the lag operator is applied twice to a process  $\{X_t : -\infty < t < \infty\}$ , we write  $L^2X_t = X_{t-2}$ . In general, for any integer  $k > 0$ ,  $L^kX_t = X_{t-k}$ . It is convenient to define  $L^0 = 1$  as the identity operator that gives  $L^0X_t = X_t$ , and to define  $L^{-k}$  as the operator that moves the sequence forward:  $L^{-k}X_t = X_{t+k}$  for any integer  $k > 0$ .

We define a *p-th order polynomial in the lag operator*  $B(L) = B_0 + B_1L + B_2L^2 + \dots + B_pL^p$ , where  $B_1, \dots, B_p$  are real numbers, as the operator that yields

$$B(L)X_t = (B_0 + B_1L + B_2L^2 + \dots + B_pL^p)X_t = B_0X_t + B_1X_{t-1} + \dots + B_pX_{t-p}.$$

When an infinite sum  $B_0X_t + B_1X_{t-1} + B_2X_{t-2} + \dots$  converges in some sense (such as convergence in  $L^2$ ), (?????? Need to use other expressions instead of  $L^2$  because we use  $L$  for the lag operator in this paragraph) we write  $B(L) = B_0 + B_1L + B_2L^2 + \dots$ , and

Masao needs to check this!

$$B(L)X_t = (B_0 + B_1L + B_2L^2 + \dots)X_t = B_0X_t + B_1X_{t-1} + B_2X_{t-2} + \dots.$$

For a vector stochastic process  $\{\mathbf{X}_t : -\infty < t < \infty\}$ , a polynomial in the lag operator  $\mathbf{B}_0 + \mathbf{B}_1L + \mathbf{B}_2L^2 + \dots + \mathbf{B}_pL^p$  for matrices  $\mathbf{B}_0, \dots, \mathbf{B}_p$  with real numbers is used in the same way, so that

$$(\mathbf{B}_0 + \mathbf{B}_1L + \mathbf{B}_2L^2 + \dots + \mathbf{B}_pL^p)\mathbf{X}_t = \mathbf{B}_0\mathbf{X}_t + \mathbf{B}_1\mathbf{X}_{t-1} + \dots + \mathbf{B}_p\mathbf{X}_{t-p}.$$

Using the lag operator,  $\mathbf{X}_t = \Phi_0\mathbf{e}_t + \Phi_1\mathbf{e}_{t-1} + \dots$  can be expressed as

$$(4.3) \quad \mathbf{X}_t = \Phi(L)\mathbf{e}_t,$$

where  $\Phi(L) = \Phi_0 + \Phi_1L + \Phi_2L^2 + \dots$ .

### 4.3 Moving Average Representation

If  $X_t$  is linearly regular and covariance stationary with mean  $\mu$ , then it has a Moving Average (MA) representation of the form  $X_t = \mu + \Phi(L)e_t$  or

$$(4.4) \quad X_t = \mu + \Phi_0e_t + \Phi_1e_{t-1} + \Phi_2e_{t-2} + \dots,$$

where  $\Phi_0 = 1$ . If  $\Phi(L)$  is a polynomial of infinite order,  $X_t$  is a moving average process of infinite order (denoted MA( $\infty$ )). If  $\Phi(L)$  is a polynomial of order  $q$ ,  $X_t$  is a moving average process of order  $q$  (denoted MA( $q$ )). In this section, we study how some properties of  $X_t$  depend on  $\Phi(L)$ .

An MA(1) process  $X_t$  has a representation  $X_t = \mu + e_t + \Phi e_{t-1}$  as in Example 2.8, where  $e_t$  is a white noise process that satisfies (2.10), and  $\mu$  and  $\Phi$  are constants. The mean, variance, and autocovariance of this process are given in Example 2.8,  $E(X_t) = \mu$ , and its  $k$ -th autocorrelation is  $\rho_k = \frac{\Phi}{1+\Phi^2}$  if  $|k| = 1$ , and  $\rho_k = 0$  if  $|k| > 1$ .

An MA( $q$ ) process  $X_t$  satisfies

$$(4.5) \quad X_t = \mu + e_t + \Phi_1 e_{t-1} + \cdots + \Phi_q e_{t-q},$$

where  $e_t$  is a white noise process that satisfies (2.10), and  $\mu$  and  $\Phi_1, \dots, \Phi_q$  are real numbers. A moving average process is covariance stationary for any  $(\Phi_1, \dots, \Phi_q)$ .<sup>1</sup>

Using (2.10), we obtain the mean of an MA( $q$ ) process:

$$(4.6) \quad E(X_t) = \mu,$$

its variance:

$$(4.7) \quad \gamma_0 = E[(X_t - \mu)^2] = \sigma^2(1 + \Phi_1^2 + \cdots + \Phi_q^2),$$

and its  $j$ -th autocovariance:

$$(4.8) \quad \begin{aligned} \gamma_j &= E[(X_t - \mu)(X_{t-j} - \mu)] \\ &= \begin{cases} \sigma^2(\Phi_j + \Phi_{j+1}\Phi_1 + \cdots + \Phi_q\Phi_{q-j}) & \text{for } |j| \leq q \\ 0 & \text{for } |j| > q \end{cases} \end{aligned}$$

Hence the  $j$ -th autocorrelation of an MA( $q$ ) process is zero when  $|j| > q$ .

When a vector stochastic process  $\{\cdots, \mathbf{X}_{-2}, \mathbf{X}_{-1}, \mathbf{X}_0, \mathbf{X}_1, \cdots, \mathbf{X}_t, \cdots\}$  can be written as

$$(4.9) \quad \mathbf{X}_t = \boldsymbol{\mu} + \boldsymbol{\Phi}_0 \mathbf{e}_t + \boldsymbol{\Phi}_1 \mathbf{e}_{t-1} + \cdots + \boldsymbol{\Phi}_q \mathbf{e}_{t-q},$$

---

<sup>1</sup>We often impose conditions on  $(\Phi_1, \dots, \Phi_q)$  as we will discuss later in this chapter.

for a white noise process  $\mathbf{e}_t$ , then  $\mathbf{X}_t$  has a  $q$ -th order (one-sided) moving average (MA( $q$ )) representation. For any  $\Phi_0, \dots, \Phi_q$ , a process with MA( $q$ ) representation is covariance stationary. As  $q$  goes to infinity, an MA( $\infty$ ) representation

$$(4.10) \quad \mathbf{X}_t = \boldsymbol{\mu} + \Phi_0 \mathbf{e}_t + \Phi_1 \mathbf{e}_{t-1} + \dots$$

is well defined and covariance stationary if  $\sum_{j=0}^{\infty} |\Phi_j^i|^2 < \infty$  for the  $i$ -th row of  $\Phi_j$ ,  $\Phi_j^i$ . In this case,  $\mathbf{X}_t$  has a moving average representation of infinite order.

## 4.4 The Wold Representation

Let  $\{\dots, \mathbf{X}_{-2}, \mathbf{X}_{-1}, \mathbf{X}_0, \mathbf{X}_1, \dots, \mathbf{X}_t, \dots\}$  be a covariance stationary  $n$ -dimensional vector process with mean zero. Let  $H_t$  be the linear information set generated by the current and past values of  $\mathbf{X}_t$ .<sup>2</sup> We use the notation,  $\hat{E}(y|\mathbf{X}_t, \mathbf{X}_{t-1}, \mathbf{X}_{t-2}, \dots)$  for  $\hat{E}(y|H_t)$ . Note that the information set grows larger over time and the sequence  $\{H_t : -\infty < t < \infty\}$  is increasing in the sense that  $H_t \subset H_{t+1}$  for all  $t$ . Let  $H_{-\infty}$  be the set of random variables that are in  $H_t$  for all  $t$ :  $H_{-\infty} = \bigcap_{n=1}^{\infty} H_{t-n}$ . Then  $0 = \mathbf{0}'\mathbf{X}_t$  is a member of  $H_t$ . Therefore, the constant zero is always a member of  $H_{-\infty}$ . The stochastic process  $\mathbf{X}_t$  is *linearly regular* if  $H_{-\infty}$  contains only the constant zero when  $H_{-\infty} = \bigcap_{n=1}^{\infty} H_{t-n}$ , in which  $H_t$  is generated by the current and past values of  $\mathbf{X}_t$ . The stochastic process  $\mathbf{X}_t$  is *linearly deterministic* if  $H_t = H_{-\infty}$  for all  $t$ . For example, if  $\mathbf{X}_t$  is an  $n$ -dimensional vector of constants, then  $\mathbf{X}_t$  is linearly deterministic.

We can now state the Wold decomposition theorem, which states that any covariance stationary process can be decomposed into linearly regular and linearly deterministic components:

---

<sup>2</sup>We only define the linear information set for a finite number of random variables. See Appendix 3.A for further explanation.

**Proposition 4.1** (*The Wold Decomposition Theorem*) Let  $\{\cdots, \mathbf{X}_{-1}, \mathbf{X}_0, \mathbf{X}_1, \cdots, \mathbf{X}_t, \cdots\}$  be a covariance stationary vector process with mean zero. Then it can be written as

$$(4.11) \quad \mathbf{X}_t = \sum_{j=0}^{\infty} \Phi_j \mathbf{e}_{t-j} + \mathbf{g}_t,$$

where  $\Phi_0 = \mathbf{I}_n$ ,  $\sum_{j=0}^{\infty} |\Phi_j^i|^2 < \infty$  for the  $i$ -th row of  $\Phi_j$ ,  $\Phi_j^i$ , and

$$(4.12) \quad \mathbf{e}_t = \mathbf{X}_t - \hat{E}(\mathbf{X}_t | \mathbf{X}_{t-1}, \mathbf{X}_{t-2}, \mathbf{X}_{t-3}, \cdots)$$

and

$$(4.13) \quad \mathbf{g}_t = \hat{E}(\mathbf{X}_t | \mathbf{H}_{-\infty}).$$

■

It can be shown that  $\sum_{j=0}^{\infty} \Phi_j \mathbf{e}_{t-j}$  is a linearly regular covariance stationary process and  $\mathbf{g}_t$  is linearly deterministic. Hence if  $\mathbf{X}_t$  is not linearly regular, it is possible to remove  $\mathbf{g}_t$  and work with a linearly regular process as long as  $\mathbf{g}_t$  can be estimated.

**Proposition 4.2** (*The Wold Representation*) Let  $\{\cdots, \mathbf{X}_{-1}, \mathbf{X}_0, \mathbf{X}_1, \cdots, \mathbf{X}_t, \cdots\}$  be a linearly regular covariance stationary vector process with mean zero. Then it can be written as

$$(4.14) \quad \mathbf{X}_t = \sum_{j=0}^{\infty} \Phi_j \mathbf{e}_{t-j},$$

where  $\Phi_0 = \mathbf{I}_n$ ,  $\sum_{j=0}^{\infty} |\Phi_j^i|^2 < \infty$  for the  $i$ -th row of  $\Phi_j$ ,  $\Phi_j^i$ , and  $\mathbf{e}_t$  is defined by (4.12). ■

The Wold representation gives a unique MA representation when the MA innovation  $\mathbf{e}_t$  is restricted to the form given by Equation (4.12). There may exist infinitely

many other MA representations when the MA innovation is not restricted to be given by (4.12) as we will discuss below.

In many macroeconomic models, stochastic processes that we observe (real GDP, interest rates, stock prices, etc.) are considered to be generated from the nonlinear function of underlying shocks. In this sense, the processes in these models are nonlinear, but Proposition 4.1 states that even a nonlinear stochastic process has a linear moving average representation as long as it is linearly regular and covariance stationary.

In order to give a sketch of a proof of the Wold Representation Theorem, consider a linearly regular stochastic process  $\{X_t\}_{-\infty}^{\infty}$  that may not be necessarily a linear function of underlying shocks. Define  $u_t = X_t - \hat{E}(X_t | H_{t-1})$ , and

$$(4.15) \quad U_t = \{z | z = bu_t \text{ for } b \in \mathbb{R}\}$$

where  $H_t$  is the linear information set generated by the current and past values of  $X_t$ . Then, we have the following relationship

$$(4.16) \quad H_t = H_{t-1} + U_t,$$

and each element of  $H_t$  is orthogonal to each element of  $U_t$ . In this case,

$$(4.17) \quad \begin{aligned} \hat{E}(h | H_t) &= \hat{E}(h | H_{t-1} + U_t) \\ &= \hat{E}(h | H_{t-1}) + \hat{E}(h | U_t) \end{aligned}$$

for any  $h$ . Because  $H_{t-1} = H_{t-2} + U_{t-1}$ , we have

$$(4.18) \quad H_t = H_{t-2} + U_t + U_{t-1},$$

and by continuing this process, we have

$$(4.19) \quad H_t = \sum_{j=0}^{\infty} U_{t-j}.$$

Youngsoo  
needs to  
check this!

Therefore,  $X_t$  can be written as

$$(4.20) \quad X_t = \hat{E}(X_t | \mathbb{H}_t) = \hat{E}(X_t | \sum_{j=0}^{\infty} U_{t-j}) = \sum_{j=0}^{\infty} \Phi_j u_{t-j},$$

which is the Wold representation of  $X_t$ .

**Example 4.1** Suppose that  $u_t$  is a Gaussian white noise with variance of 1. Let  $X_t = u_t^2 - 1$ . Then the Wold representation of  $X_t$  is  $X_t = e_t$ , where  $e_t = u_t^2 - 1$ . ■

In this example,  $X_t$  is a nonlinear transformation of a Gaussian white noise. The shock that generates  $X_t$ ,  $u_t$ , is normally distributed. However, the innovation in its Wold representation,  $e_t$ , is not normally distributed. Thus, the innovation in the Wold representation of a process can have a different distribution from the underlying shock that generates the process.

Even when the underlying shocks that generate processes are i.i.d., the innovations in the Wold representation may not be i.i.d. as in the next example.

**Example 4.2** Suppose that  $u_t$  is an i.i.d Gaussian white noise with variance of 1, so that  $E(u_t^3) = 0$ . Let  $X_t$  be generated by  $X_t = u_t + \Phi(u_{t-1}^2 - 1)$ . Then  $E(X_t X_{t-1}) = E[u_t u_{t-1} + \Phi u_{t-1}^3 - \Phi u_{t-1} + \Phi u_t u_{t-2}^2 - \Phi u_t + \Phi^2 (u_{t-1}^2 - 1)(u_{t-2}^2 - 1)] = 0$ . Hence the Wold representation of  $X_t$  is  $X_t = e_t$ , where  $e_t = u_t + \Phi(u_{t-1}^2 - 1)$ . ■

Note that the Wold representation innovation  $e_t$  in this example is serially uncorrelated, but not i.i.d. because  $e_t (= u_t + \Phi u_{t-1}^2)$  and  $e_{t-1} (= u_{t-1} + \Phi u_{t-2}^2)$  are related nonlinearly through the  $\Phi u_{t-1}^2$  and  $u_{t-1}$  terms.

The Wold representation states that any linearly regular covariance stationary process has an MA representation. Therefore, it is useful to estimate an MA representation in order to study how linear projections of future variables depend on

their current and past values. Higher order MA representations and vector MA representations are hard to estimate, however, and it is often convenient to consider AR representations and ARMA representations, which are closely related to MA representations.

## 4.5 Autoregression Representation

A process  $X_t$ , which satisfies  $B(L)X_t = \delta + e_t$  with  $B_0 = 1$  or

$$X_t + B_1X_{t-1} + B_2X_{t-2} + \cdots = \delta + e_t$$

for a white noise process  $e_t$ , is an autoregression. If  $B(L)$  is a polynomial of infinite order,  $X_t$  is an autoregression of infinite order (denoted  $\text{AR}(\infty)$ ). If  $B(L)$  is a polynomial of order  $p$ ,  $X_t$  is an autoregression of order  $p$  (denoted  $\text{AR}(p)$ ). In this section, we study how some properties of  $X_t$  depend on  $B(L)$ .

### 4.5.1 Autoregression of Order One

Consider a process  $X_t$  that satisfies

$$(4.21) \quad X_t = \delta + BX_{t-1} + e_t \quad \text{for } t \geq 1,$$

where  $e_t$  is a white noise process with variance  $\sigma^2$  and  $X_0$  is a random variable that gives an initial condition for (4.21). Such a process is called an *autoregression of order 1*, denoted by  $\text{AR}(1)$ . It is often convenient to consider (4.21) in a deviation-from-the-mean form:

$$(4.22) \quad X_t - \mu = B(X_{t-1} - \mu) + e_t \quad \text{for } t \geq 1,$$

where  $\mu = \frac{\delta}{1-B}$ . Substituting (4.22) recursively, we obtain  $X_1 - \mu = B(X_0 - \mu) + e_1$  and  $X_2 - \mu = B(X_1 - \mu) + e_2 = B^2(X_0 - \mu) + Be_1 + e_2$ , so that

$$(4.23) \quad X_t - \mu = B^t(X_0 - \mu) + B^{t-1}e_1 + B^{t-2}e_2 + \cdots + Be_{t-1} + e_t \quad \text{for } t \geq 1.$$

In this way,  $X_t$  is defined for any real number  $B$ .

Suppose that  $X_0$  is uncorrelated with  $e_1, e_2, \dots$ . When the absolute value of  $B$  is greater than or equal to one, then the variance of  $X_t$  increases over time. Hence  $X_t$  cannot be covariance stationary. In macroeconomics, the case in which  $B = 1$  is of importance, and will be discussed in detail in Chapter 13.

Consider the case where the absolute value of  $B$  is less than one. In this case,  $B^t X_0(\omega)$  becomes negligible as  $t$  goes to infinity for a fixed  $\omega$ . As seen in Example 2.9, however, the process  $X_t$  is not covariance stationary in general. Whether or not  $X_t$  is stationary depends upon the initial condition  $X_0$ . In order to choose  $X_0$ , consider an MA process

$$(4.24) \quad X_t = \mu + e_t + Be_{t-1} + B^2e_{t-2} + \cdots,$$

and choose the initial condition for the process  $X_t$  in (4.21) by

$$(4.25) \quad X_0 = \mu + e_0 + Be_{-1} + B^2e_{-2} + \cdots.$$

When this particular initial condition is chosen,  $X_t$  is covariance stationary.

With the lag operator, (4.22) can be written as

$$(4.26) \quad (1 - BL)(X_t - \mu) = e_t.$$

We define the inverse of  $(1 - BL)$  as

$$(4.27) \quad (1 - BL)^{-1} = 1 + BL + B^2L^2 + B^3L^3 + \cdots,$$

when the absolute value of  $B$  is less than one. When a process  $X_t$  has an MA representation of the form (4.24), we write

$$(4.28) \quad X_t = \mu + (1 - BL)^{-1}e_t,$$

which is the MA( $\infty$ ) representation of an AR(1) process.

### 4.5.2 The $p$ -th Order Autoregression

A  $p$ -th order autoregression satisfies

$$(4.29) \quad X_t = \delta + B_1X_{t-1} + B_2X_{t-2} + \cdots + B_pX_{t-p} + e_t \quad \text{for } t \geq 1.$$

The stability condition is that all the roots of

$$(4.30) \quad 1 - B_1z - B_2z^2 - \cdots - B_pz^p = 0$$

are larger than one in absolute value, or equivalently, all the roots of

$$(4.31) \quad z^p - B_1z^{p-1} - B_2z^{p-2} - \cdots - B_p = 0$$

are smaller than one in absolute value.

Consider, for instance, the special case of a AR(1) process with  $B_1 = 1$  and  $X_0 = 0$ :

$$(4.32) \quad X_t = X_{t-1} + e_t$$

$$(4.33) \quad = e_1 + e_2 + \cdots + e_{t-1} + e_t \quad \text{for } t \geq 1,$$

where  $E(X_t) = 0$  and  $E(X_{t-i}X_{t-j}) = \sigma^2$  for  $i = j$ . Note that  $Var(X_1) = \sigma^2$ ,  $Var(X_2) = 2\sigma^2$ ,  $\cdots$ ,  $Var(X_t) = t\sigma^2$ . Since the variance of  $X_t$  varies over time,  $X_t$  is nonstationary. Note also that its first difference is stationary since  $e_t (= X_t - X_{t-1})$

is stationary. Such a process is called *difference stationary*. When a (possibly infinite order) polynomial in the lag operator  $\Phi(L) = \Phi_0 + \Phi_1 L + \Phi_2 L^2 + \dots$  is given, we consider a complex valued function  $\Phi(z^{-1}) = \Phi_0 + \Phi_1 z^{-1} + \Phi_2 z^{-2} + \dots$  by replacing the lag operator  $L$  by a complex number  $z$ . Consider a condition

$$(4.34) \quad \Phi(z) = \Phi_0 + \Phi_1 z + \Phi_2 z^2 + \dots = 0.$$

If a complex number  $z_i$  satisfies the condition (4.34), then  $z_i$  is a *zero* of  $\Phi(z)$ . We also say that  $z_i$  is a root of the equation  $\Phi(z) = 0$ .

## 4.6 ARMA

An ARMA( $p, q$ ) process satisfies

$$(4.35) \quad X_t = \delta + B_1 X_{t-1} + B_2 X_{t-2} + \dots + B_p X_{t-p} + e_t + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \dots$$

If  $B(1) = 1 - B_1 - \dots - B_p \neq 1$ , we have the deviation-from-the-mean form

$$(4.36) \quad B(L)(X_t - \mu) = \theta(L)e_t,$$

where  $\mu = \frac{\delta}{B(1)}$ . We define the inverse of  $B(L) = B_0 + B_1 L + \dots + B_p L^p$  as the lag polynomial  $B(L)^{-1}$  such that

$$(4.37) \quad B(L)^{-1} B(L) = 1.$$

As long as  $B_0 \neq 0$ ,  $B(L)^{-1}$  exists always. However,  $B(L)^{-1} \varepsilon_t$  may or may not be defined. Provided that the  $p$ -th order polynomial  $B(z)$  satisfies stability conditions, the ARMA( $p, q$ ) process yields the MA( $\infty$ ) representation

$$(4.38) \quad X_t = \mu + \Phi(L)e_t,$$

Kyungho  
needs to  
check this!

where  $\Phi(L) = B(L)^{-1}\theta(L) = \Phi_0 + \Phi_1L + \theta_2L^2 + \dots$  and  $\sum_{j=0}^{\infty} |\theta_j|^2 \leq \infty$ .

On the other hand, if  $\theta(z)$  satisfies stability conditions that all roots of  $\theta(z) = 0$  lie outside the unit circle, then  $\theta(L)$  is invertible and the ARMA( $p, q$ ) process yields the AR( $\infty$ ) representation<sup>3</sup>

$$(4.39) \quad \theta(L)^{-1}B(L)X_t = \delta^* + e_t,$$

where  $\delta^* = \frac{\delta}{\theta(1)}$ . Therefore, if both  $B(z)$  and  $\theta(z)$  satisfy stability conditions, then the ARMA( $p, q$ ) process has both the MA( $\infty$ ) and AR( $\infty$ ) representations.

## 4.7 Fundamental Innovations

Let  $\mathbf{X}_t$  be a covariance stationary vector process with mean zero that is linearly regular. Then the Wold representation in (4.14) gives an MA representation. There are infinitely many other MA representations.

**Example 4.3** let  $u_t$  be a white noise, and  $X_t = u_t$ . Then  $X_t = u_t$  is an MA representation. Let  $u_t^* = u_{t+1}$ . Then  $X_t = u_{t-1}^*$  is another MA representation. ■

In this example, another MA representation is obtained by adopting a different dating procedure for the innovation.

It is often convenient to restrict our attention to the MA representations for which the information content of the current and past values of the innovations is the same as that of the current and past values of  $\mathbf{X}_t$ . Let

$$(4.40) \quad \mathbf{X}_t = \sum_{j=0}^{\infty} \Phi_j \mathbf{u}_{t-j} = \Phi(L)\mathbf{u}_t$$

---

<sup>3</sup>Without any loss of generality, we assume that there are no common roots of  $B(z) = 0$  and  $\theta(z) = 0$ . In such a case, we can write the ARMA( $p, q$ ) process by the ARMA( $p - m, q - m$ ) process that has no common roots, where  $m$  is the number of common roots. See Hayashi (2000, p. 382) for further discussion.

be an MA representation for  $\mathbf{X}_t$ . Let  $H_t$  be the linear information set generated by the current and past values of  $\mathbf{X}_t$ , and  $H_t^u$  be the linear information set generated by the current and past values of  $\mathbf{u}_t$ . Then  $H_t \subset H_t^u$  because of (4.40). The innovation process  $\mathbf{u}_t$  is said to be *fundamental* if  $H_t = H_t^u$ . The innovation in the Wold representation is fundamental.

In Example 4.3,  $X_t = u_t$  is a fundamental MA representation while  $X_t = u_{t-1}^*$  is not. As a result of the dating procedure used for  $X_t = u_{t-1}^*$ , the information set generated by the current and past values of  $u_t^* : \{u_t^*, u_{t-1}^*, \dots\}$  is equal to  $H_{t+1}$ , and is strictly larger than the information set generated by  $H_t$ .

The concept of fundamental innovations is closely related to the concept of invertibility. If the MA representation (4.40) is invertible, then  $\mathbf{u}_t = \Phi(L)^{-1}\mathbf{X}_t$ . Therefore,  $H_t^u \subset H_t$ . Since (4.40) implies  $H_t \subset H_t^u$ ,  $H_t = H_t^u$ . Thus if the MA representation (4.40) is invertible, then  $\mathbf{u}_t$  is fundamental.

If all the roots of  $\det[\Phi(z)] = 0$  lie outside the unit circle, then  $\Phi(L)$  is invertible, and  $\mathbf{u}_t$  is fundamental. If all the roots of  $\det[\Phi(z)] = 0$  lie on or outside the unit circle, then  $\Phi(L)$  may not be invertible, but  $\mathbf{u}_t$  is fundamental. Thus for fundamentalness, we can allow some roots of  $\det[\Phi(z)] = 0$  to lie on the unit circle.

In the univariate case, if  $X_t = \Phi(L)u_t$  and all the roots of  $\Phi(z) = 0$  lie on or outside the unit circle, then  $u_t$  is fundamental. For example, let  $X_t = u_t + \Phi u_{t-1}$ . If  $|\Phi| < 1$ , then this MA representation is invertible, and  $u_t$  is fundamental. If  $\Phi = 1$  or if  $\Phi = -1$ , then this MA representation is not invertible, but  $u_t$  is fundamental. If  $|\Phi| > 1$ , then  $u_t$  is not fundamental.

The MA representations with fundamental innovations are useful; it is easier to express projections of variables onto  $H_t$  with them than if they had non-fundamental

innovations. For example, let  $X_t$  be a univariate process with an MA(1) representation:  $X_t = u_t + \Phi u_{t-1}$ . It is natural to assume that economic agents observe  $X_t$ , but not  $u_t$ . Therefore, the economic agents' forecast for  $X_{t+1}$  can be modeled as  $\hat{E}(X_{t+1}|\mathbf{H}_t)$  rather than  $\hat{E}(X_{t+1}|\mathbf{H}_t^u)$ . If  $|\Phi| \leq 1$ ,  $u_t$  is fundamental, and  $\hat{E}(X_{t+1}|\mathbf{H}_t) = \hat{E}(X_{t+1}|\mathbf{H}_t^u) = \Phi u_t$ . On the other hand, if  $|\Phi| > 1$ ,  $u_t$  is not fundamental, and  $\hat{E}(X_{t+1}|\mathbf{H}_t) \neq \hat{E}(X_{t+1}|\mathbf{H}_t^u) = \Phi u_t$ , and there is no easy way to express  $\hat{E}(X_{t+1}|\mathbf{H}_t)$ .

## 4.8 The Spectral Density

Consider a covariance stationary process  $Y_t$  such that  $Y_t - E(Y_t)$  is linearly regular. Then  $Y_t - E(Y_t) = b(L)e_t = \sum_{j=0}^{\infty} b_j e_{t-j}$  for a square summable  $\{b_j\}$  and a white noise process  $e_t$  such that  $E(e_t^2) = 1$  and  $E(e_t e_s) = 0$  for  $t \neq s$ . Its  $k$ -th autocovariance  $\Phi(k) = E[(Y_t - E(Y_t))(Y_{t-k} - E(Y_{t-k}))']$  does not depend on date  $t$ . For a real number  $r$ , define

$$(4.41) \quad \exp(ir) = \cos(r) + i \sin(r),$$

where  $i = \sqrt{-1}$ . The spectral density of  $Y_t$ ,  $f(\lambda)$  is defined by

$$(4.42) \quad f(\lambda) = \left( \sum_{j=0}^{\infty} b_j \exp(-i\lambda j) \right) \left( \sum_{j=0}^{\infty} b_j \exp(i\lambda j) \right).$$

Then

$$(4.43) \quad f(\lambda) = \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} \Phi(k) \exp(i\lambda k)$$

for a real number  $\lambda$  ( $-\pi < \lambda < \pi$ ) when the autocovariances are absolutely summable. The spectral density is a function of  $\lambda$ , which is called the frequency. Using the

properties of the *cos* and *sin* functions and the fact that  $\Phi(k) = \Phi(-k)$ , it can be shown that

$$(4.44) \quad f(\lambda) = \frac{1}{2\pi} \Phi(0) + 2 \sum_{k=1}^{\infty} \Phi(k) \cos(\lambda k),$$

where  $f(\lambda) = f(-\lambda)$  and  $f(\lambda)$  is nonnegative for all  $\lambda$ .

Equation (4.43) gives the spectral density from the autocovariances. When the spectral density is given, the autocovariances can be calculated from the following formula:

$$(4.45) \quad \int_{-\pi}^{\pi} f(\lambda) \exp(i\lambda k) d\lambda = \Phi(k).$$

Thus the spectral density and the autocovariances contain the same information about the process. In some applications, it is more convenient to examine the spectral density than the autocovariances. For example, it requires infinite space to plot the autocovariance for  $k = 0, 1, 2, \dots$ , whereas the spectral density can be concisely plotted.

An interpretation of the spectral density is given by the special case of (4.45) in which  $k = 0$ :

$$(4.46) \quad \int_{-\pi}^{\pi} f(\lambda) d\lambda = \Phi(0).$$

This relationship suggests an intuitive interpretation that  $f(\lambda)$  is the contribution of the frequency  $\lambda$  to the variance of  $Y_t$ .

This intuition can be formalized by the *spectral representation theorem* which states that any covariance stationary process  $Y_t$  with absolutely summable autocovariances can be expressed in the form

$$(4.47) \quad Y_t = \mu + \int_0^{\pi} [\alpha(\lambda) \cos(\lambda t) + \delta(\lambda) \sin(\lambda t)] d\lambda,$$

where  $\alpha(\lambda)$  and  $\delta(\lambda)$  are random variables with mean zero for any  $\lambda$  in  $[0, \pi]$ . These variables have the further properties that for any frequencies  $0 < \lambda_1 < \lambda_2 < \lambda_3 < \lambda_4 < \pi$ , the variable  $\int_{\lambda_1}^{\lambda_2} \alpha(\lambda)$  is uncorrelated with  $\int_{\lambda_3}^{\lambda_4} \alpha(\lambda)$ , and the variable  $\int_{\lambda_1}^{\lambda_2} \delta(\lambda)$  is uncorrelated with  $\int_{\lambda_3}^{\lambda_4} \delta(\lambda)$ . For any  $0 < \lambda_1 < \lambda_2 < \pi$  and  $0 < \lambda_3 < \lambda_4 < \pi$ , the variable  $\int_{\lambda_1}^{\lambda_2} \alpha(\lambda)$  is uncorrelated with  $\int_{\lambda_3}^{\lambda_4} \delta(\lambda)$ . For such a process, the portion of the variance due to cycles with frequency less than or equal to  $\lambda_1$  is given by

$$(4.48) \quad 2 \int_0^{\lambda_1} f(\lambda) d\lambda.$$

## Exercises

**4.1** Let  $u_t$  be a white noise, and  $x_t = u_t + 0.8u_{t-1}$ . Is  $x_t$  covariance stationary? Is  $u_t$  fundamental for  $x_t$ ? Give an expression for  $\hat{E}(x_t | u_{t-1}, u_{t-2}, \dots)$  in terms of past  $u_t$ 's. Is it possible to give an expression for  $\hat{E}(x_t | x_{t-1}, x_{t-2}, \dots)$  in terms of past  $u_t$ 's? If so, give an expression. Explain your answers.

**4.2** Let  $u_t$  be a white noise, and  $x_t = u_t + 1.2u_{t-1}$ . Is  $x_t$  covariance stationary? Is  $u_t$  fundamental for  $x_t$ ? Give an expression for  $\hat{E}(x_t | u_{t-1}, u_{t-2}, \dots)$  in terms of past  $u_t$ 's. Is it possible to give an expression for  $\hat{E}(x_t | x_{t-1}, x_{t-2}, \dots)$  in terms of past  $u_t$ 's? If so, give an expression. Explain your answers.

**4.3** Let  $u_t$  be a white noise, and  $x_t = u_t + u_{t-1}$ . Is  $x_t$  covariance stationary? Is  $u_t$  fundamental for  $x_t$ ? Give an expression for  $\hat{E}(x_t | u_{t-1}, u_{t-2}, \dots)$  in terms of past  $u_t$ 's. Is it possible to give an expression for  $\hat{E}(x_t | x_{t-1}, x_{t-2}, \dots)$  in terms of past  $u_t$ 's? If so, give an expression. Explain your answers.

## References

HAYASHI, F. (2000): *Econometrics*. Princeton University Press, Princeton.

## Chapter 5

# STOCHASTIC REGRESSORS IN LINEAR MODELS

This chapter introduces the conditional Gauss-Markov Theorem, asymptotic theory, Monte Carlo, and Bootstrap as tools to evaluate estimators and tests. These tools are illustrated in the form that is convenient for most applications of structural econometrics for linear time series models in this chapter although they will be useful for nonlinear models as explained in later chapters.

In most applications in macroeconomics, regressors are stochastic, and the Gauss Markov Theorem for nonstochastic regressors do not apply. It is still possible to use the conditional Gauss Markov Theorem in some applications if a strict version of the exogeneity assumption (which will be called the strict exogeneity assumption) can be made to show that the OLS estimator is unbiased and efficient conditional on the realization of the regressors. If a normality assumption is added, it the estimator's exact small sample distributions can be obtained.

In some applications such as those of dynamic cointegrating regression explained in Chapter 14, the strict exogeneity assumption is typically made. So the conditional Gauss Markov Theorem can be used. However, in many other time series applications,

the strict exogeneity assumption is not attractive. If lagged dependent variables are included in regressors, the assumption cannot be made because it causes logical inconsistency. If the strict exogeneity assumption does not apply, then estimators are biased.

In rational expectations models, stringent distributional assumptions, such as an assumption that the disturbances are normally distributed, are unattractive. Without such assumptions, however, it is not possible to obtain the exact distributions of estimators in finite samples. For this reason, asymptotic theory describes the properties of estimators as the sample size goes to infinity.

Many researchers use asymptotic theory at initial stages of an empirical research project. Given the difficulties of obtaining the exact small sample distributions of estimators in many applications, this utilization seems to be a sound strategy. If the sample size is “large”, then asymptotic theory must be a good approximation of the true properties of estimators. The problem is that no one knows how large the sample size should be, because the answer depends on the nature of each application. After the importance of a research project is established, small sample properties of the estimators used in the project are often studied. For this purpose, Monte Carlo experiments can be used.

When asymptotic theory gives poor approximations in small sample, Bootstrap methods can be very useful. Bootstrap methods often give more accurate approximations of the exact small sample properties than asymptotic theory in applications to cross sectional data. In time series applications, there are some difficult issues that Bootstrap methods can have. This chapter explains such a difficulty that applied researchers should be aware of.

## 5.1 The Conditional Gauss Markov Theorem

Masao  
needs to  
check this!

In regressions (5.4) and (5.7),  $\mathbf{X}_t$  is *strictly exogenous in the time series sense* if  $E(e_t | \dots, \mathbf{X}_{t+2}, \mathbf{X}_{t+1}, \mathbf{X}_t, \mathbf{X}_{t-1}, \mathbf{X}_{t-2}, \dots) = 0$ . This is a very restrictive assumption that does not hold in all applications of cointegration discussed in Chapter 15. For example,  $E(e_t | \mathbf{X}_t, \mathbf{X}_{t-1}, \mathbf{X}_{t-2}, \dots) = 0$  in some applications because  $e_t$  is a forecast error. However, the forecast error is usually correlated with future values of  $\mathbf{X}_t$ . Hence the strict exogeneity assumption is violated. Nevertheless, as Choi and Ogaki (1999) argue, it is useful to observe that the Gauss Markov theorem applies to cointegrating regressions in order to understand small sample properties of various estimators for cointegrating vectors. Moreover, this observation leads to a Generalized Least Squares (GLS) correction to spurious regressions.

Let  $\sigma(\mathbf{X})$  be the smallest  $\sigma$ -field with respect to which the random variables in  $\mathbf{X}$  are measurable. We use the notation  $E[Z | \sigma(\mathbf{X})]$  to denote the usual conditional expectation of  $Z$  conditional on  $\mathbf{X}$  as defined by Billingsley (1986) for a random variable  $Z$ .  $E[Z | \sigma(\mathbf{X})]$  is a random variable, and  $E[Z | \sigma(\mathbf{X})](s)$  denotes the value of the random variable at  $s$  in  $S$  (what is  $s$ ?). It should be noted that the definition is given under the condition that  $Z$  is integrable, namely  $E(|Z|) < \infty$ .

Masao  
needs to  
check this!

This condition can be too restrictive when we define the conditional expectation of the OLS estimator in some applications as we discuss later. <sup>1</sup>

Masao  
needs to  
check this!

For this reason, we will also use a different concept of expectation conditional on  $\mathbf{X}$  that can be used when  $Z$  and  $vec(\mathbf{X})$  have probability density functions  $f_Z(z)$

Masao  
needs to  
check this!

<sup>1</sup>Loeve (1978) slightly relaxes this restriction by defining the conditional expectation for any random variable whose expectation exists (but may not be finite) with an extension of the Radon-Nikodym theorem. This definition can be used for  $E(\cdot | \sigma(X))$ , but this slight relaxation does not solve our problem which we describe later.

and  $f_X(\text{vec}(\mathbf{x}))$ , respectively. In this case, if  $f_X(\text{vec}(\mathbf{x}))$  is positive, we define the expectation of  $Z$  conditional on  $\mathbf{X}(s) = \mathbf{x}$  as

$$(5.1) \quad E[Z|\mathbf{X}(s) = \mathbf{x}] = \int_{-\infty}^{\infty} \frac{f_Z(z)}{f_X(\text{vec}(\mathbf{x}))} dz.$$

For this definition, we use the notation  $E[Z|\mathbf{X}(s) = \mathbf{x}]$ . This definition can only be used when the probability density functions exist and  $f_X(\text{vec}(\mathbf{x}))$  is positive, but the advantage of this definition for our purpose is that the conditional expectation can be defined even when  $E(Z)$  does not exist. For example let  $Z = \frac{Y}{X}$  where  $Y$  and  $X$  are independent random variables with a standard normal distribution. Then  $Z$  has the Cauchy distribution, and  $E(Z)$  does not exist. Thus,  $E[Z|\sigma(X)]$  cannot be defined.<sup>2</sup> However, we can define  $E[Z|X(s) = x]$  for all  $s$  in the probability space because the density function of  $X$  is always positive.

In the special case in which both types of conditional expectations can be defined, they coincide. More precisely, suppose that  $Z$  and  $\text{vec}(\mathbf{X})$  have probability density functions, that the probability density function of  $\text{vec}(\mathbf{X})$  is always positive, and that  $Z$  is integrable. Then  $E[Z|\sigma(\mathbf{X})](s) = E[Z|\mathbf{X}(s)]$  with probability one.

Let  $\mathbf{y} = (y_1, y_2, \dots, y_T)'$  be a  $T \times 1$  vector of random variables, and  $\mathbf{e} = (e_1, e_2, \dots, e_T)'$  be a  $T \times 1$  vector of random variables. We are concerned with a linear model of the form:

**Assumption 5.1**  $\mathbf{y} = \mathbf{X}\mathbf{b}_0 + \mathbf{e}$ ,

where  $\mathbf{b}_0$  is a  $K \times 1$  vector of real numbers. We assume that the expectation of  $\mathbf{e}$  conditional on  $\mathbf{X}$  is zero:

---

<sup>2</sup>It should be noted that we cannot argue that  $E(Z) = E(E(\frac{Y}{X}|\sigma(X))) = E(\frac{E(Y|\sigma(X))}{X}) = 0$  even though  $\frac{1}{X}$  is measurable in  $\sigma(X)$  because  $E(\frac{Y}{X}|\sigma(X))$  is not defined.

**Assumption 5.2**  $E[\mathbf{e}|\sigma(\mathbf{X})] = \mathbf{0}$ .

Since  $E[\mathbf{e}|\sigma(\mathbf{X})]$  is only defined when each element of  $\mathbf{e}$  is integrable, Assumption 5.2 implicitly assumes that  $E(\mathbf{e})$  exists and is finite. It also implies  $E(\mathbf{e}) = \mathbf{0}$  because of the law of iterated expectations. Given  $E(\mathbf{e}) = \mathbf{0}$ , a sufficient condition for Assumption 5.2 is that  $\mathbf{X}$  is statistically independent of  $\mathbf{e}$ . Since Assumption 5.2 does not imply that  $\mathbf{X}$  is statistically independent of  $\mathbf{e}$ , Assumption 5.2 is weaker than the assumption of the independent stochastic regressors. With the next assumption, we assume that  $\mathbf{e}$  is conditionally homoskedastic and  $e_t$  is not serially correlated:

**Assumption 5.3**  $E[\mathbf{e}\mathbf{e}'|\sigma(\mathbf{X})] = \sigma^2\mathbf{I}_T$ .

Let  $G = \{s \text{ in } S : \mathbf{X}(s)'\mathbf{X}(s) \text{ is nonsingular}\}$ . Since the determinant of a matrix is a continuous function of the elements of a matrix,  $G$  is a member of the  $\sigma$ -field  $\mathcal{F}?????$ .

Masao  
needs to  
check this!

For any  $s$  in  $G$ , the OLS estimator is

$$(5.2) \quad \mathbf{b}_T = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

From Assumption 5.1,  $\mathbf{b}_T = \mathbf{b}_0 + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}$ . Hence the conditional Gauss-Markov theorem can be proved when the expectation of  $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}$  and  $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}\mathbf{e}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$  can be defined. For this purpose, we consider the following two alternative assumptions:

**Assumption 5.4**  $E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}\mathbf{e}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}]$  exists and is finite.

**Assumption 5.4'**  $\mathbf{e}$  and  $\text{vec}(\mathbf{X})$  have probability density functions, and the probability density functions of  $\text{vec}(\mathbf{X})$  are positive for all  $s$  in  $G$ .

A sufficient condition for Assumption 5.4 is that the distributions of  $\mathbf{X}$  and  $\mathbf{e}$  have finite supports. Under Assumption 5.4,  $E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}]$  also exists and is finite. Hence  $E(\mathbf{b}_T|\sigma(\mathbf{X}))$  can be defined. From Assumptions 5.1-5.3,  $E(\mathbf{b}_T|\sigma(\mathbf{X})) = \mathbf{b}_0 + E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}|\sigma(\mathbf{X})] = \mathbf{b}_0$  for  $s$  in  $G$  with probability  $Pr(G)$ . Under Assumptions 5.1-5.4,  $E[(\mathbf{b}_T - \mathbf{b}_0)'(\mathbf{b}_T - \mathbf{b}_0)|\sigma(\mathbf{X})]$  can be defined, and  $E[(\mathbf{b}_T - \mathbf{b}_0)'(\mathbf{b}_T - \mathbf{b}_0)|\sigma(\mathbf{X})] = E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}\mathbf{e}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}|\sigma(\mathbf{X})] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E[\mathbf{e}\mathbf{e}'|\sigma(\mathbf{X})]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$  for  $s$  in  $G$  with probability  $Pr(G)$ . The problem with Assumption 5.4 is that it is not easy to verify Assumption 5.4 for many distributions of  $\mathbf{X}$  and  $\mathbf{e}_t$  that are often used in applications and Monte Carlo studies.

Under Assumptions 5.1-5.3 and 5.4',  $E[\mathbf{b}_T|\mathbf{X}(s)] = \mathbf{b}_0$  and  $E[(\mathbf{b}_T - \mathbf{b}_0)'(\mathbf{b}_T - \mathbf{b}_0)|\mathbf{X}(s)] = \sigma^2(\mathbf{X}(s)'\mathbf{X}(s))^{-1}$  for any  $s$  in  $G$ .

Corresponding with Assumption 5.4 and 5.4', we consider two definitions of the conditional version of the Best Linear Unbiased Estimator (BLUE). Given a set  $H$  in the  $\sigma$ -field  $\mathcal{F}$ , the *Best Linear Unbiased Estimator (BLUE) conditional on  $\sigma(\mathbf{X})$  in  $H$*  is defined as follows. An estimator  $\mathbf{b}_T$  for  $\mathbf{b}_0$  is the BLUE conditional on  $\sigma(\mathbf{X})$  in  $H$  if (1)  $\mathbf{b}_T$  is linear conditional on  $\sigma(\mathbf{X})$ , namely,  $\mathbf{b}_T$  can be written as  $\mathbf{b}_T = \mathbf{A}\mathbf{y}$  where  $\mathbf{A}$  is a  $K \times T$  matrix, and each element of  $\mathbf{A}$  is measurable  $\sigma(\mathbf{X})$ ; (2)  $\mathbf{b}_T$  is unbiased conditional on  $\sigma(\mathbf{X})$  in  $G$ , namely,  $E(\mathbf{b}_T|\sigma(\mathbf{X})) = \mathbf{b}_0$  for  $s$  in  $H$  with probability  $Pr(H)$ ; (3) for any linear unbiased estimator  $\mathbf{b}^*$  conditional on  $\mathbf{X}(s) = \mathbf{x}$  for which  $E(\mathbf{b}^*\mathbf{b}^{*'})$  exists and is finite,  $E[(\mathbf{b}_T - \mathbf{b}_0)(\mathbf{b}_T - \mathbf{b}_0)'|\mathbf{X}(s) = \mathbf{x}] \leq E[(\mathbf{b}^* - \mathbf{b}_0)(\mathbf{b}^* - \mathbf{b}_0)'|\mathbf{X}(s) = \mathbf{x}]$  in  $H$  with probability  $Pr(H)$ , namely,  $E[(\mathbf{b}^* - \mathbf{b}_0)(\mathbf{b}^* - \mathbf{b}_0)'|\mathbf{X}(s) = \mathbf{x}] - E[(\mathbf{b}_T - \mathbf{b}_0)(\mathbf{b}_T - \mathbf{b}_0)'|\mathbf{X}(s) = \mathbf{x}]$  is a positive semidefinite matrix with probability one for  $s$  in  $H$  with probability  $Pr(H)$ .

An estimator  $\mathbf{b}_T$  for  $\mathbf{b}_0$  is the BLUE conditional on  $\mathbf{X}(s) = \mathbf{x}$  in  $H$  if (1)  $\mathbf{b}_T$

is linear conditional on  $\mathbf{X}(s)$  in  $H$ , namely,  $\mathbf{b}_T$  can be written as  $\mathbf{b}_T = \mathbf{A}\mathbf{y}$  where  $\mathbf{A}$  is a  $K \times T$  matrix, and each element of  $\mathbf{A}$  is measurable  $\sigma(\mathbf{X})$ ; (2)  $\mathbf{b}_T$  is unbiased conditional on  $\mathbf{X}(s) = \mathbf{x}$  in  $H$ , namely,  $E(\mathbf{b}_T | \mathbf{X}(s) = \mathbf{x}) = \mathbf{b}_0$  for any  $s$  in  $H$ ; (3) for any linear unbiased estimator  $\mathbf{b}^*$  conditional on  $\mathbf{X}(s) = \mathbf{x}$  for which  $E(\mathbf{b}^* \mathbf{b}^{*'} | \mathbf{X}(s) = \mathbf{x})$  exists and is finite,  $E[(\mathbf{b}_T - \mathbf{b}_0)(\mathbf{b}_T - \mathbf{b}_0)' | \mathbf{X}(s) = \mathbf{x}] \leq E[(\mathbf{b}^* - \mathbf{b}_0)(\mathbf{b}^* - \mathbf{b}_0)' | \mathbf{X}(s) = \mathbf{x}]$  in  $H$ , namely,  $E[(\mathbf{b}^* - \mathbf{b}_0)(\mathbf{b}^* - \mathbf{b}_0)' | \mathbf{X}(s) = \mathbf{x}] - E[(\mathbf{b}_T - \mathbf{b}_0)(\mathbf{b}_T - \mathbf{b}_0)' | \mathbf{X}(s) = \mathbf{x}]$  is a positive semidefinite matrix for any  $s$  in  $H$ .

With these preparations, the following theorem can be stated:

**Theorem 5.1** (*The Conditional Gauss-Markov Theorem*) Under Assumptions 5.1-5.4, the OLS estimator is the BLUE conditional on  $\sigma(\mathbf{X})$  in  $G$ . Under Assumptions 5.1-5.3 and 5.4', the OLS estimator is the BLUE conditional on  $\mathbf{X}(s) = \mathbf{x}$  in  $G$ . ■

The theorem can be proved by applying any of the standard proofs of the (unconditional) Gauss-Markov theorem by replacing the unconditional expectation with the appropriate conditional expectation.

Under Assumptions 5.1-5.4, the unconditional expectation and the unconditional covariance matrix of  $\mathbf{b}_T$  can be defined. With an additional assumption that  $Pr(G) = 1$  or

**Assumption 5.5**  $\mathbf{X}'\mathbf{X}$  is nonsingular with probability one,

we obtain the following corollary of the theorem:

**Proposition 5.1** Under Assumptions 5.1-5.5, the OLS estimator is unconditionally unbiased and has the minimum unconditional covariance matrix among all linear unbiased estimators conditional on  $\sigma(\mathbf{X})$ .

**Proof** Using the law of iterated expectations,  $E(\mathbf{b}_T) = E\{E[\mathbf{b}_T|\sigma(\mathbf{X})]\} = E(\mathbf{b}_0) = \mathbf{b}_0$ , and  $E[(\mathbf{b}_T - \mathbf{b}_0)(\mathbf{b}_T - \mathbf{b}_0)'] = E\{E[(\mathbf{b}_T - \mathbf{b}_0)(\mathbf{b}_T - \mathbf{b}_0)']|\sigma(\mathbf{X})\} = \sigma^2 E[(\mathbf{X}'\mathbf{X})^{-1}]$ . For the minimum covariance matrix part, let  $\mathbf{b}^*$  be another linear unbiased estimator conditional on  $\sigma(\mathbf{X})$ . Then

$$(5.3) \quad E[(\mathbf{b}^* - \mathbf{b}_0)(\mathbf{b}^* - \mathbf{b}_0)'] = E[(\mathbf{b}_T - \mathbf{b}_0)(\mathbf{b}_T - \mathbf{b}_0)'] + \Delta,$$

where  $\Delta$  is a positive semidefinite matrix with probability one. Then  $E[(\mathbf{b}^* - \mathbf{b}_0)(\mathbf{b}^* - \mathbf{b}_0)'] - E[(\mathbf{b}_T - \mathbf{b}_0)(\mathbf{b}_T - \mathbf{b}_0)'] = [E(\mathbf{b}^*\mathbf{b}^{*'}) - \mathbf{b}_0\mathbf{b}_0'] - [E(\mathbf{b}_T\mathbf{b}_T') - \mathbf{b}_0\mathbf{b}_0'] = E[E(\mathbf{b}^*\mathbf{b}^{*'}|\sigma(\mathbf{X})) - E[E(\mathbf{b}_T\mathbf{b}_T'|\sigma(\mathbf{X}))]] = E(\Delta)$  is a positive semidefinite matrix. (?????) ■

Masao  
needs to  
check this!

A few remarks for this proposition are in order:

**Remark 5.1** Assumption 5.4 cannot be replaced by Assumption 5.4' for this proposition. Under Assumption 5.4',  $E(\mathbf{b}_T)$  and  $E[(\mathbf{b}_T - \mathbf{b}_0)(\mathbf{b}_T - \mathbf{b}_0)']$  may not exist. ■

**Remark 5.2** In this proposition, the covariance matrix of  $\mathbf{b}_T$  is  $\sigma^2 E[(\mathbf{X}'\mathbf{X})^{-1}]$ , which is different from  $\sigma^2 [E(\mathbf{X}'\mathbf{X})]^{-1}$ . This result may seem to contradict the standard asymptotic theory, but it does not. Asymptotically,  $\frac{1}{T}\mathbf{X}'\mathbf{X}$  converges almost surely to  $E[X_t X_t']$  if  $X_t$  is stationary and ergodic. Hence the limit of the covariance matrix of  $\sqrt{T}(\mathbf{b}_T - \mathbf{b}_0)$ ,  $\sigma^2 E[\{\frac{1}{T}(\mathbf{X}'\mathbf{X})\}^{-1}]$ , is equal to the asymptotic covariance matrix,  $\sigma^2 [E(X_t X_t')]^{-1}$ . ■

## 5.2 Unconditional Distributions of Test Statistics

In order to study distributions of the  $t$  ratios and  $F$  test statistics, we need an additional assumption:

**Assumption 5.6** Conditional on  $\mathbf{X}$ ,  $\mathbf{e}$  follows a multivariate normal distribution.

Given a  $1 \times K$  vector of real numbers  $\mathbf{R}$ , consider a random variable

$$(5.4) \quad N_R = \frac{\mathbf{R}(\mathbf{b}_T - \mathbf{b}_0)}{\sigma[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}]^{\frac{1}{2}}}$$

and the usual  $t$  ratio for  $\mathbf{R}\mathbf{b}_0$

$$(5.5) \quad t_R = \frac{\mathbf{R}(\mathbf{b}_T - \mathbf{b}_0)}{\hat{\sigma}[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{\frac{1}{2}}}.$$

Here  $\hat{\sigma}$  is the positive square root of  $\hat{\sigma}^2 = \frac{1}{T-K}(\mathbf{y} - \mathbf{X}\mathbf{b}_T)'(\mathbf{y} - \mathbf{X}\mathbf{b}_T)$ . With the standard argument,  $N_R$  and  $t_R$  can be shown to follow the standard normal distribution and Student's  $t$  distribution with  $T - K$  degrees of freedom conditional on  $\mathbf{X}$ , respectively, under either Assumptions 5.1-5.6 or Assumptions 5.1-5.3, 5.4', and 5.5-5.6. The following proposition is useful in order to derive the unconditional distributions of these statistics.

**Proposition 5.2** If the probability density function of a random variable  $Z$  conditional on a random vector  $\mathbf{Q}$  does not depend on the values of  $\mathbf{Q}$ , then the marginal probability density function of  $Z$  is equal to the probability density function of  $Z$  conditional on  $\mathbf{Q}$ . ■

This proposition is obtained by integrating the probability density function conditional on  $\mathbf{Q}$  over all possible values of the random variables in  $\mathbf{Q}$ . Since  $N_R$  and  $t_R$  follow the standard normal and the Student's  $t$  distribution conditional on  $\mathbf{X}$ , respectively, Proposition 5.2 implies the following proposition:

**Proposition 5.3** Under the Assumptions 5.1-5.6, or under the Assumptions 5.1-5.3, 5.4', and 5.5-5.6,  $N_R$  is the standard normal random variable and  $t_R$  is the Student's  $t$  random variable with  $T - K$  degrees of freedom. ■

Similarly, the usual  $F$  test statistics also follow (unconditional)  $F$  distributions. These results are sometimes not well understood by econometricians. For example, a standard textbook, Judge et al. (1985, p.164), states that "our usual test statistics

do not hold in finite samples” on the grounds that  $\mathbf{b}_T$ 's (unconditional) distribution is not normal. It is true that  $\mathbf{b}_T$  is a nonlinear function of  $\mathbf{X}$  and  $\mathbf{e}$ , so it does not follow a normal distribution even if  $\mathbf{X}$  and  $\mathbf{e}$  are both normally distributed. However, the usual  $t$  and  $F$  test statistics have usual (unconditional) distributions as a result of Proposition 5.2.

### 5.3 The Law of Large Numbers

If an estimator  $\mathbf{b}_T$  converges almost surely to a vector of parameters  $\mathbf{b}_0$ , then  $\mathbf{b}_T$  is *strongly consistent* for  $\mathbf{b}_0$ . If an estimator  $\mathbf{b}_T$  converges in probability to a vector of parameters  $\mathbf{b}_0$ , then  $\mathbf{b}_T$  is *weakly consistent* for  $\mathbf{b}_0$ .

Consider a univariate stationary stochastic process  $\{X_t\}$ . When  $X_t$  is stationary,  $E(X_t)$  does not depend on date  $t$ . Therefore, we often write  $E(X)$  instead of  $E(X_t)$ . Assume that  $E(|X|)$  is finite, and consider a sequence of random variables  $[Y_T : T \geq 1]$ , where  $Y_T = \frac{1}{T} \sum_{t=1}^T X_t$  is the sample mean of  $X$  computed from a sample of size  $T$ . In general, the sample mean does not converge to its unconditional expected value, but converges almost surely to an expectation of  $X$  conditional on an information set. For the sample mean to converge almost surely to its unconditional mean, we require the series to be ergodic. A stationary process  $\{X_t\}$  is said to be *ergodic* if, for any bounded functions  $f : R^{i+1} \mapsto R$  and  $g : R^{j+1} \mapsto R$ ,

$$(5.6) \quad \begin{aligned} & \lim_{T \rightarrow \infty} |E[f(X_t, \dots, X_{t+i})g(X_{t+T}, \dots, X_{t+T+j})]| \\ &= |E[f(X_t, \dots, X_{t+i})]| |E[g(X_t, \dots, X_{t+j})]|. \end{aligned}$$

Heuristically, a stationary process is ergodic if it is asymptotically independent: that is, if  $(X_t, \dots, X_{t+i})$  and  $(X_{t+T}, \dots, X_{t+T+j})$  are approximately independent for large enough  $T$ .

**Proposition 5.4** (*The strong law of large numbers*) If a stochastic process  $[X_t : t \geq 1]$  is stationary and ergodic, and if  $E(|X|)$  is finite, then  $\frac{1}{T} \sum_{t=1}^T X_t \rightarrow E(X)$  almost surely. ■

## 5.4 Convergence in Distribution and Central Limit Theorem

This section explains a definition of convergence in distribution and presents some central limit theorems. These central limit theorems are based on martingale difference sequences, and are useful in many applications of rational expectations models.

Central limit theorems establish that the sample mean scaled by  $T$  converges in distribution to a normal distribution<sup>3</sup> under various regularity conditions. The following central limit theorem by Billingsley (1961) is useful for many applications because we can apply it when economic models imply that a variable is a martingale difference sequence.

**Proposition 5.5** (*Billingsley's Central Limit Theorem*) Suppose that  $e_t$  is a stationary and ergodic martingale difference sequence adapted to  $I_t$ , and that  $E(|e|^2) < \infty$ . Assume that  $I_{t-1} \subset I_t$  for all  $t$ . Then

$$\frac{1}{\sqrt{T}} \sum_{t=1}^T e_t \xrightarrow{D} N(0, E(e^2)).$$

■

If  $e_t$  is an i.i.d. white noise, then it is a stationary and ergodic martingale difference sequence adapted to  $I_t$  which is generated from  $\{e_t, e_{t-1}, \dots\}$ . Hence Billingsley's

---

<sup>3</sup>In some central limit theorems, the limiting distribution is not normal.

Central Limit Theorem is more general than the central limit theorems for i.i.d. processes such as the Lindeberg- Levy theorem, which is usually explained in econometric text books. However, Billingsley's Central Limit Theorem cannot be applied to any serially correlated series.

A generalization of the theorem to serially correlated series is due to Gordin (1969):

**Proposition 5.6** (*Gordin's Central Limit Theorem*) Suppose that  $e_t$  is a univariate stationary and ergodic process with mean zero and  $E(|e|^2) < \infty$ , that  $E(e_t|e_{t-j}, e_{t-j-1}, \dots)$  converges in mean square to 0 as  $j \rightarrow \infty$ , and that

$$(5.7) \quad \sum_{j=0}^{\infty} [E(r_{tj}^2)]^{\frac{1}{2}} < \infty,$$

where

$$(5.8) \quad r_{tj} = E(e_t | \mathbf{I}_{t-j}) - E(e_t | \mathbf{I}_{t-j-1}),$$

where  $\mathbf{I}_t$  is the information set generated from  $\{e_t, e_{t-1}, \dots\}$ . Then  $e_t$ 's autocovariances are absolutely summable, and

$$(5.9) \quad \frac{1}{\sqrt{T}} \sum_{t=1}^T e_t \xrightarrow{D} N(0, \Omega),$$

where

$$(5.10) \quad \Omega = \lim_{T \rightarrow \infty} \sum_{j=-T+1}^{T-1} E(e_t e_{t-j}).$$

■

When  $e_t$  is serially correlated, the sample mean scaled by  $T$  still converges to a normal distribution, but the variance of the limiting normal distribution is affected by serial correlation as in (5.10).

In (5.10),  $\Omega$  is called a *long-run variance* of  $e_t$ . Intuition behind the long-run variance can be obtained by observing

$$(5.11) \quad E\left[\left(\frac{1}{\sqrt{T}} \sum_{t=1}^T e_t\right)^2\right] = \sum_{j=-T+1}^{T-1} \frac{T-|j|}{T} E(e_t e_{t-j})$$

and that the right hand side (5.11) is the Cesaro sum of  $\sum_{j=-T+1}^{T-1} E(e_t e_{t-j})$ . Thus when  $\sum_{j=-T+1}^{T-1} E(e_t e_{t-j})$  converges, its limit is equal to the limit of the right hand side of (5.11) (Apostol, 1974).

Another expression for the long-run variance can be obtained from an MA representation of  $e_t$ . Let  $e_t = \Psi(L)u_t = \Psi_0 u_t + \Psi_1 u_{t-1} + \dots$  be an MA representation. Then  $E(e_t e_{t-j}) = (\Psi_j \Psi_0 + \Psi_{j+1} \Psi_1 + \Psi_{j+2} \Psi_2 + \dots) E(u_t^2)$ , and  $\Omega = \{(\Psi_0^2 + \Psi_1^2 + \Psi_2^2 + \dots) + 2(\Psi_1 \Psi_0 + \Psi_2 \Psi_1 + \Psi_3 \Psi_2 + \dots) + 2(\Psi_2 \Psi_0 + \Psi_3 \Psi_1 + \Psi_4 \Psi_2 + \dots) + \dots\} E(u_t^2) = (\Psi_0 + \Psi_1 + \Psi_2 + \dots)^2 E(u_t^2)$ . Hence

$$(5.12) \quad \Omega = \Psi(1)^2 E(u_t^2).$$

In the next example, we consider a multi-period forecasting model. For this model, it is easy to show that Gordin's Theorem is applicable to the serially correlated forecast error.

**Example 5.1** (*The Multi-Period Forecasting Model*) Suppose that  $I_t$  is an information set generated by  $\{\mathbf{Y}_t, \mathbf{Y}_{t-1}, \mathbf{Y}_{t-2}, \dots\}$ , where  $\mathbf{Y}_t$  is a stationary and ergodic vector stochastic process. In typical applications, economic agents are assumed to use current and past values of  $\mathbf{Y}_t$  to generate their information set. Let  $X_t$  be a stationary and ergodic random variable in the information set  $I_t$  with  $E(|X_t|^2) < \infty$ . We consider an  $s$ -period ahead forecast of  $X_t$ ,  $E(X_{t+s}|I_t)$ , and the forecast error,  $e_t = X_{t+s} - E(X_{t+s}|I_t)$ .

It is easy to verify that all the conditions for Gordin's Theorem are satisfied for  $e_t$ . Moreover, because  $E(e_t|\mathbf{I}_t) = 0$  and  $e_t$  is in the information set  $\mathbf{I}_{t+s}$ ,  $E(e_t e_{t-j}) = E(E(e_t e_{t-j}|\mathbf{I}_t)) = E(e_{t-j} E(e_t|\mathbf{I}_t)) = 0$  for  $j \geq s$ . Hence  $\Omega = \lim_{j \rightarrow \infty} \sum_{-j}^j E(e_t e_{t-j}) = \sum_{j=-s+1}^{s-1} E(e_t e_{t-j})$ . ■

Hansen (1985) generalized Gordin's Central Limit Theorem to vector processes. In this book, we call the generalized theorem Gordin and Hansen's Central Limit Theorem.

**Proposition 5.7** (*Gordin and Hansen's Central Limit Theorem*) Suppose that  $\mathbf{e}_t$  is a vector stationary and ergodic process with mean zero and finite second moments, that  $E(\mathbf{e}_t | \mathbf{e}_{t-j}, \mathbf{e}_{t-j-1}, \dots)$  converges in mean square to 0 as  $j \rightarrow \infty$ , and that

$$(5.13) \quad \sum_{j=0}^{\infty} [E(\mathbf{r}'_{tj} \mathbf{r}_{tj})]^{\frac{1}{2}} < \infty,$$

where

$$(5.14) \quad \mathbf{r}_{tj} = E(\mathbf{e}_t | \mathbf{I}_{t-j}) - E(\mathbf{e}_t | \mathbf{I}_{t-j-1}),$$

where  $\mathbf{I}_t$  is the information set generated from  $\{\mathbf{e}_t, \mathbf{e}_{t-1}, \dots\}$ . Then  $\mathbf{e}_t$ 's autocovariances are absolutely summable, and

$$\frac{1}{\sqrt{T}} \sum_{t=1}^T \mathbf{e}_t \xrightarrow{D} N(\mathbf{0}, \Omega)$$

where

$$(5.15) \quad \Omega = \lim_{T \rightarrow \infty} \sum_{j=-T+1}^{T-1} E(\mathbf{e}_t \mathbf{e}'_{t-j}).$$

■

The matrix  $\Omega$  in Equation (5.15) is called the *long-run covariance matrix* of  $\mathbf{e}_t$ .

As in the univariate case, another expression for the long-run covariance can be obtained from an MA representation of  $\mathbf{e}_t$ . Let  $\mathbf{e}_t = \Psi(L)\mathbf{u}_t = \Psi_0\mathbf{u}_t + \Psi_1\mathbf{u}_{t-1} + \dots$  be an MA representation. Then  $E(\mathbf{e}_t\mathbf{e}'_{t-j}) = (\Psi_j + \Psi_{j+1} + \Psi_{j+2} + \dots)E(\mathbf{u}_t\mathbf{u}'_t)(\Psi_0 + \Psi_1 + \Psi_2 + \dots)'$ , and  $\Omega = (\Psi_0 + \Psi_1 + \Psi_2 + \dots)E(\mathbf{u}_t\mathbf{u}'_t)(\Psi_0 + \Psi_1 + \Psi_2 + \dots)'$ . Hence

$$(5.16) \quad \Omega = \Psi(1)E(\mathbf{u}_t\mathbf{u}'_t)\Psi(1)'$$

In the next example, Gordin and Hansen's Central Limit Theorem is applied to a serially correlated vector process:

**Example 5.2** Continuing Example 5.1, let  $\mathbf{Z}_t$  be a random vector with finite second moments in the information set  $I_t$ . Define  $\mathbf{f}_t = \mathbf{Z}_t e_t$ . Then  $E(\mathbf{f}_t|I_t) = E(\mathbf{Z}_t e_t|I_t) = E(\mathbf{Z}_t E(e_t|I_t)) = \mathbf{0}$ . In empirical work, it is often necessary to apply a central limit theorem to a random vector such as  $\mathbf{f}_t$ . It is easy to verify that all conditions for Gordin and Hansen's Theorem are satisfied for  $\mathbf{f}_t$ . Moreover,  $E(\mathbf{f}_t|I_t) = \mathbf{0}$  and  $\mathbf{f}_t$  is in the information set  $I_{t+s}$ , thus  $E(\mathbf{f}_t\mathbf{f}'_{t-j}) = E(E(\mathbf{f}_t\mathbf{f}'_{t-j}|I_t)) = E(E(\mathbf{f}_t|I_t)\mathbf{f}'_{t-j}) = \mathbf{0}$  for  $j \geq s$ . Hence  $\Omega = \lim_{j \rightarrow \infty} \sum_{-j}^j E(\mathbf{f}_t\mathbf{f}'_{t-j}) = \sum_{j=-s+1}^{s-1} E(\mathbf{f}_t\mathbf{f}'_{t-j})$ . ■

We assumed that the process is stationary and ergodic for the law of large numbers and central limit theorems. In most applications, this ergodic stationarity assumption is general enough. However, in some applications, such an assumption may not be convenient. For example, suppose that data of a process of interest shows an initial rapid growth and then stabilizes. It is not attractive to assume ergodic stationarity because the expected value of the process seems initially rising. In such cases, we can use an alternative assumption that the process is mixing. Mixing can be regarded as an asymptotic independence. For stationary and ergodic processes,

we used the concept of martingale difference sequence for central limit theorems. For mixing processes, the corresponding concept is mixingale processes. The concepts of mixing and mixingale are explained in Appendix A.

## 5.5 Consistency and Asymptotic Distributions of OLS Estimators

Consider a linear model,

$$(5.17) \quad y_t = \mathbf{x}'_t \mathbf{b}_0 + e_t,$$

where  $y_t$  and  $e_t$  are stationary and ergodic random variables, and  $\mathbf{x}_t$  is a  $p$ -dimensional stationary and ergodic random vector. We assume that the orthogonality conditions

$$(5.18) \quad E(\mathbf{x}_t e_t) = \mathbf{0}$$

are satisfied, and that  $E(\mathbf{x}_t \mathbf{x}'_t)$  is nonsingular.<sup>4</sup> Imagine that we observe a sample of  $(y_t, \mathbf{x}'_t)$  of size  $T$ . Proposition 5.4 shows that  $\frac{1}{T} \sum_{t=1}^T \mathbf{x}_t \mathbf{x}'_t$  converges to  $E(\mathbf{x}_t \mathbf{x}'_t)$  almost surely. Hence with probability one,  $\sum_{t=1}^T \mathbf{x}_t \mathbf{x}'_t(s)$  is nonsingular for large enough  $T$ , and the Ordinary Least Squares (OLS) estimator for (5.17) can be written as

$$(5.19) \quad \mathbf{b}_T = \left( \sum_{t=1}^T \mathbf{x}_t \mathbf{x}'_t \right)^{-1} \left( \sum_{t=1}^T \mathbf{x}_t y_t \right).$$

In order to apply the Law of Large Numbers to show that the OLS estimator is strongly consistent, rewrite (5.19) from (5.17) after scaling each element of the right side by  $T$ :

$$(5.20) \quad \mathbf{b}_T - \mathbf{b}_0 = \left( \frac{1}{T} \sum_{t=1}^T \mathbf{x}_t \mathbf{x}'_t \right)^{-1} \left( \frac{1}{T} \sum_{t=1}^T (\mathbf{x}_t e_t) \right).$$

---

<sup>4</sup>Appendix 3.A explains why these types of conditions are called orthogonality conditions.

Applying Proposition 5.4, we obtain

$$(5.21) \quad \mathbf{b}_T - \mathbf{b}_0 \rightarrow [E(\mathbf{x}_t \mathbf{x}_t')]^{-1} (E(\mathbf{x}_t e_t)) = \mathbf{0} \quad \text{almost surely.}$$

Hence the OLS estimator,  $\mathbf{b}_T$ , is a strongly consistent estimator. In order to obtain the asymptotic distribution of the OLS estimator, we make an additional assumption that a central limit theorem applies to  $\mathbf{x}_t e_t$ . In particular, assuming that Gordin and Hansen's Martingale Approximation Central Limit Theorem is applicable, we multiply both sides of (5.20) by the square root of  $T$ :

$$(5.22) \quad \sqrt{T}(\mathbf{b}_T - \mathbf{b}_0) = \left( \frac{1}{T} \sum_{t=1}^T \mathbf{x}_t \mathbf{x}_t' \right)^{-1} \left( \frac{1}{\sqrt{T}} \sum_{t=1}^T (\mathbf{x}_t e_t) \right).$$

Therefore,

$$(5.23) \quad \sqrt{T}(\mathbf{b}_T - \mathbf{b}_0) \xrightarrow{D} N(\mathbf{0}, [E(\mathbf{x}_t \mathbf{x}_t')]^{-1} \boldsymbol{\Omega} [E(\mathbf{x}_t \mathbf{x}_t')]^{-1})$$

where  $\boldsymbol{\Omega}$  is the long-run covariance matrix of  $\mathbf{x}_t e_t$ :

$$(5.24) \quad \boldsymbol{\Omega} = \sum_{j=-\infty}^{\infty} E(e_t e_{t-j} \mathbf{x}_t \mathbf{x}_{t-j}').$$

## 5.6 Consistency and Asymptotic Distributions of IV Estimators

Consider the linear model (5.17) for which the orthogonality conditions (5.18) are not satisfied. In this case, we try to find a  $p$ -dimensional stationary and ergodic random vector  $\mathbf{z}_t$ , which satisfies two types of conditions: the orthogonality condition

$$(5.25) \quad E(\mathbf{z}_t e_t) = \mathbf{0},$$

and the relevance condition that  $E(\mathbf{z}_t \mathbf{x}_t')$  is nonsingular. We define the *Linear Instrumental Variable* (IV) estimator as

$$(5.26) \quad \mathbf{b}_T = \left( \sum_{t=1}^T \mathbf{z}_t \mathbf{x}_t' \right)^{-1} \sum_{t=1}^T \mathbf{z}_t y_t.$$

Then

$$(5.27) \quad \mathbf{b}_T - \mathbf{b}_0 = \left( \frac{1}{T} \sum_{t=1}^T \mathbf{z}_t \mathbf{x}_t' \right)^{-1} \left( \frac{1}{T} \sum_{t=1}^T \mathbf{z}_t e_t \right).$$

Applying Proposition 5.4, we obtain

$$(5.28) \quad \mathbf{b}_T - \mathbf{b}_0 \rightarrow [E(\mathbf{z}_t \mathbf{x}_t')]^{-1} (E(\mathbf{z}_t e_t)) = \mathbf{0} \quad \text{almost surely.}$$

Hence the linear IV estimator,  $\mathbf{b}_T$ , is a strongly consistent estimator. Assuming that the Vector Martingale Approximation Central Limit Theorem is applicable to  $\mathbf{z}_t e_t$ ,

$$(5.29) \quad \sqrt{T}(\mathbf{b}_T - \mathbf{b}_0) \xrightarrow{D} N(\mathbf{0}, [E(\mathbf{z}_t \mathbf{x}_t')]^{-1} \boldsymbol{\Omega} [E(\mathbf{z}_t \mathbf{x}_t')]^{-1})$$

where  $\boldsymbol{\Omega}$  is the long-run covariance matrix of  $\mathbf{z}_t e_t$ :

$$(5.30) \quad \boldsymbol{\Omega} = \sum_{j=-\infty}^{\infty} E(e_t e_{t-j} \mathbf{z}_t \mathbf{z}_{t-j}').$$

## 5.7 Nonlinear Functions of Estimators

In many applications of linear models, we are interested in nonlinear functions of  $\mathbf{b}_0$ , say  $\mathbf{a}(\mathbf{b}_0)$ . This section explains the delta method, which is a convenient method to derive asymptotic properties of  $\mathbf{a}(\mathbf{b}_T)$  as an estimator for  $\mathbf{b}_0$  where  $\mathbf{b}_T$  is a weakly consistent estimator for  $\mathbf{b}_0$ . In many applications,  $\mathbf{b}_T$  is an OLS estimator or a linear IV estimator. Later????? in this book we will use the proof of the delta method to prove the asymptotic normality of the GMM estimator. (?????  $\mathbf{a}$  not bold, f is better?)

**Proposition 5.8** Suppose that  $\{\mathbf{b}_T\}$  is a sequence of  $p$ -dimensional random vectors such that  $\sqrt{T}(\mathbf{b}_T - \mathbf{b}_0) \xrightarrow{D} \mathbf{z}$  for a random vector  $\mathbf{z}$ . If  $\mathbf{a}(\cdot) : R^p \mapsto R^r$  is continuously differentiable at  $\mathbf{b}$ , then

$$\sqrt{T}[\mathbf{a}(\mathbf{b}_T) - \mathbf{a}(\mathbf{b}_0)] \xrightarrow{D} \mathbf{d}(\mathbf{b}_0)\mathbf{z},$$

Masao  
needs to  
check this!

Masao  
needs to  
check this!

where  $\mathbf{d}(\mathbf{b}_0) = \frac{\partial \mathbf{a}(\mathbf{b})}{\partial \mathbf{b}'} \Big|_{\mathbf{b}=\mathbf{b}_0}$  denotes the  $r \times p$  matrix of first derivatives evaluated at  $\mathbf{b}_0$ . In particular, if  $\mathbf{z} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$ , then

$$\sqrt{T}[\mathbf{a}(\mathbf{b}_T) - \mathbf{a}(\mathbf{b}_0)] \xrightarrow{D} N(\mathbf{0}, \mathbf{d}(\mathbf{b}_0)\boldsymbol{\Sigma}\mathbf{d}(\mathbf{b}_0)').$$

Masao  
needs to  
check this!

**Proof** ?????????????

■

## 5.8 Remarks on Asymptotic Theory

When we use asymptotic theory, we do not have to make restrictive assumptions that the disturbances are normally distributed. Serial correlation and conditional heteroskedasticity can be easily taken into account as long as we can estimate the long-run covariance matrix (which is the topic of the next chapter).

It is a common mistake to think that the linearity of the formula for the long-run covariance matrix means a linearity assumption for the process of  $\mathbf{x}_t e_t$  (for the OLS estimator) or  $\mathbf{z}_t e_t$  (for the IV estimator). It should be noted that we did not assume that  $\mathbf{x}_t e_t$  or  $\mathbf{z}_t e_t$  was generated by linear functions (i.e., a moving average process in the terminology of Chapter 4) of independent white noise processes. Even when  $\mathbf{x}_t e_t$  or  $\mathbf{z}_t e_t$  is generated from nonlinear functions of independent white noise processes, the distributions based on the long-run covariance matrices give the correct limiting distributions. This point is related to the Wold representation for nonlinear processes discussed in Chapter 4. Even when  $\mathbf{z}_t e_t$  is generated as a nonlinear process, as long as it is a linearly regular and covariance stationary process, it has the Wold representation:  $\mathbf{z}_t e_t = \boldsymbol{\Psi}(L)\mathbf{u}_t$ , and its long-run covariance matrix is given by (5.30).

## 5.9 Monte Carlo Methods

This section gives an introduction to Monte Carlo methods. An important advanced Monte Carlo method called the Markov Chain Monte Carlo (MCMC)<sup>5</sup> will be explained in Chapter 12 for the Bayesian Approach. The MCMC method is a very powerful numerical integration method that can be used for both the Bayesian statistics and the Classical statistics even though most applications of the method so far have been in the Bayesian statistics. Asymptotic theory is used to obtain approximations of the exact finite sample properties of estimators and test statistics. In many time series applications, the exact finite sample properties cannot be obtained. For example, in a regression with lagged dependent variables, we can assume neither that the regressor is nonrandom nor that the error term is strictly exogenous in the time series sense. In many applications with financial variables, the assumption that the error term in a regression is normal is inappropriate because many authors have found evidence against normality for several financial variables. Asymptotic theory gives accurate approximations when the sample size is “large,” but exactly how “large” is enough depends on each application. One method to study the quality of asymptotic approximations is the Monte Carlo simulations.

### 5.9.1 Random Number Generators

In relatively simple Monte Carlo studies, data are generated with computer programs called *pseudo-random number generators*. These programs generate sequences of values that appear to be draws from a specified probability distribution. Modern pseudo-random generators are accurate enough that we can ignore the fact that

---

<sup>5</sup>It is important to the extent that its emergence is called the MCMC revolution.

numbers generated are not exactly independent draws from a specified probability distribution for most purposes.<sup>6</sup> Hence in the rest of this appendix, phrases such as “values that appear to be” are often suppressed.

Recall that when a probability space  $\Omega$  is given, the whole history of a stochastic process  $\{e_t(s)\}_{t=1}^N$  is determined when a point in the probability space  $s$  is given. For a random number generator, we use a number called the starting *seed* to determine  $s$ . Then the random number generator automatically updates the seed each time a number is generated. It should be noted that the same sequence of numbers is generated whenever the same starting seed is given to a random number generator.

Generated random numbers are used to generate samples. From actual data, we obtain only one sample, but in Monte Carlo studies, we can obtain many samples from generated random numbers. Each time a sample is generated, we compute estimators or test statistics of interest. After replicating many samples, we can estimate small sample properties of the estimators or test statistics by studying the generated distributions of these variables and compare them with predictions of asymptotic theory.

Most programs offer random number generators for the uniform distribution and the standard normal distribution. One can produce random numbers with other distributions by transforming generated random numbers. See the Appendix for more explanations.

---

<sup>6</sup>One exception is that a pseudo-random number generator ultimately cycles back to the initial value generated and repeats the sequence when too many numbers are generated. Most modern pseudo-random number generators cycle back after millions of values are drawn, and this tendency is not a problem for most Monte Carlo studies. However, in some studies in which millions or billions of values are needed, there can be a serious problem.

### 5.9.2 Estimators

When a researcher applies an estimator to actual data without the normality assumption, asymptotic theory is used as a guide of small sample properties of the estimator. In some cases, asymptotic theory does not give a good approximation of the exact finite sample properties. A Monte Carlo study can be used to estimate the true finite sample properties. For example, the mean, median, and standard deviation of the realized values of the estimator over generated samples can be computed and reported as estimates of the true values of these statistics. For example,  $N$  independent samples are created and an estimate  $b_i$  ( $i \geq 1$ ) for a parameter  $b_0$  is calculated for the  $i$ -th sample. Then the expected value of the estimator  $E(b_i)$  can be estimated by its mean over the samples:  $\frac{1}{N} \sum_{i=1}^N b_i$ . By the strong law of large numbers, the mean converges almost surely to the expected value.

Other properties can also be reported, depending on the purpose of the study. For example, Nelson and Startz (1990) report estimated 1%, 5%, 10%, 50%, 90%, and 99% fractiles for an IV estimator and compared them with fractiles implied by the asymptotic distribution. This influential paper uses Monte Carlo simulations to study the small sample properties of IV estimator and its  $t$ -ratio when instruments are poor in the sense that the relevance condition is barely satisfied.

When the deviation from the normal distribution is of interest, the skewness and kurtosis are often estimated and reported. The *skewness* of a variable  $Y$  with mean  $\mu$  is

$$(5.31) \quad \frac{E(Y - \mu)^3}{[Var(Y)]^{\frac{3}{2}}}.$$

A variable with negative skewness is more likely to be far below the mean than it is

to be far above, and conversely a variable with positive skewness is more likely to be far above the mean than it is to be below. If  $Y$  has a symmetric distribution such as a normal distribution, then the skewness is zero. The *kurtosis* of  $Y$  is

$$(5.32) \quad \frac{E(Y - \mu)^4}{[Var(Y)]^2}.$$

If  $Y$  is normally distributed, the kurtosis is 3. If the kurtosis of  $Y$  exceeds 3, then its distribution has more mass in the tails than the normal distribution with the same variance.

### 5.9.3 Tests

When a researcher applies a test to actual data without the normality assumption, asymptotic theory is typically used. For example, the critical value of 1.96 is used for a test statistic with the asymptotic normal distribution for the significance level of 5%. The significance level and critical value based on the asymptotic distribution are called the *nominal significance level* and the *nominal critical value*, respectively. The probability of rejecting the null hypothesis when it is true is called the *size* of the test. Since the asymptotic distribution is not exactly equal to the exact distribution of the test statistic, the true size of the test based on the nominal critical value is usually either larger or smaller than the nominal significance level. This property is called the *size distortion*. If the true size is larger than the nominal significance level, the test *overrejects* the null hypothesis and is said to be *liberal*. If the true size is smaller than the nominal significance level, the test *underrejects* the null hypothesis and is said to be *conservative*. Using the distribution of the test statistic produced by a Monte Carlo simulation, one can estimate the true critical value.

The *power* of the test is the probability of rejecting the null hypothesis when

the alternative hypothesis is true. In Monte Carlo studies, two versions of the power can be reported for each point of the alternative hypothesis: the power based on the nominal critical value and the power based on the estimated true critical value. The latter is called the *size corrected power*. The power based on the nominal critical value is also of interest because it is the probability of rejecting the null hypothesis in practice if asymptotic theory is used. On the other hand, the size corrected power is more appropriate for the purpose of comparing tests. For example, a liberal test tends to have a higher power based on the nominal critical value than a conservative test. However, we cannot conclude the liberal test is better from this observation because the probability of Type I error is not equal for the two tests.

## 5.10 Bootstrap

When the asymptotic distribution of a random variable such as a parameter estimate and test statistic is unknown or unreliable, an estimation method called the *bootstrap* is used as an alternative to the asymptotic theory. The bootstrap estimates the unknown underlying probability distribution of interest using a known distribution function generated by a random sampling procedure. In this sense, the bootstrap distribution treats the random sample as if it is a good representation of the population. Under mild regularity conditions and with the sample sizes typically encountered in applied work, this method can provide as accurate an approximation as that obtained from the asymptotic theory. Moreover, often in cross-sectional applications, the bootstrap approximations can achieve the level of accuracy comparable to higher-order asymptotic approximations. When the bootstrap improves upon first-order asymptotic approximations, it is said to benefit from *asymptotic refinements*.

Asymptotic refinements are an important feature of the bootstrap in reducing or eliminating finite-sample bias of an estimator or finite-sample errors in the rejection probabilities of statistical tests. For these reasons, since its introduction by Efron (1979), the bootstrap has become a practical and increasingly popular tool in applied econometrics.<sup>7</sup>

To illustrate how the bootstrap is implemented in a simple setting, suppose you have a random sample  $\{x_1, x_2, \dots, x_T\}$  of an i.i.d. random variable  $x$  with cumulative distribution function (CDF)  $F_0$ . Let  $Q_T = Q_T(x_1, x_2, \dots, x_T)$  denote the statistic of interest, and  $G_T(q, F_0) \equiv Pr(Q_T \leq q)$  the exact, finite-sample CDF of  $Q_T$ . Because  $F_0$  is usually unknown in applications, the bootstrap method replaces  $F_0$  with its estimator  $F_T$ , and approximates  $G_T(q, F_0)$  by the bootstrap distribution  $G_T(q, F_T)$  based on which you can make inferences about  $Q_T$ .

There are two possible specifications of  $F_T$ . The *nonparametric bootstrap* uses the empirical distribution function of the data as  $F_T$ . The other approach, the *parametric bootstrap*, uses a parametric estimator of  $F_0$  as  $F_T$ . For instance, if  $x$  is assumed to be normally distributed with mean  $\mu$  and variance  $\sigma^2$ , then  $F_T$  is defined as  $N(\hat{\mu}, \hat{\sigma}^2)$  where  $\hat{\mu}$  and  $\hat{\sigma}^2$  are consistent estimates of  $\mu$  and  $\sigma^2$ , respectively.

In most applications,  $G_T(q, F_T)$  cannot be evaluated analytically, but is approximated using a Monte Carlo simulation. The steps for this procedure are as follows.

1. Draw a bootstrap sample of size  $T$ ,  $X^* = \{x_1^*, x_2^*, \dots, x_T^*\}$ , from the distribution corresponding to  $F_T$  randomly. For the nonparametric bootstrap, the observations are resampled from the original data set with replacement, with each point in the sample having the equal probability  $1/T$  of being drawn. Clearly,

---

<sup>7</sup>See, e.g., Jeong and Maddala (1993) and Horowitz (2001) for survey and details of different bootstrap methods and their theoretical justification.

some of the original data points may be included in  $X^*$  once or more than once, while others may not be included at all. For the parametric bootstrap,  $X^*$  is generated using a random number generator.

2. Using  $X^*$ , compute the bootstrap statistic  $Q_{T,1}^* \equiv Q_T(x_1^*, x_2^*, \dots, x_T^*)$ .
3. Repeat steps 1 and 2  $B$  times to obtain observations  $\{Q_{T,1}^*, \dots, Q_{T,B}^*\}$ .
4. The bootstrap distribution  $G_T(q, F_T)$  is estimated by  $G_T^*(q, F_T) = Pr(Q_T^* \leq q)$  putting mass  $1/B$  at each point of  $\{Q_{T,1}^*, \dots, Q_{T,B}^*\}$ .

The resulting bootstrap distribution  $G_T^*(q, F_T)$  is then used to compute  $p$ -values or confidence intervals, and make inferences about  $Q_T$  which is computed in conventional ways.

In order for the bootstrap distribution  $G_T(\cdot, F_T)$  to be an adequate estimator of  $G_T(\cdot, F_0)$ , it must be consistent. That is,  $G_T(\cdot, F_T)$  must converge in probability to the asymptotic CDF of  $Q_T$ ,  $G_\infty(\cdot, F_0)$ , as  $T \rightarrow \infty$ . Essentially, the conditions for the consistency of  $G_T(\cdot, F_T)$  require that  $F_T$  is a consistent estimator of  $F_0$ , and  $G_T(\cdot, F)$  is continuous in  $F$  in an appropriate sense. It then follows that  $G_T(\cdot, F_T)$  approaches  $G_T(\cdot, F_0)$  for a sufficiently large sample size.<sup>8</sup>

Although these conditions are likely to be satisfied in many cases of interest in econometrics, they can be violated in some applications. For instance, for the heavy-tailed distributions of Athreya (1987) or the unit root AR(1) model of Basawa, Mallik, McCormick, Reeves, and Taylor (1991), the standard bootstrap method results in poor approximations to the asymptotic distribution of interest. Thus, although the

---

<sup>8</sup>For a precise definition of consistency, see Appendix A. For conditions for consistency and a detailed discussion on consistency, see Section 2.1 of Horowitz (2001).

bootstrap methods serve as an attractive alternative to the asymptotic theory in many applications, it must be borne in mind that, just as with any econometric methods, they, too, cannot be used blindly.

The following example illustrates an application of the bootstrap to an autoregressive (AR) model, and shows why it requires a non-standard procedure. Consider the AR process of order 1 with an intercept and time trend,

$$x_t = \theta + \mu t + \alpha x_{t-1} + \epsilon_t \quad \text{for } 1 \leq t \leq T,$$

where  $\epsilon_t$  is i.i.d.,  $|\alpha| < 1$ , and  $x_0$  is a random variable with a stationary distribution so that  $x_t$  is stationary. Let  $\hat{\alpha}$  be the ordinary least square (OLS) estimator of the autoregressive root  $\alpha$ . The usual asymptotic theory indicates that  $T^{1/2}(\hat{\alpha} - \alpha)$  converges in distribution to a normal random variable with zero mean. On the contrary, the OLS estimator is significantly downward biased, and the exact, finite-sample distribution of  $\alpha$  is asymmetric and has fatter tails than the normal distribution.<sup>9</sup> In this case, if  $\epsilon_t$  is i.i.d. and normally distributed, then the exact, finite-sample CDF of  $\hat{\alpha}$  only depends on  $\alpha$ , and can be computed numerically using Andrews' (1993) procedure without relying on Monte Carlo or bootstrap simulations. The deviations from the prediction of the asymptotic theory are considerable especially when  $\alpha$  is close to one. For example, for the sample size of 60, the OLS estimator has downward median-biases of 0.08, 0.09, and 0.15 when  $\alpha$  is 0.7, 0.85, and 0.99, respectively. Clearly, using the asymptotic distribution leads to an inaccurate approximation to the exact, finite-sample distribution of  $\hat{\alpha}$  and hence results in misleading inferences.<sup>10</sup>

<sup>9</sup>It should be noted that the strict exogeneity assumption is violated because of the lagged dependent variable. Hence the argument for the conditional Gauss-Markov theorem cannot be applied.

<sup>10</sup>An alternative asymptotic theory called the local-to-unity asymptotic theory can be applied in this case as in Chan and Wei (1987) and Phillips (1987)

If  $\epsilon_t$  is not i.i.d. or normally distributed, the exact, finite-sample distribution is estimated using bootstrap methods. Tables of the 0.05, 0.5, and 0.95 quantiles of  $\hat{\alpha}$  can be found in Andrews (1993) for different sample sizes, AR specifications, and distributions of  $\epsilon_t$ .<sup>11</sup>

An important characteristic of the AR models with a near unit root is that the asymptotic distribution of and hence quantile functions for the test statistic depend on  $\alpha$ . Nevertheless, the conventional bootstrap approximates quantile functions by evaluating them at the point estimate  $\hat{\alpha}$  and thereby making an implicit assumption that these functions are constant, which is false in the AR models. Consequently, the standard bootstrap confidence intervals fail to provide asymptotically correct coverage probabilities.

Table 1 summarizes the 0.05, 0.5, and 0.95 true quantiles of the nonstudentized test statistic  $S_T(\alpha) = \hat{\alpha} - \alpha$  for the sample sizes of 40 and 150 over the values of  $\alpha$  from 0.70 to 1, assuming that the errors are i.i.d. and normally distributed.<sup>12</sup>

Table 1

$\alpha$	T=40			T=150		
	$q_{0.05}$	$q_{0.5}$	$q_{0.95}$	$q_{0.05}$	$q_{0.5}$	$q_{0.95}$
0.70	-0.390	-0.118	0.071	-0.144	-0.029	0.062
0.80	-0.403	-0.135	0.038	-0.137	-0.031	0.045
0.85	-0.412	-0.146	0.019	-0.133	-0.033	0.035
0.90	-0.425	-0.160	-0.002	-0.129	-0.036	0.023
0.93	-0.436	-0.172	-0.017	-0.127	-0.038	0.015
0.97	-0.457	-0.194	-0.040	-0.127	-0.045	0.000
0.99	-0.472	-0.209	-0.055	-0.133	-0.052	-0.010
1.00	-0.481	-0.218	-0.065	-0.140	-0.060	-0.018

<sup>11</sup>For a probability  $p$ , the  $p$  quantile of a random variable  $X$  is the minimum value of  $x$  for which  $Pr(X \leq x) = p$  is satisfied.

<sup>12</sup>Following Andrews (1993), we restrict the parameter space to be  $\alpha \in (-1, 1]$ . This assumption is made in order to avoid the dependence of the distribution of the OLS estimator on the initial condition.

These values are computed from table 3 in Andrews (1993) by subtracting the true value of  $\alpha$  from the quantile values in the corresponding row. It is clear from the above table that the quantile functions are varying for different values of  $\alpha$ . An appropriate bootstrap quantile function must therefore be a function of  $\alpha$  rather than  $\hat{\alpha}$ :

$$q_{0.05}^*(\alpha) \leq S_T(\alpha) \leq q_{0.95}^*(\alpha),$$

such that

$$Pr(q_{0.05}^*(\alpha) \leq S_T(\alpha) \leq q_{0.95}^*(\alpha)) = 0.90.$$

The above statement is *exact* in the sense that once we know the exact finite distribution of the quantiles for a given  $\alpha$ , then this set has the correct coverage probability. The upper and lower bounds are thus given by

$$-q_{0.95}^*(\alpha) + \hat{\alpha} \leq \alpha \leq -q_{0.05}^*(\alpha) + \hat{\alpha}.$$

Table 1 can be used to compute the median-unbiased estimator and the two-sided 90% and one-sided 95% confidence intervals for  $\alpha$ . Because the grid of  $\alpha$  values is finite, interpolation may be necessary for the values of  $\alpha$  in between those reported. To see how the table can be used in applications, suppose you have the OLS estimate  $\hat{\alpha}$  of 0.781 and the sample size  $T$  of 40. The median-unbiased estimate of  $\alpha$  is the intersection of  $S_T(\alpha)$  and  $q_{0.5}^*(\alpha)$ . That is,  $\alpha$  is such that  $\hat{\alpha} - \alpha = q_{0.5}^*(\alpha)$ . According to table 1, this occurs when  $\alpha = 0.99$  ( $0.781 - 0.99 = -0.209$ ). The lower and upper bounds of the 90% confidence interval can be found in the same way. For the lower bound, the endpoint is the value of  $\alpha$  such that  $\hat{\alpha} - \alpha = q_{0.95}^*(\alpha)$ . You see that  $\alpha + q_{0.95}^*(\alpha) = 0.771$  for  $\alpha = 0.7$ , and  $\alpha + q_{0.95}^*(\alpha) = 0.838$  for  $\alpha = 0.8$ . Because  $\hat{\alpha} = 0.781$ , the lower bound must lie between 0.7 and 0.8. By interpolation, this is 0.715. The upper bound can be found by  $\hat{\alpha} - \alpha = q_{0.05}^*(\alpha)$ . Because the parameter

space is restricted to be  $\alpha \in (-1, 1]$ , any value of  $\hat{\alpha}$  that is above 0.519 for  $T=40$  and 0.860 for  $T=150$  corresponds to the upper bound of 1. Thus, in this example,  $\hat{\alpha} > 0.519$ , and hence the upper bound of the confidence interval is 1.

This interval is equivalent to Hansen's (1999) *grid bootstrap* for the case of the i.i.d. Gaussian errors. He proposes a nonparametric bootstrap method for constructing confidence intervals for  $\alpha$  from bootstrap quantile functions of  $\alpha$ , and reports that it has improved performance over the standard bootstrap method when  $\alpha$  is close to one.

The condition under which the grid bootstrap confidence interval is first-order accurate only requires that the nuisance parameters are consistently estimated, and no restriction is imposed on the estimate of the parameter of interest. On the other hand, the consistency of the standard bootstrap confidence interval requires that the parameters are consistently estimated and the test statistic of the hypothesis has an asymptotic distribution, where the convergence to the asymptotic distribution is locally uniform in the parameter space. Thus, the conditions for the grid bootstrap are strictly less restrictive than those for the latter in the sense of first-order asymptotic coverage, suggesting that the grid bootstrap is more broadly applicable.

## Appendix

### 5.A Weakly dependence process

Weakly dependence process is a stochastic process where serial dependence exists, but it is restricted suitably so that the limit theorems, such as LLN, CLT, and FCLT, can be applied. There are many different types of weakly dependence processes

depending on its degree of serial dependence. In this section, we review some of the most commonly used ones in the nonstationary econometrics.

The reason why we study weakly dependence process for the nonstationary econometrics is that the nonstationary econometrics is also time-series econometrics, and in time-series econometrics, serial dependence exists in almost all applications. Therefore, we want our asymptotic theories for the nonstationary econometrics can also be applied to the data that has serial dependence.

### 5.A.1 Independent Process

**Definition 5.A.1** *A stochastic process  $\{X_t\}_{-\infty}^{\infty}$  is said to be independent if  $P(A \cap B) = P(A)P(B)$  for a pair of  $A \in \mathcal{F}_{-\infty}^t$  and  $B \in \mathcal{F}_{t+m}^{\infty}$  for all  $t$  and  $m$ .*

Independence implies that there is no relationship between  $X_t$  and  $X_{t'}$  for any  $t \neq t'$ , therefore each observation can be treated as an observation from a random sample. From the time series econometrics perspective, independence is the most stringent restriction on the behavior of a stochastic process. It is difficult to find a case where independence assumption is appropriate. However, it can be used as a benchmark against which asymptotic theories of other dependent processes might be compared.

### 5.A.2 Mixing Process

The idea of independence that there is no relationship between any pair of  $X_t$  and  $X_{t'}$  is rather special, especially for time series data. However, it might be reasonable to expect that the degree of dependence between  $X_t$  and  $X_{t'}$  is decreasing as the time  $t$  and  $t'$  are getting farther separated from each other. We formalize this idea by introducing the concept of mixing.

**Definition 5.A.2** A stochastic process  $\{X_t\}_{-\infty}^{\infty}$  is said to be mixing (or regular) if, for every  $B \in \mathcal{F}$ ,

$$\sup_{A \in \mathcal{F}_{-\infty}^t} |P(A \cap B) - P(A)P(B)| \rightarrow 0 \text{ as } t \rightarrow -\infty.$$

Mixing can be regarded as an asymptotic independence. Note that an independent process is also mixing. An alternative definition of mixing can be described in terms of remote event. Remote event is defined as an event contained in the remote  $\sigma$ -field,  $\mathcal{F}_{-\infty} = \bigcap_t \mathcal{F}_{-\infty}^t$ .

**Definition 5.A.3** A stochastic process  $\{X_t\}_{-\infty}^{\infty}$  is said to be mixing (or regular) if every remote event has probability 0 or 1.

Since mixing is defined by remote events as in Definition 5.A.3, it can hardly provide us with useful description of dependence between events that are widely separated in time, but not in the remote events. Therefore, for a workable theory we need the concepts of mixing coefficients. In this section, we introduce only two most important mixing coefficients,  $\alpha$ -mixing and  $\phi$ -mixing although there are several other different versions available. Let  $\mathcal{G}$  and  $\mathcal{H}$  be  $\sigma$ -subfields of  $\mathcal{F}$ . The  $\alpha$ -mixing (strong mixing) coefficient is defined by

$$\alpha(\mathcal{G}, \mathcal{H}) = \sup_{G \in \mathcal{G}, H \in \mathcal{H}} |P(G \cap H) - P(G)P(H)|,$$

the uniform mixing coefficient is defined by

$$\phi(\mathcal{G}, \mathcal{H}) = \sup_{G \in \mathcal{G}, H \in \mathcal{H}; P(G) > 0} |P(H|G) - P(H)|$$

Then, the sequence  $\{X_t\}_{-\infty}^{\infty}$  is said to be  $\alpha$ -mixing (or strong mixing) if

$$\alpha_m = \sup_t \alpha(\mathcal{F}_{-\infty}^t, \mathcal{F}_{t+m}^{\infty}) \rightarrow 0 \text{ as } m \rightarrow \infty,$$

similary, it is said to be  $\phi$ -mixing (or uniform mixing) if

$$\phi_m = \sup_t \phi(\mathcal{F}_{-\infty}^t, \mathcal{F}_{t+m}^\infty) \rightarrow 0 \text{ as } m \rightarrow \infty$$

Note that if  $\alpha_m = 0$  for all  $m$ , the sequence becomes independent. Measure of the dependence can be based on the rate of convergence at which the mixing coefficients tend to zero. The rate of convergence is quantified by that for some number  $\varphi > 0$ ,  $\alpha_m(\phi_m) \rightarrow 0$  sufficiently fast that

$$\sum_{m=1}^{\infty} \alpha_m^{\frac{1}{\varphi}} < \infty \text{ or } \sum_{m=1}^{\infty} \phi_m^{\frac{1}{\varphi}} < \infty.$$

A sequence is said to be  $\alpha$ -mixing ( $\phi$ -mixing) of size  $-\varphi_0$  if  $\alpha_m = O(m^{-\varphi})$  ( $\phi_m = O(m^{-\varphi})$ ) for some  $\varphi > \varphi_0$ .

### 5.A.3 Martingale Difference Process

Independence and mixing are conditions for every event in  $\mathcal{F}$ . Since sup is taken over all the events in  $\mathcal{F}$ , usually it is the most peculiar event that determines the properties. However, in many case, those peculiar event that determine the properties of a stochastic process may not be our main interest. Therefore, sometimes it is more useful if we confine our attention to more restricted measure of dependence, and admit more stochastic process into consideration. Martingale difference and mixingale are two key concepts.

**Definition 5.A.4** *A stochastic process  $\{X_t\}_{-\infty}^{\infty}$  is said to be a martingale difference (m.d.) sequence if  $X_t$  is integrable and*

$$E(X_t | \mathcal{F}_{-\infty}^{t-1}) = 0 \text{ a.s.}$$

Table 5.1: Dependence between  $X_t$  and  $X_{t+m}$ 

	For all $m$		As $m \rightarrow \infty$
Every events	Independent	$\Rightarrow$	Mixing
	$\Downarrow$		$\Downarrow$
1-period ahead predictability	Martingale Difference	$\Rightarrow$	Mixingale

This implies that  $\{\dots, X_{t-1}\}$  have no impact on the prediction of  $X_t$ . It can be thought that  $X_t$ 's are independent each other in terms of one-period ahead predictability.

#### 5.A.4 Mixingale Process

Although martingale difference restrict our attention to more restricted measure of dependence, namely predictability, it is still rather special in time series setting that  $X_t$  has no prediction power on  $X_{t+m}$  at all. Similarly in mixing, it might be more natural to expect that the degree of dependence between between  $X_t$  and  $X_{t+m}$  in term of predictability is getting smaller as the time  $m$  increases. Mixingale captures this idea.

**Definition 5.A.5**  $\{X_t\}_{-\infty}^{\infty}$  is said to be an  $L_p$ -mixingale if

$$\|E(X_t | \mathcal{F}_{-\infty}^{t-m})\|_p \leq \zeta_m \rightarrow 0 \text{ as } m \rightarrow \infty$$

This is the most general dependence concept for that most of asymptotic theories go through.

It can be said that mixingales are to mixing as martingale differences are to independent. Table 1 summarize the relationship among these dependence concepts.

### 5.A.5 Near-Epoch Dependent (NED) Process

**Definition 5.A.6** Let  $\{V_t\}_{-\infty}^{\infty}$  be a stochastic process on a probability space  $(\mathcal{S}, \mathcal{F}, P)$ . Define  $\sigma$ -subfields  $\mathcal{F}_s^t = \sigma(V_s, \dots, V_t)$ . A stochastic process  $\{X_t\}_{-\infty}^{\infty}$  is said to be  $L_p$ -NED on  $\{V_t\}_{-\infty}^{\infty}$  for  $p > 0$ , if for  $m \geq 0$ ,

$$\|X_t - E(X_t | \mathcal{F}_{t-m}^{t+m})\|_p \leq d_t \nu(m),$$

where  $d_t$  is a sequence of positive constants, and  $\nu(m) \rightarrow 0$  as  $m \rightarrow \infty$ .

We say that  $X_t$  is NED of size  $-\lambda$  on the process  $V_t$  if  $\nu(m) = O(m^{-\lambda-\varepsilon})$  for some  $\varepsilon > 0$ . In the application,  $V_t$  usually is a mixing process.

The near-epoch dependence concept is most useful due to the following theorem

**Theorem 5.2** Let  $\{V_t\}_{-\infty}^{\infty}$  be  $\alpha$ -mixing of size  $-a$ . If  $\{X_t\}_{-\infty}^{\infty}$  is an  $L_r$ -bounded zero-mean sequence and  $L_p$ -NED of size  $-b$  on  $V_t$  with constant  $\{d_t\}$  for  $r > p \geq 1$ , then  $\{X_t, \mathcal{F}_{-\infty}^t\}$  is an  $L_p$ -mixingale of size  $-\min\left[b, a\left(\frac{1}{p} - \frac{1}{r}\right)\right]$  with constant  $c_t \ll \max\{\|X_t\|_r, d_t\}$ .

**Theorem 5.3** Let  $\{V_t\}_{-\infty}^{\infty}$  be  $\phi$ -mixing of size  $-a$ . If  $\{X_t\}_{-\infty}^{\infty}$  is an  $L_r$ -bounded zero-mean sequence and  $L_p$ -NED of size  $-b$  on  $V_t$  with constant  $\{d_t\}$  for  $r > p \geq 1$ , then  $\{X_t, \mathcal{F}_{-\infty}^t\}$  is an  $L_p$ -mixingale of size  $-\min\left[b, a\left(1 - \frac{1}{r}\right)\right]$  with constant  $c_t \ll \max\{\|X_t\|_r, d_t\}$ .

## 5.B Functional Central Limit Theorem

The functional central limit theorem (FCLT) is a generalization of the central limit theorem (CLT) to a stochastic process; in the CLT, a sequence of distributions of

random variables converges to its limit, meanwhile, in the FCLT, a sequence of distributions of stochastic processes converges to its limit.

To see the difference, consider a sequence of stationary random variables  $u_t$  where  $E(u_t) = 0$  and  $E(u_t^2) = \sigma^2$ :

$$u_1, u_2, \dots, u_n.$$

From them, we can construct the following sequence of random variables:

$$X_n = \frac{1}{\sqrt{n}} \sum_{t=1}^n u_t.$$

Note that for every  $n$ ,  $X_n$  is a well-defined random variable, therefore it has a distribution denoted by  $F_n(x)$ . In the CLT, we are concerned about the limit of the sequence of the distributions. What the CLT imply is that for every  $x$  where  $F_\infty(x)$  is continuous, as  $n \rightarrow \infty$ ,

$$F_1(x), F_2(x), \dots, F_n(x), \dots \rightarrow F_\infty(x)$$

where  $F_\infty(x)$  is a normal distribution.

From the sequence of  $u_t$ 's, we can also construct the following sequence of random function of  $r \in [0, 1]$ :

$$X_n(r) = \frac{1}{\sqrt{n}} \sum_{t=1}^{[nr]} u_t.$$

Although it is not a simple task to define the distributions of the random functions, by abuse of notation, we can define  $F_n(x)$  be a distribution of  $X_n(r)$ . Then, what we are concerned about with the FCLT is limit of the sequence of the distributions,  $F_n(x)$ . What the FCLT imply is that for every  $x$  where  $F_\infty(x)$  is continuous, as  $n \rightarrow \infty$ ,

$$F_1(x), F_2(x), \dots, F_n(x), \dots \rightarrow F_\infty(x).$$

where  $F_\infty(x)$  is the distribution of the Wiener process. Formal definitions and theorems are given in the subsequent sub-sections.

For the notational convenience, we introduce a triangular stochastic array. Array notation is especially convenient when the points of a sample are subjected to scale transformations, depending on the whole sample. A typical example is  $\{\{X_{nt}\}_{t=1}^n\}_{n=1}^\infty$  where  $X_{nt} = \frac{X_t}{n}$ . A triangular stochastic array is a doubly-indexed collection of random variables,

$$\begin{pmatrix} X_{11} & X_{21} & X_{31} & \dots \\ X_{12} & X_{22} & X_{32} & \dots \\ \vdots & \vdots & \vdots & \\ X_{1,k_1} & \vdots & \vdots & \\ & X_{2,k_2} & \vdots & \\ & & X_{3,k_3} & \\ & & & \ddots \end{pmatrix},$$

which is compactly written as  $\{\{X_{mn}\}_{m=1}^{k_n}\}_{n=1}^\infty$ , where  $k_n$  is an increasing integer sequence.

### 5.B.1 Central Limit Theorem

Since the FCLT is a generalization of the CLT, we can understand the FCLT through the comparison with the CLT. Therefore, we review the CLT first. In below, we present two versions of the CLT: one for the martingale difference sequence, and the other for NED functions of strong mixing processes.

**Theorem 5.4** *Let  $\{U_{nt}, \mathcal{F}_{nt}\}$  be a martingale difference array with finite unconditional variances  $\{\sigma_{nt}^2\}$ , and  $\sum_{t=1}^n \sigma_{nt}^2 = 1$ . Define  $X_n = \sum_{t=1}^n U_{nt}$ . If the following assumptions holds:*

1.  $\sum_{t=1}^n U_{nt}^2 \xrightarrow{p} 1$

$$2. \max_{1 \leq t \leq n} |U_{nt}| \xrightarrow{p} 0$$

then,  $X_n \xrightarrow{d} N(0, 1)$ .

It is instructive to apply the above theorem to the i.i.d. data, which is the simplest case. Let  $u_1, u_2, \dots, u_t, \dots$  be a i.i.d. sequence with  $E(u_t) = 0$  and  $E(u_t^2) = \sigma^2$ . Also, define  $U_{nt} = \frac{u_t}{\sigma\sqrt{n}}$ , and  $\mathcal{F}_{nt} = \sigma(u_t, u_{t-1}, \dots)$ . Then,  $U_{nt}$  has the finite unconditional variance

$$\sigma_{nt}^2 = E(U_{nt}^2) = E\left(\frac{u_t^2}{\sigma^2 n}\right) = \frac{1}{n} < \infty,$$

and its sum is equal to one

$$\sum_{t=1}^n \sigma_{nt}^2 = \sum_{t=1}^n \frac{1}{n} = 1$$

Also, it can be shown that two conditions are satisfied:

1.  $\sum_{t=1}^n U_{nt}^2 = \sum_{t=1}^n \frac{u_t^2}{\sigma^2 n} = \frac{1}{n} \sum_{t=1}^n \left(\frac{u_t}{\sigma}\right)^2 \xrightarrow{p} 1$  by the LLN.
2.  $\max_{1 \leq t \leq n} |U_{nt}| = \max_{1 \leq t \leq n} \left| \frac{u_t}{\sigma\sqrt{n}} \right| = \left| \frac{u_t}{\sigma\sqrt{n}} \right| \xrightarrow{p} 0$ . Note that the last equality holds because  $u_t$  is identically distributed, and it converges to zero because any random variable is finite.

Therefore,  $X_n = \sum_{t=1}^n U_{nt} = \frac{1}{\sqrt{n}} \sum_{t=1}^n \frac{u_t}{\sigma} \xrightarrow{d} N(0, 1)$ .

**Theorem 5.5** Let  $\{\{U_{nt}\}_{t=1}^n\}_{n=1}^\infty$  be a triangular stochastic array, let  $\{\{\mathbf{V}_{nt}\}_{t=-\infty}^\infty\}_{n=1}^\infty$  be a stochastic array, and let  $\mathcal{F}_{n,t-m}^{t+m} = \sigma(\mathbf{V}_{n,s}, t-m \leq s \leq t+m)$ . Define  $X_n = \sum_{t=1}^n U_{nt}$ . If the following assumptions holds:

1.  $U_{nt}$  is  $\mathcal{F}_{n,-\infty}^t/\mathcal{B}$ -measurable, with  $E(U_{nt}) = 0$  and  $E(X_n^2) = 1$

2. There exists a positive constant array  $\{c_{nt}\}$  such that  $\sup_{n,t} \|U_{nt}/c_{nt}\|_r < \infty$  for  $r > 2$
3.  $U_{nt}$  is  $L_2$ -NED of size  $-1$  on  $\{\mathbf{V}_{nt}\}$ , which is  $\alpha$ -mixing of size  $-r/(r-2)$
4.  $\sup_n nM_n^2 < \infty$ , where  $M_n = \max_{1 \leq t \leq n} \{c_{nt}\}$

then,  $X_n \xrightarrow{d} N(0, 1)$

### 5.B.2 Functional Central Limit Theorem

In the FCLT, a sequence of distributions of stochastic processes converges to the limit. In below, we present two versions of the FCLT: one for the martingale difference sequence, and the other for NED functions of mixing processes.

**Theorem 5.6** *Let  $\{U_{nt}, \mathcal{F}_{nt}\}$  be a martingale difference array with finite unconditional variances  $\{\sigma_{nt}^2\}$ , and  $\sum_{t=1}^n \sigma_{nt}^2 = 1$ . Define  $X_n(r) = \sum_{t=1}^{[nr]} U_{nt}$  for  $r \in [0, 1]$ . If the following assumptions holds:*

1.  $\sum_{t=1}^n U_{nt}^2 \xrightarrow{p} 1$
2.  $\max_{1 \leq t \leq n} |U_{nt}| \xrightarrow{p} 0$
3.  $\lim_{n \rightarrow \infty} \sum_{t=1}^{[nr]} \sigma_{nt}^2 = r$  for all  $r \in [0, 1]$

then,  $X_n \Rightarrow W(r)$

**Theorem 5.7** *Let  $\{\{U_{nt}\}_{t=1}^{K_n}\}_{n=1}^\infty$  be a zero-mean stochastic array,  $\{\{c_{nt}\}_{t=1}^{K_n}\}_{n=1}^\infty$  be an array of positive constants, and  $\{K_n(r)\}_{n=1}^\infty$  be a sequence of integer-valued, right-continuous and increasing function of  $r \in [0, 1]$  with  $K_n(0) = 0$  for all  $n$  and  $K_n(r) - K_n(s) \rightarrow \infty$  as  $n \rightarrow \infty$  if  $r > s$ . Define  $X_n^K(r) = \sum_{t=1}^{K_n(r)} U_{nt}$ . If the following assumptions hold:*

1.  $\sup_{n,t} \left\| \frac{U_{nt}}{c_{nt}} \right\|_r < \infty$  for  $r > 2$
2.  $U_{nt}$  is  $L_2$ -NED of size  $-\gamma \in [-1, -\frac{1}{2}]$  with respect to the constants  $c_{nt}$  on an array  $\{\mathbf{V}_{nt}\}$  which is  $\alpha$ -mixing of size  $-r/(r-2)$
3.  $\sup_{r \in [0,1], \delta \in (0,1-r]} \left\{ \limsup_{n \rightarrow \infty} \frac{v_n^2(r, \delta)}{\delta} \right\} < \infty$ , where  $v_n^2(r, \delta) = \sum_{t=K_n(r)+1}^{K_n(r+\delta)} c_{nt}^2$
4.  $\max_{1 \leq i \leq K_n(1)} c_{nt} = O(K_n(1)^{\gamma-1})$ , where  $\gamma$  is defined in (2)
5.  $E(X_n^K(r)^2) \rightarrow r$  as  $n \rightarrow \infty$ , for each  $r \in [0, 1]$

then,  $X_n^K(r) \Rightarrow W(r)$

In Theorem 5.7, we use a general increasing function of  $r$ ,  $K_n(r)$ . It is instructive to consider the standard case where  $K_n(r) = [nr]$  and  $X_n(r) = \frac{1}{\sqrt{n}} \sum_{t=1}^{[nr]} u_t$ . This case is presented in the following theorem

**Theorem 5.8** *Let  $\{u_t\}$  be a stochastic process with  $E(u_t) = 0$ , uniformly  $L_r$ -bounded, and  $L_2$ -NED of size  $-\frac{1}{2}$  on an  $\alpha$ -mixing process of size  $-r/(r-2)$  for  $r > 2$ . Define  $X_n(r) = \frac{1}{\sqrt{n}} \sum_{t=1}^{[nr]} u_t$ . If the following assumption holds:*

$$E \left( \frac{1}{\sqrt{n}} \sum_{t=1}^n u_t \right)^2 \rightarrow \sigma^2 < \infty$$

then,  $X_n(r) \Rightarrow W(r)$

## 5.C Consistency of Bootstrap

**Definition:** Let  $P_T$  denote the joint probability distribution of the sample  $\{x_1, x_2, \dots, x_T\}$ .

Let  $\Phi$  denote the space of permitted distribution functions. The bootstrap estimator

$G_T(\cdot, F_T)$  is consistent if for  $\varepsilon > 0$  and  $F_0 \in \Phi$ ,

$$\lim_{T \rightarrow \infty} P_T \left( \sup_q |G_T(q, F_T) - G_\infty(q, F_0)| > \varepsilon \right) = 0$$

## 5.D Hansen's (1999) Grid Bootstrap

A sample  $X_T$  of size  $n$  is generated from a distribution  $G_T(x|\alpha, \nu) = P(X_T \leq x|\alpha, \nu)$  which depends on a parameter of interest  $\alpha \in R$  and a nuisance parameter  $\nu \in \Xi$ . Denote by  $\hat{\alpha}$  an estimate of  $\alpha$  with standard error  $s(\hat{\alpha})$ . We assume that, for each  $\alpha$ , there is some estimator  $\hat{\nu} \in \Xi$  of the nuisance parameter  $\nu$ , which may or may not depend on  $\alpha$ . Let  $S(\alpha)$  be a test statistic of the hypothesis  $H_0 : \alpha_0 = \alpha$ , and  $F_T(x|\alpha, \nu) = P(S_T(\alpha) \leq x|\alpha, \nu)$  be a distribution function of  $S(\alpha)$ . Examples of  $S(\alpha)$  include the nonstudentized estimate  $b(\alpha) = \hat{\alpha} - \alpha$  and the  $t$ -statistic  $t(\alpha) = (\hat{\alpha} - \alpha)/s(\hat{\alpha})$ . The quantile function  $q_T(\theta|\alpha, \nu)$  is the  $\theta$  quantile of the distribution of  $S_T(\alpha)$ , and satisfies

$$F_T(q_T(\theta|\alpha, \nu)|\alpha, \nu) = \theta.$$

$q_T(\theta|\alpha, \nu)$  is approximated by the *bootstrap quantile function*  $q_T^*(\theta|\alpha) = q_T(\theta|\alpha, \hat{\nu}(\alpha))$ , which is evaluated at the estimate  $\hat{\nu}(\alpha)$  and is thus random. The  $\beta$ -level grid-bootstrap confidence interval for  $\alpha$  is defined as the set

$$C_g = \{\alpha \in R : q_T^*(\theta_1|\alpha) \leq S_T(\alpha) \leq q_T^*(\theta_2|\alpha)\}$$

where  $\theta_1 = 1 - (1 - \beta)/2$  and  $\theta_2 = (1 - \beta)/2$ ; so  $\beta = \theta_2 - \theta_1$ .

In order to calculate  $C_g$ , we need to estimate the bootstrap quantile functions  $q_T^*(\theta|\alpha)$ , which are generally unknown, by simulation as follows. For a given  $\alpha$ , let  $G_T^*(x|\alpha) = G_T(x|\alpha, \hat{\nu}(\alpha))$  be the bootstrap distribution of the sample.

1. Generate random samples  $X_T^*$  from  $G_T^*(x|\alpha)$  by simulation.
2. Using  $X_T^*$ , calculate the test statistic  $S_T^*(\alpha)$ .
3. Repeat steps 1 and 2  $B$  times.

4. Sort the  $B$  simulated test statistics  $S_T^*(\alpha)$ . The  $100\theta\%$  order statistic  $\hat{q}_T^*(\theta|\alpha)$  is the simulation estimate of  $q_T^*(\theta|\alpha)$  as a function of  $\alpha$ .
5. Pick a grid  $A_G = [\alpha_1, \dots, \alpha_G]$ , and calculate  $\hat{q}_T^*(\theta|\alpha)$  at each  $\alpha \in A_G$  by simulation.
6. For a given  $\alpha$ , smooth the estimated function  $\hat{q}_T^*(\theta|\alpha)$  using the kernel estimate:

$$\tilde{q}_n^*(\theta|\alpha) = \frac{\sum_{j=1}^G K\left(\frac{\alpha-\alpha_j}{\gamma}\right) \hat{q}_n^*(\theta|\alpha_j)}{\sum_{j=1}^G K\left(\frac{\alpha-\alpha_j}{\gamma}\right)}$$

where  $K(z)$  is the Epanechnikov kernel  $K(z) = \frac{3}{4}(1-z^2)I(|z| \leq 1)$ , and  $\gamma$  is a bandwidth chosen by least-square cross-validation.

## 5.E Monte Carlo Methods with GAUSS

This appendix explains how Monte Carlo methods explained in this chapter are implemented with GAUSS, that is explained in Appendix A. The concepts and programs are similar in other computer languages such as MATLAB.

### 5.E.1 Random Number Generators

Most programs offer random number generators for the uniform distribution and the standard normal distribution. For example,

```
y=RNDN(r,c);
```

in GAUSS generates  $r \times c$  values that appear to be a realization of independent standard normal random variables that will be stored in an  $r \times c$  matrix. The starting seed for RNDN can be given by a statement

```
RNDSEED n;
```

where the value of the seed  $n$  must be in the range  $0 < n < 2^{31} - 1$ .

One can produce random numbers with other distributions by transforming generated random numbers. The following examples are some of the transformations that are often used.

**Example 5.E.1** A  $\chi^2$  random variable with  $d$  degrees of freedom can be created from  $d$  independent random variables with the standard normal distribution. If  $e_i \sim N(0, 1)$ , and if  $e_i$  is independent from  $e_j$  for  $j \neq i$ , then  $\sum_{i=1}^d e_i^2$  follows the  $\chi^2$  distribution with  $d$  degrees of freedom. ■

For example, in GAUSS one can generate a  $T \times 1$  vector with values that appear to be a realization of an i.i.d.  $\{x\}_{t=1}^T$  of random variables with the  $\chi^2$  distribution with  $d$  degrees of freedom by the following program:

```
e=RNDN(T,d);
x=sumc((e^2)');
```

**Example 5.E.2** A random variable that follows the Student's  $t$  distribution with  $d$  degrees of freedom can be generated from  $d + 1$  independent random variables with the standard normal distribution. If  $e_i \sim N(0, 1)$ , and if  $e_i$  is independent from  $e_j$  for  $j \neq i$ , then  $x = e_1 / \sqrt{\sum_{i=2}^{d+1} e_i^2 / d}$  follows the  $t$  distribution with  $d$  degrees of freedom. ■

For example, in GAUSS one can generate a  $T \times 1$  vector with values that appear to be a realization of an i.i.d.  $\{x\}_{t=1}^T$  of random variables with the  $t$  distribution with  $d$  degrees of freedom by the following program:

```
e=RNDN(T,d+1);
c=sumc((e[.,2:d+1]^2)');
x=e[.,1]/sqrt(c/d);
```

**Example 5.E.3** A  $K$ -dimensional random vector which follows  $N(\mathbf{0}, \Psi)$  for any positive definite covariance matrix  $\Psi$  can be generated from  $K$  independent random variables with the standard normal distribution. Let  $\Psi = \mathbf{P}\mathbf{P}'$  be the Cholesky decomposition of  $\Psi$ , in which  $\mathbf{P}$  is a lower triangular matrix. If  $e_i \sim N(0, 1)$ , and if  $e_i$  is independent from  $e_j$  for  $j \neq i$ , then  $\mathbf{X} = \mathbf{P}\mathbf{e} \sim N(\mathbf{0}, \Psi)$  where  $\mathbf{e} = (e_1, e_2, \dots, e_K)'$ . ■

For example, in GAUSS one can generate a  $T \times K$  matrix with values that appear to be a realization of an i.i.d.  $\{\mathbf{X}_t\}_{t=1}^T$  of  $K$ -dimensional random vectors with the  $N(0, C)$  distribution with the following program provided that the matrix  $C$  is already defined.

```
e=RNDN(T,K);
P=chol(C)';
x=eP;
```

Note that the Cholesky decomposition in GAUSS gives an upper triangular matrix. Thus, the above program transposes the matrix to a lower triangular matrix.

## 5.E.2 Estimators

### 5.E.3 A Pitfall in Monte Carlo Simulations

Common mistakes are made by many graduate students when they first use Monte Carlo simulations. These mistakes happen when they repeatedly use a random number generator to conduct simulations. These mistakes are caused by updating seeds arbitrarily in the middle of a simulation. Recall that once the starting seed is given, a random number generator automatically updates the seed whenever it creates a number. The way the seed is updated depends on the program.

The following example illustrates common mistakes in a simple form:

**Example 5.E.4** *A Monte Carlo Program with a Common Mistake (I)*

```

ss=3937841;
i=1;
vecm=zeros(100,1);
do until i>100;
    RNDSEED ss;
    y=RNDN(50,1);
    m=meanc(y);
    vecm[i]=m;
    i=i+1;
endo;

```

In this example, the programmer is trying to create 100 samples of the sample mean of a standard normal random variable  $y$  when the sample size is 50. However, exactly the same data are generated 100 times because the same starting seed is given for each replication inside the do-loop. This mistake is innocuous because it is easy to detect. The following program contains a mistake which is harder to detect:

**Example 5.E.5** *A Monte Carlo Program with a Common Mistake (II)*

```

ss=3937841;
i=1;
vecm=zeros(100,1);
do until i>100;
    RNDSEED ss+i;
    y=RNDN(50,1);
    m=meanc(y);
    vecm[i]=m;
    i=i+1;
endo;

```

The problem is that the seed is updated in an arbitrary way in each sample by giving a different starting seed. There is no guarantee that one sample is independent from the others. A correct program would put the RNDSEED statement before the do loop. For example, the RNDSEED statement inside the do loop should be removed and the statement

```
RNDSEED ss;
```

can be added after the first line.

In Monte Carlo simulations, it is also important to control the starting seed so that the simulation results can be replicated. When you publish Monte Carlo results, it is advisable to put enough information in the publication so that others can exactly replicate the results.<sup>13</sup> At the very least, a record of the information should be kept. If no RNDSEED statement is given before the RNDN command is used, GAUSS will take the starting seed from the computer clock. Then there is no way to exactly replicate these Monte Carlo results.

#### 5.E.4 An Example Program

This section describes basic Monte Carlo methods with an example program. In the following example, the sample mean is calculated as an estimator for the expected value of  $X_t$ ,  $E(X_t)$ , where  $X_t = \mu + e_t$  and  $e_t$  is drawn from the  $t$  distribution with 3 degrees of freedom. The  $t$  distribution with 3 degrees of freedom has thick tails and large????? , outlying values have high probability. Hence the  $t$  distribution is often considered a better distribution to describe some financial variables. Because  $X_t$  is not normally distributed, the standard theory for the exact finite sample properties cannot be applied. The example is concerned with the  $t$  test of the null hypothesis that  $\mu = 0$ . Because a random variable with the  $t$  distribution with 3 degrees of freedom has zero mean and a finite second moment, asymptotic theory predicts that the  $t$  test statistic of the sample mean divided by the estimated standard error approximately follows the standard normal distribution.

Masao  
needs to  
check this!

<sup>13</sup>This information is also relevant because different computer specifications and different versions of the program (such as GAUSS) can produce different results.

**Example 5.E.6** *The program.*

```

@MCMEAN.PRG @ Monte Carlo Program for the sample mean@
@This example program is a GAUSS program to calculate
the empirical size and power of the t-test for  $H_0: E(X)=0$ ,
where X follows t-distribution with 3 degrees of freedom.
The power is calculate for the case when  $E(X)=0.2$ . @

RNDSEED 382974;
output file=mc.out reset;
tend=25; @the sample size@
nor=1000; @the number of replications@
df=3; @ d.f. for the t-distribution of X@
i=1;
tn=zeros(nor,1); @used to store t-values under H0@
ta=zeros(nor,1); @used to store t-values under H1@
do until i>nor;
  nrv=RNDN(tend,df+1); @normal r.v.'s@
  crv=nrv[.,2:df+1]^2; @chi square r.v.'s@
  x0=nrv[.,1]/sqrt(sumc(crv)/df); @t distribution: used under H0@
  x1=x0+0.2; @used for H1@
  mx0=meanc(x0);
  mx1=meanc(x1);
  sighat0=sqrt((x0-mx0)'(x0-mx0)/(tend-1)); @simgahat under H0@
  sighat1=sqrt((x1-mx1)'(x1-mx1)/(tend-1)); @sigmahat under H1@
  tn[i]=meanc(x0)*sqrt(tend)/sighat0; @t-value under H0@
  ta[i]=meanc(x1)*sqrt(tend)/sighat1; @t-value under H1@
  i=i+1;
endo;
? "***** When H0 is true *****";
? "The estimated size with the nominal critical value";
? meanc(abs(tn).>1.96);
? "The estimated true 5-percent critical value";
sorttn=sortc(abs(tn),1);
etcv=sorttn[int(nor*0.95)];
? etcv;
? "***** When H1 is true *****";
? "The estimated power with the nominal critical value";
? meanc(abs(ta).>1.96);
? "The estimated size corrected power";
? meanc(abs(ta).>etcv);

```

```
output off;
```

Some features of the example are important. Before the do-loop of the replications, the program set up an output file by

```
output file=mc.out;
```

Then to avoid the common mistake explained in 5.E.3, it makes the RNDNSEEED statement before the do-loop.

It is a good idea to minimize the content inside the do-loop to speed up replications. Everything that can be done outside the do-loop should be done there. For example, the program defines variables to store the test results:

```
tn=zeros(nor,1);  
ta=zeros(nor,1);
```

In GAUSS, the do-loop can be set up as follows:

```
i=1;  
do until i>250;  
... (Program for each replication)  
i=i+1;  
endo;
```

After the do-loop, the program calculates characteristics of the generated distributions of test statistics under the null hypothesis and the alternative hypothesis such as the frequency of rejecting the null with the nominal critical value.

## Exercises

**5.1** Show that all conditions of Gordin's Central Limit Theorem are satisfied for  $e_t$  in Example 5.1.

**5.2** Show that all conditions of Gordin and Hansen's Central Limit Theorem are satisfied for  $\mathbf{f}_t$  in Example 5.2.

**5.3** Let  $e_t = \Psi(L)u_t = \Psi_0 u_t + \Psi_1 u_{t-1} + \dots$  be an MA representation. What is the long-run variance of  $f_t = (1 - L)e_t$ ?

**5.4** Explain what it means to say that “a test under-rejects in small samples” (or “a test is conservative”). When a test is conservative, which is greater, the true critical value or the nominal critical value?

**5.5** Consider the linear model

$$y_t = \mathbf{x}_t' \boldsymbol{\beta} + e_t.$$

where  $\mathbf{x}_t$  is a  $k$ -dimensional vector.

Let  $\mathbf{z}_t$  be a  $k \times 1$  vector of instrumental variables. We will adopt the following set of assumptions:

(A1)  $(e_t, \mathbf{x}_t, \mathbf{z}_t)_{t=1}^{\infty}$  is a stationary and ergodic stochastic process.

(A2)  $\mathbf{z}_t e_t$  have finite second moments.

(A3)  $E(e_t^2 | \mathbf{z}_t) = \sigma^2$ , where  $\sigma$  is a constant.

(A4)  $E(e_t | \mathbf{I}_t) = 0$  for a sequence of information sets  $(\mathbf{I}_t)_{t=1}^{\infty}$  which is increasing (i.e.,  $\mathbf{I}_t \subset \mathbf{I}_{t+1}$ ),  $\mathbf{z}_t$  and  $\mathbf{x}_t$  are in  $\mathbf{I}_t$ , and  $y_t$  is in  $\mathbf{I}_{t+1}$ .

(A5)  $E(\mathbf{z}_t \mathbf{x}_t')$  is nonsingular.

Note that  $E(e_t) = 0$  is implied by (A4) if  $\mathbf{z}_t$  includes a constant.

Note that many rational expectations models imply **(A4)**. For the following problems, prove the asymptotic properties of the instrumental variable (IV) estimator,  $\mathbf{b}_{IV}$ , for  $\boldsymbol{\beta}$  under **(A1)**-**(A5)**. Use a central limit theorem and a strong law of large numbers given in this chapter, and indicate which ones you are using and where you are using them in your proof.

- (a) Express the IV estimator  $\mathbf{b}_{IV}$  in terms of  $\mathbf{z}_t, \mathbf{x}_t$ , and  $y_t (t = 1, \dots, T)$  when  $\sum_{t=1}^T \mathbf{z}_t \mathbf{x}_t'$  is nonsingular.
- (b) Let  $\mathbf{g}_t = \mathbf{z}_t e_t$ . Prove that  $\mathbf{g}_t$  is a martingale difference sequence.
- (c) Prove that the IV estimator is consistent under **(A1)**-**(A5)**.
- (d) Prove that the IV estimator is asymptotically normally distributed. Derive the formula of the covariance matrix of the asymptotic distribution.
- (e) Explain what happens if  $y_t$  is in  $I_{t+2}$  in **(A4)**.

**5.6** Consider the linear model

$$y_t = \mathbf{x}_t' \boldsymbol{\beta} + \epsilon_t,$$

where  $\mathbf{x}_t$  is a  $k$ -dimensional vector. Following Hayashi (2000), suppose that this model satisfies the classical linear regression model assumptions for any sample size ( $n$ ) as follows:

- (A1)** Linearity:  $y_t = \mathbf{x}_t' \boldsymbol{\beta} + e_t$ .
- (A2)** Ergodic stationarity:  $\{y_t, \mathbf{x}_t\}$  is jointly stationary and ergodic.
- (A3)** Predetermined regressors:  $E(e_t \mathbf{x}_t) = \mathbf{0}$ .

- (A4) Rank condition:  $E(\mathbf{x}_t \mathbf{x}_t')$  is nonsingular (and hence finite).
- (A5)  $\mathbf{x}_t e_t$  is a martingale difference sequence with finite second moments.
- (A6) Finite fourth moments for regressors:  $E[(x_{it} x_{jt})^2]$  exists and finite for all  $i, j$  ( $= 1, 2, \dots, k$ ).
- (A7) Conditional homoskedasticity:  $E(e_t^2 | \mathbf{x}_t) = \sigma^2 > 0$ .

Further, assume that  $e_t$  is normally distributed conditional on  $\mathbf{X}$ , where  $\mathbf{X}$  is an  $n \times k$  matrix with  $\mathbf{x}_t'$  in its  $t$ -th row. Let

$$t_k = \frac{b_k - \bar{\beta}_k}{SE(b_k)} = \frac{b_k - \bar{\beta}_k}{\sqrt{s^2 [(\mathbf{X}'\mathbf{X})^{-1}]_{kk}}}$$

be the  $t$  statistic for the null hypothesis  $\beta_k = \bar{\beta}_k$ .

- (a) Prove that  $t_k$  converges in distribution to the standard normal distribution as the sample size goes to infinity. You do not have to prove that  $s^2$  is consistent  $\sigma^2$  for this question. You can assume that  $s^2$  is consistent.
- (b) Based on the asymptotic result in (a), suppose that you set the nominal size to be 5 percent and reject the null hypothesis when  $|t_k|$  is greater than 1.96. Does this test overreject or underreject. How do you know? Suppose that  $k = 3$ . Is the actual size larger than 10 percent when  $n = 4$ . What if  $n = 8, 9, 10, 11$ ? Explain.

### 5.7 Consider the linear model

$$(5.E.1) \quad \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

Let  $k \times 1$  matrix  $\mathbf{x}'_t$  be the  $t$ -th row of the regressor matrix  $\mathbf{X}$ . The model (5.E.1) can be written as

$$(5.E.2) \quad y_t = \mathbf{x}'_t \boldsymbol{\beta} + e_t$$

We will adopt the following set of assumptions:

(A1)  $(e_t, \mathbf{x}_t)_{i=t}^{\infty}$  are independent and identically distributed (i.i.d.) random vectors.

(A2)  $\mathbf{x}_t$  and  $e_t$  have finite second moments.

(A3)  $E(e_t^2 | \mathbf{x}_t) = \sigma^2$  which is a constant.

(A4)  $E(\mathbf{x}_t e_t) = 0$  for all  $t \geq 1$

(A5)  $E(\mathbf{x}_t \mathbf{x}'_t)$  is nonsingular.

Note that  $E(e_t) = 0$  is implied by (A4) if  $\mathbf{x}_t$  includes a constant.

Consider the model (5.E.1) with  $k = 1$ . Assume that  $x_t$  follows  $N(0,1)$ . Assume that  $x_t$  and  $e_t$  are independent. Under the null hypothesis  $H_0$ , the true value of  $\beta$  is 0, so that  $y_t = e_t$ .

Consider the standard  $t$  statistic,

$$(5.E.3) \quad t_1 = \frac{b - \beta}{\hat{\sigma}_1 \sqrt{\mathbf{X}'\mathbf{X}}^{-1}}$$

where  $\hat{\sigma}_1^2 = (\mathbf{Y} - \mathbf{X}b)'(\mathbf{Y} - \mathbf{X}b)/(n - k)$ . Consider another version of the  $t$  statistic

$$(5.E.4) \quad t_2 = \frac{b - \beta}{\hat{\sigma}_2 \sqrt{\mathbf{X}'\mathbf{X}}^{-1}}$$

where  $\hat{\sigma}_2^2 = (\mathbf{Y} - \mathbf{X}b)'(\mathbf{Y} - \mathbf{X}b)/n$ . Note that both  $t_1$  and  $t_2$  converge in distribution to a random variable with the standard normal distribution.

Consider two alternative assumptions for  $e_t$ .

(A6)  $e_t$  follows the  $t$  distribution with 4 degrees of freedom.

(A6')  $e_t$  follows the standard normal distribution.

Note that Assumptions 1.1 - 1.5 are satisfied with **(A6')**, so that  $t_1$  has the exact  $t$  distribution with  $n - k$  degrees of freedom.

Using GAUSS, conduct a Monte Carlo simulation with the sample size of 26 and 500 replications under Assumption **(A6)**.

- (a) Use the  $t_1$  in (5.E.3) to estimate
- (i) the true size of the  $t$  test for  $H_0 : \beta = 0$  based on the nominal significance level of 5% and the nominal critical value based on the standard normal distribution are used.
  - (ii) the true size of the  $t$  test for  $H_0 : \beta = 0$  based on the nominal significance level of 5% and the nominal critical value based on the  $t$  distribution with 25 degrees of freedom.
  - (iii) the true critical value of the  $t$  test for the 5% significance level,
  - (iv) the power of the test at  $\beta = 0.15$  based on the nominal critical value,
  - (v) the size corrected power of the test.
- (b) Use the  $t_2$  in (5.E.4) and repeat the exercises (a) – (e).

For the starting seed, use 3648xxxx, where xxxx is your birth date, such as 0912 for September 12. Submit your program and output. For each  $t$  ratio, discuss whether it is better to use the standard distribution or the  $t$  distribution critical values for this application. Also discuss whether  $t_1$  or  $t_2$  is better for this application.

### References

- ANDREWS, D. W. K. (1993): "Exactly Median-Unbiased Estimation of First Order Autoregressive/Unit Root Models," *Econometrica*, 61(1), 139–165.
- APOSTOL, T. M. (1974): *Mathematical Analysis*. Addison-Wesley, Reading, Massachusetts.

- ATHREYA, K. B. (1987): "Bootstrap of the Mean in the Infinite Variance Case," *Annals of Statistics*, 15(2), 724–731.
- BASAWA, I. V., A. K. MALLIK, W. P. MCCORMICK, J. H. REEVES, AND R. L. TAYLOR (1991): "Bootstrapping Unstable First-Order Autoregressive Processes," *Annals of Statistics*, 19(2), 1098–1101.
- BILLINGSLEY, P. (1961): "The Lindeberg-Levy Theorem for Martingales," *Proceedings of the American Mathematical Society*, 12, 788–792.
- (1986): *Probability and Measure*. Wiley, New York.
- CHAN, N. H., AND C. Z. WEI (1987): "Asymptotic Inference for Nearly Nonstationary AR(1) Processes," *Annals of Statistics*, 15(3), 1050–1063.
- CHOI, C.-Y., AND M. OGAKI (1999): "The Gauss-Markov Theorem for Cointegrating and Spurious Regressions," Working Paper No. 01-13, Department of Economics, Ohio State University.
- EFRON, B. (1979): "Bootstrap Methods: Another Look at the Jackknife," *Annals of Statistics*, 7(1), 1–26.
- GORDIN, M. I. (1969): "The Central Limit Theorem for Stationary Processes," *Soviet Mathematics-Doklady*, 10, 1174–1176.
- HANSEN, B. E. (1999): "The Grid Bootstrap and the Autoregressive Model," *Review of Economics and Statistics*, 81(4), 594–607.
- HANSEN, L. P. (1985): "A Method for Calculating Bounds on the Asymptotic Covariance Matrices of Generalized Method of Moments Estimators," *Journal of Econometrics*, 30, 203–238.
- HAYASHI, F. (2000): *Econometrics*. Princeton University Press, Princeton.
- HOROWITZ, J. L. (2001): "The Bootstrap," in *Handbook of Econometrics*, ed. by J. J. Heckman, and E. E. Leamer, vol. 5, chap. 52, pp. 3159–3228. Elsevier Science Publishers.
- JEONG, J., AND G. S. MADDALA (1993): "A Perspective on Application of Bootstrap Methods in Econometrics," in *Handbook of Statistics*, ed. by G. S. Maddala, C. R. Rao, and H. D. Vinod, vol. 11, pp. 573–610. Elsevier Science Publishers.
- JUDGE, G. G., W. E. GRIFFITHS, R. C. HILL, H. LÜTKEPOHL, AND T. LEE (1985): *The Theory and Practice of Econometrics*. Wiley, New York, 2nd edn.
- LOEVE, M. (1978): *Probability Theory II*. Springer-Verlag, New York, 4th edn.
- NELSON, C. R., AND R. STARTZ (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.
- PHILLIPS, P. C. B. (1987): "Towards a Unified Asymptotic Theory for Autoregression," *Biometrika*, 74(3), 535–547.

# Chapter 6

## ESTIMATION OF THE LONG-RUN COVARIANCE MATRIX

An estimate of the long-run covariance matrix,  $\mathbf{\Omega}$ , is necessary to calculate asymptotic standard errors for the OLS and linear IV estimators presented in Chapter 5. Estimation of the long-run covariance matrix will be important for GMM estimators introduced later in Chapter 9 and many of the estimation and testing methods for nonstationary variables. Chapter 13 shows that an estimate of the long-run variance of a random variable is also useful in estimating the importance of the random walk component of some nonstationary random variables. This chapter discusses estimation methods for the long-run covariance matrix.

Let  $\{\mathbf{u}_t : -\infty < t < \infty\}$  be a stationary and ergodic vector stochastic process with mean zero. We will discuss estimation methods of the long-run covariance matrix of  $\mathbf{u}_t$ :

$$(6.1) \quad \mathbf{\Omega} = \lim_{j \rightarrow \infty} \sum_{-j}^j E(\mathbf{u}_t \mathbf{u}'_{t-j}).$$

Depending on the application, we take different variables as  $\mathbf{u}_t$ . When  $\mathbf{\Omega}$  is used for the calculation of the asymptotic standard errors for the OLS estimator, we take

$\mathbf{u}_t = \mathbf{x}_t(y_t - \mathbf{x}_t'\mathbf{b}_0)$ . For the calculation of the asymptotic standard errors for the linear IV estimator, we take  $\mathbf{u}_t = \mathbf{z}_t(y_t - \mathbf{x}_t'\mathbf{b}_0)$ . Because  $\mathbf{b}_0$  is unknown, the sample counterpart of  $\mathbf{u}_t$ ,  $\mathbf{z}_t(y_t - \mathbf{x}_t'\mathbf{b}_T)$ , is used to estimate  $\mathbf{\Omega}$  where  $\mathbf{b}_T$  is a consistent estimator for  $\mathbf{b}_0$ . For the application in Chapter 13,  $\mathbf{u}_t$  is a random variable such as the first difference of the log real GNP minus its expected value, and the first difference minus its estimated mean is used for the sample counterpart. Thus in many applications,  $\mathbf{u}_t$  is unobservable and its sample counterpart is constructed from a consistent estimator for a parameter vector. When  $\mathbf{\Omega}_T$  is a consistent estimator for  $\mathbf{\Omega}$ ,  $\mathbf{\Omega}_T^* = f(T)\mathbf{\Omega}_T$  is also a consistent estimator as long as  $\lim_{T \rightarrow \infty} f(T) = 1$  for any real valued function  $f(T)$ . Therefore, we can consider various forms of  $f(T)$  to improve small sample properties. If  $p$  parameters are estimated to compute the sample counterpart of  $\mathbf{u}_t$ , then  $f(T) = \frac{T}{T-p}$  is a small sample degrees of freedom adjustment that is often used for each  $\mathbf{\Omega}_T$  presented in this chapter.<sup>1</sup>

## 6.1 Serially Uncorrelated Variables

This section treats the case where  $E(\mathbf{u}_t\mathbf{u}'_{t-\tau}) = \mathbf{0}$  for  $\tau \neq 0$ . Many rational expectations models imply this property. In this case,  $\mathbf{\Omega} = E(\mathbf{u}_t\mathbf{u}'_t)$  can be estimated by  $\frac{1}{T} \sum_{t=1}^T \mathbf{u}_t\mathbf{u}'_t$ . For linear IV estimators, this is White's (1980) heteroskedasticity consistent estimator. In this case,  $\mathbf{u}_t = \mathbf{z}_t(y_t - \mathbf{x}_t'\mathbf{b}_T)$  and  $\frac{1}{T} \sum_{t=1}^T \mathbf{u}_t\mathbf{u}'_t = \frac{1}{T} \sum_{t=1}^T (y_t - \mathbf{x}_t'\mathbf{b}_T)^2 \mathbf{z}_t\mathbf{z}'_t$ .

In some cases, conditional homoskedasticity is assumed in the economic model, and an econometrician may wish to impose this property on the estimate for  $\mathbf{\Omega} = E(e_t^2)E(\mathbf{z}_t\mathbf{z}'_t)$ . Then  $\frac{1}{T} \sum_{t=1}^T (y_t - \mathbf{x}_t'\mathbf{b}_T)^2 \frac{1}{T} \sum_{t=1}^T \mathbf{z}_t\mathbf{z}'_t$  with a small sample degree of

<sup>1</sup>Some other forms of small sample adjustments have been used (see, e.g., Ferson and Foerster, 1994).

freedom adjustment such as  $\frac{T}{T-p}$  is used to estimate  $\mathbf{\Omega}$ .

## 6.2 Serially Correlated Variables

This section treats the case where the disturbance is serially correlated in the context of time series analysis.

### 6.2.1 Unknown Order of Serial Correlation

In many applications, the order of serial correlation is unknown. The estimators of the long-run covariance matrix in the presence of conditional heteroskedasticity and autocorrelation are called Heteroskedasticity and Autocorrelation Consistent (HAC) estimators.

Let  $\mathbf{\Phi}(\tau) = E(\mathbf{u}_t \mathbf{u}'_{t-\tau})$ . Many HAC estimators use the sample version of  $\mathbf{\Phi}(\tau)$ ,

$$(6.2) \quad \mathbf{\Phi}_T(\tau) = \frac{1}{T} \sum_{t=\tau+1}^T \mathbf{u}_t \mathbf{u}'_{t-\tau} \quad \text{for } 0 \leq \tau \leq T-1$$

and  $\mathbf{\Phi}_T(\tau) = \mathbf{\Phi}_T(-\tau)'$  for  $\tau < 0$ . Given the data of  $\mathbf{u}_1, \dots, \mathbf{u}_T$ ,  $\mathbf{\Phi}_T(\tau)$  for a large lag  $\tau$  cannot be estimated with many observations. For example, we have only one observation for  $\mathbf{\Phi}_T(T-1)$ . Hence it is natural to put much less weight on  $\mathbf{\Phi}_T(\tau)$  with large  $\tau$  than on  $\mathbf{\Phi}_T(\tau)$  with small  $\tau$ . The weights are described by a real valued function called a kernel function. The kernel HAC estimators for  $\mathbf{\Omega}$  in the literature have the form

$$(6.3) \quad \mathbf{\Omega}_T = \sum_{\tau=-T+1}^{T-1} k\left(\frac{\tau}{S_T}\right) \mathbf{\Phi}_T(\tau),$$

where  $k(\cdot)$  is a real-valued kernel, and  $S_T$  is a band-width parameter.<sup>2</sup> Examples of

---

<sup>2</sup>These terminologies follow Andrews (1991), and are somewhat different from those in kernel estimations in other contexts.

kernels that have been used by econometricians include the following:

$$\begin{aligned}
 (6.4) \quad k(x) &= \begin{cases} 1 & \text{for } |x| \leq 1 \\ 0 & \text{otherwise} \end{cases} && \text{Truncated kernel,} \\
 k(x) &= \begin{cases} 1 - |x| & \text{for } |x| \leq 1 \\ 0 & \text{otherwise} \end{cases} && \text{Bartlett kernel,} \\
 k(x) &= \begin{cases} 1 - 6x^2 + 6|x|^3 & \text{for } |x| \leq \frac{1}{2} \\ 2(1 - |x|)^3 & \text{for } \frac{1}{2} < |x| \leq 1 \\ 0 & \text{otherwise} \end{cases} && \text{Parzen kernel,} \\
 k(x) &= \frac{25}{12\pi^2 x^2} \left( \frac{\sin(6\pi x/5)}{6\pi x/5} - \cos\left(\frac{6\pi x}{5}\right) \right) && \text{QS kernel.}
 \end{aligned}$$

The estimators of Hansen (1982) and White (1984, p.152) use the truncated kernel; the Newey and West (1987) estimator uses the Bartlett kernel; and the estimator of Gallant (1987, p.533) uses the Parzen kernel. The estimators corresponding to these kernels place zero weights on  $\Phi(\tau)$  for  $\tau \geq S_T$ , so that  $S_T - 1$  is called the lag truncation number. Andrews (1991) advocates an estimator which uses the Quadratic Spectral (QS) kernel, which does not place zero weights on any  $\Phi(\tau)$  for  $|\tau| \leq T - 1$ .<sup>3</sup>

One important problem is how to choose the bandwidth parameter  $S_T$ . Andrews (1991) provides formulas for the optimal choice of the bandwidth parameter,  $S_T^*$ , for a variety of kernels. The  $S_T^*$  is optimal in the sense of minimizing the MSE for a given positive semidefinite matrix  $\mathbf{W}$ :<sup>4</sup>

$$(6.5) \quad S_T^* = \begin{cases} 1.1447(\alpha(1)T)^{\frac{1}{3}} & \text{Bartlett kernel} \\ 2.6614(\alpha(2)T)^{\frac{1}{5}} & \text{Parzen kernel} \\ 1.3221(\alpha(2)T)^{\frac{1}{5}} & \text{QS kernel,} \end{cases}$$

<sup>3</sup>Hansen (1992) relaxes an assumption made by these authors to show the consistency of the kernel estimators.

<sup>4</sup>To be exact, the optimal bandwidth parameter minimizes the asymptotic truncated MSE. See Andrews (1991).

and

$$(6.6) \quad \begin{aligned} \alpha(q) &= \frac{2(\text{vec}\mathbf{f}^{(q)})'\mathbf{W}\text{vec}\mathbf{f}^{(q)}}{\text{tr}\mathbf{W}(\mathbf{I} + \mathbf{K}_{pp})\mathbf{f}^{(0)} \otimes \mathbf{f}^{(0)}}, \\ \mathbf{f}^{(q)} &= \frac{1}{2\pi} \sum_{j=-\infty}^{\infty} |j|^q \Phi(\tau), \end{aligned}$$

where  $\mathbf{W}$  is a  $p^2 \times p^2$  weight matrix and  $\mathbf{K}_{pp}$  is the  $p^2 \times p^2$  commutation matrix that transforms  $\text{vec}(\mathbf{A})$  into  $\text{vec}(\mathbf{A}')$ , i.e.,  $\mathbf{K}_{pp} = \sum_{i=1}^p \sum_{j=1}^p \mathbf{e}_i \mathbf{e}_j' \otimes \mathbf{e}_j \mathbf{e}_i'$ , where  $\mathbf{e}_i$  is the  $i$ -th elementary  $p$ -vector. Here  $\mathbf{f}^{(0)}$  is the spectral density at frequency zero, and the long-run covariance matrix  $\boldsymbol{\Omega}$  is equal to  $2\pi\mathbf{f}^{(0)}$ . Unfortunately, these formulas include the unknown parameters we wish to estimate. This outcome presents a serious circular problem.

Andrews proposes automatic bandwidth estimators in which these unknown parameters are estimated from the data by a parameterized model. His method involves two steps. The first step is to parameterize the model to estimate the law of motion of  $\mathbf{u}_t$ . For example, we can use an AR(1) model for each element of  $\mathbf{u}_t$  or a VAR(1) model for  $\mathbf{u}_t$ . The second step is to calculate the parameters for the optimal bandwidth parameter from the estimated law of motion. For example, we calculate the unknown parameters by assuming that the estimated AR(1) model is true. In his Monte Carlo simulations, Andrew uses an AR(1) parameterization for each term of the disturbance, which seems to work well in the models he considers. Newey and West (1994) propose an alternative method based on truncated sums of the sample autocovariances; this method avoids the use of any parametric model.

Another issue is the choice of the kernel. One serious problem with the truncated kernel is that the corresponding estimator is not guaranteed to be positive semidefinite. Andrews (1991) shows that the QS kernel is an optimal kernel in the

sense that it minimizes the asymptotic MSE among the kernel estimators that are guaranteed to be positive semidefinite. His Monte Carlo simulations show that the QS kernel and the Parzen kernel work better than the Bartlett kernel in most of the models he considers. He also finds that even the estimators based on the QS kernel and the Parzen kernel are not satisfactory in the sense that the standard errors calculated from these estimators are not accurate in small samples when the amount of autocorrelation is large.

Since the kernel HAC estimators do not seem satisfactory in many cases, Andrews and Monahan (1992) propose an estimator based on VAR prewhitening. The intuition behind this proposition is that the kernel HAC estimators only take care of the MA components of  $\mathbf{u}_t$  and cannot handle the AR components well in small samples. The first step in the VAR prewhitening method is to run a VAR of the form

$$(6.7) \quad \mathbf{u}_t = \mathbf{A}_1 \mathbf{u}_{t-1} + \mathbf{A}_2 \mathbf{u}_{t-2} + \cdots + \mathbf{A}_p \mathbf{u}_{t-p} + \mathbf{v}_t.$$

Note that the model (6.7) need not be a true model in any sense. The estimated VAR is used to form an estimate  $\hat{\mathbf{v}}_t$  and a kernel HAC estimator is applied to the estimated  $\hat{\mathbf{v}}_t$  to estimate the long-run variance of  $\mathbf{v}_t$ ,  $\mathbf{\Omega}_T^*$ . The estimator based on the QS kernel with the automatic bandwidth parameter can be used to find  $\hat{\mathbf{v}}_t$  for example. Then the sample counterpart of the formula

$$(6.8) \quad \mathbf{\Omega} = [\mathbf{I} - \sum_{\tau=1}^p \mathbf{A}_\tau]^{-1} \mathbf{\Omega}^* [\mathbf{I} - \sum_{\tau=1}^p \mathbf{A}'_\tau]^{-1}$$

is used to form an estimate of  $\mathbf{\Omega}$ . Andrews and Monahan use the VAR of order one in their Monte Carlo simulations. Their results suggest that the prewhitened kernel HAC estimator performs better than the non-prewhitened kernel HAC estimators for

the purpose of calculating the standard errors of estimators.<sup>5</sup>

In a recent paper, den Haan and Levin (1996) propose a HAC estimator based on (6.7) without using any kernel estimation, which is called the Vector Autoregression Heteroskedasticity and Autocorrelation Consistent (VARHAC) estimator. This estimator has an advantage over any estimator that involves kernel estimation in that the circular problem associated with estimating the optimal bandwidth parameter can be avoided. For the VARHAC estimator, a usual method is to choose the order of AR such as the AIC is applied to (6.7). Then the sample counterpart of (6.8) with  $\mathbf{\Omega}^* = E(\mathbf{v}_t \mathbf{v}_t')$  is used to estimate  $\mathbf{\Omega}$ . Their Monte Carlo evidence indicates that the VARHAC estimator performs better than the non-prewhitened and prewhitened kernel estimators in many cases. On the other hand, Cochrane (1988) basically argues that kernel estimators are better than VARHAC estimators for his purpose of estimating the random walk component as discussed in Chapter 13. Thus, it seems necessary to compare VARHAC estimators with other estimators in different contexts for various applications.

In sum, existing Monte Carlo evidence for estimation of  $\mathbf{\Omega}$  recommends VAR prewhitening and either the QS or Parzen kernel estimator together with Andrews' (1991) automatic bandwidth parameter if a kernel HAC estimator is to be utilized. Though the QS kernel estimator may be preferred to the Parzen kernel estimator because of its asymptotic optimality, it takes more time to calculate the QS kernel estimators than the Parzen kernel estimators. This difference may be important when estimation is repeated many times. The VARHAC estimator of den Haan and Levin (1996) seems to have important advantages over estimators involving kernel

---

<sup>5</sup>Park and Ogaki's (1991) Monte Carlo simulations suggest that the VAR prewhitening improves estimators of  $\mathbf{\Omega}$  in the context of cointegrating regressions.

estimation, even though it is a relatively new method, and has more Monte Carlo evidence for various applications.

### 6.2.2 Known Order of Serial Correlation

In some applications, the order of serial correlation is known in the sense that the economic model implies a particular order. Assume that the order of serial correlation is known to be  $s$ .

In this case, there exist the zero restrictions on the autocovariances that  $\Phi(\tau) = \mathbf{0}$  for  $|\tau| > s$ . Imposing these zero restrictions on the estimator of  $\Omega$  leads to a more efficient estimator.<sup>6</sup> Since  $\Omega = \sum_{\tau=-s}^s \Phi(\tau)$  in this case, a natural estimator is

$$(6.9) \quad \Omega_T = \sum_{\tau=-s}^s \Phi_T(\tau),$$

which is the truncated kernel estimator.

Hansen and Hodrick (1980) study a multi-period forecasting model that leads to  $s \geq 1$ . They use (6.9) with conditional homoskedasticity imposed (as discussed at the end of Section 6.1 above). Their method of calculating the standard errors for linear regressions is known as Hansen-Hodrick correction.

A possible problem with the estimator (6.9) is that  $\Omega_T$  is not guaranteed to be positive semidefinite if  $s \geq 1$ . In applications, researchers often encounter cases where  $\Omega_T$  is invertible but is not positive semidefinite. If this happens,  $\Omega_T$  should not be used to form the optimal GMM estimator (e.g., Newey and West, 1987). There exist at least two ways to handle this problem. One way is to use Eichenbaum, Hansen, and Singleton's (1988) modified Durbin method. The first step of this method is

---

<sup>6</sup>In some applications, the order of serial correlation may be different for different terms of  $\mathbf{u}_t$ . The econometrician may wish to impose these restrictions.

to estimate the VAR (6.7) for a large  $p$  by solving the Yule Walker equations. The second step is to estimate an MA( $s$ ) representation

$$(6.10) \quad \mathbf{u}_t = \mathbf{B}_1 \mathbf{v}_{t-1} + \dots + \mathbf{B}_s \mathbf{v}_{t-s} + \mathbf{e}_t,$$

by regressing the estimated  $\mathbf{u}_t$  on estimated lagged  $\mathbf{v}_t$ 's. Then the sample counterpart of

$$(6.11) \quad \mathbf{\Omega} = (\mathbf{I} + \mathbf{B}_1 + \dots + \mathbf{B}_s) E(\mathbf{e}_t \mathbf{e}_t') (\mathbf{I} + \mathbf{B}_1 + \dots + \mathbf{B}_s)'$$

is used to form an estimate of  $\mathbf{\Omega}$  that imposes the zero restrictions. This method is not reliable when the number of elements in  $\mathbf{u}_t$  is large relative to the sample size because too many parameters in (6.7) need to be estimated. The number of elements in  $\mathbf{u}_t$  need to be kept as small as possible when using this method.

Another method uses one of the kernel HAC estimators (or VAR prewhitened kernel estimators if  $s$  is large) that is guaranteed to be positive semidefinite. When employing this method, the zero restrictions should *not* be imposed even though  $\mathbf{\Phi}(\tau)$  is known to be zero for  $|\tau| > s$ . In order to illustrate this method in a simple example, consider the case where  $s = 1$  and Newey and West's (1987) Bartlett kernel estimator is used. Then

$$(6.12) \quad \mathbf{\Omega}_T = \sum_{\tau=-\ell}^{\ell} \frac{S_T - |\tau|}{S_T} \mathbf{\Phi}_T(\tau),$$

where  $\ell = S_T - 1$  is the lag truncation number. If  $\ell = 1$  is used to impose the zero restrictions, then  $\mathbf{\Omega}_T$  converges to  $\mathbf{\Phi}(0) + \frac{1}{2}\mathbf{\Phi}(1) + \frac{1}{2}\mathbf{\Phi}(-1)$ , which is not equal to  $\mathbf{\Omega} = \mathbf{\Phi}(0) + \mathbf{\Phi}(1) + \mathbf{\Phi}(-1)$ . Thus  $\ell$  must increase as  $T$  increases to obtain a consistent estimator. On the other hand, if  $\ell > 1$  is used and the zero restrictions are imposed by setting  $\mathbf{\Phi}_T(\tau)$  in (6.6) equal to zero for  $|\tau| > 1$ , then the resulting estimator is no longer guaranteed to be positive semidefinite.

In this chapter, we focused on consistent estimators for the long-run covariance matrix. Recently, some researchers have pointed out that we may not need to have consistent estimators for some purposes such as computing standard errors for regression estimators or computing Wald tests. For example, small sample properties of Wald tests computed from inconsistent estimates of the long-run covariance matrix may be better for some data generating processes. See Kiefer, Vogelsang, and Bunzel (2000), Kiefer and Vogelsang (2002a,b), and Müller (2004).

## Exercises

**6.1** (*The Multi-Period Forecasting Model*) Suppose that  $I_t$  is an information set generated by  $\{\mathbf{Y}_t, \mathbf{Y}_{t-1}, \mathbf{Y}_{t-2}, \dots\}$ , where  $\mathbf{Y}_t$  is a stationary and ergodic vector stochastic process. Economic agents are assumed to use current and past  $\mathbf{Y}_t$  to generate their information set. Let  $X_t$  be a stationary and ergodic random variable in the information set  $I_t$  with  $E(|X_t|^2) < \infty$ . We consider a 3-period forecast of  $X_t$ ,  $E(X_{t+3}|I_t)$ , and the forecast error,  $e_t = X_{t+3} - E(X_{t+3}|I_t)$ .

- (a) Give an expression for the long-run variance of  $e_t$ . Which methods do you suggest to use in order to estimate the long-run variance?
- (b) Let  $\mathbf{Z}_t$  be a random vector with finite second moments in the information set  $I_t$ . Define  $\mathbf{f}_t = \mathbf{Z}_t e_t$ . Give an expression for the long covariance matrix of  $\mathbf{f}_t$ . Which methods do you suggest to use in order to estimate the long-run variance?

## References

- ANDREWS, D. W. K. (1991): "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation," *Econometrica*, 59(3), 817–858.
- ANDREWS, D. W. K., AND J. C. MONAHAN (1992): "An Improved Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimator," *Econometrica*, 60(4), 953–966.

- COCHRANE, J. H. (1988): “How Big is the Random Walk in GNP?,” *Journal of Political Economy*, 96(5), 893–920.
- DEN HAAN, W. J., AND A. LEVIN (1996): “Inference From Parametric And Non-Parametric Covariance Matrix Estimation Procedures,” NBER Technical Working Paper No. 195.
- EICHENBAUM, M., L. P. HANSEN, AND K. J. SINGLETON (1988): “A Time Series Analysis of Representative Agent Models of Consumption and Leisure Choice under Uncertainty,” *Quarterly Journal of Economics*, 103, 51–78.
- FERSON, W. E., AND S. R. FOERSTER (1994): “Finite Sample Properties of the Generalized Methods of Moments in Tests of Conditional Asset Pricing Models,” *Journal of Financial Economics*, 36, 29–55.
- GALLANT, A. R. (1987): *Nonlinear Statistical Models*. John Wiley and Sons, New York.
- HANSEN, B. E. (1992): “Consistent Covariance Matrix Estimation for Dependent Heterogeneous Processes,” *Econometrica*, 60(4), 967–972.
- HANSEN, L. P. (1982): “Large Sample Properties of Generalized Method of Moments Estimators,” *Econometrica*, 50(4), 1029–1054.
- HANSEN, L. P., AND R. J. HODRICK (1980): “Forward Exchange Rates as Optimal Predictors of Future Spot Rates,” *Journal of Political Economy*, 88, 829–853.
- KIEFER, N. M., AND T. J. VOGELSANG (2002a): “Heteroskedasticity-Autocorrelation Robust Standard Errors Using the Bartlett Kernel Without Truncation,” *Econometrica*, 70, 2093–2095.
- (2002b): “Heteroskedasticity-Autocorrelation Robust Testing Using Bandwidth Equal to Sample Size,” *Econometric Theory*, 18, 1350–1366.
- KIEFER, N. M., T. J. VOGELSANG, AND H. BUNZEL (2000): “Simple Robust Testing of Regression Hypotheses,” *Econometrica*, 68(3), 695–714.
- MÜLLER, U. K. (2004): “A Theory of Robust Long-Run Variance Estimation,” Manuscript, Princeton University.
- NEWBY, W. K., AND K. D. WEST (1987): “A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix,” *Econometrica*, 55(3), 703–708.
- (1994): “Automatic Lag Selection in Covariance Matrix Estimation,” *Review of Economic Studies*, 61, 631–653.
- PARK, J. Y., AND M. OGAKI (1991): “VAR Prewhitening to Estimate Short-Run Dynamics: On Improved Method of Inference in Cointegrated Models,” RCER Working Paper No. 281.
- WHITE, H. (1980): “A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity,” *Econometrica*, 48(4), 817–838.
- (1984): *Asymptotic Theory For Econometricians*. Academic Press, New York.

# Chapter 7

## TESTING LINEAR FORECASTING MODELS

Some economic models imply that a linear function of a variable  $X_t$  is a forecaster of  $Y_{t+1}$  in the sense that

$$(7.1) \quad E(Y_{t+1} | I_t) = a + bX_t,$$

where  $a$  and  $b$  are constants, and  $I_t$  is an information set. Typically,  $I_t$  is the information set available to economic agents at date  $t$ , and includes the current and past values of  $X_t$  and  $Y_t$ . Equation (7.1) is called a linear forecasting model. In some cases, a linear function of a variable  $X_t$  is a forecaster of  $Y_{t+s}$  :

$$(7.2) \quad E(Y_{t+s} | I_t) = a + bX_t.$$

If  $Y_{t+s}$  is in  $I_{t+s}$ , (7.2) is a multi-period linear forecasting model. In this chapter, we discuss some standard methods to test these linear forecasting models.

### 7.1 Forward Exchange Rates

In usual market transactions (called *spot transactions*), transactions are carried out immediately. In forward contracts, two parties agree to carry out transactions at a

specified future date. In foreign exchange forward contracts, a party agrees to deliver specified units of a currency to another party who agrees to pay a specified price.

Let  $F_{t,1}$  be the forward exchange rate at date  $t$  of a foreign currency to be delivered at date  $t+1$ : at date  $t$  a contract is made in which  $F_{t,1}$  units of the domestic currency is promised to be paid when one unit of the foreign currency is delivered at date  $t+1$ . Let  $S_t$  be the spot exchange rate at date  $t$  which is expressed as the price of one unit of the foreign currency in terms of the domestic currency. Assume that the domestic investors are risk neutral. For now, assume that risk neutrality is defined about gambles involving the domestic currency. Given that preferences are defined over goods rather than currencies, risk neutrality should be defined about gambles involving goods. The assumption of risk neutrality over the domestic currency leads to Siegel's (1972) Paradox as discussed below. Under this assumption,

$$(7.3) \quad F_{t,1} = E(S_{t+1} | I_t)$$

should hold in equilibrium, where  $I_t$  is the information set available at date  $t$ . To see this relation suppose that  $F_{t,1} > E(S_{t+1} | I_t)$ . Then the domestic investors' expected profit is positive when they sell the foreign currency with forward contracts. The supply of foreign currency will be infinite, and therefore equilibrium cannot be attained. If  $F_{t,1} < E(S_{t+1} | I_t)$ , then the domestic investors' expected profit is positive when they buy the foreign currency with forward contracts.

Let  $F_{t,s}$  be the forward exchange rate at date  $t$  of a foreign currency to be delivered at date  $t+s$ . Then with a similar argument,

$$(7.4) \quad F_{t,s} = E(S_{t+s} | I_t)$$

should hold. Relation (7.4) is implied by the uncovered interest parity (UIP) if we

assume the covered interest parity (CIP)<sup>1</sup>, and Relation (7.3) is a special case of UIP when  $s = 1$ .

Given (7.4), a natural way to test UIP is to consider a regression

$$(7.5) \quad S_{t+s} = a + bF_{t,s} + e_t.$$

Then (7.4) implies that  $E(e_t|I_t) = 0$  when  $a = 0$  and  $b = 1$ . Since  $F_{t,s}$  is in  $I_t$ , if  $S_{t+s}$  and  $F_{t,s}$  are stationary, then the asymptotic theory of OLS in Chapter 5 applies to this regression. In the data of exchange rates, it is often observed that the first difference of  $\ln(S_{t+s})$ , the first difference  $\ln(F_{t,s})$ , and  $\ln(S_{t+s}) - \ln(F_{t,s})$  appear to be stationary. However,  $S_{t+s}$  and  $F_{t,s}$  do not appear to be stationary, hence the asymptotic theory of Chapter 5 does not apply to (7.5). One solution found in the literature is to apply cointegration to (7.5) or the log version of (7.5).<sup>2</sup>

We consider transforming the data to obtain a regression with stationary variables. For this purpose, we first take the natural log of both sides of (7.4) to obtain an approximate relation

$$(7.6) \quad \ln(F_{t,s}) = a + E(\ln(S_{t+s})|I_t),$$

where  $a$  is a constant. This relation is an approximation because the log of the expected value of a random variable is not the expected value of the log of the variable.

The approximation error for (7.6) can be significant and may lead to the rejection of the model when the exchange rate is conditionally heteroskedastic even when UIP holds. This problem may be serious because conditional heteroskedasticity

---

<sup>1</sup>CIP says that  $1 + i_{t+s} = \frac{F_{t,s}}{S_t}(1 + i_{t,s}^*)$  while UIP says that  $1 + i_{t+s} = \frac{E(S_{t+s}|I_t)}{S_t}(1 + i_{t,s}^*)$ , where  $i_{t,s}$  and  $i_{t,s}^*$  are interest rates on domestic deposit and foreign deposit, respectively.

<sup>2</sup>This solution is, however, problematic for the purpose of testing UIP as discussed in Chapter 14.

is detected for most exchange rate series especially when high frequency data (e.g., weekly or daily) are used. To illustrate this point, assume that  $S_{t+s}$  is log normally distributed conditional on  $I_t$ , then  $\ln(E(S_{t+s}|I_t)) = E(\ln(S_{t+s})|I_t) + \frac{1}{2}Var(\ln(S_{t+s})|I_t)$ .

Hence (7.6) is exact with  $a = \frac{1}{2}Var(\ln(S_{t+s}))$  if  $\ln(S_{t+s})$  is conditionally homoskedastic. However, if  $\ln(S_{t+s})$  is conditionally heteroskedastic,  $Var(\ln(S_{t+s})|I_t)$  is not constant. Hence (7.6) is an approximation, and the approximation error is more important for data with stronger conditional heteroskedasticity effects.

Assuming that this approximation error is negligible, we consider a regression

$$(7.7) \quad \ln(S_{t+s}) - \ln(F_{t,s}) = a + \mathbf{X}'_t \mathbf{b} + e_t,$$

where  $\mathbf{X}_t$  is a stationary random vector that is in  $I_t$ . Then (7.6) implies that  $\mathbf{b} = \mathbf{0}$  and  $E(e_t|I_t) = 0$  (note that  $a \neq 0$  here). Assuming that  $\ln(S_{t+s}) - \ln(F_{t,s})$  is stationary, the asymptotic theory in Chapter 5 applies to (7.7). For example,  $\mathbf{X}_t$  is a vector of lagged values of  $\ln(S_{t+s}) - \ln(F_{t,s})$ :  $\mathbf{X}_t = (\ln(S_t) - \ln(F_{t-s,s}), \ln(S_{t-1}) - \ln(F_{t-1-s,s}), \dots, \ln(S_{t-k}) - \ln(F_{t-k-s,s}))'$ . UIP can be tested by testing the null hypothesis  $H_0 : \mathbf{b} = \mathbf{0}$ .

The assumption of risk neutrality over the domestic currency leads to *Siegel's Paradox*. Assume that foreign investors are risk neutral over their currency. Then the same argument made for (7.3) for the domestic investors imply

$$(7.8) \quad \frac{1}{F_{t,1}} = E\left(\frac{1}{S_{t+1}}|I_t\right).$$

Since  $\frac{1}{X}$  is a convex function, (7.3) and (7.8) cannot hold at the same time. This property is known as Siegel's Paradox.<sup>3</sup>

---

<sup>3</sup>Because preferences are defined over goods, risk neutrality should be defined over goods. Siegel's Paradox is a result of defining risk neutrality over currencies. In order to illustrate this point, imagine

## 7.2 The Euler Equation

Consider an economy with a single good, in which the current and past values of a random vector  $\mathbf{X}_t$  generate the information set  $I_t$ , which is available to the economic agents. The random vector  $\mathbf{H}_t = [\mathbf{X}'_0, \mathbf{X}'_1, \dots, \mathbf{X}'_t]'$  summarizes  $I_t$ . Let  $Prob(\mathbf{H}_t)$  denote the probability of  $\mathbf{H}_t$ . For simplicity, we assume that the economy ends at date  $T$ , and that there exist  $N$  possible values of  $\mathbf{H}_T$ . With this notation,  $\mathbf{H}_T$  plays the role of the state of the world  $s$  in Chapter 2, and  $\mathbf{H}_t$  specifies the subset in the partition of  $S$  at date  $t$ . The history notation is more convenient for the purpose of this section to ensure that consumption is in the information available at date  $t$ .

We assume that the representative consumer maximizes the lifetime utility function

$$(7.9) \quad U = \sum_{t=0}^T \sum_{\mathbf{H}_t} Prob(\mathbf{H}_t) \beta^t u(C_t(\mathbf{H}_t)),$$

where  $\beta$  is a discount factor,  $u(\cdot)$  is the utility function, and  $C_t(\mathbf{H}_t)$  is the consumption at date  $t$  with history  $\mathbf{H}_t$ . As a bench mark case, we assume that there exists a complete set of contingent security markets at date 0. Assuming that there are  $N$  states of the world, and the contingent security for one unit of  $C_t(\mathbf{H}_t)$  costs  $P_t(\mathbf{H}_t)$  in terms of the good at date 0, the lifetime budget constraint is

$$(7.10) \quad \sum_{t=0}^T \sum_{\mathbf{H}_t} P_t(\mathbf{H}_t) C_t(\mathbf{H}_t) = \sum_{t=0}^T \sum_{\mathbf{H}_t} P_t(\mathbf{H}_t) C_t^e(\mathbf{H}_t),$$

where  $C_t^e(\mathbf{H}_t)$  is the endowment. Let  $\lambda$  be the Lagrange multiplier for the budget constraint (7.10). Then the first order conditions for the consumer's maximization

---

that there are two consumption goods in the world economy: a good purchased with the domestic currency, and another good purchased with the foreign currency. The real version of (7.3), expressed in terms of the domestic good, and the real version of (7.8), expressed in terms of the foreign good, are not subject to Siegel's Paradox (see, e.g. Frankel, 1979, 1980). Engel (1984) empirically tests the absence of expected real profits from forward market speculation and shows that Siegel's paradox is not empirically important in this case.

problem include

$$(7.11) \quad \beta^t \text{Prob}(\mathbf{H}_t) \text{mu}(C_t(\mathbf{H}_t)) = \lambda P_t(\mathbf{H}_t),$$

where  $\text{mu}(\cdot)$  is the derivative of the utility function (marginal utility). Hence

$$(7.12) \quad \frac{\beta^{t+1} \text{Prob}(\mathbf{H}_{t+1}) \text{mu}(C_{t+1}(\mathbf{H}_{t+1}))}{\beta^t \text{Prob}(\mathbf{H}_t) \text{mu}(C_t(\mathbf{H}_t))} = \frac{P_{t+1}(\mathbf{H}_{t+1})}{P_t(\mathbf{H}_t)},$$

which we call the *state-by-state intertemporal first order condition*. This type of condition is useful in testing for complete risk sharing as we will discuss in Chapter 17.

The first order condition (7.12) does not necessarily hold when markets are incomplete. We derive the asset pricing equation and Euler equation, which can be shown to hold for some incomplete market models, from this first order condition. For this purpose, imagine that a security pays off  $D_{t+1}(\mathbf{H}_{t+1})$  units of the good at date  $t+1$  when the history  $\mathbf{H}_{t+1}$  is realized. Let  $V_t(\mathbf{H}_t)$  be the price of the security in terms of the good at date  $t$  when the history  $\mathbf{H}_t$  is realized. Then an arbitrage condition gives

$$(7.13) \quad V_t(\mathbf{H}_t) = \frac{\sum_{\mathbf{H}_{t+1}|\mathbf{H}_t} P_{t+1}(\mathbf{H}_{t+1}) D_{t+1}(\mathbf{H}_{t+1})}{P_t(\mathbf{H}_t)},$$

where the summation in the numerator sums up all  $\mathbf{H}_{t+1}$ 's that follow  $\mathbf{H}_t$ . The numerator is the price of the security in terms of the good at date 0, and the denominator is the price of the good at date  $t$ , so that the security price is expressed in terms of the good at date  $t$ . Substituting (7.12) into (7.13) yields

$$(7.14) \quad V_t(\mathbf{H}_t) = \frac{\sum_{\mathbf{H}_{t+1}|\mathbf{H}_t} \beta \text{prob}(\mathbf{H}_{t+1}) \text{mu}_{t+1} D_{t+1}(\mathbf{H}_{t+1})}{\text{prob}(\mathbf{H}_t) \text{mu}_t},$$

where  $\text{mu}_t$  denotes  $\text{mu}(C_t(\mathbf{H}_t))$ . Noting that  $\frac{\text{Prob}(\mathbf{H}_{t+1})}{\text{Prob}(\mathbf{H}_t)}$  is the probability of  $\mathbf{H}_{t+1}$

conditional on  $\mathbf{H}_t$ , we can rewrite (7.14) as

$$(7.15) \quad V_t = \frac{E(\beta mu_{t+1} D_{t+1} | \mathbf{I}_t)}{mu_t},$$

which we call the *asset pricing equation*.

Dividing both sides of the asset pricing equation (7.15) by  $V_t$  yields

$$(7.16) \quad \frac{E(\beta mu_{t+1} R_{t+1} | \mathbf{I}_t)}{mu_t} = 1,$$

which is called the *Euler equation* where  $R_{t+1} = \frac{D_{t+1}}{V_t}$  is the real gross asset return. It should be noted that the asset pricing equation and the Euler equation hold for any asset while the state-by-state intertemporal first order condition only holds for the contingent securities since  $P_t(\mathbf{H}_t)$  in (7.12) is the price of contingent security rather than any other security.

### 7.3 The Martingale Model of Consumption

Consider a bond that pays one unit of the good at date  $t+1$  without any uncertainty, which we call the real risk free bond. Let  $R_{t+1}^f$  be the real gross asset return on the real risk free bond. Then  $R_{t+1}^f - 1$  is the real interest rate. Assume that the real interest rate is constant, and that  $\beta R_{t+1}^f = 1$ . Then the Euler equation (7.16) implies

$$(7.17) \quad E(mu_{t+1} | \mathbf{I}_t) = mu_t.$$

Therefore, under these assumptions, the marginal utility is a martingale adapted to  $\mathbf{I}_t$ . This implication is testable when the intra-period utility function is parameterized, so that  $mu_t$  is related to consumption.

Hall (1978) assumes that the intra-period utility function is quadratic:

$$(7.18) \quad u(C_t) = -\alpha(C_t - \gamma)^2,$$

where  $\alpha$  and  $\gamma$  are positive constants. Then  $mu_t = -2\alpha(C_t - \gamma)$ , and (7.17) implies

$$(7.19) \quad E(C_{t+1}|\mathbf{I}_t) = C_t.$$

Thus Euler equation implies that consumption is a martingale adapted to  $\mathbf{I}_t$ . Therefore, this model is called the *martingale model of consumption*. With an additional assumption that consumption is conditionally homoskedastic, (7.19) implies that consumption is a random walk. For this reason, some authors prefer to call this model the *random walk model of consumption*.

This martingale (or random walk) hypothesis can be tested by applying OLS to

$$(7.20) \quad C_{t+1} - C_t = a + \mathbf{X}'_t \mathbf{b} + e_t$$

where  $\mathbf{X}_t$  is a stationary random vector which is in  $\mathbf{I}_t$ . Then (7.19) implies that  $a = 0$ ,  $\mathbf{b} = \mathbf{0}$ , and  $E(e_t|\mathbf{I}_t) = 0$ .

## 7.4 The Linearized Euler Equation

It should be noted that the random walk model of consumption is derived under the assumptions of a quadratic utility function and a constant real interest rate. These assumptions are not attractive. There exists some evidence that real interest rates are not even stationary (see, Rose, 1988). The quadratic utility function has an implication that both absolute and relative risk aversion coefficients increase with consumption. The intertemporal elasticity of substitution is the reciprocal of the relative risk aversion coefficient for the time-separable expected utility function, and the quadratic utility function implies that the elasticity decreases as consumption increases. These implications are counterintuitive to most people upon introspection, and there is empirical evidence against them (see Chapter 17?????????).

Masao  
needs to  
check this!

Most researchers agree that the isoelastic utility function,

$$u(C) = \frac{1}{1-\alpha}[C^{1-\alpha} - 1]$$

is more reasonable than the quadratic utility function. For this utility function, the relative risk aversion coefficient is  $\alpha$  (a constant), and the absolute risk aversion coefficient decreases with consumption. The intertemporal elasticity of substitution is  $\frac{1}{\alpha}$ . With this utility function,  $mu_t = C_t^{-\alpha}$ , and (7.16) implies

$$(7.21) \quad E(\beta R_{t+1} C_{t+1}^{-\alpha} | I_t) = C_t^{-\alpha}.$$

With an assumption that  $R_t$  and  $C_t$  are jointly log normally distributed conditional on  $I_t$ , we obtain

$$(7.22) \quad E(\ln(R_{t+1}) - \alpha \ln(C_{t+1}) | I_t) = -\ln(\beta) - \frac{1}{2} \text{Var}(\ln(R_{t+1} C_{t+1}^{-\alpha}) | I_t) - \alpha \ln(C_t).$$

Further assuming that  $\ln(R_{t+1} C_{t+1}^{-\alpha})$  is conditionally homoskedastic with respect to  $I_t$ , we obtain the linearized version of the Euler equation (7.21):

$$(7.23) \quad E(\ln(R_{t+1}) - \alpha \ln(C_{t+1}) | I_t) = b - \alpha \ln(C_t),$$

where  $b = -\ln(\beta) - \frac{1}{2} \text{Var}(\ln(R_{t+1} C_{t+1}^{-\alpha}) | I_t)$  is a constant. Note that the linearized Euler equation (7.23) holds for any asset return under the stated assumptions.

With an additional assumption that the real interest rate is constant as in Section 7.3, we can obtain a result similar to the random walk hypothesis. In this case, (7.23) implies

$$(7.24) \quad E(\ln(C_{t+1}) | I_t) = c + \ln(C_t),$$

where  $c = -\frac{b}{\alpha} + \frac{1}{\alpha} \ln(R_{t+1})$ . As in the previous section, we can test this model by applying OLS to

$$(7.25) \quad \ln(C_{t+1}) - \ln(C_t) = c + \mathbf{X}'_t \mathbf{b} + e_t$$

where  $\mathbf{X}_t$  is a stationary random vector which is in  $I_t$ . Equation (7.24) implies that  $\mathbf{b} = \mathbf{0}$ , and  $E(e_t|I_t) = 0$ .

The linearized Euler equation (7.23) has been used by many researchers without the additional assumption of the constant real interest rate. Hansen and Singleton (1983) apply the maximum likelihood estimation method to (7.23). Hall (1988) estimates the intertemporal elasticity of substitution from

$$(7.26) \quad \ln(C_{t+1}) - \ln(C_t) = d + \frac{1}{\alpha} \ln(R_{t+1}) + e_t.$$

Equation (7.23) implies that  $d = -\frac{b}{\alpha}$  and  $E(e_t|I_t) = 0$ . Since  $\ln(R_{t+1})$  is not in  $I_t$ , OLS cannot be applied to (7.26). Any stationary variable in  $I_t$ , however, can be used as an instrumental variable for (7.26). Hansen and Singleton (1996) also apply an IV method to (7.23).

## 7.5 Optimal Taxation

The method to derive the martingale property of consumption can be applied to other optimization problems. A good example is the optimal taxation model tested by Barro (1981), Sahasakul (1986), Kingston (1984), Mankiw (1987), and Bizer and Durlauf (1990) among others.

Assume that the government minimizes the following quadratic cost function at date  $t$ :

$$(7.27) \quad E_t \sum_{j=0}^{\infty} \beta^j (c_0 \tau_{t+j} + \frac{c_1}{2} \tau_{t+j}^2),$$

subject to the budget constraint

$$(7.28) \quad B_{t+1} = R[B_t + g_t - \tau_t], \quad B_t \text{ bounded for all } t$$

by choosing  $\{\tau_{t+j}, B_{t+j}\}_{j=0}^{\infty}$ . Here  $\{g_t\}_{t=j}^{\infty}$  is a stochastic process describing the ratio of government spending to GDP,  $\tau_t$  is the tax collected as a percentage of GDP,  $B_t$  is the real value of a one-period risk free bond to be repaid at date  $t+1$  as a percentage of GDP, and  $R$  is the gross real interest rate, which is assumed to be constant. We assume that  $\beta R = 1$ .

The Euler equation for the maximization problem is

$$(7.29) \quad E(\tau_{t+1} | \mathbf{I}_t) = \tau_t.$$

As in the consumption case, this martingale hypothesis can be tested by applying OLS to

$$(7.30) \quad \tau_{t+1} - \tau_t = a + \mathbf{X}'_t \mathbf{b} + e_t,$$

where  $\mathbf{X}_t$  is a stationary random vector which is in  $\mathbf{I}_t$ . Then (7.29) implies that  $a = 0$ ,  $\mathbf{b} = \mathbf{0}$ , and  $E(e_t | \mathbf{I}_t) = 0$ . Barro (1981), Kingston (1984), and Mankiw (1987) have found that movements of U.S. tax rates over time are roughly consistent with the martingale hypothesis. On the other hand, Sahasakul (1986) reports that U.S. tax rates are predictably related to wars and recessions, which is evidence against the martingale hypothesis. Bizer and Durlauf (1990) report evidence against the hypothesis based on a frequency- domain based test (see Section 16.3??????? below).

Masao  
needs to  
check this!

## 7.6 Tests of Forecast Accuracy

Tests of forecast accuracy can be used to test economic models. A prominent example is tests for exchange rate models in the literature that started by Meese and Rogoff (1983) who compared predictions of exchange rate models with predictions of the random walk model.

### 7.6.1 The Monetary Model of Exchange Rates

A multi-period forecasting formulation in Mark (1995) is motivated by the the monetary model of the monetary model of Frenkel (1976), Mussa (1976), and Bilson (1978).

The monetary model implies the present value relationship

$$(7.31) \quad s_t = (1 - \beta)E\left(\sum_{i=1}^{\infty} \beta^i f_{t+i} | I_t\right).$$

where  $s_t$  is the log exchange rate,  $f_t = m_t - m_t^* - \gamma(y_t - y_t^*)$  where  $m_t$  is the log domestic money supply,  $y_t$  is the log domestic income. We call  $f_t$  fundamentals. Here,  $\gamma$  is the income elasticity of money demand,  $\beta = \alpha/(1 + \alpha)$  where  $\alpha$  is the interest semi-elasticity of money demand. If  $f_t$  is a driftless random walk, then the present value relationship implies  $s_t = f_t$ , and the log exchange rate is a random walk. However, deviations from the log exchange rate from the fundamentals are known to be persistent. These considerations motivated Mark to investigate the projection of the k-period-ahead change in the log exchange rate on its current deviation from the fundamental value

$$(7.32) \quad s_{t+k} - s_t = a + bX_t + e_{1t},$$

where  $X_t = f_t - s_t$  and  $e_{1t}$  is a forecast error. On the other hand, if the log exchange rate is a driftless random walk,  $a = b = 0$ , then

$$(7.33) \quad s_{t+k} - s_t = e_{2t},$$

where  $e_{2t}$  be the forecast error of the random walk model. Tests described in this section compare forecast accuracy based on the differences in mean squared prediction

errors (MSPEs) from these two models. Unlike the previous works that had found that the economic model does not improve forecast accuracy over the random walk model, Mark (1995) found evidence in favor of the economic model for long-horizon changes (large values of  $k$ ). However, Kilian (1999) pointed out problems with Mark's (1995) bootstrap procedure. With a corrected bootstrap procedure, Kilian found no evidence of increased long-horizon predictability. In panel data that combine time series of 19 industrialized countries, Mark and Sul (2001) found evidence of increased long-horizon predictability. Thus the evidence is mixed for the monetary model compared with the random walk model.

Engel and West (2005) showed analytically that in present value models such as Equation (7.31), the log exchange rate manifests near-random walk behavior if the first difference of fundamentals is stationary and  $\beta$  is near one. Their result helps explain that fundamentals provide little help in predicting changes in the log exchange rates.

## 7.7 The Taylor Rule Model of Exchange Rates

Many recent papers have explored various aspects of exchange rate models with the Taylor Rule (see, e.g., Mark, 2005; Engel and West, 2005, 2006; Clarida and Waldman, 2008; Kim and Ogaki, 2009). Molodtsova, Nikolsko-Rzhevskyy, and Papell (2008) and Molodtsova and Papell (2009) find strong evidence of exchange rate predictability using the Taylor rule model. In this model, under the assumption that uncovered interest parity holds, exchange rate movements are related to the differential of short-term nominal interest rates between two countries. In each country, the nominal interest rate is in turn set by the central bank that follows a policy rule proposed

by Taylor (1993). According to Taylor's original specification, the home central bank adjusts the nominal interest rate in response to changes in the domestic inflation and output gap:

$$\tilde{i}_t = \pi_t + \phi(\pi_t - \tilde{\pi}) + \gamma y_t + \tilde{r},$$

where  $\tilde{i}_t$  is the target rate of the short-term nominal interest rate,  $\pi_t$  is inflation,  $\tilde{\pi}$  is the inflation target,  $y_t$  is the output gap, and  $\tilde{r}$  is the equilibrium level of the real interest rate. The Taylor rule for small open economies may include the real exchange rate  $s_t$  (Clarida, Galí, and Gertler, 1998):

$$(7.34) \quad \tilde{i}_t = \pi_t + \phi(\pi_t - \tilde{\pi}) + \gamma y_t + \delta s_t + \tilde{r}.$$

Empirical studies (e.g., Clarida, Galí, and Gertler, 1998, 2000) find that central banks engage in interest rate smoothing so that the observed nominal rate  $i_t$  is a partial adjustment of its lagged value and the target rate:

$$(7.35) \quad i_t = (1 - \rho)\tilde{i}_t + \rho i_{t-1} + v_t.$$

Suppose the foreign central bank follows an analogous policy rule:

$$(7.36) \quad i_t^* = (1 - \rho^*)\tilde{i}_t^* + \rho^* i_{t-1}^* + v_t^*.$$

Taking the difference between the policy reaction function of home country (7.35) and that of foreign country (7.36) yields the interest rate differential:

$$(7.37) \quad i_t - i_t^* = \beta + \beta_\pi \pi_t - \beta_\pi^* \pi_t^* + \beta_y y_t - \beta_y^* y_t^* + \beta_s s_t + \beta_s^* s_t + \rho i_{t-1} - \rho^* i_{t-1}^* + \eta_t,$$

where  $\eta_t = v_t - v_t^*$ ,  $\beta = (\tilde{r} - \phi\tilde{\pi})(1 - \rho)$ ,  $\beta_\pi = (1 + \phi)(1 - \rho)$ ,  $\beta_y = \gamma(1 - \rho)$ , and  $\beta_s = \delta(1 - \rho)$ . Analogous definitions apply for foreign coefficients denoted by a star. Note that since  $s_t = -s_t^*$ , we have  $\beta_s s_t + \beta_s^* s_t^*$ .

Assume that uncovered interest rate parity holds:  $\Delta e_{t+1} = i_t - i_t^*$  where  $e_t$  is the log of the nominal exchange rate defined as the domestic currency price of foreign currency. Equating the UIP condition with the interest rate differential (7.37) yields the Taylor rule model of exchange rates:

$$(7.38) \quad \Delta e_{t+1} = \beta + \beta_\pi \pi_t - \beta_\pi^* \pi_t^* + \beta_y y_t - \beta_y^* y_t^* + \beta_s s_t + \beta_s^* s_t + \rho i_{t-1} - \rho^* i_{t-1}^* + \eta_t.$$

Molodtsova and Papell (2009) evaluate out-of-sample forecasts of one-month-ahead exchange rate movements using the Taylor-rule model for the monthly U.S. exchange rates against 12 OECD countries. The data spans from March 1973 to June 2006 (December 1998 for the European Monetary Union countries). The predictive performance of the model is evaluated using the CW test statistics for the null hypothesis that the exchange rate follows a random walk against the alternative hypothesis that it is predictable by the model (7.38).

They find that the Taylor rule model exhibits strong evidence of short-term exchange rate predictability, especially when the real exchange rates are excluded from equation (7.38). For that specification, the model outperforms the random walk for 10 out of 12 currencies at the 10% significance level - four of them at the 1% level and additional six at the 5% level - using one of the three output gap specifications they consider (the linear trend, the quadratic trend, and the HP-filter). By contrast, using the same dataset and the inference method, they find much less evidence of exchange rate predictability with conventional models of exchange rates (the UIP model of Clark and West, 2006; the monetary model of Mark, 1995; and the PPP model of Mark and Sul, 2001). Even after combining the results from these three models, they find statistically significant evidence of exchange rate predictability at the 5% level for only 3 of the 12 currencies and for an additional currency at the 10%

level.

### 7.7.1 Diebold and Mariano (1995)

One of the commonly used methods for testing forecast accuracy is the test of equal accuracy proposed by Diebold and Mariano (1995, the DM test).

Consider two competing forecast series  $y_{1t}$  and  $y_{2t}$  of the time series  $y_t$ , with associated forecast errors  $e_{1t}$  and  $e_{2t}$ ,  $t = 1, \dots, T$ , respectively. The DM test is applicable to a wide variety of accuracy measures. Here, as in many applications, we compare forecast accuracy based on the differences in mean squared prediction errors (MSPEs) from the two series. In this case, the DM test evaluates the null hypothesis that the population mean of the MSPE differences is 0,  $E(e_{1t}^2 - e_{2t}^2) = 0$ , against the alternative hypothesis  $E(e_{1t}^2 - e_{2t}^2) \neq 0$ .

Let  $\bar{d}$  denote the sample mean of the MSPE differential:

$$\bar{d} = \frac{1}{T} \sum_{t=1}^T (e_{1t}^2 - e_{2t}^2).$$

If the MSPE differential is covariance stationary and short memory, then  $\sqrt{T}\bar{d}$  is asymptotically normally distributed with mean zero and variance  $2\pi f_d(0)$  where  $f_d(0) = \frac{1}{2\pi} \sum_{\tau=-\infty}^{\infty} \gamma_d(\tau)$  is the spectral density of the MSPE differential at frequency 0, and  $\gamma_d(\tau) = E[(e_{1t}^2 - e_{2t}^2)(e_{1t-\tau}^2 - e_{2t-\tau}^2)]$  is the autocovariance of the MSPE differential. The DM statistic is given by

$$DM = \frac{\bar{d}}{\sqrt{\frac{2\pi \hat{f}_d(0)}{T}}},$$

where  $2\pi \hat{f}_d(0)$  is a consistent estimator of  $2\pi f_d(0)$ . It is obtained by a weighted sum of the sample autocovariances,

$$2\pi \hat{f}_d(0) = \sum_{\tau=-(T-1)}^{(T-1)} 1 \left( \frac{\tau}{S(T)} \right) \hat{\gamma}_d(\tau),$$

where  $1(\tau/S(T))$  is the lag window,  $S(T)$  is the truncation lag, and  $\hat{\gamma}_d(\tau) = \frac{1}{T} \sum_{t=|\tau|+1}^T (d_t - \bar{d})(d_{t-|\tau|} - \bar{d})$  with  $d_t \equiv e_{1t}^2 - e_{2t}^2$ . Diebold and Mariano (1995) suggest the use of the uniform lag window,

$$1\left(\frac{\tau}{S(T)}\right) = \begin{cases} 1 & \text{for } \left|\frac{\tau}{S(T)}\right| \leq 1, \\ 0 & \text{otherwise} \end{cases}$$

and the truncating lag  $S(T) = (k - 1)$  since optimal  $k$ -step-ahead forecast errors are at most  $(k - 1)$ -dependent.

### 7.7.2 Clark and West (2006) and Clark and West (2007)

Now suppose we wish to compare out-of-sample forecast accuracy of two nested models. One example as in the exchange rate forecasting literature above is the case where a linear econometric model (model 2) is compared to a random walk model (model 1):

$$\text{Model 1: } y_t = e_t$$

$$\text{Model 2: } y_t = \mathbf{X}'_t \boldsymbol{\beta} + e_t,$$

where  $e_t$  in both models is a zero mean martingale difference which may be conditionally heteroskedastic.

Let  $T + 1$  be the sample size of  $y_t$  and  $\mathbf{X}_t$  which is divided into two subsamples  $T + 1 = R + P$ . For illustration, suppose we are comparing one-period-ahead forecasts,  $y_{t+1}$ .<sup>4</sup> Model 2 is estimated using data prior to  $t$  to generate  $P$  predictions for  $y_{t+1}$ ,  $t = R, R + 1, \dots, T$ . The out-of-sample MSPEs of the two models are,

$$\begin{aligned} \text{Model 1: } \hat{\sigma}_1^2 &\equiv P^{-1} \sum_{t=T-P+1}^T y_{t+1}^2, \\ \text{Model 2: } \hat{\sigma}_2^2 &\equiv P^{-1} \sum_{t=T-P+1}^T (y_{t+1} - \mathbf{X}'_{t+1} \hat{\boldsymbol{\beta}}_t)^2. \end{aligned}$$

---

<sup>4</sup>For multi-horizon predictions, see Clark and West (2006).

Recall that the DM test is based on the assumption that the difference in sample MSPEs from two models is asymptotically zero. However, Clark and West (2006) and Clark and West (2007) show that this is not the case when the two models are nested. To see this, write:

$$(7.39) \quad \hat{\sigma}_1^2 - \hat{\sigma}_2^2 = 2 \left( P^{-1} \sum_{t=T-P+1}^T y_{t+1} \mathbf{X}'_{t+1} \hat{\boldsymbol{\beta}}_t \right) - \left[ P^{-1} \sum_{t=T-P+1}^T (\mathbf{X}'_{t+1} \hat{\boldsymbol{\beta}}_t)^2 \right].$$

Under the null hypothesis of equal predictive accuracy,  $y_t$  follows a martingale difference ( $\boldsymbol{\beta} = \mathbf{0}$ ) as in model 1. Therefore,  $y_{t+1} = e_{t+1}$  and  $E e_{t+1} \mathbf{X}'_{t+1} \hat{\boldsymbol{\beta}}_t = 0$ , and thus the first term in equation (16.11) is expected to be approximately zero. However, the second term is  $-P^{-1} \sum_{t=T-P+1}^T (\mathbf{X}'_{t+1} \hat{\boldsymbol{\beta}}_t)^2 < 0$ , and thus the MSPE from model 2 is expected to be greater than that of model 1:

$$\hat{\sigma}_1^2 - \hat{\sigma}_2^2 \xrightarrow{p} -E(\mathbf{X}'_{t+1} \hat{\boldsymbol{\beta}}_t)^2 < 0.$$

The DM statistics, while appropriate for non-nested models, do not adjust for this shift, and result in non-normal test statistics when the models are nested. Therefore, hypothesis tests based on standard normal critical values are usually poorly sized, failing to reject the null hypothesis when it should (McCracken, 2004; Clark and McCracken, 2001, 2005). This is particularly problematic for tests of out-of-sample predictability of financial data for which the null hypothesis is a random walk.

Clark and West (2006) and Clark and West (2007) propose an asymptotically normal test for two nested models that properly adjusts the difference in MSPEs by a consistent estimate of  $E(\mathbf{X}'_{t+1} \hat{\boldsymbol{\beta}}_t)^2$ . This test is applicable when  $\boldsymbol{\beta}_t$  is estimated from rolling regressions using data from  $t - R + 1$  to  $t$ .<sup>5</sup>

<sup>5</sup>Clark and West (2007) consider a general parametric specification of the null model (model 1) that is smaller than the alternative model (model 2). Thus, model 2 reduces to model 1 if some of the parameters in model 2 are zero.

The bias-adjusted difference of the sample mean MSPEs is given by,

$$\begin{aligned}\bar{f} &\equiv \hat{\sigma}_1^2 - \left[ \hat{\sigma}_2^2 - P^{-1} \sum_{t=T-P+1}^T (\mathbf{X}'_{t+1} \hat{\boldsymbol{\beta}}_t)^2 \right] \\ &= P^{-1} \sum_{t=T-P+1}^T \hat{f}_{t+1}.\end{aligned}$$

where  $\hat{f}_{t+1} \equiv y_{t+1}^2 - [(y_{t+1} - \mathbf{X}'_{t+1} \hat{\boldsymbol{\beta}}_t)^2 - (\mathbf{X}'_{t+1} \hat{\boldsymbol{\beta}}_t)^2]$ . Under some mild conditions,  $\sqrt{P}\bar{f}$  is asymptotically normally distributed with mean zero and variance  $V \equiv 4E(y_{t+1} \mathbf{X}'_{t+1} \hat{\boldsymbol{\beta}}_t)^2$ .

The adjusted test statistic is

$$CW = \frac{\bar{f}}{\sqrt{\hat{V}/P}},$$

where  $\hat{V} \equiv 4P^{-1} \sum_{t=T-P+1}^T (y_{t+1} \mathbf{X}'_{t+1} \hat{\boldsymbol{\beta}}_t)^2 = P^{-1} \sum_{t=T-P+1}^T (\hat{f}_{t+1} - \bar{f})^2$  is a consistent estimator of  $V$ . Clark and West (2006) present simulation results showing that inferences of the CW statistics using normal critical values are properly sized. Note that the alternative hypothesis of this test is that  $y_t$  is linearly predictable ( $\boldsymbol{\beta} \neq \mathbf{0}$ ) as in model 2, implying that the population MSPE of model 2 is smaller than that of model 1. Therefore, this test is one-sided, and the null hypothesis is rejected when the CW test statistic is significantly positive.

## References

- BARRO, R. J. (1981): "On the Predictability of Tax-Rate Changes," NBER Working Paper No. 636.
- BILSON, J. F. O. (1978): "Rational Expectations and the Exchange Rate," in *The Economics of Exchange Rates: Selected Studies*, ed. by J. A. Frenkel, and H. G. Johnson, pp. 75–96. Addison-Wesley, Reading, MA.
- BIZER, D. S., AND S. N. DURLAUF (1990): "Testing the Positive Theory of Government Finance," *Journal of Monetary Economics*, 26, 123–141.
- CLARIDA, R. H., J. GALÍ, AND M. GERTLER (1998): "Monetary Policy Rules in Practice: Some International Evidence," *European Economic Review*, 42(6), 1033–1067.
- (2000): "Monetary Policy Rules and Macroeconomic Stability: Evidence and Some Theory," *Quarterly Journal of Economics*, 115(1), 147–180.

- CLARIDA, R. H., AND D. WALDMAN (2008): “Is Bad News About Inflation Good News for the Exchange Rate? And, If So, Can That Tell Us Anything about the Conduct of Monetary Policy?,” in *Asset Prices and Monetary Policy*, ed. by J. Y. Campbell, pp. 371–396. The University of Chicago Press.
- CLARK, T. E., AND M. W. MCCrackEN (2001): “Tests of Equal Forecast Accuracy and Encompassing for Nested Models,” *Journal of Econometrics*, 105(1), 85–110.
- (2005): “Evaluating Direct Multistep Forecasts,” *Econometric Reviews*, 24(4), 369–404.
- CLARK, T. E., AND K. D. WEST (2006): “Using Out-of-Sample Mean Squared Prediction Errors to Test the Martingale Difference Hypothesis,” *Journal of Econometrics*, 135(1–2), 155–186.
- (2007): “Approximately Normal Tests for Equal Predictive Accuracy in Nested Models,” *Journal of Econometrics*, 138(1), 291–311.
- DIEBOLD, F. X., AND R. S. MARIANO (1995): “Comparing Predictive Accuracy,” *Journal of Business and Economic Statistics*, 13(3), 253–263.
- ENGEL, C., AND K. D. WEST (2005): “Exchange Rates and Fundamentals,” *Journal of Political Economy*, 113(3), 485–517.
- (2006): “Taylor Rules and the Deutschmark-Dollar Real Exchange Rate,” *Journal of Money, Credit, and Banking*, 38(5), 1175–1194.
- ENGEL, C. M. (1984): “Testing for the Absence of Expected Real Profits from Forward Market Speculation,” *Journal of International Economics*, 17, 299–308.
- FRANKEL, J. A. (1979): “The Diversifiability of Exchange Risk,” *Journal of International Economics*, 9, 379–393.
- (1980): “Tests of Rational Expectations in the Forward Exchange Market,” *Southern Economic Journal*, 46, 1083–1101.
- FRENKEL, J. A. (1976): “A Monetary Approach to the Exchange Rate: Doctrinal Aspects and Empirical Evidence,” *Scandinavian Journal of Economics*, 78(2), 200–224.
- HALL, R. E. (1978): “Stochastic Implications of the Life Cycle-Permanent Income Hypothesis: Theory and Evidence,” *Journal of Political Economy*, 86, 971–987.
- (1988): “Intertemporal Substitution in Consumption,” *Journal of Political Economy*, 96, 339–357.
- HANSEN, L. P., AND K. J. SINGLETON (1983): “Stochastic Consumption, Risk Aversion, and the Temporal Behavior of Asset Returns,” *Journal of Political Economy*, 91, 249–265.
- (1996): “Efficient Estimation of Linear Asset-Pricing Models with Moving Average Errors,” *Journal of Business and Economic Statistics*, 14, 53–68.
- KILIAN, L. (1999): “Exchange Rates and Monetary Fundamentals: What Do We Learn from Long-Horizon Regressions?,” *Journal of Applied Econometrics*, 14(5), 491–510.
- KIM, H., AND M. OGAKI (2009): “Purchasing Power Parity and the Taylor Rule,” Working Paper No. 09-03, Department of Economics, Ohio State University.

- KINGSTON, G. H. (1984): "Efficient Timing of Income Taxes," *Journal of Public Economics*, 24(2), 271–280.
- MANKIW, N. G. (1987): "The Optimal Collection of Seigniorage: Theory and Evidence," *Journal of Monetary Economics*, 20(2), 327–341.
- MARK, N. C. (1995): "Exchange Rates and Fundamentals: Evidence on Long-Horizon Predictability," *American Economic Review*, 85, 201–218.
- (2005): "Changing Monetary Policy Rules, Learning, and Real Exchange Rate Dynamics," NBER Working Paper No. 11061.
- MARK, N. C., AND D. SUL (2001): "Nominal Exchange Rates and Monetary Fundamentals: Evidence from a Small Post-Bretton Woods Panel," *Journal of International Economics*, 53(1), 29–52.
- MCCRACKEN, M. W. (2004): "Parameter Estimation and Tests of Equal Forecast Accuracy between Non-nested Models," *International Journal of Forecasting*, 20(3), 503–514.
- MEESE, R. A., AND K. ROGOFF (1983): "Empirical Exchange Rate Models of the Seventies: Do They Fit Out of Sample?," *Journal of International Economics*, 14(1–2), 3–24.
- MOLODTSOVA, T., A. NIKOLSKO-RZHEVSKYY, AND D. H. PAPELL (2008): "Taylor Rules with Real-Time Data: A Tale of Two Countries and One Exchange Rate," *Journal of Monetary Economics*, 55(S1), S63–S79.
- MOLODTSOVA, T., AND D. H. PAPELL (2009): "Out-of-Sample Exchange Rate Predictability with Taylor Rule Fundamentals," *Journal of International Economics*, 77(2), 167–180.
- MUSSA, M. (1976): "The Exchange Rate, the Balance of Payments, and Monetary and Fiscal Policy under a Regime of Controlled Floating," *Scandinavian Journal of Economics*, 78(2), 229–248.
- ROSE, A. K. (1988): "Is the Real Interest Rate Stable?," *Journal of Finance*, 43(5), 1095–1112.
- SAHASAKUL, C. (1986): "The U.S. Evidence on Optimal Taxation over Time," *Journal of Monetary Economics*, 18(3), 251–275.
- SIEGEL, J. J. (1972): "Risk, Interest Rates and the Forward Exchange," *Quarterly Journal of Economics*, 86(2), 303–309.
- TAYLOR, J. B. (1993): "Discretion versus Policy Rules in Practice," *Carnegie-Rochester Conference Series on Public Policy*, 39, 195–214.

## Chapter 8

# VECTOR AUTOREGRESSION TECHNIQUES

This chapter discusses econometric techniques for vector autoregressions (VAR). In most cases, the variables in VAR are assumed to be stationary.<sup>1</sup>

Let  $\mathbf{y}_t$  be an  $n$ -dimensional vector stochastic process that is covariance stationary. Because  $\mathbf{y}_t$  is covariance stationary, it has a Wold representation:

$$(8.1) \quad \mathbf{y}_t = \boldsymbol{\mu} + \boldsymbol{\epsilon}_t + \boldsymbol{\Psi}_1 \boldsymbol{\epsilon}_{t-1} + \boldsymbol{\Psi}_2 \boldsymbol{\epsilon}_{t-2} + \cdots = \boldsymbol{\mu} + \boldsymbol{\Psi}(L)\boldsymbol{\epsilon}_t,$$

where  $\boldsymbol{\Psi}(L) = \mathbf{I}_n + \sum_{s=1}^{\infty} \boldsymbol{\Psi}_s L^s$  and  $L$  is the lag operator. Assuming that  $\boldsymbol{\Psi}(L)$  is invertible,  $\mathbf{y}_t$  has a VAR representation. Assuming that the VAR representation is of order  $p$ :

$$(8.2) \quad \mathbf{A}(L)\mathbf{y}_t = \boldsymbol{\delta}_\epsilon + \boldsymbol{\epsilon}_t,$$

---

<sup>1</sup>A VAR model may include nonstationary variables. Chapter 16 treats the case where some of the variables in VAR are difference stationary and cointegrated, terms that will be introduced later. When the difference stationary variables are not cointegrated, we can take the first difference to make them stationary for VAR.

where

$$\begin{aligned}
 (8.3) \quad \delta_\epsilon &= \Psi(1)^{-1}\boldsymbol{\mu} = \mathbf{A}(1)\boldsymbol{\mu}, \\
 \mathbf{A}(L) &= \Psi(L)^{-1} = \mathbf{I}_n - \sum_{i=1}^p \mathbf{A}_i L^i, \\
 \boldsymbol{\epsilon}_t &= \mathbf{y}_t - \hat{E}(\mathbf{y}_t | \mathbf{y}_{t-1}, \mathbf{y}_{t-2}, \mathbf{y}_{t-3}, \dots)
 \end{aligned}$$

and

$$(8.4) \quad E(\boldsymbol{\epsilon}_t \boldsymbol{\epsilon}_t') = \boldsymbol{\Sigma}_\epsilon.$$

Here  $\hat{E}(\cdot | \mathbf{y}_{t-1}, \mathbf{y}_{t-2}, \mathbf{y}_{t-3}, \dots)$  is defined to be the linear projection operator onto the linear space spanned by a constant (say, 1) and  $\mathbf{y}_{t-1}, \mathbf{y}_{t-2}, \mathbf{y}_{t-3}, \dots$ . In virtually all applications,  $\boldsymbol{\Sigma}_\epsilon$  is not diagonal. However, the Seemingly Unrelated Regression Estimator (SUR) coincides with the OLS estimator for (8.2) because the regressors are identical for all regressions when OLS is applied to each row of (8.2).

## 8.1 OLS Estimation

The VAR (8.2) gives a system of regression equations. It may appear that the SUR estimator should be used to estimate these equations because the error terms are contemporaneously correlated. However, the OLS and SUR estimators coincide because the regressors are the same for all equations. Hence, we can estimate each equation by OLS.

It is often convenient to use a matrix expression to write the OLS estimators for the VAR system. For this purpose, rewrite (8.2) by stacking it from  $t = 1, \dots, T$  after transpose:

$$(8.5) \quad \mathbf{Y} = \mathbf{XB} + \mathbf{U}$$

where

$$\mathbf{Y} = \begin{bmatrix} \mathbf{y}'_1 \\ \vdots \\ \mathbf{y}'_t \\ \vdots \\ \mathbf{y}'_T \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & \mathbf{y}'_{1-1} \cdots \mathbf{y}'_{1-p} \\ \vdots \\ 1 & \mathbf{y}'_{t-1} \cdots \mathbf{y}'_{t-p} \\ \vdots \\ 1 & \mathbf{y}'_{T-p} \cdots \mathbf{y}'_{T-p} \end{bmatrix}, \mathbf{B} = \begin{bmatrix} \boldsymbol{\delta}'_\epsilon \\ \mathbf{A}'_1 \\ \vdots \\ \mathbf{A}'_p \end{bmatrix}, \text{ and } \mathbf{U} = \begin{bmatrix} \boldsymbol{\epsilon}'_1 \\ \vdots \\ \boldsymbol{\epsilon}'_t \\ \vdots \\ \boldsymbol{\epsilon}'_T \end{bmatrix}.$$

In order to apply OLS techniques, express (8.5) in its vector form:

$$(8.6) \quad \mathbf{y} = (\mathbf{I}_n \otimes \mathbf{X})\mathbf{b} + \mathbf{u},$$

where  $\mathbf{y} = \text{vec}(\mathbf{Y})$ ,  $\mathbf{b} = \text{vec}(\mathbf{B})$ ,  $\mathbf{u} = \text{vec}(\mathbf{U})$ , and  $E(\mathbf{u}\mathbf{u}') = \boldsymbol{\Sigma}_\epsilon \otimes \mathbf{I}_T$ . Applying OLS techniques, we get

$$(8.7) \quad \hat{\mathbf{b}} = (\mathbf{I}_n \otimes (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{y}$$

and

$$(8.8) \quad \text{var}(\hat{\mathbf{b}}) = \boldsymbol{\Sigma}_\epsilon \otimes (\mathbf{X}'\mathbf{X})^{-1}.$$

In many applications, we express the asymptotic variance in (8.8) using the notation  $\mathbf{a} = \text{vec}(\boldsymbol{\delta}_\epsilon \mathbf{A}_1 \cdots \mathbf{A}_p)$ . Let  $\mathbf{K}_{rc}$  be the  $rc \times rc$  dimensional commutation matrix that has the property of  $\text{vec}(\mathbf{M}') = \mathbf{K}_{rc}\text{vec}(\mathbf{M})$  for any  $r \times c$  matrix  $\mathbf{M}$ . Then, we can show that

$$(8.9) \quad \hat{\mathbf{a}} = \mathbf{K}_{(np+1)n}\hat{\mathbf{b}}$$

and

$$(8.10) \quad \text{var}(\hat{\mathbf{a}}) = (\mathbf{X}'\mathbf{X})^{-1} \otimes \boldsymbol{\Sigma}_\epsilon.$$

## 8.2 Granger Causality

Let  $\mathbf{y}_t = (x_t, y_t)'$  be a two dimensional covariance stationary process. We say that  $y$  fails to *Granger-cause*  $x$  if for all  $s > 0$ ,

$$(8.11) \quad \hat{E}(x_{t+s}|x_t, x_{t-1}, \dots, y_t, y_{t-1}, \dots) = \hat{E}(x_{t+s}|x_t, x_{t-1}, \dots).$$

We also say that  $y$  is not *linearly informative* about future  $x$ , or  $x$  is exogenous in the time series sense with respect to  $y$ .

One can test the null hypothesis that  $y$  fails to Granger-cause  $x$  by applying the OLS to

$$(8.12) \quad x_t = \delta_{\epsilon_1} + a_{1,11}x_{t-1} + \dots + a_{p,11}x_{t-p} + a_{1,12}y_{t-1} + \dots + a_{p,12}y_{t-p} + \epsilon_{1t}.$$

If  $y$  fails to Granger-cause  $x$ , then  $a_{i,12} = 0$  for  $i = 1, \dots, p$  in (8.12). Conversely, if  $a_{i,12} = 0$  for  $i = 1, \dots, p$  in (8.12), then

$$(8.13) \quad \hat{E}(x_{t+1}|x_t, x_{t-1}, \dots, y_t, y_{t-1}, \dots) = \delta_{\epsilon_1} + a_{1,11}x_t + \dots + a_{p,11}x_{t-p+1}$$

and

$$(8.14) \quad \begin{aligned} & \hat{E}(x_{t+2}|x_t, x_{t-1}, \dots, y_t, y_{t-1}, \dots) \\ &= \delta_{\epsilon_1} + a_{1,11}\hat{E}(x_{t+1}|x_t, x_{t-1}, \dots, y_t, y_{t-1}, \dots) + a_{2,11}x_t + \dots + a_{p,11}x_{t-p+2}. \end{aligned}$$

Repeating this argument, we see that  $y$  fails to Granger-cause  $x$ . Hence we test the null hypothesis

$$(8.15) \quad H_0 : a_{i,12} = 0 \text{ for } i = 1, \dots, p$$

in (8.12) in order to test for Granger causality.

The result that  $y$  fails to Granger-cause  $x$  if and only if (8.15) holds in (8.12) can be used to find restrictions on the VAR representation for  $\mathbf{y} = (x, y)'$ . Suppose that  $y$  fails to Granger-cause  $x$ , and  $x$  Granger-causes  $y$ .<sup>2</sup> Let the VAR representation of  $\mathbf{y}$  be given by (8.2). Then the restrictions (8.15) hold if and only if  $\mathbf{A}_i$  is lower triangular for each  $i$ :

$$(8.16) \quad \mathbf{A}_i = \begin{bmatrix} a_{i,11} & 0 \\ a_{i,21} & a_{i,22} \end{bmatrix}.$$

Hence  $y$  fails to Granger-cause  $x$ , and  $x$  Granger-causes  $y$  if and only if the VAR representation for  $\mathbf{y} = (x, y)'$  given by (8.2) satisfies the restrictions that  $\mathbf{A}_i$  is lower triangular for each  $i$  as in (8.16).

Suppose that an econometrician finds evidence for the hypothesis that  $y$  fails to Granger-cause  $x$ , but  $x$  Granger-causes  $y$  (i.e., the null hypothesis that  $y$  fails to Granger-cause  $x$  cannot be rejected, but the null hypothesis that  $x$  fails to Granger-cause  $y$  can be rejected). For example, researchers have found some evidence that real GDP fails to Granger-cause the money supply, and the money supply Granger-causes real GDP. This type of finding is consistent with some economic models which predict that a decrease in the money supply causes real GDP to fall.

It should be noted, however, that Granger-causality relationships can be very different from causal relationships when economic variables respond to future expected values of other variables as in the rational expectations models. Hence Granger-causality test results must be interpreted with caution.

For example, consider the present value model of a stock price:

$$(8.17) \quad p_t = E\left(\sum_{i=1}^{\infty} b^i d_{t+i} | I_t\right).$$

---

<sup>2</sup>Having defined the meaning of “ $x$  fails to Granger-cause  $y$ ,” we define “ $x$  Granger-causes  $y$ ,” to mean “ $x$  does not fail to Granger-cause  $y$ .”

where  $p_t$  is the stock price and  $d_t$  is the dividend. In order to illustrate the point in a simple example, assume that

$$(8.18) \quad d_t = u_t + \delta u_{t-1} + v_t,$$

where  $u_t$  and  $v_t$  are normal i.i.d. and are independent of each other. Here, the mean of the log of the dividend is normalized to be zero. Then

$$(8.19) \quad E_t(d_{t+i}) = \begin{cases} \delta u_t & \text{for } i = 1 \\ 0 & \text{for } i > 1 \end{cases},$$

which implies  $p_t = b\delta u_t$ . Therefore,  $\delta u_{t-1} = b^{-1}p_{t-1}$ . Hence, the VAR representation for  $\mathbf{y}_t = (p_t, d_t)'$  is

$$(8.20) \quad \begin{bmatrix} p_t \\ d_t \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ b^{-1} & 0 \end{bmatrix} \begin{bmatrix} p_{t-1} \\ d_{t-1} \end{bmatrix} + \begin{bmatrix} b\delta u_t \\ u_t + v_t \end{bmatrix}.$$

Since the VAR coefficient matrix is lower triangular, the dividend fails to Granger-cause the stock price, and the stock price Granger-causes the dividend in this example.

Since the changes in the future expected dividends cause the stock price to change in the present value model, the causal relationship is the opposite of the Granger-causality relationship. This result occurs because the stock price responds to the future expected values of the dividends in the present value model. When future dividends are expected to rise, the current stock price rises. Hence, the stock price tends to move before the dividend moves. This result does not mean that the stock price causes the dividend to move, but can mean that the stock price Granger-causes the dividend as in the example. In this sense, Granger “causality” is a misnomer.<sup>3</sup> It is safer to interpret Granger causality test results in terms of linear informativeness.

---

<sup>3</sup>Leamer (1985) suggests to use the word “precedence” instead of “causality”. He argues that what is tested in “Granger Causality” is whether one variable regularly precedes another and that “precedence” is not sufficient for causality.

An example with this interpretation is Stock and Watson's (1989) application to search for economic variables that forecast business cycle movements.

### 8.3 The Impulse Response Function

Consider a moving average representation

$$\mathbf{y}_t = \boldsymbol{\mu} + \boldsymbol{\Psi}_0^* \boldsymbol{\epsilon}_t^* + \boldsymbol{\Psi}_1^* \boldsymbol{\epsilon}_{t-1}^* + \boldsymbol{\Psi}_2^* \boldsymbol{\epsilon}_{t-2}^* + \cdots = \boldsymbol{\mu} + \boldsymbol{\Psi}^*(L) \boldsymbol{\epsilon}_t^*.$$

Let  $y_{it}$  be the  $i$ -th element of  $\mathbf{y}_t$ ,  $\epsilon_{jt}^*$  be the  $j$ -th element of  $\boldsymbol{\epsilon}_t^*$ , and  $\psi_{s,ij}^*$  be the  $(i, j)$ -th element of  $\boldsymbol{\Psi}_s^*$ . If  $\epsilon_{jt}^*$  is increased by one unit while holding all the other elements of  $\boldsymbol{\epsilon}_{t+\tau}^*$  constant for all positive and negative  $\tau$ , then  $y_{i,t+s}$  will increase by  $\psi_{s,ij}^*$  for  $s > 0$ . In this sense,

$$(8.21) \quad \frac{\partial y_{i,t+s}}{\partial \epsilon_{jt}^*} = \psi_{s,ij}^*,$$

or, using matrix notation,

$$(8.22) \quad \frac{\partial \mathbf{y}_{t+s}}{\partial \boldsymbol{\epsilon}_t^{*'}} = \boldsymbol{\Psi}_s^*,$$

A plot of  $\psi_{s,ij}^*$  for  $s = 1, 2, \dots$  is the *impulse response function* of  $y_i$  with respect to  $\epsilon_{jt}^*$ .

One convenient way to estimate the impulse response function is to choose the Wold representation (8.1):

$$\mathbf{y}_t = \boldsymbol{\mu} + \boldsymbol{\epsilon}_t + \boldsymbol{\Psi}_1 \boldsymbol{\epsilon}_{t-1} + \boldsymbol{\Psi}_2 \boldsymbol{\epsilon}_{t-2} + \cdots = \boldsymbol{\mu} + \boldsymbol{\Psi}(L) \boldsymbol{\epsilon}_t,$$

estimate the VAR representation by applying OLS to each row of  $\mathbf{y}_t$ , and simulate the estimated VAR representation to obtain an estimate of  $\boldsymbol{\Psi}_s$ .

There exist two difficulties in interpreting the impulse response function. The first difficulty is that  $\boldsymbol{\Sigma}_\epsilon = E(\boldsymbol{\epsilon}_t \boldsymbol{\epsilon}_t')$  is not diagonal. This property means that the

other elements of  $\boldsymbol{\epsilon}_t$  tend to move with  $\epsilon_{jt}$  when  $\epsilon_{jt}$  changes. Hence, it is not very meaningful to consider the effect of changes in  $\epsilon_{jt}$  on  $y_{i,t+s}$  while holding the other elements of  $\boldsymbol{\epsilon}_t$  constant. Computing an orthogonalized impulse response function is one method to avoid this difficulty. We assume that  $\boldsymbol{\Sigma}_\epsilon$  is positive definite. Then, given the ordering of variables in  $\mathbf{y}_t$ , there exists a unique lower triangular matrix  $\boldsymbol{\Phi}_0$  with 1's along the principal diagonal and a unique diagonal matrix  $\boldsymbol{\Lambda}$  with positive entries along the principal diagonal such that

$$(8.23) \quad \boldsymbol{\Sigma}_\epsilon = \boldsymbol{\Phi}_0 \boldsymbol{\Lambda} \boldsymbol{\Phi}_0'$$

Let

$$(8.24) \quad \mathbf{e}_t = \boldsymbol{\Phi}_0^{-1} \boldsymbol{\epsilon}_t.$$

Then  $E(\mathbf{e}_t \mathbf{e}_t') = \boldsymbol{\Phi}_0^{-1} \boldsymbol{\Sigma}_\epsilon (\boldsymbol{\Phi}_0^{-1})' = \boldsymbol{\Lambda}$  which is diagonal. Since

$$(8.25) \quad \boldsymbol{\epsilon}_t = \boldsymbol{\Phi}_0 \mathbf{e}_t,$$

$\mathbf{y}_t$  has an MA representation in terms of  $\mathbf{e}_t$ :

$$(8.26) \quad \mathbf{y}_t = \boldsymbol{\mu} + \boldsymbol{\Phi}_0 \mathbf{e}_t + \boldsymbol{\Psi}_1 \boldsymbol{\Phi}_0 \mathbf{e}_{t-1} + \boldsymbol{\Psi}_2 \boldsymbol{\Phi}_0 \mathbf{e}_{t-2} + \cdots = \boldsymbol{\mu} + \boldsymbol{\Phi}(L) \mathbf{e}_t,$$

where  $\boldsymbol{\Phi}(L) = \sum_{s=0}^{\infty} \boldsymbol{\Phi}_s L^s$  and  $\boldsymbol{\Phi}_s = \boldsymbol{\Psi}_s \boldsymbol{\Phi}_0$ . Let  $e_{jt}$  be the  $j$ -th element of  $\mathbf{e}_t$  and  $\phi_{s,ij}$  be the  $(i, j)$ -th element of  $\boldsymbol{\Phi}_s$ . Then (8.26) implies that

$$(8.27) \quad \frac{\partial y_{i,t+s}}{\partial e_{jt}} = \phi_{s,ij}.$$

A plot of (8.27) as a function of  $s \geq 0$  is an *orthogonalized impulse response function*.

The sample counterparts of  $\boldsymbol{\Psi}_s$  and  $\boldsymbol{\Phi}_0$  can be used to estimate the orthogonalized impulse response function. For example, the Cholesky factorization, which

GAUSS can be used to compute, of the estimate of  $\Sigma_\epsilon$  can be used to estimate  $\Phi_0$ . If  $\mathbf{P}$  is the Cholesky factorization of  $\Sigma_\epsilon$ , then  $\mathbf{P} = \Phi_0 \Lambda^{\frac{1}{2}}$ , and the principal diagonal of  $\mathbf{P}$  is the principal diagonal of  $\Lambda^{\frac{1}{2}}$ . Hence,  $\Phi_0 = \mathbf{P} \Lambda^{-\frac{1}{2}}$ . This formula can be used to construct a sample counterpart of  $\Phi_0$ .

The second difficulty in interpreting the impulse response function is that it is not possible to interpret  $\epsilon_t$  or  $\mathbf{e}_t$  as shocks to the economy without imposing any economic structure to the VAR representation. For example, if the first element in  $\mathbf{y}_t$  is the money supply, it is tempting to interpret the first element of  $\epsilon_t$  as the money supply shock which represents random changes in the money supply. With this interpretation, one can learn about how endogenous variables respond to the money supply shock by examining the impulse response functions. However, without any economic model,  $\epsilon_t$  is simply the forecast error when the linear forecasting rule is used with the past values of  $\mathbf{y}_t$  as the information set. In some linear rational expectations models,  $\epsilon_t$  is simply the difference between the economic agents' forecast and the linear forecast based on the past values of  $\mathbf{y}_t$ . When the economic agents use a nonlinear forecasting rule with a larger information set, their forecast can be very different from  $\hat{E}(\mathbf{y}_t | \mathbf{y}_{t-1}, \mathbf{y}_{t-2}, \dots)$ . In these models, it is not clear what we learn from the impulse response functions. Section 8.5 will discuss structural models that provide economically meaningful shocks with various restrictions. Under the recursive assumptions introduced in Section 8.5, the orthogonalized impulse response function discussed above can be used to compute impulse response functions of the structural shocks. In the majority of the VAR applications, the recursive assumptions are used. Under other assumptions, alternative methods are used to compute impulse response functions for the structural shocks as explained below.

We provide three traditional methods of computing confidence intervals of impulse responses: asymptotic normal approximation (see, e.g., Lütkepohl, 1990), bootstrap (see, e.g., Runkle, 1987; Kilian, 1998), and Monte Carlo integration (see, e.g., Doan, 1992; Sims and Zha, 1999). All three methods are asymptotically valid in stationary models but not the same in small samples. Kilian (1998) shows from his Monte Carlo simulation that bootstrap-after-bootstrap method performs better than others in small samples, while Sims and Zha (1999) argue that the Bayesian intervals have a firmer theoretical foundation and show how to obtain correct intervals for over-identified models.

## 8.4 Forecast error decomposition

Denoting the  $h$ -step forecast error by

$$(8.28) \quad \begin{aligned} \mathbf{y}_{t+h} - \hat{E}_t \mathbf{y}_{t+h} &= \sum_{s=0}^{\infty} \Psi_s (\boldsymbol{\epsilon}_{t+h-s} - \hat{E}_t \boldsymbol{\epsilon}_{t+h-s}) \\ &= \sum_{s=0}^{h-1} \Psi_s \boldsymbol{\epsilon}_{t+h-s}, \end{aligned}$$

the forecast error variance is computed from the diagonal components of

$$(8.29) \quad E(\mathbf{y}_{t+h} - \hat{E}_t \mathbf{y}_{t+h})^2 = \sum_{s=0}^{h-1} \Psi_s \Sigma_{\epsilon} \Psi_s'.$$

In particular, the forecast error variance of the  $i$ -th variable,  $y_{i,t+h}$ , is defined by

$$(8.30) \quad \sum_{s=0}^{h-1} \Psi_{s,i} \Sigma_{\epsilon} \Psi_{s,i}'.$$

where  $\Psi_{s,i}$  denotes the  $i$ -th row of  $\Psi_s$ .

The same two difficulties concerning the interpretation of the impulse response function exist for the forecast variance decomposition. As with the impulse response

function, the recursive assumptions have been employed in many VAR applications so that the orthogonalized shocks  $\mathbf{e}_t$  in (8.24) are structural shocks.

The contribution of orthogonalized shocks to forecast error variance of the  $h$ -step forecast is defined by the diagonal components of

$$(8.31) \quad \sum_{s=0}^{h-1} \Phi_s \Lambda \Phi_s'.$$

In particular, the contribution of the  $j$ -th orthogonalized shock,  $e_j$ , to the forecast error variance of the  $i$ -th variable,  $y_{i,t+h}$ , is<sup>4</sup>

$$(8.32) \quad \sum_{s=0}^{h-1} (\phi_{s,ij})^2 d_{jj},$$

where  $d_{jj}$  is the variance of the  $j$ -th orthogonalized shock. The sample counterparts of  $\Phi$  and  $d_{jj}$  can be used to estimate this contribution.

Finally, dividing (8.32) by (8.30) yields the fraction of the  $h$ -step forecast error variance of the  $i$ -th variable attributed to the  $j$ -th orthogonalized shock.

## 8.5 Structural VAR Models

This section discusses structural economic models in which the orthogonalized impulse response functions are meaningful. A class of structural models can be written in the following form of a structural dynamic model:

$$(8.33) \quad \mathbf{B}_0 \mathbf{y}_t = \boldsymbol{\delta} + \mathbf{B}_1 \mathbf{y}_{t-1} + \mathbf{B}_2 \mathbf{y}_{t-2} + \cdots + \mathbf{B}_p \mathbf{y}_{t-p} + \mathbf{e}_t$$

where  $\mathbf{B}_i$  is a  $n \times n$  matrix, and  $\boldsymbol{\delta}$  is a  $n \times 1$  vector. Here  $\mathbf{B}_0$  is a nonsingular matrix of real numbers with 1's along its principal diagonal, and  $\mathbf{e}_t$  is a stationary  $n$ -dimensional

---

<sup>4</sup>By virtue of the assumption that orthogonalized shocks are mutually uncorrelated, we can separate the contribution of each orthogonalized shock.

vector of random variables with  $E(\mathbf{e}_t | \mathbf{y}_{t-1}, \mathbf{y}_{t-2}, \dots) = \mathbf{0}$ . This structural model is related to its reduced form with  $\mathbf{e}_t = \mathbf{B}_0 \boldsymbol{\epsilon}_t$ ,  $\boldsymbol{\delta} = \mathbf{B}_0 \boldsymbol{\delta}_\epsilon$ ,  $\mathbf{B}_i = \mathbf{B}_0 \mathbf{A}_i$  for  $i = 1, \dots, p$ . In many applications, it is assumed that the shocks are mutually uncorrelated so that the covariance matrix of  $\mathbf{e}_t$  is diagonal.

**Example 8.1** Consider a model of money demand. Let  $m_t$  be the real money balance,  $m_t^d$  be the desired real money balance, and  $i_t$  be the nominal interest rate:

$$(8.34) \quad m_t^d = \beta_0 + \beta_1 i_t.$$

Suppose that the actual money holdings are slowly adjusted toward the desired level so that

$$(8.35) \quad m_t - m_t^d = \alpha(m_{t-1} - m_{t-1}^d) + e_t^d,$$

where  $0 < \alpha < 1$ , and  $e_t^d$  is a money demand shock. Substituting (8.34) into (8.35) yields

$$(8.36) \quad m_t = \beta_0(1 - \alpha) + \alpha m_{t-1} + \beta_1 i_t - \alpha \beta_1 i_{t-1} + e_t^d.$$

Imagine that the central bank determines the money supply at date  $t$  so that  $i_t$  is at a desired level given by the right hand side of the following equation:

$$(8.37) \quad i_t = \gamma_0 + \gamma_1 m_{t-1} + \gamma_2 i_{t-1} + e_t^s,$$

where  $e_t^s$  is a money supply shock. Then, when we choose  $(i_t, m_t)'$  as  $\mathbf{y}_t$ , this money demand model is of the form (8.33):

$$(8.38) \quad \begin{bmatrix} 1 & 0 \\ -\beta_1 & 1 \end{bmatrix} \begin{bmatrix} i_t \\ m_t \end{bmatrix} = \begin{bmatrix} \gamma_0 \\ \beta_0(1 - \alpha) \end{bmatrix} + \begin{bmatrix} \gamma_2 & \gamma_1 \\ -\alpha\beta_1 & \alpha \end{bmatrix} \begin{bmatrix} i_{t-1} \\ m_{t-1} \end{bmatrix} + \begin{bmatrix} e_t^s \\ e_t^d \end{bmatrix}. \blacksquare$$

In general,  $\mathbf{B}_0$  in (8.33) is not diagonal because some variables in  $\mathbf{y}_t$  are affected by other variables in  $\mathbf{y}_t$  as well as lagged values of the variables in  $\mathbf{y}_t$ . In Example 8.1,  $m_t$  is affected by  $i_t$  as well as lagged values of  $i_{t-1}$  and  $m_{t-1}$ .

In many structural models, it is reasonable to assume that the covariance matrix of  $\mathbf{e}_t$  is diagonal. In Example 8.1,  $e_t^d$  is the money demand shock and  $e_t^s$  is the money supply shock, and these shocks should be uncorrelated. In this case, the impulse response functions with respect to the elements of  $\mathbf{e}_t$  can be interpreted without any problem and are of interest. We will assume that  $\mathbf{\Lambda} = E(\mathbf{e}_t \mathbf{e}_t')$  is diagonal for the rest of this chapter.

When the reduced form VAR (8.2) is estimated, various restrictions can be imposed on  $\mathbf{B}_0$  to compute the impulse response functions of  $\mathbf{e}_t$ . For example, suppose that  $\mathbf{B}_0$  is known. Let  $\mathbf{\Phi}_0 = \mathbf{B}_0^{-1}$ ,  $\Psi_s = \partial \mathbf{y}_{t+s} / \partial \boldsymbol{\epsilon}_t'$  be the impulse response function with respect to  $\boldsymbol{\epsilon}_t$ , and  $\Phi_{0,j}$  be the  $j$ -th column of  $\mathbf{\Phi}_0$ . By the same argument used for the orthogonalized impulse response function,  $\Psi_s \Phi_{0,j}$  gives the impulse response function with respect to  $e_{jt}$ .

In most models,  $\mathbf{B}_0$  is unknown. A restriction on  $\mathbf{B}_0$  often used in applications is that it is a lower triangular matrix. Example 8.1 satisfies this restriction. In the example,  $i_t$  is determined by  $i_{t-1}$  and  $m_{t-1}$  and is not affected by  $m_t$ . Note that  $\mathbf{B}_0$  would not be lower triangular if  $\mathbf{y}_t$  were defined to be  $(m_t, i_t)'$  rather than  $(i_t, m_t)'$ . Thus the order of the variables in  $\mathbf{y}_t$  is important. In general,  $\mathbf{B}_0$  is lower triangular when the model has a *recursive structure*:  $y_{1t}$  is determined when the past values of  $\mathbf{y}_t$  are given,  $y_{2t}$  is determined by  $y_{1t}$  and the past values of  $\mathbf{y}_t$ ,  $y_{3t}$  is determined by  $y_{1t}, y_{2t}$ , and the past values of  $\mathbf{y}_t$ .

When  $\mathbf{B}_0$  is lower triangular,  $\mathbf{B}_0^{-1}$  is a lower triangular matrix and has 1's along the principal diagonal. It is known that when a positive matrix  $\Sigma_\epsilon$  is given, there exists a unique lower triangular matrix  $\Phi_0$  which has ones along the principal diagonal such that  $\Sigma_\epsilon = \Phi_0 \Lambda \Phi_0'$ . Hence  $\mathbf{B}_0$  can be computed by the Cholesky factorization using  $\mathbf{B}_0 = \Phi_0^{-1}$ . Thus, the standard method of computing the orthogonalized impulse response function yields the impulse response function with respect to  $\mathbf{e}_t$  when  $\mathbf{B}_0$  is lower triangular. On the other hand, when  $\mathbf{B}_0$  is not lower triangular the Choleski decomposition cannot be used, and ML or GMM estimation is often used as discussed in Section 8.6.3.

## 8.6 Identification

In order to identify  $\mathbf{B}_0$ , we need at least  $n^2$  restrictions. In most cases, we assume that structural shocks are mutually uncorrelated. This orthogonality condition implies the variance-covariance matrix of structural disturbances is diagonal and gives  $\frac{n(n-1)}{2}$  restrictions. Second, we impose a normalization condition that the diagonal components of  $\mathbf{B}_0$  are 1's, which yields  $n$  restrictions.<sup>5</sup> Structural VAR varies depending on how the additional  $\frac{n(n-1)}{2}$  conditions are imposed for identification.

### 8.6.1 Short-Run Restrictions for Structural VAR

The simplest model originating with Sims (1980) assumes that  $\mathbf{B}_0$  is lower triangular. This structure is called recursive assumptions. This gives  $\frac{n(n-1)}{2}$  necessary conditions so that the model is just identified as shown below. Letting  $\Phi_0 = \mathbf{B}_0^{-1}$ , it follows

---

<sup>5</sup>Instead, we can consider an alternative normalization condition that the variance-covariance matrix of structural disturbances is an identity matrix. This change does not affect the main results.

from  $\mathbf{e}_t = \mathbf{B}_0\boldsymbol{\epsilon}_t$  that

$$(8.39) \quad \boldsymbol{\Phi}_0\boldsymbol{\Lambda}\boldsymbol{\Phi}_0' = \boldsymbol{\Sigma}_\epsilon,$$

where  $\boldsymbol{\Phi}_0$  is also a lower triangular matrix. Let  $\mathbf{P}$  be a lower triangular matrix of the Cholesky decomposition of  $\boldsymbol{\Sigma}_\epsilon$  so that  $\mathbf{P}\mathbf{P}' = \boldsymbol{\Sigma}_\epsilon$ . From  $\boldsymbol{\Phi}_0\boldsymbol{\Lambda}^{\frac{1}{2}} = \mathbf{P}$ , it follows that

$$(8.40) \quad \boldsymbol{\Phi}_0 = \mathbf{P}\boldsymbol{\Lambda}^{-\frac{1}{2}},$$

where  $\boldsymbol{\Lambda} = [\text{diag}(\mathbf{P})]^2$ .

Typically, researchers decide the order of variables to use from the type of restrictions, but do not use a tightly specified economic model to derive these restrictions in applications. Instead, impulse responses estimated from recursive assumptions are compared with implications of economic models. Some researchers make a more explicit connection between estimated impulse responses and an economic model. Rotemberg and Woodford (1999) minimize a distance measure between impulse responses estimated from recursive assumptions and impulse responses implied by a monetary model by choosing parameters of the model. Their monetary model incorporates an optimum monetary policy rule that is similar to the rule proposed by Taylor (1993).

Blanchard and Watson (1986) consider the case where  $\mathbf{B}_0$  is not lower triangular. As their four-variable model includes eight unknown parameters in  $\mathbf{B}_0$ , they use a priori theoretical and empirical information about the private sector behavior and policy reaction functions on two of the parameters, and impose four zero restrictions to achieve identification on the remaining six ( $=\frac{n(n-1)}{2}$ ) unknown parameters. Given these restrictions, their model is just identified. From  $\mathbf{e}_t = \mathbf{B}_0\boldsymbol{\epsilon}_t$  it follows that

$$(8.41) \quad \boldsymbol{\Lambda} = \mathbf{B}_0\boldsymbol{\Sigma}_\epsilon\mathbf{B}_0',$$

which yields unique solutions for  $\mathbf{B}_0$  and  $\mathbf{\Lambda}$ . Gordon and Leeper (1994) use full information maximum likelihood estimation to study liquidity effects in their over-identified model. To identify their model, they impose conventional exclusion restrictions and plausible informational assumptions from a traditional view of monetary policy and private sector behavior, such as which variables enter demand and supply for the reserve market.

Bernanke (1986) considers a model that allows more than one structural shock in an equation. The structural form is

$$(8.42) \quad \mathbf{B}(L)\mathbf{y}_t = \mathbf{F}\boldsymbol{\epsilon}_t.$$

Assume that  $\mathbf{B}_0$  is not lower triangular but that there are  $\frac{n(n-1)}{2}$  unknown parameters in  $\mathbf{B}_0$  and  $\mathbf{F}$ . From  $\mathbf{F}\boldsymbol{\epsilon}_t = \mathbf{B}_0\boldsymbol{\epsilon}_t$  it follows that

$$(8.43) \quad \mathbf{\Lambda} = \mathbf{F}^{-1}\mathbf{B}_0\boldsymbol{\Sigma}_\epsilon\mathbf{B}_0'\mathbf{F}^{-1'},$$

which yields the unique solutions for  $\mathbf{B}_0$ ,  $\mathbf{F}$  and  $\mathbf{\Lambda}$ .

### 8.6.2 Identification of block recursive systems

Christiano, Eichenbaum, and Evans (1999) provide a theoretical background and illustrate identification of block recursive systems. Partitioning  $\mathbf{y}_t$  into three blocks is convenient to illustrate the block recursive structure:

$$(8.44) \quad \mathbf{y}_t = \begin{bmatrix} \mathbf{y}_{1t} \\ s_t \\ \mathbf{y}_{2t} \end{bmatrix},$$

where  $\mathbf{y}_t$  is a vector of  $n(= n_1 + 1 + n_2)$  variables of interest,  $s_t$  is a monetary policy variable,  $\mathbf{y}_{1t}$  includes  $n_1$  variables which are in the information set when the Fed implements a monetary policy, and  $\mathbf{y}_{2t}$  contains  $n_2$  variables which are excluded from

the information set. Alternatively,  $\mathbf{y}_{1t}$  does not respond to a monetary policy shock contemporaneously, while  $\mathbf{y}_{2t}$  does. The block recursive assumption imposes zero restrictions on the following partitioned  $\mathbf{B}_0$ :

$$(8.45) \quad \mathbf{B}_0 = \begin{bmatrix} b_{11} & 0 & 0 \\ (n_1 \times n_1) & (n_1 \times 1) & (n_1 \times n_2) \\ b_{21} & b_{22} & 0 \\ (1 \times n_1) & (1 \times 1) & (1 \times n_2) \\ b_{31} & b_{32} & b_{33} \\ (n_2 \times n_1) & (n_2 \times 1) & (n_2 \times n_2) \end{bmatrix}$$

Two zero restrictions,  $b_{12} = b_{13} = 0$ , are required for the monetary policy shock to be orthogonal to other structural shocks, while the restriction  $b_{23} = 0$  implies the assumption that the Fed does not have information about variables in  $y_{2t}$  when it makes a monetary policy decision.

The following property may help explain the block recursive system:

$$(8.46) \quad \begin{bmatrix} \mathbf{B}_{11} & \mathbf{0} \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{B}_{11}^{-1} & \mathbf{0} \\ -\mathbf{B}_{22}^{-1}\mathbf{B}_{21}\mathbf{B}_{11}^{-1} & \mathbf{B}_{22}^{-1} \end{bmatrix}$$

The block recursive structure gives sufficient conditions to identify a monetary policy shock, and the ordering within  $y_{1t}$  and  $y_{2t}$  does not affect the results if one is interested in the effects of a monetary policy shock. Instead, the ordering across two groups might affect the results substantially.

### 8.6.3 Two-step ML estimation

When  $\mathbf{B}_0$  is not lower triangular, maximum likelihood estimation or GMM estimation can be used once the structural model is identified as discussed in the following section. As VAR models involve a large number of parameters, two-step estimation is often used. The reduced form VAR model is estimated in the first step, and ML or GMM estimation is used in the second step focusing on the relation of  $\mathbf{B}_0 \hat{\Sigma}_\epsilon \mathbf{B}'_0 = \mathbf{\Lambda}$  to

estimate  $\mathbf{B}_0$  and  $\mathbf{\Lambda}$  from the first step estimate of  $\mathbf{\Sigma}_\epsilon$ . The two-step ML estimation is discussed by Giannini (1992) in detail, while two-step GMM estimation is used by Bernanke and Mihov (1998).

Suppose that the model is identified with short-run economic restrictions:

$$(8.47) \quad \text{vec}(\mathbf{B}_0) = \mathbf{S}_b \mathbf{b}_s + \mathbf{s}_b,$$

where  $\mathbf{b}_s$  be a  $n_s$  ( $\leq \frac{n(n+1)}{2}$ ) dimensional vector of free parameters in  $\mathbf{B}_0$ , and the restrictions are expressed by an  $n^2 \times n_s$  matrix of  $\mathbf{S}$  and  $n^2 \times 1$  vector of  $\mathbf{s}_b$ .

Then the following are used in the second step for ML estimation:

(a) Likelihood function:

$$(8.48) \quad L(\mathbf{B}_0) = T \log |\mathbf{B}_0| - \frac{T}{2} \text{trace}(\mathbf{B}'_0 \mathbf{B}_0 \hat{\mathbf{\Sigma}})$$

(b) Gradient:

$$(8.49) \quad \mathbf{g}(\mathbf{B}_0) = T[\text{vec}(\mathbf{B}'_0{}^{-1}) - (\hat{\mathbf{\Sigma}} \otimes \mathbf{I}_{n_2}) \text{vec}(\mathbf{B}_0)]$$

(c) Information matrix:

$$(8.50) \quad \mathbf{I}_T(\mathbf{B}_0) = 2T(\mathbf{B}_0^{-1} \otimes \mathbf{I}_{n_2}) \mathbf{N}_{n_2} (\mathbf{B}'_0{}^{-1} \otimes \mathbf{I}_{n_2})$$

(d) Score algorithm:

$$(8.51) \quad \mathbf{b}_{s,i+1} = \mathbf{b}_{s,i} + [\mathbf{I}_T(\mathbf{b}_{s,i})]^{-1} \mathbf{g}(\mathbf{b}_{s,i}),$$

where  $\mathbf{g}(\mathbf{b}_s) = \mathbf{S}'_b \mathbf{g}(\mathbf{B}_0)$ ,  $\mathbf{I}_T(\mathbf{b}_s) = \mathbf{S}'_b \mathbf{I}_T(\mathbf{B}_0) \mathbf{S}_b$ , and  $i$  denotes the iteration step.

In addition, if  $\mathbf{B}_0$  is over-identified, the over-identifying restrictions can be tested using

$$(8.52) \quad LRT = 2(L(\hat{\mathbf{\Sigma}}) - L(\hat{\mathbf{B}}_{0,ML})),$$

where  $L(\hat{\mathbf{\Sigma}}) = -\frac{T}{2} \log |\hat{\mathbf{\Sigma}}| - \frac{nT}{2}$ , and LRT is asymptotically  $\chi^2_{(q)}$ -distributed, where  $q$  is the number of over-identification.

## Appendix

This appendix provides three traditional methods of computing confidence intervals of impulse responses, which are widely used as standard tools for economic analysis in the applied VAR literature (see, e.g., Baillie, 1987; Runkle, 1987).

### 8.A Asymptotic Interval Method

Let  $\boldsymbol{\theta} = (\mathbf{a}', \boldsymbol{\sigma}')'$ , where  $\mathbf{a} = \text{vec}(\mathbf{A}_1, \mathbf{A}_2, \dots, \mathbf{A}_p)$  and  $\boldsymbol{\sigma} = \text{vech}(\boldsymbol{\Sigma})$ .<sup>6</sup> It is well known that  $\boldsymbol{\theta}$  is asymptotically normally distributed

$$\sqrt{T}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma}_\theta),$$

where

$$\boldsymbol{\Sigma}_\theta = \begin{bmatrix} \boldsymbol{\Sigma}_a & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_\sigma \end{bmatrix} = \begin{bmatrix} [E(\mathbf{x}_t \mathbf{x}_t')]^{-1} \otimes \boldsymbol{\Sigma} & \mathbf{0} \\ \mathbf{0} & 2\mathbf{D}_n^+(\boldsymbol{\Sigma} \otimes \boldsymbol{\Sigma})\mathbf{D}_n^{+'} \end{bmatrix},$$

$\mathbf{x}_t = [\mathbf{y}'_{t-1}, \mathbf{y}'_{t-2}, \dots, \mathbf{y}'_{t-p}]'$ , and  $\mathbf{D}_n^+$  is the Moore-Penrose inverse of  $\mathbf{D}_n$ . Refer to Hamilton (1994) for its derivation and extended discussion.

In addition to impulse responses derived in the text, it is often of interest to trace the accumulated responses

$$\boldsymbol{\Psi}_{ci} = \sum_{j=0}^i \boldsymbol{\Psi}_j, \quad \boldsymbol{\Phi}_{ci} = \boldsymbol{\Psi}_{ci} \boldsymbol{\Phi}_0$$

and the total accumulated responses

$$\boldsymbol{\Psi}(1) = \sum_{j=0}^{\infty} \boldsymbol{\Psi}_j = \mathbf{A}(1)^{-1}, \quad \boldsymbol{\Phi}(1) = \boldsymbol{\Psi}(1) \boldsymbol{\Phi}_0.$$

---

<sup>6</sup>Note that we define  $\mathbf{a}$  slightly differently from Section 8.1 which includes the constant term  $\boldsymbol{\delta}_\epsilon$ .

Let  $\boldsymbol{\ell}_p$  be the  $p$ -dimensional vector with ones and denote

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_1 & \mathbf{A}_2 & \cdots & \mathbf{A}_{p-1} & \mathbf{A}_p \\ \mathbf{I}_n & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_n & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{I}_n & \mathbf{0} \end{bmatrix} \text{ and } \mathbf{J}_{np} = \begin{bmatrix} \mathbf{I}_n & \vdots & \mathbf{0}_{n \times n(p-1)} \end{bmatrix}.$$

Consider a VAR model with short-run restrictions of the form  $\text{vec}(\mathbf{B}_0) = \mathbf{S}_b \mathbf{b}_s + \mathbf{s}_b$  and define  $\mathbf{G}_{\phi\sigma} = \mathbf{G}_{\phi b} \mathbf{G}_{\phi\phi b}^+$  where  $\mathbf{G}_{\phi b} = \begin{bmatrix} -\mathbf{B}_0'^{-1} \otimes \mathbf{B}_0^{-1} & \vdots & \mathbf{I}_n \otimes \mathbf{B}_0^{-1} \end{bmatrix} \mathbf{S}_b$  and  $\mathbf{G}_{\phi\phi b} = 2\mathbf{D}_n^+ \begin{bmatrix} -\mathbf{B}_0^{-1} \mathbf{B}_0'^{-1} \otimes \mathbf{B}_0^{-1} & \vdots & \mathbf{B}_0^{-1} \otimes \mathbf{B}_0^{-1} \end{bmatrix} \mathbf{S}_b$ . With this notation, we obtain the asymptotic distributions of the impulse responses in the next proposition. See Lütkepohl (1990) for just-identified recursive VARs and Jang (2004) for more generalized VARs including non-recursive and over-identified models.

**Proposition 8.A.1** *Suppose  $\sqrt{T}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma}_\theta)$  and  $\text{vec}(\mathbf{B}_0) = \mathbf{S}_b \mathbf{b}_s + \mathbf{s}_b$ .*

*Then*

$$(a) \quad \sqrt{T} \text{vec}(\hat{\boldsymbol{\Psi}}_i - \boldsymbol{\Psi}_i) \xrightarrow{d} N(\mathbf{0}, \mathbf{G}_{\Psi ai} \boldsymbol{\Sigma}_a \mathbf{G}'_{\Psi ai}), \quad i = 1, 2, \dots,$$

*where*

$$\mathbf{G}_{\Psi ai} = \frac{\partial \text{vec}(\boldsymbol{\Psi}_i)}{\partial \mathbf{a}'} = \sum_{j=0}^{i-1} \mathbf{J}_{np} (\mathbf{A}')^{i-1-j} \otimes \boldsymbol{\Psi}_j;$$

$$(b) \quad \sqrt{T} \text{vec}(\hat{\boldsymbol{\Psi}}_{ci} - \boldsymbol{\Psi}_{ci}) \xrightarrow{d} N(\mathbf{0}, \mathbf{G}_{\Psi cai} \boldsymbol{\Sigma}_a \mathbf{G}'_{\Psi cai}), \quad i = 1, 2, \dots,$$

*where*

$$\mathbf{G}_{\Psi cai} = \frac{\partial \text{vec}(\boldsymbol{\Psi}_{ci})}{\partial \mathbf{a}'} = \sum_{j=0}^i \mathbf{G}_{\Psi aj};$$

$$(c) \quad \sqrt{T} \text{vec}(\hat{\boldsymbol{\Psi}}(1) - \boldsymbol{\Psi}(1)) \xrightarrow{d} N(\mathbf{0}, \mathbf{G}_{\Psi 1a} \boldsymbol{\Sigma}_a \mathbf{G}'_{\Psi 1a})$$

*where*

$$\mathbf{G}_{\Psi 1a} = \frac{\partial \text{vec}(\boldsymbol{\Psi}(1))}{\partial \mathbf{a}'} = \boldsymbol{\ell}_p' \otimes \boldsymbol{\Psi}(1)' \otimes \boldsymbol{\Psi}(1);$$

$$(d) \sqrt{T} \text{vec}(\hat{\Phi}_i - \Phi_i) \xrightarrow{d} N(\mathbf{0}, \mathbf{G}_{\Phi_{ai}} \Sigma_a \mathbf{G}'_{\Phi_{ai}} + \mathbf{G}_{\Phi_{\sigma i}} \Sigma_{\sigma} \mathbf{G}'_{\Phi_{\sigma i}}), \quad i = 0, 1, 2, \dots,$$

where

$$\begin{aligned} \mathbf{G}_{\Phi_{ai}} &= \frac{\partial \text{vec}(\Phi_i)}{\partial \mathbf{a}'} = \begin{cases} \mathbf{0}, & i = 0 \\ (\Phi'_0 \otimes \mathbf{I}_n) \mathbf{G}_{\Psi_{ai}}, & i = 1, 2, \dots \end{cases} \quad \text{and} \\ \mathbf{G}_{\Phi_{\sigma i}} &= \frac{\partial \text{vec}(\Phi_i)}{\partial \boldsymbol{\sigma}'} = (\mathbf{I}_{n_2} \otimes \Psi_i) \mathbf{G}_{\phi\sigma}; \end{aligned}$$

$$(e) \sqrt{T} \text{vec}(\hat{\Phi}_{ci} - \Phi_{ci}) \xrightarrow{d} N(\mathbf{0}, \mathbf{G}_{\Phi_{cai}} \Sigma_a \mathbf{G}'_{\Phi_{cai}} + \mathbf{G}_{\Phi_{\sigma ci}} \Sigma_{\sigma} \mathbf{G}'_{\Phi_{\sigma ci}}), \quad i = 0, 1, 2, \dots,$$

where

$$\begin{aligned} \mathbf{G}_{\Phi_{cai}} &= \frac{\partial \text{vec}(\Phi_{ci})}{\partial \mathbf{a}'} = \sum_{j=0}^i \mathbf{G}_{\Phi_{aj}} \quad \text{and} \\ \mathbf{G}_{\Phi_{\sigma ci}} &= \frac{\partial \text{vec}(\Phi_{ci})}{\partial \boldsymbol{\sigma}'} = \sum_{j=0}^i \mathbf{G}_{\Phi_{\sigma j}}; \end{aligned}$$

$$(f) \quad \sqrt{T} \text{vec}(\hat{\Phi}(1) - \Phi(1)) \xrightarrow{d} N(\mathbf{0}, \mathbf{G}_{\Phi_{1a}} \Sigma_a \mathbf{G}'_{\Phi_{1a}} + \mathbf{G}_{\Phi_{1\sigma}} \Sigma_{\sigma} \mathbf{G}'_{\Phi_{1\sigma}}),$$

where

$$\begin{aligned} \mathbf{G}_{\Phi_{1a}} &= \frac{\partial \text{vec}(\Phi_i)}{\partial \mathbf{a}'} = (\Phi'_0 \otimes \mathbf{I}_n) \mathbf{G}_{\Psi_{1a}} \quad \text{and} \\ \mathbf{G}_{\Phi_{1\sigma}} &= \frac{\partial \text{vec}(\Phi_i)}{\partial \boldsymbol{\sigma}'} = (\mathbf{I}_{n_2} \otimes \Psi(1)) \mathbf{G}_{\phi\sigma}. \end{aligned}$$

**Proof** (a)–(c) See Lütkepohl (1990) Proposition 1.

(d)–(f) See Jang (2004) Theorem 3.2.

## 8.B Bias-Corrected Bootstrap Method

Kilian (1998) suggests the following algorithm for the bias-corrected bootstrap (bootstrap after bootstrap) method:

1. Estimate the  $\text{VAR}(p)$  in equation (8.2) and generate 1000 bootstrap replications

$\hat{\mathbf{a}}^*$  from

$$\hat{\mathbf{A}}(L) \mathbf{y}_t^* = \hat{\boldsymbol{\delta}}_{\epsilon} + \boldsymbol{\epsilon}_t^*,$$

using standard nonparametric bootstrap techniques.

2. Approximate the bias term  $\boldsymbol{\lambda} = E(\hat{\mathbf{a}} - \mathbf{a})$  by  $\boldsymbol{\lambda}^* = E^*(\hat{\mathbf{a}}^* - \hat{\mathbf{a}})$ , which suggests  $\hat{\boldsymbol{\lambda}} = \bar{\mathbf{a}}^* - \hat{\mathbf{a}}$  for the bias estimate where  $\bar{\mathbf{a}}^*$  is the mean of the bootstrap sample of  $\hat{\mathbf{a}}^*$ .
3. Adjust  $\hat{\mathbf{a}}$  for stationarity correction to avoid pushing stationary impulse responses into the nonstationary region.
  - (i) Compute  $m(\hat{\mathbf{a}})$ , the modulus of the largest root of the companion matrix associated with  $\hat{\mathbf{a}}$ .
  - (ii) If  $m(\hat{\mathbf{a}}) \geq 1$ , set  $\tilde{\mathbf{a}} = \hat{\mathbf{a}}$  without any adjustments.
  - (iii) Otherwise, construct the bias-corrected coefficient estimate  $\tilde{\mathbf{a}} = \hat{\mathbf{a}} - \hat{\boldsymbol{\lambda}}$ . If  $m(\tilde{\mathbf{a}}) \geq 1$ , let  $\hat{\boldsymbol{\lambda}}_1 = \hat{\boldsymbol{\lambda}}$  and  $\nu_1 = 1$ . Define  $\hat{\boldsymbol{\lambda}}_{j+1} = \nu_j \hat{\boldsymbol{\lambda}}_j$  and  $\nu_{j+1} = \nu_j - 0.01$ . Set  $\tilde{\mathbf{a}} = \tilde{\mathbf{a}}_j$  after iterating on  $\tilde{\mathbf{a}}_j = \hat{\mathbf{a}} - \hat{\boldsymbol{\lambda}}_j$  for  $j = 1, 2, \dots$  until  $m(\tilde{\mathbf{a}}) < 1$ .
4. Substitute  $\tilde{\mathbf{a}}$  for  $\hat{\mathbf{a}}$  and generate 2000 new bootstrap replications  $\hat{\mathbf{a}}^*$  from

$$\tilde{\mathbf{A}}(L)\mathbf{y}_t^* = \tilde{\boldsymbol{\delta}}_\epsilon + \boldsymbol{\epsilon}_t^*,$$

using standard nonparametric bootstrap techniques.

5. Compute  $\tilde{\mathbf{a}}^*$  from  $\hat{\mathbf{a}}^*$  and  $\hat{\boldsymbol{\lambda}}^*$  with the adjustment of  $\hat{\mathbf{a}}^*$  for stationarity correction as described in Step 3.
6. Compute the  $\alpha$  and  $1 - \alpha$  percentile intervals of impulse responses generated with  $\tilde{\mathbf{a}}^*$  and  $\hat{\boldsymbol{\sigma}}^*$ .

## 8.C Monte Carlo Integration

Consider the VAR system in the form of (8.5). Assuming that  $u_t$  is i.i.d. and normally distributed, Zellner (1971) finds that  $\Sigma_\epsilon$  follows the Normal-inverse Wishart posterior distribution, with the prior,  $f(\mathbf{b}, \Sigma_\epsilon) \sim |\Sigma_\epsilon|^{-\frac{n+1}{2}}$ :

$$(8.C.1) \quad \Sigma_\epsilon^{-1} \sim \text{Wishart}((T\hat{\Sigma}_\epsilon)^{-1}, T) \quad \text{with given } \hat{\Sigma}_\epsilon$$

and

$$(8.C.2) \quad \mathbf{b} \sim N(\hat{\mathbf{b}}, \Sigma_\epsilon \otimes (\mathbf{X}'\mathbf{X})^{-1}).$$

Doan (1992) and Sims and Zha (1999) suggest the following parametric Monte Carlo integration method for computing impulse responses:

1. Estimate (16.17) and let  $\hat{\mathbf{b}}$  and  $\hat{\Sigma}$  be these estimates.
2. Let  $\mathbf{A}$  be a lower triangular matrix of Choleski decomposition of  $(\mathbf{X}'\mathbf{X})^{-1}$ .
3. Let  $\mathbf{S}^{-1}$  be a lower triangular matrix of Choleski decomposition of  $\hat{\Sigma}_\epsilon^{-1}$ .
4. Generate  $n \times T$  random numbers,  $\mathbf{w}_b$ , from the normal distribution,  $N(0, \frac{1}{T})$ .
5. Generate  $(n(p-1) + r + 1) \times n$  random numbers,  $\mathbf{u}_b$ , from the standard normal distribution,  $N(0, 1)$ .
6. Let  $\mathbf{r}_b = \mathbf{w}_b' \mathbf{S}^{-1}$ , and get  $\Sigma_b^{-1} = \mathbf{r}_b' \mathbf{r}_b$ .
7. Let  $\mathbf{S}_b$  be a lower triangular matrix of Choleski decomposition of  $\Sigma_b$ .
8. Let  $\mathbf{b} = \hat{\mathbf{b}} + \mathbf{e}_b$ , in which  $\mathbf{e}_b = \mathbf{A} \mathbf{u}_b \mathbf{S}_b'$ . Then,  $\mathbf{b} \sim N(\hat{\mathbf{b}}, \Sigma_b \otimes (\mathbf{X}'\mathbf{X})^{-1})$ .
9. Draw impulse responses,  $\mathbf{ir}_b$ , as described in Section 16.3.3.

10. Repeat 4 ~ 9,  $B$  times, and calculate 95% upper and lower bands of impulse responses using

$$(8.C.3) \quad Upper = \frac{1}{B} \sum_{b=1}^B \mathbf{ir}_b + 2 \left( \frac{1}{B} \sum_{b=1}^B \mathbf{ir}_b^2 - \left( \frac{1}{B} \sum_{b=1}^B \mathbf{ir}_b \right)^2 \right)^{\frac{1}{2}}$$

and

$$(8.C.4) \quad Lower = \frac{1}{B} \sum_{b=1}^B \mathbf{ir}_b - 2 \left( \frac{1}{B} \sum_{b=1}^B \mathbf{ir}_b^2 - \left( \frac{1}{B} \sum_{b=1}^B \mathbf{ir}_b \right)^2 \right)^{\frac{1}{2}}.$$

## Exercises

**8.1** Let  $y_t$  and  $m_t$  be detrended log GDP and log money supply, respectively. Assume that  $\mathbf{z}_t = (y_t, m_t)'$  is a covariance stationary process with a  $p$ -th order VAR representation.

- (a) Define the concept, “ $y$  fails to Granger-cause  $m$ ”.
- (b) How do you test the hypothesis that log GDP fails to Granger-cause log money supply?
- (c) Imagine that you find empirical evidence that  $y$  fails to Granger-cause  $m$ , and  $m$  Granger-causes  $y$ . Discuss why this evidence can be consistent with a model in which money is neutral in the short run (money is neutral when changes in the level of money supply cannot affect any real economic variable such as real GDP).
- (d) Define the orthogonalized impulse response function. Let

$$(8.E.1) \quad \mathbf{B}_0 \mathbf{z}_t = \boldsymbol{\delta} + \mathbf{B}_1 \mathbf{z}_{t-1} + \mathbf{B}_2 \mathbf{z}_{t-2} + \cdots + \mathbf{B}_p \mathbf{z}_{t-p} + \mathbf{e}_t$$

be a structural model for  $\mathbf{z}_t$ , where  $\mathbf{B}_i$  is a  $n \times n$  matrix, and  $\boldsymbol{\delta}$  is a  $n \times 1$  vector. Here  $\mathbf{B}_0$  is a nonsingular matrix of real numbers with 1's along its principal diagonal, and  $\mathbf{e}_t$  is a stationary  $n$ -dimensional vector of normally distributed i.i.d. random variables. Discuss conditions for  $\mathbf{B}_0$  under which the orthogonalized impulse response function represents the effects of each element of  $\mathbf{e}_t$  on  $\mathbf{z}_{t+s}$ .

**8.2** True or False. Briefly explain your answers.

- (a) OLS estimation is equivalent to SUR estimation for a reduced-form VAR model because the regressors are identical.
- (b) OLS estimation is equivalent to SUR estimation for a structural-form VAR model because structural disturbances are uncorrelated.
- (c) In a recursive VAR model,  $e_1 = \epsilon_1$ .
- (d) In a recursive VAR model, impulse responses to  $e_1$  are the same as those of  $\epsilon_1$ .
- (e) In a recursive VAR model,  $e_n = \epsilon_n$ .
- (f) In a recursive VAR model, impulse responses to  $e_n$  are the same as those of  $\epsilon_n$ .

### References

- BAILLIE, R. T. (1987): "Inference in Dynamic Models Containing 'Surprise' Variables," *Journal of Econometrics*, 35, 101–117.
- BERNANKE, B. S. (1986): "Alternative Explanations of the Money-Income Correlation," *Carnegie-Rochester Conference Series on Public Policy*, 25, 49–100.
- BERNANKE, B. S., AND I. MIHOV (1998): "Measuring Monetary Policy," *Quarterly Journal of Economics*, 113(3), 869–902.
- BLANCHARD, O. J., AND M. W. WATSON (1986): "Are Business Cycles All Alike?," in *The American Business Cycle: Continuity and Change*, ed. by R. J. Gordon, vol. 25 of *National Bureau of Economic Research Studies in Business Cycles*, pp. 123–182. University of Chicago Press, Chicago.

- CHRISTIANO, L. J., M. EICHENBAUM, AND C. L. EVANS (1999): "Monetary Policy Shocks: What Have We Learned and to What End?," in *Handbook of Macroeconomics*, ed. by J. Taylor, and M. Woodford, vol. 1, chap. 2, pp. 65–148. Elsevier Science.
- DOAN, T. A. (1992): *RATS User's Manual, Version 4*. Estima, Evanston, IL.
- GIANNINI, C. (1992): *Topics in Structural VAR Econometrics*. Springer Verlag, New York.
- GORDON, D. B., AND E. M. LEEPER (1994): "The Dynamic Impacts of Monetary Policy: An Exercise in Tentative Identification," *Journal of Political Economy*, 102(6), 1228–1247.
- HAMILTON, J. D. (1994): *Time Series Analysis*. Princeton University Press, Princeton.
- JANG, K. (2004): "Generalized Two-Step Maximum Likelihood Estimation of Structural Vector Autoregressive Models Partially Identified with Short-Run restrictions," Manuscript.
- KILIAN, L. (1998): "Small-Sample Confidence Intervals for Impulse Response Functions," *Review of Economics and Statistics*, 80(2), 218–230.
- LEAMER, E. (1985): "Vector Autoregressions for Causal Inference," in *Carnegie-Rochester Conference Series On Public Policy: Understanding Monetary Regimes*, ed. by K. Brunner, and A. H. Meltzer, vol. 22, pp. 255–304. Elsevier.
- LÜTKEPOHL, H. (1990): "Asymptotic Distributions of Impulse Response Functions and Forecast Error Variance Decompositions of Vector Autoregressive Models," *Review of Economics and Statistics*, 72(1), 116–125.
- ROTEMBERG, J. J., AND M. WOODFORD (1999): "Interest Rate Rules in an Estimated Sticky Price Model," in *Monetary Policy Rules*, ed. by J. B. Taylor, vol. 31 of *NBER-Business Cycles Series*, chap. 2, pp. 57–126. The University of Chicago Press, Chicago, With Comment by Martin Feldstein.
- RUNKLE, D. E. (1987): "Vector Autoregressions and Reality," *Journal of Business and Economic Statistics*, 5, 437–442.
- SIMS, C. A. (1980): "Macroeconomics and Reality," *Econometrica*, 48, 1–48.
- SIMS, C. A., AND T. ZHA (1999): "Error Bands for Impulse Responses," *Econometrica*, 67(5), 1113–1156.
- STOCK, J. H., AND M. W. WATSON (1989): "New Indexes of Leading and Coincident Economic Indicators," in *NBER Macroeconomics Annual*, pp. 351–394.
- TAYLOR, J. B. (1993): "Discretion versus Policy Rules in Practice," *Carnegie-Rochester Conference Series on Public Policy*, 39, 195–214.
- ZELLNER, A. (1971): *An Introduction to Bayesian Inference in Econometrics*. Wiley, New York.

# Chapter 9

## GENERALIZED METHOD OF MOMENTS

### 9.1 Asymptotic Properties of GMM Estimators

#### 9.1.1 Moment Restriction and GMM Estimators

To motivate GMM estimation, consider Hansen and Singleton's (1982) Consumption-Based Capital Asset Pricing Model (C-CAPM). A representative agent maximizes

$$(9.1) \quad \sum_{t=1}^{\infty} \beta^t E(U(c_t) | I_0)$$

subject to a budget constraint. Hansen and Singleton (1982) use an isoelastic intraperiod utility function

$$(9.2) \quad U(c_t) = \frac{1}{1-\alpha} (c_t^{1-\alpha} - 1),$$

where  $c_t$  is real consumption at date  $t$ ,  $\beta$  is a discount factor and  $\alpha > 0$  is the reciprocal of the intertemporal elasticity of substitution ( $\alpha$  is also the relative risk aversion coefficient for consumption in this model). The standard Euler equation for the optimization problem is

$$(9.3) \quad \frac{E[\beta c_{t+1}^{-\alpha} R_{t+1} | I_t]}{c_t^{-\alpha}} = 1,$$

where  $R_{t+1}$  is the gross real return of an asset and  $I_t$  is an information set available at time  $t$ . This Euler equation can be rearranged as

$$(9.4) \quad E[\beta(\frac{C_{t+1}}{C_t})^{-\alpha} R_{t+1} - 1 | I_t] = 0.$$

Let  $\mathbf{z}_t$  be a vector of variables whose values are known at time  $t$ . Then  $\mathbf{z}_t \in I_t$  and

$$(9.5) \quad E[\mathbf{z}_t \{ \beta(\frac{C_{t+1}}{C_t})^{-\alpha} R_{t+1} - 1 \} | I_t] = \mathbf{0}.$$

By the law of iterative expectations, we obtain the orthogonality conditions to be used in GMM estimation,

$$(9.6) \quad E[\mathbf{z}_t \{ \beta(\frac{C_{t+1}}{C_t})^{-\alpha} R_{t+1} - 1 \}] = \mathbf{0}.$$

Let  $\{\mathbf{x}_t : t = 1, 2, \dots\}$  be a stationary and ergodic vector stochastic process,  $\mathbf{b}_0$  be a  $p$ -dimensional vector of the parameters to be estimated, and  $f(\mathbf{x}_t, \mathbf{b})$  a  $q$ -dimensional vector of functions. We refer to  $\mathbf{u}_t = f(\mathbf{x}_t, \mathbf{b}_0)$  as the disturbance of GMM. Consider the (unconditional) moment restrictions

$$(9.7) \quad E(f(\mathbf{x}_t, \mathbf{b}_0)) = \mathbf{0}.$$

For example, in the Hansen and Singleton (1982) case,  $\mathbf{x}_t = (\frac{C_{t+1}}{C_t}, R_{t+1}, \mathbf{z}_t)'$ ,  $\mathbf{b}_0 = (\beta, \alpha)'$ , and  $f(\mathbf{x}_t, \mathbf{b}_0) = \mathbf{z}_t \{ \beta(\frac{C_{t+1}}{C_t})^{-\alpha} R_{t+1} - 1 \}$ .

Suppose that a law of large numbers can be applied to  $f(\mathbf{x}_t, \mathbf{b})$  for all admissible  $\mathbf{b}$ , so that the sample mean of  $f(\mathbf{x}_t, \mathbf{b})$  converges to its population mean:

$$(9.8) \quad \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T f(\mathbf{x}_t, \mathbf{b}) = E(f(\mathbf{x}_t, \mathbf{b}))$$

with probability one (or, in other words, almost surely). The basic idea of GMM estimation is to mimic the moment restrictions in (9.7) by minimizing a quadratic

form of the sample means

$$(9.9) \quad J_T(\mathbf{b}) = \left\{ \frac{1}{T} \sum_{t=1}^T f(\mathbf{x}_t, \mathbf{b}) \right\}' \mathbf{W}_T \left\{ \frac{1}{T} \sum_{t=1}^T f(\mathbf{x}_t, \mathbf{b}) \right\}$$

with respect to  $\mathbf{b}$ , where  $\mathbf{W}_T$  is a positive semidefinite matrix that satisfies

$$(9.10) \quad \lim_{T \rightarrow \infty} \mathbf{W}_T = \mathbf{W}_0$$

with probability one for a positive definite matrix  $\mathbf{W}_0$ . The matrices  $\mathbf{W}_T$  and  $\mathbf{W}_0$  are both referred to as the distance or weighting matrix. The GMM estimator,  $\mathbf{b}_T$ , is the solution of the minimization problem in (9.9). Under fairly general regularity conditions, the GMM estimator  $\mathbf{b}_T$  is a consistent estimator for arbitrary distance matrices.<sup>1</sup> The selection of the distance matrix which yields an (asymptotically) efficient GMM estimator is discussed below in Section 9.1.3.

### 9.1.2 Asymptotic Distributions of GMM Estimators

Suppose that a central limit theorem applies to the disturbance of GMM,  $\mathbf{u}_t = f(\mathbf{x}_t, \mathbf{b}_0)$ , so that  $\frac{1}{\sqrt{T}} \sum_{t=1}^T \mathbf{u}_t$  has an (asymptotic) normal distribution with mean zero and the covariance matrix  $\mathbf{\Omega}$  in large samples.<sup>2</sup> If  $\mathbf{u}_t$  is serially uncorrelated,  $\mathbf{\Omega} = E(\mathbf{u}_t \mathbf{u}_t')$ . If  $\mathbf{u}_t$  is serially correlated,

$$(9.11) \quad \mathbf{\Omega} = \lim_{j \rightarrow \infty} \sum_{-j}^j E(\mathbf{u}_t \mathbf{u}_{t-j}')$$

Some authors refer to  $\mathbf{\Omega}$  as the long run covariance matrix of  $\mathbf{u}_t$ . Let  $\mathbf{\Gamma} = E\left(\frac{\partial f(\mathbf{x}_t, \mathbf{b}_0)}{\partial \mathbf{b}'}\right)$  be the expectation of the  $q \times p$  matrix of the derivatives of  $f(\mathbf{x}_t, \mathbf{b}_0)$  with respect to  $\mathbf{b}$  and assume that  $\mathbf{\Gamma}$  has full column rank. Under suitable regularity conditions,

<sup>1</sup>Some regularity conditions that are important for applied researchers will be discussed in Section 9.3

<sup>2</sup>An advantage of the GMM estimation is that a strong distributional assumption such that  $\mathbf{u}_t$  is normally distributed is not necessary.

$\sqrt{T}(\mathbf{b}_T - \mathbf{b}_0)$  converges in distribution to a normal distribution with mean zero and the covariance matrix

$$(9.12) \quad Cov(\mathbf{W}_0) = (\mathbf{\Gamma}'\mathbf{W}_0\mathbf{\Gamma})^{-1}(\mathbf{\Gamma}'\mathbf{W}_0\mathbf{\Omega}\mathbf{W}_0\mathbf{\Gamma})(\mathbf{\Gamma}'\mathbf{W}_0\mathbf{\Gamma})^{-1}.$$

### 9.1.3 Optimal Choice of the Distance Matrix

When the number of moment conditions ( $q$ ) is equal to the number of parameters to be estimated ( $p$ ), the system is just identified. In the case of a just identified system, the GMM estimator does not depend on the choice of distance matrix. When  $q > p$ , there exist overidentifying restrictions and different GMM estimators are obtained for different distance matrices. In this case, one may choose the distance matrix that results in an (asymptotically) efficient GMM estimator. Hansen (1982) shows that the covariance matrix (9.12) is minimized when  $\mathbf{W}_0 = \mathbf{\Omega}^{-1}$ .<sup>3</sup> With this choice of the distance matrix,  $\sqrt{T}(\mathbf{b}_T - \mathbf{b}_0)$  has an approximately normal distribution with mean zero and the covariance matrix

$$(9.13) \quad Cov(\mathbf{\Omega}^{-1}) = (\mathbf{\Gamma}'\mathbf{\Omega}^{-1}\mathbf{\Gamma})^{-1}$$

in large samples.

Let  $\mathbf{\Omega}_T$  be a consistent estimator of  $\mathbf{\Omega}$ . Then  $\mathbf{W}_T = \mathbf{\Omega}_T^{-1}$  is used to obtain  $\mathbf{b}_T$ . The resulting estimator is called the optimal or efficient GMM estimator. It should be noted, however, that it is optimal given  $f(\mathbf{x}_t, \mathbf{b})$ . In the context of instrumental variable estimation, this means that instrumental variables are given. The optimal selection of instrumental variables is discussed below in Section 9.7. Let  $\mathbf{\Gamma}_T$  be a consistent estimator of  $\mathbf{\Gamma}$ . Then the standard errors of the optimal GMM estimator

---

<sup>3</sup>The covariance matrix is minimized in the sense that  $Cov(\mathbf{W}_0) - Cov(\mathbf{\Omega}^{-1})$  is a positive semidefinite matrix for any positive definite matrix  $\mathbf{W}_0$ .

$\mathbf{b}_T$  are calculated as square roots of the diagonal elements of  $\frac{1}{T}(\mathbf{\Gamma}'_T \mathbf{\Omega}_T^{-1} \mathbf{\Gamma}_T)^{-1}$ . The appropriate method for estimating  $\mathbf{\Omega}$  depends on the model. This problem is discussed in Chapter 6. It is usually easier to estimate  $\mathbf{\Gamma}$  by  $\mathbf{\Gamma}_T = \frac{1}{T} \sum_{t=1}^T \frac{\partial f(\mathbf{x}_t, \mathbf{b}_T)}{\partial \mathbf{b}'}$  than to estimate  $\mathbf{\Omega}$ . In linear models, or in some simple nonlinear models, analytical derivatives are readily available. In nonlinear models, numerical derivatives are often used.

### 9.1.4 A Chi-Square Test for the Overidentifying Restrictions

In the case where there are overidentifying restrictions ( $q > p$ ), a chi-square statistic can be used to test the overidentifying restrictions. One application of this test is to test the validity of the moment conditions implied by Euler equations for optimizing problems of economic agents. This application is discussed in Section 9.5. Hansen (1982) shows that  $T$  times the minimized value of the objective function,  $TJ_T(\mathbf{b}_T)$ , has an (asymptotic) chi-square distribution with  $q-p$  degrees of freedom if  $\mathbf{W}_0 = \mathbf{\Omega}^{-1}$  in large samples. This test is sometimes called Hansen's  $J$  test.<sup>4</sup>

Kyungho  
needs to  
check this!

If we reject the overidentifying restrictions based on Hansen's  $J$  test, it can be interpreted in two different ways. If a model implies the moment restrictions, for example, Euler equation approach, rejection of  $J$  test means that the model is rejected. However, if instrumental variables are chosen with common sense, rejection of  $J$  test means that instrumental variables are inappropriately chosen.

## 9.2 Special Cases

This section shows how linear regressions and nonlinear instrumental variable estimation are embedded in the GMM framework above.

<sup>4</sup>See Newey (1985) for an analysis of the asymptotic power properties of this chi-square test.

### 9.2.1 Ordinary Least Squares

Consider a linear model,

$$(9.14) \quad y_t = \mathbf{x}'_{2t} \mathbf{b}_0 + \epsilon_t,$$

where  $y_t$  and  $\epsilon_t$  are stationary and ergodic random variables,  $\mathbf{x}_{2t}$  is a  $p$ -dimensional stationary and ergodic random vector. OLS estimation can be embedded in the GMM framework by letting  $\mathbf{x}_t = (y_t, \mathbf{x}'_{2t})'$ ,  $f(\mathbf{x}_t, \mathbf{b}) = \mathbf{x}_{2t}(y_t - \mathbf{x}'_{2t} \mathbf{b})$ ,  $\mathbf{u}_t = \mathbf{x}_{2t} \epsilon_t$ , and  $p = q$ . Thus, the moment conditions (9.7) become the orthogonality conditions:

$$(9.15) \quad E(\mathbf{x}_{2t} \epsilon_t) = \mathbf{0}.$$

Since this is the case in a just identified system, the distance matrix  $\mathbf{W}_0$  does not matter. Note that the OLS estimator minimizes  $\sum_{t=1}^T (y_t - \mathbf{x}'_{2t} \mathbf{b})^2$  while the GMM estimator minimizes  $(\sum_{t=1}^T \mathbf{x}_{2t} (y_t - \mathbf{x}'_{2t} \mathbf{b}))' (\sum_{t=1}^T \mathbf{x}_{2t} (y_t - \mathbf{x}'_{2t} \mathbf{b}))$ . In this case, the GMM estimator coincides with the OLS estimator. To see this, note that  $(\sum_{t=1}^T \mathbf{x}_{2t} (y_t - \mathbf{x}'_{2t} \mathbf{b}))' (\sum_{t=1}^T \mathbf{x}_{2t} (y_t - \mathbf{x}'_{2t} \mathbf{b}))$  can be minimized by setting  $\mathbf{b}_T$  so that  $\sum_{t=1}^T f(\mathbf{x}_t, \mathbf{b}) = \mathbf{0}$  in the case of a just identified system. This result implies that  $\sum_{t=1}^T \mathbf{x}_{2t} y_t = (\sum_{t=1}^T \mathbf{x}_{2t} \mathbf{x}'_{2t}) \mathbf{b}_T$ . Thus, as long as  $\sum_{t=1}^T \mathbf{x}_{2t} \mathbf{x}'_{2t}$  is invertible,  $\mathbf{b}_T = (\sum_{t=1}^T \mathbf{x}_{2t} \mathbf{x}'_{2t})^{-1} \sum_{t=1}^T \mathbf{x}_{2t} y_t$ . Hence, the GMM estimator  $\mathbf{b}_T$  coincides with the OLS estimator.

### 9.2.2 Linear Instrumental Variables Regressions

Consider the linear model (9.14) and let  $\mathbf{z}_t$  be a  $q$ -dimensional random vector of instrumental variables. Then instrumental variable regressions are embedded in the GMM framework by letting  $\mathbf{x}_t = (y_t, \mathbf{x}'_{2t}, \mathbf{z}'_t)'$ ,  $f(\mathbf{x}_t, \mathbf{b}) = \mathbf{z}_t (y_t - \mathbf{x}'_{2t} \mathbf{b})$ , and  $\mathbf{u}_t = \mathbf{z}_t \epsilon_t$ .

Thus, the moment conditions become the orthogonality conditions

$$(9.16) \quad E(\mathbf{z}_t \epsilon_t) = \mathbf{0}.$$

In the case of a just identified system ( $q = p$ ), the instrumental variable regression estimator  $(\sum_{t=1}^T \mathbf{z}_t \mathbf{x}'_{2t})^{-1} \sum_{t=1}^T \mathbf{z}_t y_t$  coincides with the GMM estimator. For the case of an overidentified system ( $q > p$ ), the two-stage least-squares estimators and the three-stage least-squares estimators (for multiple regressions) can be interpreted as optimal GMM estimators when  $\epsilon_t$  is serially uncorrelated and conditionally homoskedastic.<sup>5</sup>

### 9.2.3 Linear GMM estimator

Consider the linear regression model (9.14). Let  $\mathbf{z}_t$  be a  $q$ -dimensional random vector of instrumental variables,  $\mathbf{x}_t = (y_t, \mathbf{x}'_{2t}, \mathbf{z}'_t)'$ ,  $f(\mathbf{x}_t, \mathbf{b}) = \mathbf{z}_t(y_t - \mathbf{x}'_{2t} \mathbf{b})$ , and  $\mathbf{u}_t = \mathbf{z}_t \epsilon_t$ .

For the case of an overidentified system ( $q > p$ ), the linear GMM estimator,  $\mathbf{b}_T$ , is the solution of the minimization problem (9.9), where

$$(9.17) \quad \begin{aligned} \frac{1}{T} \sum_{t=1}^T f(\mathbf{x}_t, \mathbf{b}) &= \frac{1}{T} \sum_{t=1}^T \mathbf{z}_t (y_t - \mathbf{x}'_{2t} \mathbf{b}) \\ &= \frac{1}{T} \sum_{t=1}^T \mathbf{z}_t y_t + \frac{1}{T} \sum_{t=1}^T (-\mathbf{z}_t \mathbf{x}'_{2t}) \mathbf{b} \\ &\equiv \mathbf{s}_{zy} + \mathbf{\Gamma}_T \mathbf{b}, \end{aligned}$$

$\mathbf{s}_{zy}$  ( $q \times 1$ ) is the corresponding vector of sample moments of  $E(\mathbf{z}_t y_t)$  and  $\mathbf{\Gamma}_T$  ( $q \times p$ ) is the corresponding vector of sample moments of  $E(\frac{\partial f(\mathbf{x}_t, \mathbf{b}_0)}{\partial \mathbf{b}'})$ . The first order condition for the minimization problem with respect to  $\mathbf{b}$  is

$$(9.18) \quad \mathbf{\Gamma}'_T \mathbf{W}_T \mathbf{\Gamma}_T \mathbf{b} = -\mathbf{\Gamma}'_T \mathbf{W}_T \mathbf{s}_{zy},$$

---

<sup>5</sup>This interpretation can be seen by examining the first order condition for the minimization problem (9.9).

where  $\mathbf{W}_T$  is a  $(q \times q)$  positive semidefinite matrix satisfying equation (11.2). The linear GMM estimator,  $\mathbf{b}_T$ , can be obtained by multiplying both sides by the inverse of  $\mathbf{\Gamma}'_T \mathbf{W}_T \mathbf{\Gamma}_T$ :

$$(9.19) \quad \mathbf{b}_T(\mathbf{W}_T) = -(\mathbf{\Gamma}'_T \mathbf{W}_T \mathbf{\Gamma}_T)^{-1} \mathbf{\Gamma}'_T \mathbf{W}_T \mathbf{s}_{zy}.$$

When there is a system of multiple linear equations, the multiple-equation GMM estimator can be obtained. Moreover, under the assumption of conditional homoskedasticity, the three-stage least-squares estimators can be shown to be a special case of multiple-equation GMM estimators. (For more detailed explanation, see Hayashi, 2000).

#### 9.2.4 Nonlinear Instrumental Variables Estimation

GMM is often used in the context of nonlinear instrumental variable (NLIV) estimation. Chapter 10 presents some examples of applications based on the Euler equation approach. Let  $g(\mathbf{x}_{1t}, \mathbf{b})$  be a  $k$ -dimensional vector of functions and  $\epsilon_t = g(\mathbf{x}_{1t}, \mathbf{b}_0)$ . Suppose that there exist conditional moment restrictions,  $E[\epsilon_t | \mathbf{I}_t] = 0$ . Here it is assumed that  $\mathbf{I}_t \subset \mathbf{I}_{t+1}$  for any  $t$ . Let  $\mathbf{z}_t$  be a  $q \times k$  matrix of random variables that are in the information set  $\mathbf{I}_t$ .<sup>6</sup> By the law of iterative expectations, we obtain the unconditional moment restrictions:

$$(9.20) \quad E[\mathbf{z}_t g(\mathbf{x}_{1t}, \mathbf{b}_0)] = \mathbf{0}.$$

Thus, we let  $\mathbf{x}_t = (\mathbf{x}'_{1t}, \mathbf{z}'_t)'$  and  $f(\mathbf{x}_t, \mathbf{b}) = \mathbf{z}_t g(\mathbf{x}_{1t}, \mathbf{b})$  in this case. Hansen (1982) points out that the NLIV estimators discussed by Amemiya (1974), Jorgenson and

---

<sup>6</sup>In some applications,  $\mathbf{z}_t$  is a function of  $\mathbf{b}$ . This property does not cause any problems as long as the resulting  $f(\mathbf{x}_t, \mathbf{b})$  can be written as a function of  $\mathbf{b}$  and a stationary random vector  $\mathbf{x}_t$ .

Laffont (1974), and Gallant (1977) can be interpreted as optimal GMM estimators when  $\epsilon_t$  is serially uncorrelated and conditionally homoskedastic.

Hansen and Singleton (1982) Consumption-Based Capital Asset Pricing Model (C-CAPM) can be an example of NLIV interpretation of GMM estimation. The Euler equation is

$$(9.21) \quad \frac{E[\beta c_{t+1}^{-\alpha} R_{t+1} | I_t]}{c_t^{-\alpha}} = 1,$$

where  $R_{t+1}$  is the gross real return of any asset.<sup>7</sup> The observed  $c_t$  they use is obviously nonstationary, although the specific form of nonstationarity is not clear (difference stationary or trend stationary, for example). Hansen and Singleton use  $\frac{c_{t+1}}{c_t}$  in their econometric formulation, which is assumed to be stationary.<sup>8</sup> Then we let  $\mathbf{b}_0 = (\beta, \alpha)'$ ,  $\mathbf{x}_{1t} = (\frac{c_{t+1}}{c_t}, R_{t+1})'$ , and  $g(\mathbf{x}_{1t}, \mathbf{b}_0) = \beta(\frac{c_{t+1}}{c_t})^{-\alpha} R_{t+1} - 1$ .<sup>9</sup> Stationary variables in  $I_t$ , such as the lagged values of  $\mathbf{x}_t$ , are used for instrumental variables  $\mathbf{z}_t$ . In this case,  $\mathbf{u}_t$  is in  $I_{t+1}$ , and hence  $\mathbf{u}_t$  is serially uncorrelated.

### 9.3 Important Assumptions

This section discusses two assumptions under which large sample properties of GMM estimators are derived. These two assumptions are important in the sense that applied researchers have encountered cases where, unless special care is taken, these assumptions are obviously violated.

---

<sup>7</sup>This asset pricing equation can be applied to any asset returns. For example, Mark (1985) applies the Hansen-Singleton model in asset returns in foreign exchange markets.

<sup>8</sup>In the following, assumptions about trend properties of equilibrium consumption are made. The simplest model in which these assumptions are satisfied is a pure exchange economy, with the trend assumptions imposed on endowments.

<sup>9</sup>When multiple asset returns are used,  $g(\mathbf{x}_t, \mathbf{b})$  becomes a vector of functions.

### 9.3.1 Stationarity

In Hansen (1982),  $\mathbf{x}_t$  is assumed to be (strictly) stationary. Among other things, this assumption implies that when they exist, the unconditional moments  $E(\mathbf{x}_t)$  and  $E(\mathbf{x}_t \mathbf{x}'_{t+\tau})$  cannot depend on  $t$  for any  $\tau$ . Thus, this assumption rules out deterministic trends, autoregressive unit roots, and unconditional heteroskedasticity. On the other hand, conditional moments  $E(\mathbf{x}_{t+\tau} | I_t)$  and  $E(\mathbf{x}_{t+\tau} \mathbf{x}'_{t+\tau+s} | I_t)$  can depend on  $I_t$ . Thus, the stationarity assumption does *not* rule out the possibility that  $\mathbf{x}_t$  has conditional heteroskedasticity. It should be noted that it is not enough for  $\mathbf{u}_t = f(\mathbf{x}_t, \mathbf{b}_0)$  to be stationary. It is required that  $\mathbf{x}_t$  is stationary, so that  $f(\mathbf{x}_t, \mathbf{b})$  is stationary for all admissible  $\mathbf{b}$ , not just for  $\mathbf{b} = \mathbf{b}_0$  (see Section ?????????? for an example in which  $f(\mathbf{x}_t, \mathbf{b}_0)$  is stationary but  $f(\mathbf{x}_t, \mathbf{b})$  is not for other values of  $\mathbf{b}$ ).

Masao  
needs to  
check this!

Gallant (1987) and Gallant and White (1988) show that the GMM strict stationarity assumption can be relaxed to allow for unconditional heteroskedasticity. This property does *not* mean that  $\mathbf{x}_t$  can exhibit nonstationarity by having deterministic trends or autoregressive unit roots. Some of their regularity conditions are violated by these popular forms of nonstationarity. Recent papers by Andrews and McDermott (1995) and Dwyer (1995) show that the stationarity assumption can be further relaxed for some forms of nonstationarity. However, the long-run covariance matrix estimation procedure often needs to be modified to apply their asymptotic theory. For this reason, the strict stationarity assumption is emphasized in the context of time series applications rather than the fact that this assumption can be relaxed.

Since many macroeconomic variables exhibit nonstationarity, unless a researcher is careful this assumption can be easily violated in applications. As will be explained in Subsection 9.4.2, nonstationarity in the form of trend stationarity can be treated

with ease. In order to treat another popular form of nonstationarity, unit-root nonstationarity, researchers have used transformations such as first differences or growth rates of variables (see Chapter 10 for examples).

### 9.3.2 Identification

Another important assumption of Hansen (1982) is related to identification. Let

$$(9.22) \quad J_0(\mathbf{b}) = \{E[f(\mathbf{x}_t, \mathbf{b})]\}'\mathbf{W}_0\{E[f(\mathbf{x}_t, \mathbf{b})]\}.$$

The identification assumption is that  $\mathbf{b}_0$  is the unique minimizer of  $J_0(\mathbf{b})$ . Since  $J_0(\mathbf{b}) \geq 0$  and  $J_0(\mathbf{b}_0) = 0$ ,  $\mathbf{b}_0$  is a minimizer. Hence, this assumption requires  $J_0(\mathbf{b})$  to be strictly positive for any other  $\mathbf{b}$ . This assumption is obviously violated if  $f(\mathbf{x}_t, \mathbf{b}) \equiv \mathbf{0}$  for some  $\mathbf{b}$  that does not have any economic meaning (see Chapter 10 for examples). Even when this assumption is not violated, if values of  $J_0(\mathbf{b})$  are close to zero for parameter values around the unique minimizer and for other parameter values, then we have *weak identification* problem. This problem will be discussed later in this chapter.

## 9.4 Extensions

This section explains econometric methods that are closely related to the basic GMM framework.

### 9.4.1 Sequential Estimation

This subsection discusses sequential estimation (or two step estimation). Consider a system

$$(9.23) \quad f(\mathbf{x}_t, \mathbf{b}) = \begin{bmatrix} f_1(\mathbf{x}_t, \mathbf{b}_1) \\ f_2(\mathbf{x}_t, \mathbf{b}_1, \mathbf{b}_2) \end{bmatrix},$$

where  $\mathbf{b} = (\mathbf{b}'_1, \mathbf{b}'_2)'$ ,  $\mathbf{b}_i$  is a  $p_i$ -dimensional vector of parameters, and  $f_i$  is a  $q_i$ -dimensional vector of functions. Although it is possible to estimate  $\mathbf{b}_1$  and  $\mathbf{b}_2$  simultaneously, it may be computationally convenient to estimate  $\mathbf{b}_1$  from  $f_1(\mathbf{x}_t, \mathbf{b}_1)$  first, and then estimate  $\mathbf{b}_2$  from  $f_2(\mathbf{x}_t, \mathbf{b}_1, \mathbf{b}_2)$  in a second step (see, e.g., Barro, 1976; Atkeson and Ogaki, 1996, for examples of empirical applications). In general, the asymptotic distribution of the estimator of  $\mathbf{b}_2$  is affected by the estimation of  $\mathbf{b}_1$  (see, e.g., Newey, 1984; Pagan, 1984, 1986). A GMM computer program for sequential estimation can be used to calculate the correct standard errors that take into account these effects from estimating  $\mathbf{b}_1$ . If there are overidentifying restrictions in the system, an econometrician may wish to choose the second step distance matrix in an efficient way. The choice of the second step distance matrix is analyzed by Hansen, Heaton, and Ogaki (1992).

Suppose that the first step estimator  $\mathbf{b}_T^1$  minimizes

$$(9.24) \quad J_{1T}(\mathbf{b}_1) = \left\{ \frac{1}{T} \sum_{t=1}^T f_1(\mathbf{x}_t, \mathbf{b}_1) \right\}' \mathbf{W}_{1T} \left\{ \frac{1}{T} \sum_{t=1}^T f_1(\mathbf{x}_t, \mathbf{b}_1) \right\}$$

and that the second step estimator minimizes

$$(9.25) \quad J_{2T}(\mathbf{b}_2) = \left\{ \frac{1}{T} \sum_{t=1}^T f_2(\mathbf{x}_t, \mathbf{b}_{1T}, \mathbf{b}_2) \right\}' \mathbf{W}_{2T} \left\{ \frac{1}{T} \sum_{t=1}^T f_2(\mathbf{x}_t, \mathbf{b}_{1T}, \mathbf{b}_2) \right\},$$

where  $\mathbf{W}_{iT}$  is a positive definite matrix that converges to  $\mathbf{W}_{i0}$  with probability one.

Let  $\mathbf{\Gamma}_{ij}$  be the  $q_i \times p_j$  matrix  $E\left(\frac{\partial f_i}{\partial \mathbf{b}_j}\right)$  for  $i = 1, 2$  and  $j = 1, 2$ .

Given an arbitrary  $\mathbf{W}_{10}$ , the optimal choice of the second step distance matrix is  $\mathbf{W}_{20} = \mathbf{\Omega}^{*-1}$ , where

$$(9.26) \quad \mathbf{\Omega}^* = \begin{bmatrix} -\mathbf{\Gamma}_{21}(\mathbf{\Gamma}_{11} \mathbf{W}_{10} \mathbf{\Gamma}_{11})^{-1} \mathbf{\Gamma}_{11} \mathbf{W}_{10}, & \mathbf{I} \\ \mathbf{I} & \mathbf{I} \end{bmatrix} \mathbf{\Omega} \begin{bmatrix} -\mathbf{\Gamma}_{21}(\mathbf{\Gamma}_{11} \mathbf{W}_{10} \mathbf{\Gamma}_{11})^{-1} \mathbf{\Gamma}_{11} \mathbf{W}_{10} \\ \mathbf{I} \end{bmatrix}.$$

With this choice of  $\mathbf{W}_{20}$ ,  $\frac{1}{\sqrt{T}} \sum_{t=1}^T (\mathbf{b}_{2T} - \mathbf{b}_{20})$  has an (asymptotic) normal distribution

with mean zero and the covariance matrix

$$(9.27) \quad (\mathbf{\Gamma}'_{22}\mathbf{\Omega}^{*-1}\mathbf{\Gamma}_{22})^{-1}$$

and  $TJ_{2T}(\mathbf{b}_{2T})$  has an (asymptotic) chi-square distribution with  $q_2 - p_2$  degrees of freedom. It should be noted that if  $\mathbf{\Gamma}_{21} = \mathbf{0}$ , then the effect of the first step estimation can be ignored because  $\mathbf{\Omega}^* = \mathbf{\Omega}_{22} = E(f_2(\mathbf{x}_t, \mathbf{b}_0)f_2(\mathbf{x}_t, \mathbf{b}_0)')$ .

### 9.4.2 GMM with Deterministic Trends

This subsection discusses how GMM can be applied to time series with deterministic trends (see Eichenbaum and Hansen, 1990; Ogaki, 1988, 1989, for empirical examples).

Suppose that  $\mathbf{x}_t$  is trend stationary rather than stationary. In particular, let

$$(9.28) \quad \mathbf{x}_t = d(t, \mathbf{b}_{10}) + \mathbf{x}_t^*,$$

where  $d(t, \mathbf{b}_{10})$  is a function of deterministic trends such as time polynomials and  $\mathbf{x}_t^*$  is detrended  $\mathbf{x}_t$ . Assume that  $\mathbf{x}_t^*$  is stationary with  $E(\mathbf{x}_t^*) = \mathbf{0}$  and that there are  $q_2$  moment conditions

$$(9.29) \quad E(f_2(\mathbf{x}_t^*, \mathbf{b}_{10}, \mathbf{b}_{20})) = \mathbf{0}.$$

Let  $\mathbf{b} = (\mathbf{b}'_1, \mathbf{b}'_2)'$ ,  $f_1(\mathbf{x}_t, \mathbf{b}_1) = \mathbf{x}_t - d(t, \mathbf{b}_1)$  and  $f(\mathbf{x}_t, \mathbf{b}) = [f_1(\mathbf{x}_t, \mathbf{b}_1)', f_2(\mathbf{x}_t^*, \mathbf{b}_1, \mathbf{b}_2)']'$ .

Then GMM can be applied to  $f(\mathbf{x}_t, \mathbf{b})$  to estimate  $\mathbf{b}_1$  and  $\mathbf{b}_2$  simultaneously.

### 9.4.3 Other GMM Estimators

Several alternative estimators have been developed to deal with the poor small sample performance and weak identification problem of GMM.

One of them is the *continuous-updating estimator* provided by Hansen, Heaton, and Yaron (1996). It is obtained from changing the weighting matrix with each choice

of the parameter instead of taking it as given in each step of GMM estimation. An advantage of this estimator is that it is invariant to how the moment conditions are scaled.

Others use the *information theoretic approach* to circumvent the need for estimating a weighting matrix in a two step GMM. They include the empirical likelihood estimator (see, e.g., Kitamura and Stutzer, 1997; Imbens, 1997, 2002; Imbens and Spady, 2002) and exponential tilting estimator (see, e.g., Imbens, Spady, and Johnson, 1998). These estimators are based on minimization of the Kullback-Leibler Information Criterion distance to estimate parameters and to test the over-identifying restrictions (see, e.g., Golan, 2002 for a recent explanation of information econometrics).

## 9.5 Hypothesis Testing and Specification Tests

This section discusses specification tests and Wald, Lagrange Multiplier (LM), and likelihood ratio type statistics for hypothesis testing. Gallant (1987), Newey and West (1987), and Gallant and White (1988) have considered these three test statistics, and Eichenbaum, Hansen, and Singleton (1988) considered the likelihood ratio type test for GMM (or a more general estimation method that includes GMM as a special case).

Consider  $s$  nonlinear restrictions

$$(9.30) \quad H_0 : R(\mathbf{b}_0) = \mathbf{r},$$

where  $R$  is a  $s \times 1$  vector of functions. The null hypothesis  $H_0$  is tested against the alternative of  $R(\mathbf{b}_0) \neq \mathbf{r}$ . Let  $\mathbf{\Lambda} = \frac{\partial R}{\partial \mathbf{b}'}|_{\mathbf{b}_0}$  and  $\mathbf{\Lambda}_T$  be a consistent estimator for  $\mathbf{\Lambda}$ . It is assumed that  $\mathbf{\Lambda}$  is of rank  $s$ . If the restrictions are linear, then  $R(\mathbf{b}_0) = \mathbf{\Lambda}\mathbf{b}_0$  and

$\mathbf{\Lambda}$  is known. Let  $\mathbf{b}_T^u$  be an unrestricted GMM estimator and  $\mathbf{b}_T^r$  be a GMM estimator that is restricted by (9.30). It is assumed that  $\mathbf{W}_0 = \mathbf{\Omega}^{-1}$  is used for both estimators.

The Wald test statistic is

$$(9.31) \quad T(R(\mathbf{b}_T^u) - \mathbf{r})'[\mathbf{\Lambda}_T(\mathbf{\Gamma}_T'\mathbf{\Omega}_T^{-1}\mathbf{\Gamma}_T)^{-1}\mathbf{\Lambda}_T']^{-1}(R(\mathbf{b}_T^u) - \mathbf{r}),$$

where  $\mathbf{\Omega}_T$ ,  $\mathbf{\Gamma}_T$ , and  $\mathbf{\Lambda}_T$  are estimated from  $\mathbf{b}_T^u$ . The Lagrange multiplier test statistic is

$$(9.32) \quad LM_T = \frac{1}{T} \sum_{t=1}^T f(\mathbf{x}_t, \mathbf{b}_T^r)' \mathbf{\Omega}_T^{-1} \mathbf{\Gamma}_T \mathbf{\Lambda}_T' (\mathbf{\Lambda}_T \mathbf{\Lambda}_T')^{-1} [\mathbf{\Lambda}_T (\mathbf{\Gamma}_T' \mathbf{\Omega}_T^{-1} \mathbf{\Gamma}_T)^{-1} \mathbf{\Lambda}_T']^{-1} (\mathbf{\Lambda}_T \mathbf{\Lambda}_T')^{-1} \mathbf{\Lambda}_T \mathbf{\Gamma}_T' \mathbf{\Omega}_T^{-1} \sum_{t=1}^T f(\mathbf{x}_t, \mathbf{b}_T^r),$$

where  $\mathbf{\Omega}_T$ ,  $\mathbf{\Gamma}_T$ , and  $\mathbf{\Lambda}_T$  are estimated from  $\mathbf{b}_T^r$ . Note that in linear models  $LM_T$  is equal to (9.31), where  $\mathbf{\Omega}_T$ ,  $\mathbf{\Gamma}_T$ , and  $\mathbf{\Lambda}_T$  are estimated from  $\mathbf{b}_T^r$  rather than  $\mathbf{b}_T^u$ . (?????)

Masao  
needs to  
check this!

Need to reword) The likelihood ratio type test statistic is

$$(9.33) \quad T(J_T(\mathbf{b}_T^r) - J_T(\mathbf{b}_T^u)),$$

which is  $T$  times the difference between the minimized value of the objective function when the parameters are restricted and the minimized value of the objective function when the parameters are unrestricted. It is important that the same estimator for  $\mathbf{\Omega}$  is used for both unrestricted and restricted estimation for the likelihood ratio type test statistic. Under a set of regularity conditions, all three test statistics have asymptotic chi-square distributions with  $s$  degrees of freedom. The null hypothesis is rejected when these statistics are larger than the critical values obtained from chi-square distributions.

Existing Monte Carlo evidence suggests that the small sample distributions of the Lagrange multiplier test and the likelihood ratio type test are better approxi-

mated by their asymptotic distributions than those of the Wald test (see Gallant, 1987). Another disadvantage of the Wald test is that in general, the test result for nonlinear restrictions depends on the parameterization (see, e.g., Gregory and Veall, 1985; Phillips and Park, 1988).

Though the chi-square test for the overidentifying restrictions discussed in Section 9.1 has been frequently used as a specification test in applications of GMM, other specification tests applicable to GMM are available. These include tests developed by Singleton (1985), Andrews and Fair (1988), Hoffman and Pagan (1989), Andrews (1991), Ghysels and Hall (1990a,b,c), Hansen (1990), and Dufour, Ghysels, and Hall (1994). Some of these tests are discussed by Hall (1993).

## 9.6 Numerical Optimization

For nonlinear models, it is usually necessary to apply a numerical optimization method to compute a GMM estimator by numerically minimizing the criterion function,  $J_T(\mathbf{b})$ . The Newton-Raphson method (see, e.g., Hamilton, 1994, Chapter 5) is often used with an approximation method to calculate the Hessian matrix. A problem with the Newton-Raphson method and other practical numerical optimization methods is that global optimization is not guaranteed. The GMM estimator is defined as a global minimizer of a GMM criterion function, and the proof of its asymptotic properties depends on this assumption. Therefore, the use of a local optimization method can result in an estimator that is not necessarily consistent and asymptotically normal.

If the criterion function and parameter space are convex, then the criterion function has a unique local minimum, which is also the global minimum. In this case, a local optimization algorithm started at any parameter values should be able

to reach an approximate global minimum.

For nonconvex problems, however, there can be many local minima. For such problems, an algorithm called multi-start is often used for GMM applications. In this algorithm, one starts a local optimization algorithm from initial values of the parameters to converge to a local minimum, and then one repeats the process a number of times with different initial values. The estimator is taken to be the parameter values that correspond to the smallest value of the criterion function obtained during the multi-start process.

It should be noted that this multi-start algorithm is used for a given distance matrix. When the two stage or iterative GMM estimators are used, a different distance matrix is used in each stage, and hence a different criterion function is minimized. In most GMM programs, one needs to save the distance matrix in a file in order to apply the multi-start algorithm in each stage.

A problem with the multi-start algorithm, however, is that it does not necessarily find the global optimum. Therefore, the estimator it delivers is not necessarily consistent and asymptotically normal. Andrews (1997) proposes a simple stopping-rule procedure that overcomes this difficulty.

## 9.7 The Optimal Choice of Instrumental Variables

In the NLIV model discussed in Section 9.2, there are infinitely many possible instrumental variables because any variable in  $I_t$  can be used as an instrument. Hansen (1985) characterizes an efficiency bound (that is, a greatest lower bound) for the asymptotic covariance matrices of the alternative GMM estimators and optimal instruments that attain the bound. Since it can be time consuming to obtain op-

timal instruments, an econometrician may wish to compute an estimate of the efficiency bound to assess efficiency losses from using ad hoc instruments. Hansen (1985) also provides a method for calculating this bound for models with conditionally homoskedastic disturbance terms with an invertible MA representation.<sup>10</sup> Hansen, Heaton, and Ogaki (1988) extend this method to models with conditionally heteroskedastic disturbances and models with an MA representation that is not invertible.<sup>11</sup> Hansen and Singleton (1996) calculate these bounds and optimal instruments for a continuous time financial economic model.

## 9.8 Small Sample Properties

In most cases, the exact small sample properties cannot be derived for GMM estimators. Monte Carlo simulations have been conducted to study them for various nonlinear and linear models. Tauchen (1986) shows that GMM estimators and test statistics have reasonable small sample properties for data produced by simulations for a C-CAPM. Ferson and Foerster (1994) find similar results for a model of expected returns of assets as long as GMM is iterated for estimation of  $\Omega$ . Kocherlakota (1990) uses preference parameter values of  $\beta = 1.139$  and  $\alpha = 13.7$  (in Section 9.1) in his simulations for a C-CAPM that is similar to the Tauchen's (1986) model. While these parameter values do not violate any theoretical restrictions for existence of an equilibrium, they are much larger than the estimates of these preference parameters by Hansen and Singleton (1982) and others. Kocherlakota (1990) shows that GMM estimators for these parameters are biased downward and the chi-square test for the

---

<sup>10</sup>Hayashi and Sims' (1983) estimator is applicable to this example.

<sup>11</sup>Heaton and Ogaki (1991) provide an algorithm to calculate efficiency bounds for a continuous time financial economic model based on the Hansen, Heaton, and Ogaki's (1988) method.

overidentifying restrictions tends to reject the null too frequently compared with its asymptotic size. Mao (1990) reports that the chi-square test overrejects for more conventional values of these preference parameters in his Monte Carlo simulations.

Tauchen (1986) investigates small sample properties of Hansen's (1985) optimal instrumental variable GMM estimators. He finds that the optimal estimators do not perform well in small samples as compared to GMM estimators with ad hoc instruments. Tauchen (1986) and Kocherlakota (1990) recommend a small number of instruments rather than a large number of instruments when ad hoc instruments are used.

In some applications, scaling factors are another factor to affect finite sample GMM estimates. For example, Ni (1997) demonstrates that finite sample estimates are sensitive to scaling factors, and some seemingly reasonable scaling factors systematically lead to spurious estimates. However, Hansen, Heaton, and Yaron's (1996) continuous updating estimator is not affected by scaling factors.

Arellano and Bond (1991) report Monte Carlo results on GMM estimators for dynamic panel data models. They report that the GMM estimators have substantially smaller variances than commonly used Anderson and Hsiao's (1981) estimators in their Monte Carlo experiments. They also report that the small sample distributions of the serial-correlation tests they study are well approximated by their asymptotic distributions.

A very important small sample problem is weak identification, which we will discuss in the next section.

## 9.9 Weak Identification

In many applications, the identification condition holds but is almost violated in the sense that the values of the objective function evaluated at certain parameter values other than the true values are very close to the minimized value. In such applications, we have a *weak* problem. In the context of linear IV or NLIV estimation, this is called the *weak instrument variables problem*.

Nelson and Startz (1990) perform Monte Carlo simulations to investigate small sample properties of linear instrumental variables regressions. They show that instrumental variables estimators have poor sample properties when the instruments are weakly correlated with explanatory variables. In particular, they find that the chi-square test tends to reject the null too frequently compared with its asymptotic distribution, and that  $t$ -ratios tend to be too large when the instrument is poor. Their results for  $t$ -ratios may seem counterintuitive because one might expect that the consequence of having a poor instrument would be a large standard error and a low  $t$ -ratio. Staiger and Stock (1997) show that when the instruments are weakly correlated with the endogenous regressors, conventional asymptotic distribution theory fails even if the sample size is large. These results may be expected to carry over to NLIV estimation.

In the context of two stage least squares, Staiger and Stock (1997) suggest that first stage F-statistics, which tests the hypothesis that the instruments do not enter the first stage regression, should be reported at a minimum. Stock and Yogo (2005) advocates a pre-test rule to only use two stage least squares  $t$  statistics when the first stage F statistic exceeds ten. One strategy which continually changes the instruments until the F-statistics is significant is criticized by Hall, Rudebusch, and Wilcox (1996)

as it tends to make matters worse in the Monte Carlo simulations.

## 9.10 Identification Robust Methods

When GMM has a weak identification problem, the conventional GMM asymptotics fails to provide reliable inferences. One solution is to use identification robust methods, which does not rely on the identification assumption. These methods can be applied without using a pre-test rule such as Stock and Yogo's (2005). If confidence intervals or regions of parameters generated by identification robust methods are large, that indicates the presence of the weak identification problem.

Let  $\boldsymbol{\theta}$  denote a  $p$ -dimensional vector of parameters to be estimated. Consider the  $k$  dimensional vector of moment restrictions

$$(9.34) \quad E(f_t(\boldsymbol{\theta})) = 0$$

for  $t = 1, \dots, T$  which is assumed to be uniquely satisfied at  $\theta_0$ . The objective function for the CUE is:

$$(9.35) \quad Q(\boldsymbol{\theta}) = \left( \frac{1}{\sqrt{T}} \sum_{t=1}^T f_t(\boldsymbol{\theta}) \right)' \hat{V}_{ff}(\boldsymbol{\theta})^{-1} \left( \frac{1}{\sqrt{T}} \sum_{t=1}^T f_t(\boldsymbol{\theta}) \right)$$

where  $\hat{V}_{ff}(\boldsymbol{\theta})$  is a consistent estimator of the  $k \times k$  covariance matrix  $V_{ff}(\boldsymbol{\theta})$  of the moment vector.

In addition to the moment vector  $f_t(\boldsymbol{\theta})$ , consider also its derivative with respect to  $\boldsymbol{\theta}$ :

$$q_t(\boldsymbol{\theta}) = \text{vec} \left( \frac{\partial f_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} \right)$$

and  $q_T = \frac{1}{T} \sum_{t=1}^T q_t(\boldsymbol{\theta})$ .

We assume that in the large sample,  $f_t(\boldsymbol{\theta})$  and  $q_t(\boldsymbol{\theta})$  satisfy

$$\frac{1}{\sqrt{T}} \sum_{t=1}^T \begin{pmatrix} f_t(\boldsymbol{\theta}) - E(f_t(\boldsymbol{\theta})) \\ q_t(\boldsymbol{\theta}) - E(q_t(\boldsymbol{\theta})) \end{pmatrix} \xrightarrow{d} \begin{pmatrix} \phi_f(\boldsymbol{\theta}) \\ \phi_\theta(\boldsymbol{\theta}) \end{pmatrix}$$

where  $(\phi_f(\theta)' \quad \phi_\theta(\theta)')'$  is a  $k(p+1)$  dimensional normally distributed random process with mean zero and positive semi-definite  $k(p+1) \times k(p+1)$  dimensional covariance matrix

$$V(\theta) = \lim_{T \rightarrow \infty} \text{var} \left( \begin{array}{c} \frac{1}{\sqrt{T}} \sum_{t=1}^T f_t(\theta) \\ \frac{1}{\sqrt{T}} \sum_{t=1}^T q_t(\theta) \end{array} \right) = \begin{pmatrix} V_{ff}(\theta) & V_{f\theta}(\theta) \\ V_{\theta f}(\theta) & V_{\theta\theta}(\theta) \end{pmatrix},$$

with  $V_{\theta f}(\theta) = V_{f\theta}(\theta)' = (V_{\theta f,1}(\theta)' \cdots V_{\theta f,p}(\theta)')'$ ,  $V_{\theta\theta}(\theta) = V_{\theta\theta,ij}(\theta)$ ,  $i, j = 1, \dots, p$  and  $V_{ff}(\theta)$ ,  $V_{\theta f,i}(\theta)$ ,  $V_{\theta\theta,ij}(\theta)$  are  $k \times k$  dimensional matrices for  $i, j = 1, \dots, p$ .

The derivative estimator  $q_T(\theta)$  is correlated with the average moment vector  $f_T(\theta)$  since  $V_{\theta f}(\theta) \neq 0$ . The weak instrument robust statistics therefore use an alternative estimator of the derivative of the unconditional expectation of the Jacobian that is asymptotically uncorrelated with  $f_T(\theta)$ :

$$\hat{D}_T(\theta_0) = [q_{1,T}(\theta_0) - \hat{V}_{\theta f,1}(\theta_0) \hat{V}_{ff}(\theta_0)^{-1} f_T(\theta_0) \cdots \\ q_{p,T}(\theta_0) - \hat{V}_{\theta f,p}(\theta_0) \hat{V}_{ff}(\theta_0)^{-1} f_T(\theta_0)],$$

where  $\hat{V}_{\theta f,i}(\theta)$  are  $k_f \times k_f$  estimators of the covariance matrices  $V_{\theta f,i}(\theta)$ ,  $i = 1, \dots, p$ ,  $\hat{V}_{\theta f}(\theta) = (\hat{V}_{\theta f,1}(\theta)' \cdots \hat{V}_{\theta f,p}(\theta)')'$  and  $q_T(\theta_0) = (q'_{1,T}(\theta_0) \cdots q'_{p,T}(\theta_0))'$ .

The weak instrument robust statistics can be used for hypothesis testing on both subsets and the entire vector of the parameters. Let  $\theta = (\alpha' : \beta)'$ , with  $\alpha$  and  $\beta$  being  $p_\alpha$  and  $p_\beta$  dimensional vectors, respectively, such that  $p_\alpha + p_\beta = p$ . For tests on the entire set of parameters, consider  $\beta = \theta$ . Below, we introduce four statistics that test the hypothesis  $H_0 : \beta = \beta_0$ .

- The  $S$ -statistic of Stock and Wright (2000):

$$S(\beta_0) = Q(\tilde{\alpha}(\beta_0), \beta_0),$$

where  $\tilde{\alpha}(\beta_0)$  is the CUE of  $\alpha$  given that  $\beta = \beta_0$ . This is the CUE objective function (16.64).

- The score or Lagrange Multiplier statistic:

$$LM(\beta_0) = f_T(\tilde{\alpha}(\beta_0), \beta_0)' \hat{V}_{ff}(\tilde{\alpha}(\beta_0), \beta_0)^{-\frac{1}{2}} P_{\hat{V}_{ff}(\tilde{\alpha}(\beta_0), \beta_0)^{-\frac{1}{2}} \hat{D}_T(\tilde{\alpha}(\beta_0), \beta_0)} \hat{V}_{ff}(\tilde{\alpha}(\beta_0), \beta_0)^{-\frac{1}{2}} f_T(\tilde{\alpha}(\beta_0), \beta_0),$$

where  $P_A \equiv A(A'A)^{-1}A'$  for a full rank matrix  $A$ . This can be considered as the inverse of the conditional information matrix (Kleibergen, 2007).

- The over-identification statistic:

$$SL(\beta_0) = S(\beta_0) - LM(\beta_0)$$

- The conditional likelihood ratio statistic:

$$CLR(\beta_0) = \frac{1}{2} \left[ S(\beta_0) - rk(\beta_0) + \sqrt{\{S(\beta_0) + rk(\beta_0)\}^2 - 4SL(\beta_0)rk(\beta_0)} \right]$$

where  $rk(\beta_0)$  is a statistic that tests for a lower rank value of  $J(\tilde{\alpha}(\beta_0), \beta_0)$  and is a

function of  $\hat{D}_T(\tilde{\alpha}(\beta_0), \beta_0)$  and  $\hat{V}_{\theta\theta.f}(\tilde{\alpha}(\beta_0), \beta_0) = \hat{V}_{\theta\theta}(\tilde{\alpha}(\beta_0), \beta_0) - \hat{V}_{\theta f}(\tilde{\alpha}(\beta_0), \beta_0) \hat{V}_{ff}(\tilde{\alpha}(\beta_0), \beta_0)^{-1}$

$$rk(\beta_0) = \min_{\phi \in R^{p-1}} T(1 \ \phi)' \hat{D}_T(\tilde{\alpha}(\beta_0), \beta_0)' \left[ W' \hat{V}_{\theta\theta.f}(\tilde{\alpha}(\beta_0), \beta_0) W \right]^{-1} \hat{D}_T(\tilde{\alpha}(\beta_0), \beta_0) (1 \ \phi)'$$

where  $W = (I_{p\alpha} \ \phi')' \otimes I_k$ . The CLR statistic is a GMM extension of the conditional likelihood ratio statistic of Moreira (2003) for the linear instrumental variables regression model with one included endogenous variable.

Confidence sets for the parameter(s)  $\beta$  are obtained by inverting each of the identification-robust statistics (Zivot, Startz, and Nelson, 1998). The  $(1 - \alpha)100\%$  confidence bounds coincide with the intersection of the  $1 - \alpha$  value of the test statistic with the  $(1 - \alpha)$  line. A  $(1 - \alpha)100\%$  level confidence set thus constructed contains

all the values of  $\beta_0$  for which the corresponding test of the hypothesis  $H_0 : \beta = \beta_0$  does not reject  $H_0$  at the  $\alpha\%$  level of significance. When testing for more than one parameter jointly, these are the  $1 - \alpha$  contours of the graph of the function  $1 - p(\boldsymbol{\theta})$  where  $p(\boldsymbol{\theta})$  is the  $p$ -value of a test of a joint null hypothesis on a vector of parameters  $\boldsymbol{\theta}$ . Projection based confidence sets for an element of  $\boldsymbol{\theta}$  can be obtained from these joint confidence sets by projecting the widest range of the contours to the corresponding axis.

Kleibergen and Mavroeidis (2009) use the four test statistics to conduct inference on the parameters of the New Keynesian Phillips Curve (NKPC). They find evidence that forward-looking dynamics in inflation are statistically significant and dominate backward-looking dynamics. However, the confidence intervals for the backward-looking dynamics are too wide to draw any conclusion on its significance. Moreover, even though the slope of the NKPC is estimated to be positive, it is not significantly different from zero in any of the tests. These results confirm those of several authors who have reported empirical evidence that the NKPC is relatively flat and that its GMM estimation suffers from the weak identification problem (Mavroeidis, 2005; Nason and Smith, 2008). Kleibergen and Mavroeidis (2009) also find that, overall, the LR statistic is at least as powerful as other tests in the Monte Carlo simulations, and that it also yields the smallest confidence sets in their empirical applications.

## Appendix

### 9.A Asymptotic Theory for GMM

This Appendix reviews proofs for the asymptotic properties of GMM.

### 9.A.1 Asymptotic Properties of Extremum Estimators

Many estimators are formed by minimizing or maximizing objective functions. These estimators are called extremum estimators, or optimization estimators. A GMM estimator is a special case of an extremum estimator. In this section, we prove the consistency of extremum estimators. The next section applies the results to GMM. Given  $(S, \mathcal{F}, Pr)$ , let  $\mathbf{x}$  be a  $m$ -vector of random variables,  $\mathbf{b}$  be a  $p$ -vector of parameters, and  $J_T(\mathbf{x}, \mathbf{b})$  be a sequence of real valued functions. We will often denote  $J_T(\mathbf{x}, \mathbf{b})$  by  $J_T(\mathbf{b})$ . For GMM,  $\mathbf{x}$  will be taken as  $\mathbf{x} = (\mathbf{x}'_1, \mathbf{x}'_2, \dots, \mathbf{x}'_T)'$ , so that  $m$  is  $T$  times the dimension of  $\mathbf{x}_t$ . Thus, we allow  $m$  to be a function of  $T$ . The parameter  $\mathbf{b}$  is a member of a set  $\mathcal{B} \subset \mathcal{R}^p$ , and  $\mathcal{B}$  is called the parameter space.

An important condition for the consistency of extremum estimators relies on the concept of almost sure uniform convergence. Consider a sequence of functions  $g_T : \mathcal{R}^r \times \mathcal{B} \mapsto \mathcal{R}^q$ , such that  $g_T : (\cdot, \mathbf{b})$  is measurable for each  $\mathbf{b}$  in  $\mathcal{B}$  and  $f(\mathbf{z}, \mathbf{b})$  is continuous on  $\mathcal{B}$  for each  $\mathbf{z}$  in  $\mathcal{R}^r$ . Then  $g_T$  converges to a nonstochastic function  $g_0(\mathbf{b})$  *almost surely uniformly* in  $\mathbf{b} \in \mathcal{B}$  if there exists  $F \in \mathcal{F}$  with  $Pr(F) = 1$ , such that given any  $\epsilon > 0$ , for each  $s$  in  $F$  there exists an integer  $T(s, \epsilon)$  such that for all  $T > T(s, \epsilon)$ ,  $\sup_{\mathcal{B}} |g_T(\mathbf{x}(s), \mathbf{b}) - g_0(\mathbf{b})| < \epsilon$ . Here  $|\cdot|$  denotes the Euclidean norm. In this section, we will require that the sequence of real-valued functions  $J_T(\mathbf{x}, \mathbf{b})$  converges to a nonstochastic function  $J_0(\mathbf{b})$  almost surely uniformly in  $\mathbf{b} \in \mathcal{B}$ . In the next section, we will require a sequence of vector-valued functions converges almost surely uniformly.

Consider the following set of assumptions:

**Assumption 9.A.1** The parameter space  $\mathcal{B}$  is a compact set in  $\mathcal{R}^p$ .

**Assumption 9.A.2**  $J_T(\mathbf{x}, \mathbf{b})$  is continuous in  $\mathbf{b} \in \mathcal{B}$  for all  $\mathbf{x}$  and is a measurable function of  $\mathbf{x}$  for all  $\mathbf{b} \in \mathcal{B}$ .

**Assumption 9.A.3**  $J_T(\mathbf{x}, \mathbf{b})$  converges to a nonstochastic function  $J_0(\mathbf{b})$  almost surely uniformly in  $\mathbf{b} \in \mathcal{B}$ .

**Assumption 9.A.4**  $J_0(\mathbf{b})$  attains a unique global minimum at  $\mathbf{b}_0$ .

Since  $\mathcal{B}$  is a subset in  $\mathcal{R}^k$ , Assumption 9.A.1 is equivalent to assuming that  $\mathcal{B}$  is closed and bounded. Define an extreme estimator,  $\mathbf{b}_T$ , as a value that satisfies

$$J_T(\mathbf{b}_T) = \min_{\mathbf{b} \in \mathcal{B}} J_T(\mathbf{b}).$$

A complication is that the minimizer may not be unique, and it is not easy to prove that  $\mathbf{b}_T$  can be chosen in such a way that  $\mathbf{b}_T(\mathbf{x})$  is measurable. Different solutions to this problem are possible. Here, we have adopted a set of assumptions that are stronger than the assumptions in Theorem 4.1.1 of Amemiya (1985) for the weak consistency of extremum estimators. Amemiya (1985) states that if  $\mathbf{b}_T$  is not unique, it is possible to choose a value in such a way that  $\mathbf{b}_T(\mathbf{x})$  is a measurable function of  $\mathbf{x}$ . Assuming that  $\mathbf{b}_T(\mathbf{x})$  is chosen this way, we can prove the strong consistency of extremum estimators.

**Theorem 9.A.1** (*Strong consistency of extremum estimators*) If Assumptions 9.A.1 - 9.A.4 are satisfied, then  $\mathbf{b}_T$  converges almost surely to  $\mathbf{b}_0$ .

**Proof** Given any  $\epsilon > 0$ , let  $\eta(\epsilon)$  an open ball with the center  $\mathbf{b}_0$  and the radius  $\epsilon$ . If  $\eta(\epsilon)^c \cap \mathcal{B}$  is empty for all  $\epsilon$ , the result is trivial. Suppose that  $\eta(\epsilon)^c \cap \mathcal{B}$  is nonempty. Since  $\eta(\epsilon)^c \cap \mathcal{B}$  is compact and  $J_0(\mathbf{b})$  is continuous under our assumptions,  $\min_{\mathbf{b} \in \eta(\epsilon)^c \cap \mathcal{B}} J_0(\mathbf{b})$  exists. Denote

$$\delta(\epsilon) = \min_{\mathbf{b} \in \eta(\epsilon)^c \cap \mathcal{B}} J_0(\mathbf{b}) - J_0(\mathbf{b}_0).$$

Since  $J_T(\mathbf{x}, \mathbf{b})$  converges almost surely uniformly to  $J_0(\mathbf{b})$ , there exists  $F \in \mathcal{F}$ ,  $Pr(F) = 1$  such that for each  $s$  in  $F$  and all  $T > T(s, \delta(\epsilon))$ ,  $|J_T(\mathbf{b}) - J_0(\mathbf{b})| < \frac{\delta(\epsilon)}{2}$ . For  $\mathbf{b} = \mathbf{b}_T$ , we have

$|J_T(\mathbf{b}_T) - J_0(\mathbf{b}_T)| < \frac{\delta(\epsilon)}{2}$ , and hence  $J_0(\mathbf{b}_T) < J_T(\mathbf{b}_T) + \frac{\delta(\epsilon)}{2}$ . For  $\mathbf{b} = \mathbf{b}_0$ , we have  $|J_T(\mathbf{b}_0) - J_0(\mathbf{b}_0)| < \frac{\delta(\epsilon)}{2}$  or  $J_T(\mathbf{b}_0) < J_0(\mathbf{b}_0) + \frac{\delta(\epsilon)}{2}$ . Since  $\mathbf{b}_T$  minimizes  $J_T(\mathbf{b})$  on  $\mathcal{B}$ ,  $J_T(\mathbf{b}_T) < J_0(\mathbf{b}_0) + \frac{\delta(\epsilon)}{2}$ . Therefore,  $J_0(\mathbf{b}_T) < J_0(\mathbf{b}_0) + \delta(\epsilon)$  for each  $s$  in  $F$  and all  $T > T(s, \delta(\epsilon))$ . It follows that  $\mathbf{b}_T \in \eta(\epsilon)$  for each  $s$  in  $F$  and all  $T > T(s, \delta(\epsilon))$ . Since  $\epsilon$  is arbitrary and  $Pr(F) = 1$ , it follows that  $\mathbf{b}_T$  converges to  $\mathbf{b}_0$  almost surely. ■

## 9.A.2 Consistency of GMM Estimators

In this section, we apply Theorem 9.A.1 to GMM estimators. We construct the objective function  $J_T(\mathbf{x}, \mathbf{b})$  from a stationary ergodic stochastic process  $\mathbf{x}_t$  and a function  $f : \mathcal{R}^r \times \mathcal{B} \mapsto \mathcal{R}^q$  where  $q$  is greater than or equal to  $p$ . We will often denote  $f(\mathbf{x}_t, \mathbf{b})$  by  $f_t(\mathbf{b})$  or  $f(\mathbf{b})$ . We retain Assumption 9.A.1, and impose conditions on  $\mathbf{x}_t$  and  $f$  to ensure Assumptions 9.A.2 - 9.A.4 are satisfied.

**Assumption 9.A.5**  $\{\mathbf{x}_t : t \geq 1\}$  is an  $r$ -vector stationary and ergodic process.

**Assumption 9.A.6**  $f(\cdot, \mathbf{b})$  is measurable for each  $\mathbf{b}$  in  $\mathcal{B}$  and  $f(\mathbf{z}, \mathbf{b})$  is continuous on  $\mathcal{B}$  for each  $\mathbf{z}$  in  $\mathcal{R}^r$ .

**Assumption 9.A.7**  $E(|f(\mathbf{x}_1, \mathbf{b})|)$  exists and is finite for all  $\mathbf{b} \in \mathcal{B}$  and  $E(f(\mathbf{x}_1, \mathbf{b}_0)) = \mathbf{0}$ .

Since  $\mathbf{x}_t$  is stationary and ergodic,  $f(\mathbf{x}_t, \mathbf{b})$  is also stationary and ergodic for each  $\mathbf{b}$ . Therefore, Assumption 9.A.7 can be stated with any  $\mathbf{x}_t$  instead of  $\mathbf{x}_1$ .

Consider the following set of assumptions:

**Assumption 9.A.8**  $\frac{1}{T} \sum_{t=1}^T f(\mathbf{x}_t, \mathbf{b})$  converges almost surely uniformly to  $E(f(\mathbf{b}))$  in  $\mathcal{B}$ .

Since  $f(\mathbf{x}_t, \mathbf{b})$  is stationary and ergodic with finite first moments for each  $\mathbf{b}$ ,  $\frac{1}{T} \sum_{t=1}^T f(\mathbf{x}_t, \mathbf{b})$  converges almost surely to  $E(f(\mathbf{b}))$  for each  $\mathbf{b}$  in  $\mathcal{B}$ . Assumption 9.A.8 assumes that

this convergence is uniform. A sufficient condition for this assumption will be given in the next section.

**Assumption 9.A.9**  $E(f(\mathbf{b}))$  has a unique zero value at  $\mathbf{b}_0$ .

**Assumption 9.A.10** The sequence of random positive semidefinite matrices  $\{\mathbf{W}_T : T \geq 1\}$  converges almost surely to a nonstochastic positive definite matrix  $\mathbf{W}_0$ .

Let  $\mathbf{x} = (\mathbf{x}'_1, \mathbf{x}'_2, \dots, \mathbf{x}'_T)'$ ,  $J_T(\mathbf{x}, \mathbf{b}) = \{\frac{1}{T} \sum_{t=1}^T f(\mathbf{x}_t, \mathbf{b})\}' \mathbf{W}_T \{\frac{1}{T} \sum_{t=1}^T f(\mathbf{x}_t, \mathbf{b})\}$ , and  $J_0(\mathbf{b}) = E(f(\mathbf{x}_1))' \mathbf{W}_0 E(f(\mathbf{x}_1))$ . Define a GMM,  $\mathbf{b}_T$ , as a value that satisfies

$$J_T(\mathbf{b}_T) = \min_{\mathbf{b} \in \mathcal{B}} J_T(\mathbf{b}).$$

As in Section 9.A.1, it is understood that if  $\mathbf{b}_T$  is not unique, we appropriately choose a value in such a way that  $\mathbf{b}_T(\mathbf{x})$  is a measurable function of  $\mathbf{x}$ .

**Theorem 9.A.2** (*Strong consistency of GMM estimators*) If Assumption 9.A.1, 9.A.5 - 9.A.10 are satisfied,  $\mathbf{b}_T$  converges almost surely to  $\mathbf{b}_0$ . ■

It is easy to verify that Assumptions 9.A.5 - 9.A.10 imply Assumptions 9.A.2 - 9.A.4. Therefore, Theorem 9.A.1 implies Theorem 9.A.2.

### 9.A.3 A Sufficient Condition for the Almost Sure Uniform Convergence

We directly assumed the uniform convergence in Assumption 9.A.8. It is very difficult to confirm that this assumption is satisfied in most econometric models. Hence, it is important to investigate sufficient conditions for Assumption 9.A.8 Hansen (1982) provides an important sufficient condition based on a concept called the first moment continuity of  $f$ . This section proves that the first moment continuity implies Assumption 9.A.8.

The following notation is used for our continuity restriction:

$$(9.A.1) \quad \text{Mod}_f(\delta, \mathbf{b}) = \sup\{|f(\mathbf{b}) - f(\mathbf{b}^*)| : \mathbf{b}^* \in \mathcal{B} \text{ and } |\mathbf{b} - \mathbf{b}^*| < \delta\}.$$

where  $|\cdot|$  denotes the Euclidean norm. Since  $\mathcal{B}$  is separable, a dense sequence  $\{\mathbf{b}_j : j \geq 1\}$  can be used in place of  $\mathcal{B}$  in evaluating the *supremum*. In this case,  $\text{Mod}_f(\delta, \mathbf{b})$  is a random variable for each positive value of  $\delta$  and each  $\mathbf{b}$  in  $\mathcal{B}$ . Also,  $\text{Mod}_f(\delta, \mathbf{b}) \geq \text{Mod}_f(\delta^*, \mathbf{b})$  if  $\delta$  is greater than  $\delta^*$ . Since  $f(\cdot, \mathbf{b})$  is continuous,

$$(9.A.2) \quad \lim_{\delta \rightarrow 0} \text{Mod}_f(\delta, \mathbf{b}) = 0 \text{ for all } \mathbf{b} \in \mathcal{B} \text{ and all } \mathbf{b} \in \mathcal{B}.$$

A function  $f$  is *first-moment continuous* if for each  $\mathbf{b} \in \mathcal{B}$ ,

$$(9.A.3) \quad \lim_{\delta \rightarrow 0} E[\text{Mod}_f(\delta, \mathbf{b})] = 0.$$

A necessary and sufficient condition for  $f$  to be first-moment continuous is that for each  $\mathbf{b} \in \mathcal{B}$ , there exists  $\delta > 0$  such that

$$E[\text{Mod}_f(\delta, \mathbf{b})] < \infty.$$

It is trivial to see that this condition is necessary. This condition is sufficient because  $\text{Mod}_f(\delta, p)$  is decreasing in  $\delta$ : the Dominated Convergence Theorem and (9.A.2) imply the first-moment continuity of  $f$ .

**Assumption 9.A.8'**  $f$  is first-moment continuous.

**Proposition 9.A.1** Under Assumptions 9.A.1, 9.A.5 - 9.A.7, Assumption 9.A.8' implies that Assumption 9.A.8 is satisfied. Therefore, Assumption 9.A.8 for Theorem 9.A.2 can be replaced by Assumption 9.A.8'.

The proof of this proposition given here is a modified version of the proof of a closely related theorem in Hansen, Heaton, and Ogaki (1992). The proof is long and technical but is presented here for the econometric theory-oriented readers. We prepare for the proof by proving three lemmas. To prove this proposition, we use (i) pointwise continuity of  $E(f)$ , (ii) a pointwise Law of Large Numbers for  $\frac{1}{T} \sum f_t(\mathbf{b})$  for each  $\mathbf{b}$  in  $\mathcal{B}$ , and (iii) a pointwise Law of Large Numbers for  $\frac{1}{T} \sum_t Mod_f(\delta, \mathbf{b})$  for each  $\mathbf{b}$  in  $\mathcal{B}$  and positive  $\delta$ . As will be established in Lemma 9.A.1, (i) yields an approximation of the form:

**Approximation 9.A.1** There is positive-valued function  $\delta^*(\mathbf{b}, j)$  satisfying

$$(9.A.4) \quad |E[f(\mathbf{b}^*)] - E[f(\mathbf{b})]| < \frac{1}{j}$$

for all  $\mathbf{b}^* \in \mathcal{B}$  such that  $|\mathbf{b} - \mathbf{b}^*| < \delta^*(\mathbf{b}, j)$ . ■

As will be demonstrated in Lemma 9.A.2, (ii) provides an approximation of the form:

**Approximation 9.A.2** There is an integer-valued function  $T^*(s, \mathbf{b}, j)$  and an indexed set  $\Lambda^*(\mathbf{b}) \in \mathcal{F}$  such that  $Pr\{\Lambda^*(\mathbf{b})\} = 1$  and

$$(9.A.5) \quad \left| \frac{1}{T} \sum_{t=1}^T [f_t(s, \mathbf{b})] - E[f(\mathbf{b})] \right| < \frac{1}{j}$$

for all  $T \geq T^*(s, \mathbf{b}, j)$ , and  $s \in \Lambda^*(\mathbf{b})$ . ■

As will be shown in Lemma 9.A.3, (iii) yields an approximation of the form:

**Approximation 9.A.3** There exists an integer-valued function  $T^+(s, \mathbf{b}, j)$ , a positive function  $\delta^+(\mathbf{b}, j)$ , and an indexed set  $\Lambda^+(\mathbf{b}) \in \mathcal{F}$  such that  $Pr\{\Lambda^+(\mathbf{b})\} = 1$  and

$$(9.A.6) \quad \left| \frac{1}{T} [f(\mathbf{b}) - f(\mathbf{b}^*)] \right| < \frac{1}{j}$$

for all  $\mathbf{b}^* \in \mathcal{B}$  such that  $|\mathbf{b} - \mathbf{b}^*| < \delta^+(\mathbf{b}, j)$ ,  $T \geq T^+(s, \mathbf{b}, j)$ , and  $s \in \Lambda^+(\mathbf{b})$ . ■

Although the statements of these approximations require some cumbersome notation, we use this notation to monitor when sets and numbers depend on the underlying parameter values and approximation criteria ( $\mathbf{b}$  and  $j$ ). We will prove this theorem by showing that the assumption of a compact parameter space can be used to obtain an approximation that is uniform over the parameter space.

We now consider formally these inequalities. Lemma 9.A.1 establishes the continuity of  $E(f)$ .

**Lemma 9.A.1** If Assumptions 9.A.1, 9.A.6, 9.A.7, 9.A.8' are satisfied, then so is inequality (9.A.4).

**Proof** Since  $f$  is first-moment continuous, there is a function  $\delta^*(\mathbf{b}, j)$  such that

$$(9.A.7) \quad E[\text{Mod}_f[\delta^*(\mathbf{b}, j), \mathbf{b}]] < \frac{1}{j}.$$

Note, however, that

$$(9.A.8) \quad \begin{aligned} |Ef(\mathbf{b}^*) - Ef(\mathbf{b})| &\leq E|f(\mathbf{b}^*) - f(\mathbf{b})| \\ &\leq E\{\text{Mod}_f[\delta^*(\mathbf{b}, j), \mathbf{b}]\} \\ &< \frac{1}{j} \end{aligned}$$

for all  $\mathbf{b}^* \in \mathcal{B}$  such that  $|\mathbf{b} - \mathbf{b}^*| < \delta^*(\mathbf{b}, j)$ . ■

For each element  $\mathbf{b}$  in  $\mathcal{B}$ ,  $f(\mathbf{b})$  is a random variable with a finite absolute first moment. Thus the Law of Large Numbers applies pointwise as stated in the following lemma.

**Lemma 9.A.2** If Assumptions 9.A.1, 9.A.6, and 9.A.7 are satisfied, then so is inequality (9.A.5).

**Proof** Since  $\mathbf{x}_t$  is stationary and ergodic,  $\{\frac{1}{T} \sum_{t=1}^T [f(\mathbf{b})] : T \geq 1\}$  converges to  $E[f(\mathbf{b})]$  on a set  $\Lambda^*(\mathbf{b}) \in \mathcal{F}$  satisfying  $Pr\{\Lambda^*(\mathbf{b})\} = 1$ . ■

The Law of Large Numbers also applies to time series averages of  $Mod_f(\delta, \mathbf{b})$ . Since the mean of  $Mod_f(\delta, \mathbf{b})$  can be made arbitrarily small by choosing  $\delta$  to be small, we can control the local variation of time series averages of the random function  $f$ .

**Lemma 9.A.3** If Assumptions 9.A.1, 9.A.5, 9.A.6, and 9.A.8' are satisfied, then so is inequality (9.A.5).

**Proof** Since  $f$  is first-moment continuous,  $Mod_f(\frac{1}{n}, \mathbf{b})$  has a finite first moment for some positive integer  $n$ . Since  $\mathbf{x}_t$  is stationary and ergodic,  $\{\frac{1}{T} \sum_{t=1}^T [Mod_f(\frac{1}{j}, \mathbf{b})] : T \geq 1\}$  converges to  $E[Mod_f(\frac{1}{j}, \mathbf{b})]$  on a set  $\Lambda^+(\mathbf{b}, j)$  satisfying  $Pr\{\Lambda^+(\mathbf{b}, j)\} = 1$  for  $j \geq n$ . Let

$$\Lambda^+(\mathbf{b}) = \bigcap_{j \geq n} \Lambda^+(\mathbf{b}, j).$$

Then  $\Lambda^+(\mathbf{b})$  is measurable and  $Pr\{\Lambda^+(\mathbf{b})\} = 1$ .

For each  $j$ , choose  $\frac{1}{\delta^+(\mathbf{b}, j)}$  to equal some integer greater than or equal to  $n$  such that

$$(9.A.9) \quad E\{Mod_f[\delta^+(\mathbf{b}, j), \mathbf{b}]\} < \frac{1}{2j}.$$

Since  $\{\frac{1}{T} \sum_{t=1}^T \{Mod_f[\delta^+(\mathbf{b}, j), \mathbf{b}]\} : T \geq 1\}$  converges almost surely to  $E\{Mod_f[\delta^+(\mathbf{b}, j), \mathbf{b}]\}$  on  $\Lambda^+(\mathbf{b})$ , there exists an integer-valued function  $T^+(s, \mathbf{b}, j)$  such that

$$(9.A.10) \quad \left| \frac{1}{T} \sum_{t=1}^T \{Mod_f[\delta^+(\mathbf{b}, j), \mathbf{b}]\} - E\{Mod_f[\delta^+(\mathbf{b}, j), \mathbf{b}]\} \right| < \frac{1}{2j}$$

for  $T \geq T^+(s, \mathbf{b}, j)$ . Therefore,  $\frac{1}{T} \sum_{t=1}^T \{Mod_f[\delta^+(\mathbf{b}, j), \mathbf{b}]\} < \frac{1}{j}$ . Since  $\frac{1}{T} |\sum_{t=1}^T [f_t(\mathbf{b}) - f_t(\mathbf{b}^*)]| \leq \frac{1}{T} \sum_{t=1}^T \{Mod_f[\delta^+(\mathbf{b}, j), \mathbf{b}]\}$ ,

$$(9.A.11) \quad \frac{1}{T} \left| \sum_{t=1}^T [f_t(\mathbf{b}) - f_t(\mathbf{b}^*)] \right| < \frac{1}{j}$$

for all  $\mathbf{b}^* \in \mathcal{B}$  such that  $|\mathbf{b} - \mathbf{b}^*| < \delta^+(\mathbf{b}, j)$ ,  $T \geq T^+(s, \mathbf{b}, j)$ ,  $s \in \Lambda^+(\mathbf{b})$ , and  $j \geq 1$ . ■

We now combine the conclusions from Lemmas 9.A.1 - 9.A.3 to prove Proposition 9.A.1. The idea is to exploit that fact that  $\mathcal{B}$  is compact to move from pointwise to uniform convergence. Notice that in inequalities (9.A.4) - (9.A.6),  $\Lambda^+$ ,  $\Lambda^*$ ,  $T^+$  and  $T^*$  all depend on  $\mathbf{b}$ . In the following proof, we will use compactness to show how the dependence on the parameter value can be eliminated.

**Proof of Proposition 9.A.1** In the proof of this proposition, we use notation given in (9.A.4) - (9.A.6). Let

$$(9.A.12) \quad O(\mathbf{b}, n) = \{\mathbf{b}^* \in \mathcal{B} : |\mathbf{b} - \mathbf{b}^*| < \min\{\delta^*(\mathbf{b}, n), \delta^+(\mathbf{b}, n)\}\}.$$

Then for each  $n \geq 1$ ,

$$(9.A.13) \quad \mathcal{B} = \bigcup_{\mathbf{b} \in \mathcal{B}} O(\mathbf{b}, n).$$

Since  $\mathcal{B}$  is compact

$$(9.A.14) \quad \mathcal{B} = \bigcup_{J \geq 1}^{N(n)} O(\mathbf{b}_J, n),$$

where  $N(n)$  is integer-valued and  $\{\mathbf{b}_j : j \geq 1\}$  is a sequence in  $\mathcal{B}$ . Let

$$(9.A.15) \quad \Lambda \equiv \bigcap_{j \geq 1} [\Lambda^*(\mathbf{b}_j) \cap \Lambda^+(\mathbf{b}_j)].$$

Then  $\Lambda \in \mathcal{B}$  and  $Pr(\Lambda) = 1$ . Let

$$(9.A.16) \quad T(s, n) \equiv \max\{T^*(s, \mathbf{b}_1, n), T^*(s, \mathbf{b}_2, n), \dots, T^*[s, \mathbf{b}_{N(n)}, n], \\ T^+(s, \mathbf{b}_1, n), T^+(s, \mathbf{b}_2, n), \dots, T^+[s, \mathbf{b}_{N(n)}, n]\}.$$

For  $T \geq T(s, n)$ , inequalities (9.A.4)-(9.A.6) imply that

$$(9.A.17) \quad \left| \frac{1}{T} \sum_{t=1}^T [f_t(\mathbf{b})] - E[f(\mathbf{b})] \right| \\ \leq \frac{1}{T} \left| \sum_{t=1}^T [f_t(\mathbf{b}_j)] - \sum_{t=1}^T [f(\mathbf{b}_j)] \right| + \left| \frac{1}{T} \sum_{t=1}^T [f_t(\mathbf{b}_j)] - E[f(\mathbf{b}_j)] \right| + |E[f(\mathbf{b}_j)] - E[f(\mathbf{b})]| \\ < \frac{3}{n},$$

where  $\mathbf{b}_j$  is chosen so that  $\mathbf{b} \in O(\mathbf{b}_j, n)$  for some  $1 \leq j \leq N(n)$ . Therefore,  $\frac{1}{T} \sum_{t=1}^T f_t$  converges almost surely uniformly to  $E(f)$ . ■

## 9.A.4 Asymptotic Distributions of GMM Estimators

This section proves the asymptotic normality of GMM estimators and then discusses the optimal GMM estimators. It is possible to utilize the asymptotic normality results for general extremum estimators such as Amemiya's (1985) Theorem 4.1.3 here. However, unlike with the consistency results, it is more convenient to exploit the particular structure of the GMM objective function for this proof.

Consider the following set of assumptions:

**Assumption 9.A.11**  $\{\mathbf{b}_T : T \geq 1\}$  converges almost surely to  $\mathbf{b}_0$ .

**Assumption 9.A.12**  $\mathbf{b}_0 \in \mathcal{B}^\circ \subset \mathcal{B} \subset \mathcal{R}^p$ .

**Assumption 9.A.13**  $f(\cdot, \mathbf{b})$  is continuously differentiable with respect to  $\mathbf{b}$  on  $\mathcal{B}^\circ$  and the derivative  $Df(\cdot, \mathbf{b})$  has finite first moments and is first moment continuous on  $\mathcal{B}^\circ$ .

**Assumption 9.A.14**  $\{\mathbf{W}_T : T \geq 1\}$  converges almost surely to a nonsingular matrix  $\mathbf{W}_0$  of real numbers.

**Assumption 9.A.15**  $\{\mathbf{x}_t : t \geq 1\}$  is stationary and ergodic.

**Assumption 9.A.16**  $\frac{1}{\sqrt{T}} \sum_{t=1}^T f_t(\mathbf{b}_0) \xrightarrow{D} N(\mathbf{0}, \mathbf{\Omega})$ , where  $\mathbf{\Omega} = \sum_{j=-\infty}^{\infty} E(f_t(\mathbf{x}_t, \mathbf{b}_0) f_{t-j}(\mathbf{x}_t, \mathbf{b}_0)')$ .

**Assumption 9.A.17**  $E(Df(\mathbf{x}_t, \mathbf{b}_0))$  has rank  $p$ .

We denote  $E(Df(\mathbf{x}_1, \mathbf{b}_0))$  by  $\mathbf{\Gamma}$  and  $\frac{1}{T} \sum_{t=1}^T Df(\mathbf{x}_t, \mathbf{b}_T)$  by  $\mathbf{\Gamma}_T$ .

**Theorem 9.A.3** (*Asymptotic normality of GMM estimators*) If Assumptions 9.A.11 - 9.A.17 are satisfied, then

$$\sqrt{T}(\mathbf{b}_T - \mathbf{b}_0) \xrightarrow{D} N(\mathbf{0}, (\mathbf{\Gamma}'\mathbf{W}_0\mathbf{\Gamma})^{-1}\mathbf{\Gamma}'\mathbf{W}_0\mathbf{\Omega}\mathbf{W}_0\mathbf{\Gamma}(\mathbf{\Gamma}'\mathbf{W}_0\mathbf{\Gamma})^{-1}).$$

**Proof** Assumptions 9.A.11 and 9.A.12 imply there exists  $F \in \mathcal{F}$ ,  $Pr(F) = 1$  such that for any  $s$  in  $F$  there exists an integer  $T(s)$  such that  $\mathbf{b}_T \in \mathcal{B}^\circ$  for all  $T \geq T(s)$ . Going forward, we assume that  $\mathbf{b}_T \in \mathcal{B}^\circ$ . The first order condition for the minimization of the objective function is

$$(9.A.18) \quad \mathbf{\Gamma}'_T \mathbf{W}_T \left\{ \frac{1}{T} \sum_{t=1}^T f(\mathbf{x}_t, \mathbf{b}_T) \right\} = \mathbf{0}.$$

Given  $\mathbf{x}_t$ , applying the Mean Value Theorem to each row of  $\frac{1}{T} \sum_{t=1}^T f(\mathbf{x}_t, \mathbf{b}_T)$ , we obtain

$$(9.A.19) \quad \frac{1}{T} \sum_{t=1}^T f(\mathbf{x}_t, \mathbf{b}_T) = \frac{1}{T} \sum_{t=1}^T f(\mathbf{x}_t, \mathbf{b}_0) + \mathbf{\Gamma}_T^* (\mathbf{b}_T - \mathbf{b}_0),$$

where  $\mathbf{\Gamma}_T^*$  is formed by evaluating each row of  $\frac{1}{T} \sum_{t=1}^T Df(\mathbf{x}_t, \mathbf{b})$  at an intermediate vector between  $\mathbf{b}_T$  and  $\mathbf{b}_0$ . Assumptions 9.A.11 - 9.A.13, and 9.A.15 imply that  $\mathbf{\Gamma}_T^*$  converges almost surely to  $\mathbf{\Gamma}$ . Combining (9.A.18) and (9.A.19), we obtain

$$(9.A.20) \quad \mathbf{\Gamma}'_T \mathbf{W}_T \mathbf{\Gamma}_T^* (\mathbf{b}_T - \mathbf{b}_0) = -\mathbf{\Gamma}'_T \mathbf{W}_T \left\{ \frac{1}{T} \sum_{t=1}^T f(\mathbf{x}_t, \mathbf{b}_0) \right\}.$$

$\mathbf{\Gamma}'_T \mathbf{W}_T \mathbf{\Gamma}_T^*$  converges almost surely to  $\mathbf{\Gamma}' \mathbf{W}_0 \mathbf{\Gamma}$ , which is nonsingular. Hence, for sufficiently large  $T$ ,  $\mathbf{\Gamma}'_T \mathbf{W}_T \mathbf{\Gamma}_T^*$  is nonsingular with probability one. When  $\mathbf{\Gamma}'_T \mathbf{W}_T \mathbf{\Gamma}_T^*$  is nonsingular

$$(9.A.21) \quad \sqrt{T}(\mathbf{b}_T - \mathbf{b}_0) = -(\mathbf{\Gamma}'_T \mathbf{W}_T \mathbf{\Gamma}_T^*)^{-1} \mathbf{\Gamma}'_T \mathbf{W}_T \left\{ \frac{1}{\sqrt{T}} \sum_{t=1}^T f(\mathbf{x}_t, \mathbf{b}_0) \right\}.$$

Since  $(\mathbf{\Gamma}'_T \mathbf{W}_T \mathbf{\Gamma}_T^*)^{-1} \mathbf{\Gamma}'_T \mathbf{W}_T$  converges almost surely to  $(\mathbf{\Gamma}' \mathbf{W}_0 \mathbf{\Gamma})^{-1} \mathbf{\Gamma}' \mathbf{W}_0$ , Assumption 9.A.16 implies the conclusion.  $\blacksquare$

We use the following two propositions to prove that the GMM estimator with  $\mathbf{W}_0 = \mathbf{\Omega}^{-1}$  is the optimal GMM estimator when  $\mathbf{\Omega}$  is nonsingular.

**Proposition 9.A.2** Let  $\mathbf{A}$  be a  $q \times p$  matrix of rank  $p$ , then  $\mathbf{M} = \mathbf{I}_q - \mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'$  is idempotent with rank  $q - p$ .  $\blacksquare$

**Proposition 9.A.3** Let  $\mathbf{A}$  and  $\mathbf{C}$  be symmetric nonsingular matrices of the same size. Then  $\mathbf{A} \geq \mathbf{C} \geq \mathbf{0}$  implies  $\mathbf{A}^{-1} \leq \mathbf{C}^{-1}$ .  $\blacksquare$

Note that Proposition 9.A.2 implies that  $\mathbf{M}$  is positive semidefinite.

**Assumption 9.A.18**  $\mathbf{\Omega}$  is nonsingular.

Let  $Cov(\mathbf{W}_0) = (\mathbf{\Gamma}' \mathbf{W}_0 \mathbf{\Gamma})^{-1} \mathbf{\Gamma}' \mathbf{W}_0 \mathbf{\Omega} \mathbf{W}_0 \mathbf{\Gamma} (\mathbf{\Gamma}' \mathbf{W}_0 \mathbf{\Gamma})^{-1}$ .  $Cov(\mathbf{W}_0)$  is the covariance matrix of the GMM estimator associated with  $\mathbf{W}_0$ . In particular,  $Cov(\mathbf{\Omega}^{-1}) = (\mathbf{\Gamma}' \mathbf{\Omega}^{-1} \mathbf{\Gamma})^{-1}$ .

**Theorem 9.A.4 (Optimal GMM Estimators)** Suppose that Assumptions 9.A.11 - 9.A.18 are satisfied. Then  $Cov(\mathbf{\Omega}^{-1}) \leq Cov(\mathbf{W}_0)$  for any  $p \times p$  positive definite matrix  $\mathbf{W}_0$ .

**Proof** Since  $\Omega^{-1}$  is positive definite, there exists a nonsingular  $p \times p$  matrix  $\Lambda$  such that  $\Omega^{-1} = \Lambda' \Lambda$ . Then  $\Omega = \Lambda^{-1}(\Lambda')^{-1}$ . Let  $\mathbf{A}_1 = \Lambda \Gamma$  and  $\mathbf{A}_2 = \Lambda'^{-1} \mathbf{W}_0 \Gamma$ . Since  $\mathbf{I} - \mathbf{A}_2(\mathbf{A}_2' \mathbf{A}_2)^{-1} \mathbf{A}_2'$  is positive semidefinite by Proposition 9.A.2, we have

$$(9.A.22) \quad \mathbf{A}_1' \mathbf{A}_1 \geq \mathbf{A}_1' \mathbf{A}_2 (\mathbf{A}_2' \mathbf{A}_2)^{-1} \mathbf{A}_2' \mathbf{A}_1.$$

From Proposition 9.A.3, we obtain

$$(9.A.23) \quad (\mathbf{A}_1' \mathbf{A}_1)^{-1} \leq (\mathbf{A}_1' \mathbf{A}_2)^{-1} \mathbf{A}_2' \mathbf{A}_2 (\mathbf{A}_2' \mathbf{A}_1)^{-1}.$$

Since  $Cov(\Omega^{-1}) = (\Gamma' \Omega^{-1} \Gamma)^{-1} = (\mathbf{A}_1' \mathbf{A}_1)^{-1}$  and  $Cov(\mathbf{W}_0) = (\mathbf{A}_1' \mathbf{A}_2)^{-1} \mathbf{A}_2' \mathbf{A}_2 (\mathbf{A}_2' \mathbf{A}_1)^{-1}$ , the conclusion follows from this inequality. ■

The next theorem gives the asymptotic distribution of Hansen's  $J$  test statistic for the overidentifying restrictions.

**Theorem 9.A.5** (*Hansen's  $J$  test*) Suppose that Assumptions 9.A.11 - 9.A.18 are satisfied and that  $\mathbf{W}_0 = \Omega^{-1}$ . Then  $TJ_T$  converges in distribution to a chi-square random variable with  $q - p$  degrees of freedom.

**Proof** From (9.A.19), and Theorem 9.A.2,

$$(9.A.24) \quad \frac{1}{\sqrt{T}} \sum_{t=1}^T f(\mathbf{x}_t, \mathbf{b}_T) \xrightarrow{D} N(\mathbf{0}, \mathbf{V})$$

where  $\mathbf{V} = [\mathbf{I}_q - \Gamma(\Gamma' \Omega^{-1} \Gamma)^{-1} \Gamma'] \Omega [\mathbf{I} - (\Gamma' \Omega^{-1} \Gamma)^{-1} \Gamma']$ . As in the proof of Theorem 9.A.4, let  $\Lambda$  be a nonsingular  $p \times p$  matrix such that  $\Omega^{-1} = \Lambda' \Lambda$ . Then  $\Omega = \Lambda^{-1}(\Lambda')^{-1}$ , and

$$(9.A.25) \quad \frac{1}{\sqrt{T}} \Lambda \sum_{t=1}^T f(\mathbf{x}_t, \mathbf{b}_T) \xrightarrow{D} N(\mathbf{0}, \mathbf{M})$$

where  $\mathbf{M} = \Lambda [\Omega - \Gamma(\Gamma' \Omega^{-1} \Gamma)^{-1} \Gamma'] \Lambda' = \mathbf{I} - \Lambda \Gamma (\Gamma' \Omega^{-1} \Gamma)^{-1} \Gamma' \Lambda'$  is a symmetric, idempotent matrix. The trace of  $\mathbf{M}$  is  $q - p$  because  $tr(\mathbf{M}) = tr(\mathbf{I}_q) - tr\{\Lambda \Gamma (\Gamma' \Omega^{-1} \Gamma)^{-1} \Gamma' \Lambda'\} = tr(\mathbf{I}_q) - tr\{\Gamma' \Lambda' \Lambda \Gamma (\Gamma' \Omega^{-1} \Gamma)^{-1}\} = tr(\mathbf{I}_q) - tr(\mathbf{I}_p) = q - p$ . Therefore, there exists a matrix  $\mathbf{F}$  such that  $\mathbf{F}' \mathbf{F} = \mathbf{F} \mathbf{F}' = \mathbf{I}$ , and

$$(9.A.26) \quad \mathbf{M} = \mathbf{F} \begin{bmatrix} \mathbf{I}_{q-p} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{F}'.$$

Hence, if  $\mathbf{y} \sim N(\mathbf{0}, \mathbf{M})$ , then  $\mathbf{y}' \mathbf{y} = \mathbf{y}' \mathbf{F} \mathbf{F}' \mathbf{y} \sim \chi^2(q - p)$ . Since  $\mathbf{y}' \mathbf{y}$  is a continuous function mapping  $\mathcal{R}^q$  into  $\mathcal{R}$ ,

$$(9.A.27) \quad \left\{ \frac{1}{\sqrt{T}} \sum_{t=1}^T f(\mathbf{x}_t, \mathbf{b}_T)' \right\} \Omega_T^{-1} \left\{ \frac{1}{\sqrt{T}} \sum_{t=1}^T f(\mathbf{x}_t, \mathbf{b}_T) \right\} \xrightarrow{D} \chi^2(q - p)$$

where  $\Omega_T$  is a weakly consistent estimator for  $\Omega$ . ■

## 9.B The Conditional Likelihood Ratio Statistic

The conditional likelihood ratio (CLR) statistic can be used for an identification robust method to solve weak identification problems as explained in the text. The CLR statistic was proposed by Moreira (2003) for the linear IV regression models and later extended to GMM by Kleibergen (2005).

Kleibergen (2005) proposes a GMM Lagrange multiplier statistic (the K statistic) whose asymptotic  $\chi^2$  distribution holds in a wider set of circumstances such as the presence of weak identification problem. The K statistic replaces the sample average of the derivatives of the moments in the Newey and West's (1987) GMM LM statistic with a Jacobian estimator based on the continuous updating estimator (CUE) of Hansen, Heaton, and Yaron (1996). The CUE,  $\hat{\theta}$ , is obtained by minimizing the objective function,  $Q(\theta)$ , and continuously altering the covariance matrix as  $\hat{\theta}$  is changed in the minimization. Because of the correlation between the Jacobian estimator and the average moment vector, the limiting behavior of the Newey-West GMM LM statistic depends on nuisance parameters when, for example, the expected Jacobian is zero. The Jacobian estimator based on the CUE in the K statistic avoids this problem for it is asymptotically uncorrelated with the average moment vector (Brown and Newey, 1998; Donald and Newey, 2000). Given the dataset  $Y = [Y_1 \dots Y_T]'$ , the K statistic for testing  $H_0 : \theta = \theta_0$  is

$$K(\theta_0) = \frac{1}{4T} \left( \frac{\partial Q(\theta)}{\partial \theta'} \Big|_{\theta_0} \right) \left[ \hat{D}_T(\theta_0, Y)' \hat{V}_{ff}(\theta_0)^{-1} \hat{D}_T(\theta_0, Y) \right]^{-1} \left( \frac{\partial Q(\theta)}{\partial \theta'} \Big|_{\theta_0} \right)'$$

where  $\hat{D}_T(\theta_0, Y)$  is the CUE Jacobian estimator, and  $\hat{V}_{ff}$  the positive definite covariance matrix of the vector function  $f_T(\theta, Y)$ , and has a  $\chi^2(m)$  limiting distribution

under  $H_0$  and necessary assumptions.

By construction, the K statistic is equal to zero around the values of  $\theta$  for which the objective function attains its minimum, maximum, or is at an inflection point. While the moment conditions are satisfied for the values of  $\theta$  where the objective function is minimal and the CUE is obtained, they are not satisfied at the maximal value and inflection points, and thus the K statistic suffers from a spurious decline in power for such values of  $\theta$ . In order to appropriately account for this spurious behavior of the K statistic, Kleibergen suggests applying a GMM extension of Moreira's (2003) conditional likelihood ratio statistic for linear instrumental variables regressions (Kleibergen, 2004). The K statistic is combined with a J statistic,

$$J(\theta_0) = \frac{1}{T} f_T(\theta_0, Y)' \hat{V}_{ff}(\theta_0)^{-1/2} M_{\hat{V}_{ff}(\theta_0)^{-1/2} \hat{D}_T(\theta_0, Y)} \hat{V}_{ff}(\theta_0)^{-1/2} f_T(\theta_0, Y)$$

which tests the validity of the moment equations and is asymptotically independent of the K statistic.<sup>12</sup> For these values of  $\theta$  where the objective function is at its maxima or a reflection point, the J statistic has discriminatory power because it tests the validity of the moment equations,  $H_m : E(f_t(\theta_0)) = 0$ , while the K statistic tests  $H_0 : \theta = \theta_0$  given that the moment equations hold (Kleibergen, 2004).

The resulting test statistic (the GMM-M statistic) which accounts for the spurious power decline is

$$\text{GMM-M}(\theta_0) = \frac{1}{2} \left\{ K(\theta_0) + J(\theta_0) - \text{rk}(\theta_0) + \sqrt{[K(\theta_0) + J(\theta_0) + \text{rk}(\theta_0)]^2 - 4J(\theta_0)\text{rk}(\theta_0)} \right\}$$

where  $\text{rk}(\theta_0)$  is a statistic for testing the hypothesis of a lower rank value of  $J_\theta(\theta_0)$ ,  $H_r : \text{rank}(J_\theta(\theta_0)) = m - 1$  as in Cragg and Donald (1996), Cragg and Donald (1997), Kleibergen and Paap (2006), and Robin and Smith (2000). The GMM-M( $\theta_0$ ) leads

<sup>12</sup> $M_A = I_T - P_A$  where  $P_A = A(A'A)^{-1}A'$  for a full rank matrix  $A$ .

to inference that is centered around  $\hat{\theta}$  when  $\text{GMM-M}(\hat{\theta}) = 0$ . This occurs when  $\text{rk}(\hat{\theta})$  exceeds  $J(\hat{\theta})$  which puts a condition on the rank statistic  $\text{rk}(\theta_0)$  to be used in the GMM-M statistic.

A confidence set for  $\theta$  can be obtained by specifying sequences of  $n$  increasing values for every element of  $\theta$  and creating an  $m$ -dimensional grid that contains  $n^m$  different values of  $\theta_0$ . The statistic of interest (i.e., the J, K, or GMM-M statistic) can then be computed for each of these  $n^m$  different values of  $\theta_0$ . All elements in the specified grid for which the asymptotic  $p$ -value of the statistic of interest exceeds  $\alpha$  are in the  $(1 - \alpha)100\%$  asymptotic confidence set.

## 9.C A Procedure for Hansen's $J$ Test (GMM.EXP)

Hansen's  $J$  test proceeds as follows:

- (i) Check whether the number of moment restrictions is greater than that of the estimated parameters (the corresponding condition in the program is `NMR > KGM`).
- (ii) Choose an appropriate method to estimate the long-run covariance matrix,  $\Omega_T$ . See chapter 6 for details (the corresponding variable to specify the method is `CALWFLAG`).
- (iii) Set the maximum number of iterations to estimate the optimal weighting matrix,  $W_T = \Omega_T^{-1}$  (the default is `MAXITEGM = 5`).
- (iv) Define the objective function (the corresponding part in the program to define the GMM disturbance is the `HU` procedure.)
- (v) If the test statistic value (`CHI` in the output) is greater than the critical value for

the significance level you have in mind, say 5%, then reject the null hypothesis that the over-identification restrictions are satisfied.

## Exercises

The following problems are on econometric theory and require materials in Appendix 9.A.

**9.1** (*The Minimum Distance Estimation*) Assume that the following set of assumptions is satisfied.

(A1)  $\mathbf{p}_T$  converges almost surely to a  $k$ -dimensional vector  $\mathbf{p}_0$  of real numbers.

(A2)  $\sqrt{T}(\mathbf{p}_T - \mathbf{p}_0)$  converges in distribution to a normally distributed random vector with mean zero and a nonsingular covariance matrix  $\Sigma$ .

(A3)  $\Sigma_T$  converges almost surely to  $\Sigma$ .

(A4)  $\mathbf{p}_0 = \phi(\mathbf{q}_0)$  where  $\phi$  is a continuously differentiable function that maps  $Q \subset \mathbb{R}^h$  into  $\mathbb{R}^k$ . The parameter space  $Q$  is assumed to be compact. Let  $D\phi(\mathbf{q})$  be the  $k \times h$  matrix of the derivative of  $\phi$ , then  $\mathbf{D}_0 = D\phi(\mathbf{q}_0)$  is assumed to be of rank  $h$ .

(A5)  $\mathbf{p}_0 \neq \phi(\mathbf{q})$  for all  $\mathbf{q}$  in  $Q$  except for  $\mathbf{q} = \mathbf{q}_0$ .

Consider estimating  $\mathbf{q}_0$  by minimizing

$$(9.E.1) \quad J_T(\mathbf{q}) = \{[\phi(\mathbf{q}) - \mathbf{p}_T]\}' \mathbf{W}_T \{[\phi(\mathbf{q}) - \mathbf{p}_T]\}$$

over  $Q$ , where  $\mathbf{W}_T$  is a positive semidefinite  $k \times k$  random matrix that converges almost surely to a positive definite matrix of real numbers  $\mathbf{W}$ . Let  $\mathbf{q}_T$  be the minimizer. The

estimator  $\mathbf{q}_T$  is called the minimum distance estimator. Suppose that the sequence of minimizers converges almost surely to  $\mathbf{q}_0$ .

- (a) Prove that  $\mathbf{q}_T$  is strongly consistent for  $\mathbf{q}_0$  by applying Theorem 9.A.1 attached at the end. Hint: (i) You do not need the first moment continuity to prove the almost sure uniform convergence. (ii) Define a norm for a matrix  $\mathbf{w}$  by  $|\mathbf{w}| = |\text{vec}(\mathbf{w})|$ . Then  $|\mathbf{wz}| \leq |\mathbf{w}||\mathbf{z}|$  for two conformable matrices  $\mathbf{w}$  and  $\mathbf{z}$ .
- (b) Derive the asymptotic distribution of the estimators as a function of  $\mathbf{W}$ .
- (c) Derive the greatest lower bound for the asymptotic covariance matrices of members of this family of estimators, using Propositions 9.A.2 and 9.A.3 attached at the end. What is the optimal  $\mathbf{W}$ ?
- (d) Let  $\mathbf{q}_T$  be the minimum distance estimator associated with the optimum distance matrix  $\mathbf{W}$  in  $\mathcal{B}$ . Show that the minimized value of  $TJ_T(\mathbf{q}_T)$  converges in distribution to a  $\chi^2$  random variable. What is the degree of freedom of this  $\chi^2$  test statistic?
- (e) Consider the model

$$(9.E.2) \quad y_t = \mathbf{x}'_t \mathbf{p}_0 + \epsilon_t,$$

where  $y_t$ ,  $\mathbf{x}_t$  are a stationary and ergodic random variable and a 2-dimensional random vector with finite second moments, respectively. Suppose that  $E(\mathbf{x}_t \epsilon_t) = \mathbf{0}$ , and  $\frac{1}{\sqrt{T}} \sum_{t=1}^T \mathbf{x}_t \epsilon_t$  converges in distribution to  $N(\mathbf{0}, \mathbf{\Omega})$ . Suppose that economic theory imposes the restriction  $p_{02} = (p_{01})^3$ , where  $p_{0i}$  is the  $i$ -th element of  $\mathbf{p}_0$ . Discuss how you estimate this model, imposing the restriction using the minimum distance procedure you studied in the earlier parts of this problem,

assuming that you get the initial estimator,  $\mathbf{p}_T$ , by unconstrained OLS. In particular, discuss how do you obtain an estimator for  $\Sigma$ ,  $\Sigma_T$ , and how you attain the bound you derived.

- (f) Derive the asymptotic variance of your efficient minimum distance estimator you studied in (e) in terms of  $\Omega$  and  $p_{01}$ .

**9.2** In the case of a just identified system ( $q = p$ ), show that the instrumental variable regression estimator  $(\sum_{t=1}^T \mathbf{z}_t \mathbf{x}'_{2t})^{-1} \sum_{t=1}^T \mathbf{z}_t y_t$  coincides with the GMM estimator.

**9.3** All files needed for this problem are in the GMM-CCR package. You need to use GMM and KPRGMM. Modify INDIVIS.G program (you will need to make minor modifications to the `bgm`, `nw`, the `nf`, `fc`, `fx`, `fe` in PROC INDIVIS, `mm` in PROC MOMETNTS, `dm` in PROC DATAMOM, and PROC HU procedures) as follows:

Use  $f_t = i_t$  only.

Estimate only 9 parameters ( $\theta$ ,  $A_a$ ,  $\rho_a$ ,  $\sigma_a$ ,  $A_y$ ,  $\log \gamma$ ,  $\delta$ ,  $\alpha$ , and  $\sigma_i$ ).

- (a) Compute GMM estimates and standard errors of the above nine parameters.
- (b) Compute the model moment of investment ( $\sigma_i$ ) with its standard errors, and the data moment of investment ( $\sigma_i$ ) with its standard errors
- (c) Report the Wald test statistics and p-value to compare these two numbers.

## References

AMEMIYA, T. (1974): "The Nonlinear Two-Stage Least -Squares Estimator," *Journal of Econometrics*, 2, 105–110.

——— (1985): *Advanced Econometrics*. Harvard University Press, Cambridge, Massachusetts.

- ANDERSON, T. W., AND C. HSIAO (1981): "Estimation of Dynamic-Models with Error-Components," *Journal of the American Statistical Association*, 76(375), 598–606.
- ANDREWS, D. W. K. (1991): "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation," *Econometrica*, 59(3), 817–858.
- (1997): "A Stopping Rule for the Computation of Generalized Method of Moments Estimators," *Econometrica*, 65, 913–931.
- ANDREWS, D. W. K., AND R. C. FAIR (1988): "Inference in Nonlinear Econometric Models with Structural Change," *Review of Economic Studies*, 55(4), 615–639.
- ANDREWS, D. W. K., AND C. J. McDERMOTT (1995): "Nonlinear Econometric Models with Deterministically Trending Variables," *Review of Economic Studies*, 62, 343–360.
- ARELLANO, M., AND S. BOND (1991): "Some Tests of Specification for Panel Data: Monte-Carlo Evidence and an Application to Employment Equations," *Review of Economic Studies*, 58(2), 277–297.
- ATKESON, A., AND M. OGAKI (1996): "Wealth-Varying Intertemporal Elasticities of Substitution: Evidence from Panel and Aggregate Data," *Journal of Monetary Economics*, 38, 507–534.
- BARRO, R. J. (1976): "Rational Expectations and the Role of Monetary Policy," *Journal of Monetary Economics*, 2, 1–32.
- BROWN, B. W., AND W. K. NEWEY (1998): "Efficient Semiparametric Estimation of Expectations," *Econometrica*, 66(2), 453–464.
- CRAGG, J. G., AND S. G. DONALD (1996): "On the Asymptotic Properties of LDU-Based Tests of the Rank of a Matrix," *Journal of the American Statistical Association*, 91, 1301–1309.
- (1997): "Inferring the Rank of a Matrix," *Journal of Econometrics*, *Journal of Econometrics*, 76(1–2), 223–250.
- DONALD, S. G., AND W. K. NEWEY (2000): "A Jackknife Interpretation of the Continuous Updating Estimator," *Economics Letters*, 67(3), 239–243.
- DUFOUR, J.-M., E. GHYSELS, AND A. HALL (1994): "Generalized Predictive Tests and Structural Change Analysis in Econometrics," *International Economic Review*, 35(1), 199–229.
- DWYER, M. (1995): "Essays in Nonlinear, Nonstationary, Time Series Econometrics," Ph.D. thesis, University of Rochester.
- EICHENBAUM, M., AND L. P. HANSEN (1990): "Estimating Models with Intertemporal Substitution Using Aggregate Time Series Data," *Journal of Business and Economic Statistics*, 8, 53–69.
- EICHENBAUM, M., L. P. HANSEN, AND K. J. SINGLETON (1988): "A Time Series Analysis of Representative Agent Models of Consumption and Leisure Choice under Uncertainty," *Quarterly Journal of Economics*, 103, 51–78.
- FERSON, W. E., AND S. R. FOERSTER (1994): "Finite Sample Properties of the Generalized Methods of Moments in Tests of Conditional Asset Pricing Models," *Journal of Financial Economics*, 36, 29–55.

- GALLANT, A. R. (1977): "Three-stage Least-squares Estimation for a System of Simultaneous, Nonlinear, Implicit Equations," *Journal of Econometrics*, 5, 71–88.
- (1987): *Nonlinear Statistical Models*. John Wiley and Sons, New York.
- GALLANT, A. R., AND H. WHITE (1988): *A Unified Theory of Estimation and Inference for Nonlinear Dynamic Models*. Basil Blackwell, New York.
- GHYSELS, E., AND A. HALL (1990a): "Are Consumption-Based Intertemporal Capital Asset Pricing Models Structural?," *Journal of Econometrics*, 45, 121–139.
- (1990b): "A Test for Structural Stability of Euler Conditions Parameters Estimated via the Generalized Method of Moments Estimator," *International Economic Review*, 31, 355–364.
- (1990c): "Testing Nonnested Euler Conditions with Quadratic-Based Methods of Approximation," *Journal of Econometrics*, 46, 273–308.
- GOLAN, A. (2002): "Information and Entropy Econometrics—Editor's View," *Journal of Econometrics*, 107(1–2), 1–15.
- GREGORY, A. W., AND M. R. VEALL (1985): "Formulating Wald Tests of Nonlinear Restrictions," *Econometrica*, 53(6), 1465–1468.
- HALL, A. (1993): "Some Aspects of Generalized Method of Moments Estimation," in *Handbook of Statistics*, ed. by G. S. Maddala, C. R. Rao, and H. D. Vinod, vol. 11, chap. 15, pp. 393–417. Elsevier Science Publishers.
- HALL, A. R., G. D. RUDEBUSCH, AND D. W. WILCOX (1996): "Judging Instrument Relevance in Instrumental Variables Estimation," *International Economic Review*, 37(2), 283–298.
- HAMILTON, J. D. (1994): *Time Series Analysis*. Princeton University Press, Princeton.
- HANSEN, B. E. (1990): "Lagrange Multiplier Tests for Parameter Instability in Non-Linear Models," Manuscript, University of Rochester.
- HANSEN, L. P. (1982): "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50(4), 1029–1054.
- (1985): "A Method for Calculating Bounds on the Asymptotic Covariance Matrices of Generalized Method of Moments Estimators," *Journal of Econometrics*, 30, 203–238.
- HANSEN, L. P., J. C. HEATON, AND M. OGAKI (1988): "Efficiency Bounds Implied by Multiperiod Conditional Moment Restrictions," *Journal of the American Statistical Association*, 83(403), 863–871.
- (1992): "Lecture notes on GMM," Lecture notes.
- HANSEN, L. P., J. C. HEATON, AND A. YARON (1996): "Finite-Sample Properties of Some Alternative GMM Estimators," *Journal of Business and Economic Statistics*, 14(3), 262–280.
- HANSEN, L. P., AND K. J. SINGLETON (1982): "Generalized Instrumental Variables Estimation of Nonlinear Rational Expectations Models," *Econometrica*, 50(5), 1269–1286.
- (1996): "Efficient Estimation of Linear Asset-Pricing Models with Moving Average Errors," *Journal of Business and Economic Statistics*, 14, 53–68.

- HAYASHI, F. (2000): *Econometrics*. Princeton University Press, Princeton.
- HAYASHI, F., AND C. A. SIMS (1983): “Nearly Efficient Estimation of Time Series Models with Predetermined, but Not Exogenous Instruments,” *Econometrica*, 51, 783–798.
- HEATON, J. C., AND M. OGAKI (1991): “Efficiency Bound Calculations for a Time Series Model with Conditional Heteroskedasticity,” *Economics Letters*, 35, 167–171.
- HOFFMAN, D. L., AND A. R. PAGAN (1989): “Post-Sample Prediction Tests for Generalized-Method of Moments Estimators,” *Oxford Bulletin of Economics and Statistics*, 51(3), 333–343.
- IMBENS, G. W. (1997): “One-Step Estimators for Over-Identified Generalized Method of Moments Models,” *Review of Economic Studies*, 64(3), 359–383.
- (2002): “Generalized Method of Moments and Empirical Likelihood,” *Journal of Business and Economic Statistics*, 20(4), 493–506.
- IMBENS, G. W., AND R. H. SPADY (2002): “Confidence Intervals in Generalized Method of Moments Models,” *Journal of Econometrics*, 107(1–2), 87–98.
- IMBENS, G. W., R. H. SPADY, AND P. JOHNSON (1998): “Information Theoretic Approaches to Inference in Moment Condition Models,” *Econometrica*, 66(2), 333–358.
- JORGENSON, D. W., AND J.-J. LAFFONT (1974): “Efficient Estimation of Nonlinear Simultaneous Equations with Additive Disturbances,” *Annals of Economic and Social Measurement*, 3, 615–640.
- KITAMURA, Y., AND M. STUTZER (1997): “An Information-Theoretic Alternative to Generalized Method of Moments Estimation,” *Econometrica*, 65(4), 861–874.
- KLEIBERGEN, F. (2004): “Testing Subsets of Structural Parameters in the Instrumental Variables Regression Model,” *Review of Economics and Statistics*, 86(1), 418–423.
- (2005): “Testing Parameters in GMM without Assuming That They Are Identified,” *Econometrica*, 73(4), 1103–1124.
- (2007): “Generalizing Weak Instrument Robust IV Statistics towards Multiple Parameters, Unrestricted Covariance Matrices and Identification Statistics,” *Journal of Econometrics*, 139(1), 181–216.
- KLEIBERGEN, F., AND S. MAVROEIDIS (2009): “Weak Instrument Robust Tests in GMM and the New Keynesian Phillips Curve,” *Journal of Business and Economic Statistics*, 27(3), 293–311.
- KLEIBERGEN, F., AND R. PAAP (2006): “Generalized Reduced Rank Tests Using the Singular Value Decomposition,” *Journal of Econometrics*, 133(1), 97–126.
- KOCHERLAKOTA, N. R. (1990): “On Tests of Representative Consumer Asset Pricing Models,” *Journal of Monetary Economics*, 26, 285–304.
- MAO, C. S. (1990): “Hypothesis Testing and Finite Sample Properties of Generalized Method of Moments Estimators: A Monte Carlo Study,” Working Paper No. 90–12, Federal Reserve Bank of Richmond.
- MARK, N. C. (1985): “On Time Varying Risk Premia in the Foreign Exchange Market: An Econometric Analysis,” *Journal of Monetary Economics*, 16, 3–18.

- MAVROEIDIS, S. (2005): "Identification Issues in Forward-Looking Models Estimated by GMM with an Application to the Phillips Curve," *Journal of Money, Credit, and Banking*, 37(3), 421–448.
- MOREIRA, M. J. (2003): "A Conditional Likelihood Ratio Test for Structural Models," *Econometrica*, 71(4), 1027–1048.
- NASON, J. M., AND G. W. SMITH (2008): "Identifying the New Keynesian Phillips Curve," *Journal of Applied Econometrics*, 23(5), 525–551.
- NELSON, C. R., AND R. STARTZ (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.
- NEWBY, W. K. (1984): "A Method of Moments Interpretation of Sequential Estimators," *Economics Letters*, 14(2–3), 201–206.
- (1985): "Generalized-Method of Moments Specification Testing," *Journal of Econometrics*, 29(3), 229–256.
- NEWBY, W. K., AND K. D. WEST (1987): "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, 55(3), 703–708.
- NI, S. (1997): "Scaling Factors in Estimation of Time-Nonseparable Utility Functions," *Review of Economics and Statistics*, 79(2), 234–240.
- OGAKI, M. (1988): "Learning about Preferences from Time Trends," Ph.D. thesis, University of Chicago.
- (1989): "Information in Deterministic Trends about Preferences," Manuscript.
- PAGAN, A. R. (1984): "Econometric Issues in the Analysis of Regressions with Generated Regressors," *International Economic Review*, 25, 221–247.
- (1986): "Two Stage and Related Estimators and Their Applications," *Review of Economic Studies*, 53, 517–538.
- PHILLIPS, P. C. B., AND J. Y. PARK (1988): "Asymptotic Equivalence of Ordinary Least Squares and Generalized Least Squares in Regressions with Integrated Regressors," *Journal of the American Statistical Association*, 83, 111–115.
- ROBIN, J.-M., AND R. J. SMITH (2000): "Tests of Rank," *Econometric Theory*, 16(2), 151–175.
- SINGLETON, K. J. (1985): "Testing Specifications of Economic Agents' Intertemporal Optimum Problems in the Presence of Alternative Models," *Journal of Econometrics*, 30(1–2), 391–413.
- STAIGER, D., AND J. H. STOCK (1997): "Instrumental Variables Regression with Weak Instruments," *Econometrica*, 65(3), 557–586.
- STOCK, J. H., AND J. H. WRIGHT (2000): "GMM with Weak Identification," *Econometrica*, 68(5), 1055–1096.
- STOCK, J. H., AND M. YOGO (2005): "Testing for Weak Instruments in Linear IV Regression," in *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, ed. by D. W. K. Andrews, and J. H. Stock, pp. 80–108. Cambridge University Press, Cambridge.

TAUCHEN, G. (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.

ZIVOT, E., R. STARTZ, AND C. R. NELSON (1998): “Valid Confidence Intervals and Inference in the Presence of Weak Instruments,” *International Economic Review*, 39(4), 1119–1144.

# Chapter 10

## EMPIRICAL APPLICATIONS OF GMM

GMM estimation has been frequently applied to rational expectations models. This chapter discusses examples of these applications. The main purpose is not to provide a survey of the literature but to illustrate applications. Problems that researchers have encountered in applying GMM are discussed as well as procedures they have used to address these problems. In this chapter, the notation for the NLIV model of Section 9.2 will be used.

### 10.1 Euler Equation Approach

Hansen and Richard (1987) show that virtually all asset pricing models can be written as

$$(10.1) \quad v_t = E[m_{t+1}d_{t+1}|\mathbf{I}_t]$$

where  $v_t$  is the asset price at date  $t$ ,  $m_{t+1}$  is the intertemporal marginal rate of substitution (IMRS) between date  $t$  and date  $t+1$ , and  $d_{t+1}$  is the payoff of an asset at date  $t+1$ . Each asset pricing model specifies a different IMRS.

Hansen and Singleton (1982) specify the IMRS by

$$(10.2) \quad m_{t+1} = \beta \left( \frac{c_{t+1}}{c_t} \right)^{-\alpha}$$

and measure  $c_t$  by real nondurable consumption expenditures or real nondurable and service consumption expenditures. Hansen and Singleton (1984) find that the chi-square test for the overidentifying restrictions rejects their model especially when nominal risk free bond returns and stock returns are used simultaneously.<sup>1</sup> Their finding is consistent with the Mehra and Prescott's (1985) equity premium puzzle. When the model is rejected, the chi-square test statistic does not provide much guidance as to what causes the rejection. Hansen and Jagannathan (1991) develop a diagnostic that could provide such guidance.

Brown and Gibbons (1985) use the same specification of the IMRS but propose to measure it from asset returns data rather than consumption data. An advantage of this measurement is that asset returns data are measured without measurement errors and are free from the time aggregation problem in contrast to consumption data.

They assume that  $E\left(\frac{c_{t+1}}{c_t} | I_t\right)$  is a constant that does not depend on  $I_t$ . For example, this assumption is satisfied if consumption is a martingale, in which case  $E\left(\frac{c_{t+1}}{c_t} | I_t\right) = 1$ . Then  $E\left(\frac{c_{t+\tau}}{c_t} | I_t\right) = E\left[\left(\frac{c_{t+\tau}}{c_{t+\tau-1}}\right)\left(\frac{c_{t+\tau-1}}{c_{t+\tau-2}}\right) \cdots \left(\frac{c_{t+1}}{c_t}\right) | I_t\right]$  is a constant that does not depend on  $I_t$ . Therefore,  $E\left[\beta^\tau \left(\frac{c_{t+\tau}}{c_t}\right)^{-\alpha} | I_t\right] = k_\tau$  is a constant that does not depend on  $I_t$ .

Now consider a security that pays off  $c_{t+\tau}$  as its payoff for  $\tau = 1, 2, 3, \dots$ . Then

---

<sup>1</sup>Cochrane (1989) points out that the utility that the representative consumer loses by deviating from the optimal consumption path is very small in the Hansen-Singleton model and in the Hall's (1978) model. In this sense, the Hansen-Singleton test and Hall's test may be too sensitive to economically small deviations caused by small costs of information and transactions.

the price of the security at date  $t$  will be

$$(10.3) \quad v_t = E\left[\sum_{\tau=1}^{\infty} \beta^{\tau} \left(\frac{c_{t+\tau}}{c_t}\right)^{-\alpha} c_{t+\tau} \mid I_t\right] = \left(\sum_{\tau=1}^{\infty} k_{\tau}\right) c_t.$$

Hence, the gross rate of return from holding this security from date  $t$  to date  $t+1$ ,  $R_{t+1}^m$ , is

$$(10.4) \quad R_{t+1}^m = \frac{v_{t+1} + c_{t+1}}{v_t} = k \frac{c_{t+1}}{c_t}$$

where  $k = (1 + \sum_{\tau=1}^{\infty} k_{\tau}) / \sum_{\tau=1}^{\infty} k_{\tau}$ . Hence the IMRS can be measured by  $R_{t+1}^m$ :

$$(10.5) \quad m_{t+1} = \beta \left(\frac{c_{t+1}}{c_t}\right)^{-\alpha} = \beta^* (R_{t+1}^m)^{-\alpha}.$$

where  $\beta^* = \beta k^{\alpha}$ . The Euler equation is

$$(10.6) \quad E(\beta^* (R_{t+1}^m)^{-\alpha} R_{t+1} \mid I_t) = 1$$

for any asset return  $R_{t+1}$ . To apply GMM, let  $\mathbf{b} = (\beta^*, \alpha)'$ ,  $\mathbf{x}_t = (R_{t+1}^m, R_{t+1})'$ , and  $g(\mathbf{x}_t, \mathbf{b}) = \beta^* (R_{t+1}^m)^{-\alpha} R_{t+1} - 1$  in the notation for the NLIV model.

Brown and Gibbons (1985) measure  $R_{t+1}^m$  by the New York Stock Exchange value weighted return. Even though the value weighted return is precisely measured, it is not exactly equal to  $R_{t+1}^m$  in the model because the value weighted average of the New York Stock Exchange stocks does not pay aggregate consumption as its payoff. This problem is closely related to the Roll's (1977) critique for tests of Capital Asset Pricing Models which use the value weighted returns as the market return.

Even though the Euler equation holds for any asset return, the identification assumption for GMM fails to hold when we choose  $R_{t+1}$  in (10.6) to be  $R_{t+1}^m$ . With this choice,  $g(\mathbf{x}_t, \mathbf{b}) = 0$  when  $\beta^* = 1$  and  $\alpha = 1$ .

## 10.2 Habit Formation and Durability

Many researchers have considered the effects of time-nonseparability in preferences on asset pricing. Let us replace (9.2) by

$$(10.7) \quad U(c_t, c_{t-1}, c_{t-2}, \dots) = \frac{1}{1-\alpha} (s_t^{1-\alpha} - 1),$$

where  $s_t$  is the service flow from consumption purchases. Purchases of consumption and service flows are related by

$$(10.8) \quad s_t = a_0 c_t + a_1 c_{t-1} + a_2 c_{t-2} + \dots$$

Depending on the values of the  $a_\tau$ 's, the model (10.7) leads to a model with habit formation and/or durability. For example, this type of specification for time-nonseparability has been used to model durability by Mankiw (1985), Hayashi (1982), Dunn and Singleton (1986), Eichenbaum, Hansen, and Singleton (1988), Eichenbaum and Hansen (1990), and Ogaki and Reinhart (1998a,b), and used to model habit formation by Ferson and Constantinides (1991), Ferson and Harvey (1992), Cooley and Ogaki (1996), and Ogaki and Park (1997).<sup>2</sup> Heaton (1993, 1995) used it to model a combination of durability and habit formation. Constantinides (1990) argues that habit formation could help solve the equity premium puzzle. He shows how the intertemporal elasticity of substitution and the relative risk aversion coefficient depend on the parameters  $a_\tau$  and  $\alpha$  in a habit formation model.

In this section, we discuss applications by Ferson and Constantinides (1991), Cooley and Ogaki (1996), and Ogaki and Park (1997) to illustrate econometric formulations for habit formation models. We will discuss more about applications for

---

<sup>2</sup>These papers found evidence in favor of habit formation with aggregate consumption data, but Dynan (2000) finds no evidence for habit formation in household level panel data for food.

durable goods in later sections. In their models, it is assumed that  $a_\tau = 0$  for  $\tau \geq 2$ . Let us normalize  $a_0$  to be one, so that  $\mathbf{b} = (\beta, \alpha, a_1)'$ . The asset pricing equation takes the form

$$(10.9) \quad \frac{E[\beta(s_{t+1}^{-\alpha} + \beta a_1 s_{t+2}^{-\alpha})R_{t+1} | \mathbf{I}_t]}{E[s_t^{-\alpha} + \beta a_1 s_{t+1}^{-\alpha} | \mathbf{I}_t]} = 1.$$

Then let  $\epsilon_t^0 = \beta(s_{t+1}^{-\alpha} + \beta a_1 s_{t+2}^{-\alpha})R_{t+1} - (s_t^{-\alpha} + \beta a_1 s_{t+1}^{-\alpha})$ . Though Euler equation (10.9) implies that  $E(\epsilon_t^0 | \mathbf{I}_t) = 0$ , this property cannot be used as the disturbance for GMM because both of the two regularity assumptions discussed in Section 9.3 are violated. These violations are caused by the nonstationarity of  $c_t$  and by the three sets of trivial solutions,  $\alpha = 0$  and  $1 + \beta a_1 = 0$ ;  $\beta = 0$  and  $\alpha = \infty$ ; and  $\beta = 0$  and  $a_1 = \infty$  with  $\alpha > 0$ . Ferson and Constantinides (1991) solve both of these problems by defining  $\epsilon_t = \frac{\epsilon_t^0}{s_t^{-\alpha}}$ . Since  $s_t^{-\alpha}$  is in  $\mathbf{I}_t$ ,  $E(\epsilon_t | \mathbf{I}_t) = 0$ . The disturbance is a function of  $\frac{s_{t+\tau}}{s_t}$  ( $\tau = 1, 2$ ) and  $R_{t+1}$ . When  $\frac{c_{t+1}}{c_t}$  and  $R_t$  are assumed to be stationary,  $\frac{s_{t+\tau}}{s_t}$  and the disturbance can be written as a function of stationary variables.

One problem that researchers have encountered in these applications is that  $c_{t+1} + a_1 c_t$  may be negative when  $a_1$  is close to minus one. In a nonlinear search for  $\mathbf{b}_T$  or in calculating numerical derivatives, a GMM computer program will stall if it tries a value of  $a_1$  that makes  $c_{t+1} + a_1 c_t$  negative for any  $t$ . Atkeson and Ogaki (1996) have encountered similar problems in estimating fixed subsistence levels from panel data. One way to avoid this problem is to program the function  $f(\mathbf{x}_t, \mathbf{b})$ , so that the program returns very large numbers as the values of  $f(\mathbf{x}_t, \mathbf{b})$  when non-admissible parameter values are used. However, it is necessary to ignore these large values of  $f(\mathbf{x}_t, \mathbf{b})$  when calculating numerical derivatives. This process can be done by suitably modifying programs that calculate numerical derivatives.<sup>3</sup>

<sup>3</sup>A GMM User Guide (see Ogaki, 1993b) explains these modifications for Hansen/Heaton/Ogaki

The model presented in this section is the linear specification of habit formation. More recent theoretical work often adopts the nonlinear specification of habit formation as in Campbell and Cochrane (1999, 2000) and Menzly, Santos, and Veronesi (2004), among others. The model presented in this section is also a model of internal habit formation. In models of external habit formation, the habit depends on the consumption of some exterior reference group. In the Abel's (1990) model of catching up with Jones, the habit depends on per capita aggregate consumption. Campbell and Cochrane (1999, 2000), Li (2001), and Menzly, Santos, and Veronesi (2004) study models of external habit formation. Chen and Ludvigson (2004) use the sieve minimum distance estimator developed by Newey and Powell (2003) and Ai and Chen (2003) for approximating an unknown function to empirically evaluate various specifications of habit including linear/nonlinear and internal/external habit formation. The sieve minimum distance estimator is implemented in the GMM framework.

### 10.3 State-Nonseparable Preferences

Epstein and Zin (1991) estimate a model with state-nonseparable preference specification in which the life-time utility level  $v_t$  at period  $t$  is defined recursively by

$$(10.10) \quad V_t = \{c_t^{1-\alpha} + \beta E[V_{t+1}^{1-\alpha} | \mathcal{I}_t]\}^{\frac{1-\rho}{1-\alpha}},$$

where  $\alpha > 0$  and  $\rho > 0$ . The asset pricing equation for this model is

$$(10.11) \quad E[\beta^* (R_{t+1}^m)^\eta (\frac{c_{t+1}}{c_t})^\theta R_{t+1}] = 1,$$

for any asset return  $R_{t+1}$ , where  $\beta^* = \beta^{\frac{1-\alpha}{1-\rho}}$ ,  $\eta = \frac{\rho-\alpha}{1-\rho}$ ,  $\theta = -\rho \frac{1-\alpha}{1-\rho}$ , and  $R_{t+1}^m$  is the (gross) return of the optimal portfolio ( $R_{t+1}^m$  is the return from period  $t$  to  $t+1$  of

---

GMM package.

a security that pays  $c_t$  every period forever). They use the value-weighted return of shares traded on the New York Stock Exchange as  $R_{t+1}^m$ . Thus, the Roll's (1977) critique of CAPM is relevant here as discussed.

Even though (10.11) holds for  $R_{t+1} = R_{t+1}^m$ , the identification assumption discussed in Section 9.3 is violated for this choice of  $R_{t+1}$  because there exists a trivial solution,  $(\beta^*, \eta, \theta) = (1, -1, 0)$ , for  $g(\mathbf{x}_t, \mathbf{b}) = \mathbf{0}$ . When multiple returns that include  $R_{t+1}^m$  are used simultaneously, then the whole system can satisfy the identification assumption, but the GMM estimators for this partially unidentified system are likely to have bad small sample properties. A similar problem arises when  $R_{t+1}$  does not include  $R_{t+1}^m$  but includes multiple equity returns whose linear combination is close to  $R_{t+1}^m$ . It should be noted that Epstein and Zin avoid these problems by carefully choosing returns to be included as  $R_{t+1}$  in their system.

## 10.4 Time Aggregation

The use of consumption data for C-CAPM is subject to a time aggregation problem (see, e.g., Hansen and Sargent, 1983a,b) because consumers can make decisions at intervals much finer than the observed frequency of the data and because the observed data consist of average consumption over a period of time.

In linear models for which the disturbance before time aggregation is a martingale difference, time aggregation means that the disturbance has an MA(1) structure and the instrumental variables need to be lagged an additional period. See, e.g., Grossman, Melino, and Shiller (1987), Hall (1988), and Hansen and Singleton (1996) for applications to C-CAPM and Heaton (1993) and Christiano, Eichenbaum, and Marshall (1991) for applications to Hall (1978) type permanent income models.

In nonlinear models for which the disturbance before time aggregation is a martingale difference, time aggregation has more complicated effects. Allowing the disturbance to have an MA(1) structure and letting instrumental variables lagged an additional period do not completely eliminate the effects caused by time aggregation. Nevertheless, these methods are often used to mitigate time aggregation problems in applications (see, e.g., Epstein and Zin, 1991; Ogaki and Reinhart, 1998b).

For nonlinear models, one way to use GMM to take into account the full effects of time aggregation is to combine GMM with simulations. For example, Heaton (1995) uses the method of simulated moments (MSM) for his nonlinear asset pricing model with time-nonseparable preferences in taking time aggregation into account. Bossaerts (1988), Duffie and Singleton (1993), McFadden (1989), Pakes and Pollard (1989), Lee and Ingram (1991), and Pearson (1991), among others, have studied asymptotic properties of MSM.

## 10.5 Multiple-Goods Models

Mankiw, Rotemberg, and Summers (1985), Dunn and Singleton (1986), Eichenbaum, Hansen, and Singleton (1988), Eichenbaum and Hansen (1990), and Osano and Inoue (1991), among others, have estimated versions of multiple-good C-CAPM. Basic economic formulations of these multiple-good models will be illustrated in the context of a simple model with one durable good and one nondurable good.

Let us replace (9.2) by Houthakker's (1960) addilog utility function that Miron (1986), Ogaki (1988, 1989), and Osano and Inoue (1991) among others have estimated:

$$(10.12) \quad U(c_t, d_t) = \frac{1}{1-\alpha}(c_t^{1-\alpha} - 1) + \frac{\theta}{1-\eta}(k_t^{1-\eta} - 1),$$

where  $c_t$  is nondurable consumption and  $k_t$  is household capital stock from purchases

of durable consumption good  $d_t$ .<sup>4</sup> The stock of durables is assumed to depreciate at a constant rate  $1 - a$ , where  $0 \leq a < 1$ :

$$(10.13) \quad k_t = ak_{t-1} + d_t.$$

Alternatively,  $k_t$  can be considered as a service flow in (10.8) with  $a_\tau = a^\tau$ . When  $\alpha \neq \eta$ , preferences are not quasi homothetic. In practice, the data for  $k_t$  is constructed from data for an initial stock  $k_0$ , and for  $d_t$  for  $t = 1, \dots, T$ . Let  $p_t$  be the intratemporal relative price of durable and nondurable consumption. Then the intraperiod first order condition that equates the relative price with the marginal rate of substitution is

$$(10.14) \quad p_t = \frac{\theta E(\sum_{\tau=1}^{\infty} \beta^\tau a^\tau k_{t+\tau}^{-\eta} | \mathbf{I}_t)}{c_t^{-\alpha}}.$$

Assume that  $\frac{d_{t+1}}{d_t}$  is stationary. Then  $\frac{k_{t+\tau}}{d_t}$  is stationary for any  $\tau$  because  $\frac{k_{t+\tau}}{d_t} = \sum_{i=0}^{\tau-1} a^i \frac{d_{t+\tau-i}}{d_t}$ . From (10.14),

$$(10.15) \quad \frac{p_t c_t^{-\alpha}}{d_t^{-\eta}} = \theta E[\sum_{\tau=1}^{\infty} \beta^\tau a^\tau (\frac{k_{t+\tau}}{d_t})^{-\eta} | \mathbf{I}_t].$$

Assume that the variables in  $\mathbf{I}_t$  are stationary.<sup>5</sup> Then (10.15) implies that the  $p_t \frac{c_t^{-\alpha}}{d_t^{-\eta}}$  is stationary because the right hand side of (10.15) is stationary. Taking natural logs, we conclude that  $\ln(p_t) - \alpha \ln(c_t) + \eta \ln(d_t)$  is stationary. This restriction is called the stationarity restriction.

From (10.14), define

$$(10.16) \quad \epsilon_t^0 = p_t c_t^{-\alpha} - (1 - \beta a F)^{-1} \theta k_t^{-\eta},$$

<sup>4</sup>Since the addilog utility function is not quasi-homothetic in general, the distribution of initial wealth affects the utility function of the representative consumer. The existence of a representative consumer under complete markets is discussed by Ogaki (1990) for general concave utility functions and by Atkeson and Ogaki (1996) for extended addilog utility functions.

<sup>5</sup>If  $\mathbf{I}_t$  includes nonstationary variables, assume that the right hand side of (10.14) is the same as the expectation conditioned on the stationary variables in  $\mathbf{I}_t$ .

where  $F$  is the forward operator. The first order condition (10.14) implies that  $E(\epsilon_t^0 | I_t) = 0$ . One problem is that  $\epsilon_t^0$  involves  $k_{t+\tau}$  for  $\tau$  from 0 to infinity, so that  $\epsilon_t^0$  cannot be used as the disturbance for GMM. To solve this problem, define  $\epsilon_t = (1 - \beta a F)\epsilon_t^0$ . Note that  $\epsilon_t$  involves only  $c_t, c_{t+1}, p_t, p_{t+1}$ , and  $k_t$  and that  $E[\epsilon_t | I_t] = 0$ . Hence  $\epsilon_t$  forms the basis of GMM. The only remaining problem is attaining stationarity. One might think it is enough to divide  $\epsilon_t$  by  $k_t^{-\eta}$ , so that the resulting  $\epsilon_t$  is stationary as implied by the stationarity restriction. It should be noted that it is *not* enough for  $\epsilon_t = g(\mathbf{x}_t, \mathbf{b}_0)$  to be stationary, rather it is also necessary for  $g(\mathbf{x}_t, \mathbf{b})$  to be stationary for  $\mathbf{b} \neq \mathbf{b}_0$ . Hence if  $\alpha$  and  $\eta$  are unknown and  $c_t$  or  $d_t$  is difference stationary, GMM cannot be applied to the first order condition (10.14).<sup>6</sup> Ogaki (1988, 1989) assumes that  $c_t$  and  $d_t$  are trend stationary and applies the method of Section 10.2 above to utilize the detrended version of  $\epsilon_t$ . In these applications, the restrictions on the trend coefficients and the curvature parameters  $\alpha$  and  $\eta$  implied by the stationarity restriction are imposed on the GMM estimators. Imposing the stationarity restrictions also lead to more reasonable point estimates for  $\alpha$  and  $\eta$ .

Eichenbaum, Hansen, and Singleton (1988) and Eichenbaum and Hansen (1990) use the Cobb-Douglas utility function, so that  $\alpha$  and  $\eta$  are known to be one.<sup>7</sup> They allow preferences to be nonseparable across goods and time-nonseparable, but the stationarity restriction is shown to hold. In this case, the stationarity restriction implies that  $p_t \frac{c_t^{-1}}{k_t^{-1}}$  is stationary. This transformation does not involve any unknown parameters. Hence, this transformation is used to apply GMM to their intraperiod first order conditions.

---

<sup>6</sup>Cointegrating regressions can be used for this case as explained below.

<sup>7</sup>Also see Ogaki (1988) for a discussion of the stationarity restriction implied by the Cobb-Douglas utility function.

## 10.6 Seasonality

Miron (1986) augments the Hansen and Singleton's (1982) model by including deterministic seasonal taste shifters and argues that the empirical rejection of C-CAPM by Hansen and Singleton (1982) and others might be attributable to the use of seasonally adjusted data.<sup>8</sup> Although this is theoretically possible, English, Miron, and Wilcox (1989) find that seasonally unadjusted quarterly data reject asset pricing equations at least as strongly as seasonally adjusted data.<sup>9</sup> Ogaki (1988) also finds similar empirical results for seasonally unadjusted and adjusted data in the system that involves both asset pricing equations and intraperiod first order conditions.

Singleton (1988) argues that the inclusion of taste shifters in C-CAPM is essentially equivalent to directly studying consumption data with deterministic seasonality removed. This finding results because we do not obtain much identifying information from seasonal fluctuations about preferences if most of the seasonal fluctuations come from seasonal taste shifts.<sup>10</sup> On the other hand, seasonal fluctuations may contain useful identifying information about the production functions if production functions are relatively stable over the seasonal cycle. Braun and Evans (1998) utilize such identifying information.

Ferson and Harvey (1992) construct seasonally unadjusted monthly data and estimate a C-CAPM with time nonseparable preferences. They find that seasonal habit persistence is empirically significant. Heaton (1993) also finds evidence for

---

<sup>8</sup>It should be noted that a deterministic seasonal dummy can be viewed as an artificial stationary and ergodic stochastic process (see, e.g., Ogaki, 1988, pp. 26–27). Hence, GMM can be applied to models with deterministic seasonal taste shifts.

<sup>9</sup>Hoffman and Pagan (1989) also obtain similar results.

<sup>10</sup>Beaulieu and Miron (1991) cast doubt on the view that negative output growth in the first quarter (see, e.g., Barsky and Miron, 1989) is caused by negative technology seasonal by observing negative output growth in the Southern Hemisphere.

Masao  
needs to  
check this!

seasonal habit formation in Hall (1978) type permanent income models.<sup>11</sup>

## 10.7 Monetary Models

In some applications, monetary models are estimated by applying GMM to Euler equations and/or intratemporal first order conditions. Singleton (1985), Ogaki (1988), Finn, Hoffman, and Schlagenhaut (1990), Bohn (1991), and Sil (1992) estimate cash-in-advance models, Poterba and Rotemberg (1987), Eckstein and Leiderman (1989), and Finn, Hoffman, and Schlagenhaut (1990), Imrohoroglu (1991) estimate money-in-the-utility-function (MIUF) models, and Marshall (1992) estimates a transactions-cost monetary model.

Cash-in-advance models involve only minor variations on the asset pricing equation (10.1) as long as the cash-in-advance constraints are binding and  $c_t$  is a cash good (in the terminology of Lucas and Stokey, 1987). However, nominal prices of consumption, nominal consumption, and nominal asset returns are aligned over time in a different way in monetary models than they are in the Hansen and Singleton's (1982) model. Information available to agents at time  $t$  is also considered in a different way. As a result, instrumental variables are lagged one period more than in the Hansen-Singleton model, and  $\mathbf{u}_t$  has an MA(1) structure (time aggregation has the same effects in linear models as discussed above). There is some tendency for the chi-square test statistics for the overidentifying restrictions to be more favorable for the timing conventions suggested by cash-in-advance models (see Finn, Hoffman, and Schlagenhaut, 1990; Ogaki, 1988). Ogaki (1988) focuses on monetary distortions in relative prices for a cash good and a credit good and does not find monetary

---

<sup>11</sup>See Ghysels (1990, especially Section I.3) for a survey of the economic and econometric issues of seasonality.

distortions in the U.S. data he examines.

## 10.8 Calculating Standard Errors for Estimates of Standard Deviation, Correlation, and Autocorrelation

In many macroeconomic applications, researchers report estimates of standard deviations, correlations, and autocorrelations of economic variables. It is possible to use a GMM program to calculate standard errors for these estimates, in which the serial correlation of the economic variables is taken into account (see, e.g., Backus, Gregory, and Zin, 1989; Backus and Kehoe, 1992).

For example, let  $x_t$  and  $y_t$  be economic variables of interest that are assumed to be stationary. Let  $\mathbf{x}_t = (x_t, y_t)$  and  $f(\mathbf{x}_t, \mathbf{b}) = (x_t, x_t^2, y_t, y_t^2, x_t y_t, x_t x_{t-1})' - \mathbf{b}$ , where  $f(\mathbf{x}_t, \mathbf{b})$  is a disturbance defined at time  $t$  and a quadratic form of its sample average is the objective function to be minimized in GMM estimation. Then the parameters to be estimated are the population moments;  $\mathbf{b}_0 = (E(x_t), E(x_t^2), E(y_t), E(y_t^2), E(x_t y_t), E(x_t x_{t-1}))$ . Applying GMM to  $f(\mathbf{x}_t, \mathbf{b})$ , one can obtain an estimate of  $\mathbf{b}_0$ ,  $\mathbf{b}_T$ , and an estimate of covariance matrix of  $T^{\frac{1}{2}}(\mathbf{b}_T - \mathbf{b}_0)$ .<sup>12</sup> In most applications, the order of serial correlation of  $(x_t, x_t^2, y_t, y_t^2, x_t y_t, x_t x_{t-1})'$  is unknown, and its long-run covariance matrix,  $\mathbf{\Omega}$ , can be estimated by any method in Chapter 6 (such as Andrews and Monahan's prewhitened QS kernel estimation method).

Standard deviations, correlations, and autocorrelations are nonlinear functions of  $\mathbf{b}_0$ . Hence, one can use the delta method to calculate the standard errors of the

---

<sup>12</sup>The covariance matrix  $Cov(\mathbf{\Omega}^{-1})$  is defined in (9.13). In this particular example,  $Cov(\mathbf{\Omega}^{-1})$  coincides with the long-run variance of  $f(\mathbf{x}_t, \mathbf{b}_0)$  because the derivative of  $f(\mathbf{x}_t, \mathbf{b}_0)$  is an identity. More generally, if more moment conditions are added to make the system overidentified, then  $Cov(\mathbf{\Omega}^{-1})$  will be different from the long-run covariance matrix.

estimates of these statistics. Let  $a(\mathbf{b}_0)$  be the statistic of interest. Continuing the example above, imagine that a researcher is estimating the standard deviation of  $x_t$ . Then  $a(\mathbf{b}_0) = \sqrt{\text{var}(x_t)} = (E(x_t^2) - E(x_t)^2)^{\frac{1}{2}} = (b_{02} - b_{01}^2)^{\frac{1}{2}}$ , where  $b_{01} = E(x_t)$ ,  $b_{02} = E(x_t^2)$  and  $a(\mathbf{b}_T)$  is a consistent estimator of  $a(\mathbf{b}_0)$ . If we apply the delta method explained in Proposition 5.8,  $\sqrt{T}(a(\mathbf{b}_T) - a(\mathbf{b}_0))$  has an approximate normal distribution with the variance  $\mathbf{d}(\mathbf{b}_0)\text{Cov}(\boldsymbol{\Omega}^{-1})\mathbf{d}(\mathbf{b}_0)'$  in large samples, where  $\mathbf{d}(\mathbf{b}_0)$  is the derivative of  $a(\cdot)$  evaluated at  $\mathbf{b}_0$ .

There is a pitfall that should be avoided in setting the GMM moment conditions in these applications. The parameters can enter the GMM moment conditions in nonlinear ways, but the sample moments should not. For example, it may be tempting to estimate the variance of  $x_t$  in the above example by setting the moment condition to be  $b - (x_t - \bar{x})^2$  where  $b$  is the variance to be estimated and  $\bar{x}$  is the sample mean. However, because the sample mean enters the GMM moment condition in a nonlinear way,  $E(b - (x_t - \bar{x})^2)$  is not equal to zero. This pitfall can be easily avoided by estimating  $E(x)$  and  $E(x^2)$  as in the example above.

An example of a problematic application with this type of the pitfall can be found in Section 5 of Ambler, Cardia, and Zimmermann (2004). In estimating a pair of correlations, their estimate is a solution to the problem of minimizing

$$(10.17) \quad \left\{ \frac{1}{T} \sum_{t=1}^T (\boldsymbol{\rho} - \bar{\boldsymbol{\rho}}_t) \right\}' \mathbf{W}_T \left\{ \frac{1}{T} \sum_{t=1}^T (\boldsymbol{\rho} - \bar{\boldsymbol{\rho}}_t) \right\}$$

where the parameter  $\boldsymbol{\rho}$  is a  $2 \times 1$  vector of the population correlations of four variables (say  $x_{it}$  for  $i = 1, 2, 3, 4$ ), and  $\bar{\boldsymbol{\rho}}_t$  is a  $2 \times 1$  vector whose first element is given by

$$(10.18) \quad \bar{\rho}_{1t} = \frac{(x_1 - \bar{x}_1)(x_2 - \bar{x}_2)}{\bar{\sigma}_1 \bar{\sigma}_2}$$

and whose second element is given by

$$(10.19) \quad \bar{\rho}_{2t} = \frac{(x_3 - \bar{x}_3)(x_4 - \bar{x}_4)}{\bar{\sigma}_3 \bar{\sigma}_4}.$$

Here  $\bar{x}_i$  is the sample mean and  $\bar{\sigma}_i$  is the sample variance of  $x_i$ . This set-up resembles that of GMM, but cannot be embedded in the standard GMM framework. This is because the sample mean and the sample variance enter the moment conditions in nonlinear ways.

## 10.9 Dynamic Stochastic General Equilibrium Models and GMM Estimation

Real Business Cycle Models and other Dynamic Stochastic General Equilibrium (DSGE) models can be estimated and tested by GMM. These models are often simulated and the results are evaluated without considering sampling errors. GMM gives a simple method to take into account sampling errors. Such a method was originally developed by Christiano and Eichenbaum (1992). A survey by Burnside (1999) describes how GMM estimation is used for real business cycle models and explains how to use the programs written by the author. Recent applications of GMM to DSGE models include Alexopoulos (2004) and Aguiar and Gopinath (2007). In this section, we explain a method used by Burnside, Eichenbaum, and Rebelo (1993) using a simpler model than these authors used. This method uses results in King, Plosser, and

Rebelo (1988a,b) that show how the model parameters are related to the moments of economic variables.

Consider a social planner's problem:

$$(10.20) \quad \begin{aligned} & \max_{C_t, K_t} E_0 \sum_{t=0}^{\infty} \beta^t U(C_t) \\ \text{s.t. } & Y_t = A_t K_{t-1}^\alpha = C_t + I_t \\ & I_t = K_t - (1 - \delta)K_{t-1} \\ & \ln A_t = \rho \ln A_{t-1} + \epsilon_t, \end{aligned}$$

where  $C_t$  is consumption,  $K_t$  is a capital stock,  $Y_t$  is output,  $I_t$  is investment,  $1 - \delta$  is a depreciation rate, and  $A_t$  represents the level of technology. We do not include labor to simplify the model for the pedagogical purpose. Using the budget constraint given by

$$(10.21) \quad C_t = A_t K_{t-1}^\alpha - K_t + (1 - \delta)K_{t-1}$$

the first order condition becomes

$$(10.22) \quad -U'(C_t) + \beta E_t U'(C_{t+1})(\alpha A_{t+1} K_t^{\alpha-1} + 1 - \delta) = 0.$$

In a steady state, we have  $\epsilon_t = 0$  so that  $A_t = 1$ . We can also take out the expectation as  $C_t$ ,  $K_t$ , and  $Y_t$  are constants. Thus, (10.22) implies

$$(10.23) \quad \beta(\alpha K^{\alpha-1} + 1 - \delta) = 1.$$

From (10.23) we can calculate the steady state solutions

$$(10.24) \quad \begin{aligned} K^* &= \left( \frac{1}{\alpha} \left( \frac{1}{\beta} - (1 - \delta) \right) \right)^{\frac{1}{\alpha-1}} \\ C^* &= K^{*\alpha} - \delta K^* \\ Y^* &= K^{*\alpha}. \end{aligned}$$

For solutions in other states, we need to take a log linearization using  $y = \ln x$  and

$$\begin{aligned}
(10.25) \quad f(x) &= f(e^y) \\
&= f(e^{y_0}) + \frac{\partial f(e^{y_0})}{\partial y} (y - y_0) \\
&= f(x_0) + \frac{\partial f(x_0)}{\partial x} \frac{1}{\partial \ln x_0 / \partial x} (\ln x - \ln x_0) \\
&= f(x_0) + f'(x_0) x_0 (\ln x - \ln x_0)
\end{aligned}$$

Plug (10.25) into (10.22), then

$$\begin{aligned}
(10.26) \quad U'(C_0) + U''(C_0)C_0\hat{C}_t &= \beta U'(C_0)(\alpha A_0 K_0^{\alpha-1} + 1 - \delta) \\
&+ \beta E_t U''(C_0)C_0(\alpha A_0 K_0^{\alpha-1} + 1 - \delta)\hat{C}_{t+1} \\
&+ \beta E_t U'(C_0)\alpha A_0 K_0^{\alpha-1}\hat{A}_{t+1} \\
&+ \beta E_t U''(C_0)\alpha(\alpha - 1)A_0 K_0^{\alpha-1}\hat{K}_t
\end{aligned}$$

where  $\hat{C}_t = \ln C_t - \ln C_0$ ,  $\hat{A}_t = \ln A_t - \ln A_0$ , and  $\hat{K}_t = \ln K_t - \ln K_0$ . By the property of the steady state, constant terms are cancelled out so that

$$\begin{aligned}
(10.27) \quad U''(C_0)C_0\hat{C}_t &= \beta E_t U''(C_0)C_0(\alpha A_0 K_0^{\alpha-1} + 1 - \delta)\hat{C}_{t+1} \\
&+ \beta E_t U'(C_0)\alpha A_0 K_0^{\alpha-1}\hat{A}_{t+1} + \beta E_t U''(C_0)\alpha(\alpha - 1)A_0 K_0^{\alpha-1}\hat{K}_t,
\end{aligned}$$

where  $E_t \hat{A}_{t+1} = \rho \hat{A}_t$ . Thus, this equation can be simplified by

$$(10.28) \quad \hat{C}_t = \tilde{A}_c E_t \hat{C}_{t+1} + \tilde{A}_k \hat{K}_t + \tilde{A}_a \hat{A}_t.$$

Since this equation contains two control variables, we further simplify it by replacing

$\hat{C}_t$  with the following log linearization of (10.21):

$$(10.29) \quad \hat{C}_t = A_0 K_0^\alpha \hat{A}_t + \alpha A_0 K_0^\alpha \hat{K}_{t-1} - K_0 \hat{K}_t + (1 - \delta)K_0 \hat{K}_{t-1}.$$

and finally we get

$$(10.30) \quad E_t \hat{K}_{t+1} + A_1 \hat{K}_t + A_2 \hat{K}_{t-1} = A_3 \hat{A}_t.$$

Let  $L^{-1}x_t$  denote  $E_t x_{t+1}$ , then (10.30) can be expressed by

$$(10.31) \quad (1 - B_1 L^{-1})(1 - B_2 L) \hat{K}_t = B_3 \hat{A}_t$$

or

$$(10.32) \quad \begin{aligned} (1 - B_2 L) \hat{K}_t &= (1 - B_1 L^{-1}) B_3 \hat{A}_t \\ &= B_3 \sum_{i=0}^{\infty} E_t B_1^i \hat{A}_{t+i} \\ &= B_3 \sum_{i=0}^{\infty} B_1^i \rho^i \hat{A}_t \\ &= B_3 \sum_{i=0}^{\infty} B_1^i \rho^i \hat{A}_t. \end{aligned}$$

Thus, the solution of the model is given by

$$(10.33) \quad \hat{K}_t = C_{11} \hat{K}_{t-1} + C_{12} \hat{A}_t.$$

We can also get the solution for  $C_t$  by plugging (10.33) into (10.29):

$$(10.34) \quad \hat{C}_t = C_{21} \hat{K}_{t-1} + C_{22} \hat{A}_t.$$

In general, we can always express the solutions of the model by

$$(10.35) \quad \begin{aligned} x_{t+1} &= \gamma_{xx} x_t + \gamma_{xz} z_t \\ \lambda_t &= \gamma_{\lambda x} x_t + \gamma_{\lambda z} z_t \\ u_t &= \gamma_{ux} x_t + \gamma_{uz} z_t, \end{aligned}$$

where  $x_t$  is a vector of state variables ( $K_{t-1}$ ),  $\lambda_t$  is a costate variable,  $u_t$  is a vector of control variables ( $C_t$ ), and  $z_t$  is a vector of exogenous variables ( $A_t$ ). Let the law

of motion for the exogenous variables be

$$(10.36) \quad z_t = \pi z_{t-1} + \epsilon_t,$$

then we get

$$(10.37) \quad \begin{aligned} \begin{bmatrix} x_{t+1} \\ z_{t+1} \end{bmatrix} &= \begin{bmatrix} \gamma_{xx} & \gamma_{xz} \\ 0 & \pi \end{bmatrix} \begin{bmatrix} x_t \\ z_t \end{bmatrix} + \begin{bmatrix} 0 \\ \epsilon_{t+1} \end{bmatrix} \\ &= Ms_t + \hat{\epsilon}_{t+1} \end{aligned}$$

or

$$(10.38) \quad s_{t+1} = Ms_t + \hat{\epsilon}_{t+1},$$

where  $s_t = (x_t, z_t)'$ . Let  $f_t$  be other variables of interest characterized by  $f_t = F_c u_t + F_x x_t + F_z z_t$ , then

$$(10.39) \quad \begin{aligned} \begin{bmatrix} \lambda_t \\ u_t \\ f_t \end{bmatrix} &= \begin{bmatrix} \gamma_{\lambda x} & \gamma_{\lambda z} \\ \gamma_{ux} & \gamma_{uz} \\ F_c \gamma_{ux} + F_x & F_c \gamma_{uz} + F_z \end{bmatrix} s_t \\ &= Hs_t. \end{aligned}$$

Therefore, provided with the parameters in the first order conditions and those in the law of motion for the exogenous variables, we can compute  $M$  and  $H$ . GMM is used to estimate the parameters. Once  $M$  and  $H$  are derived, we can compute the impulse response function and the autocovariance implied by the model. By taking an MA representation of (10.38), the  $h$ -step impulse response function of the  $i - th$  variable of  $(\lambda_t, u_t, f_t)'$  on the  $j - th$  shock of  $\hat{\epsilon}_t$  is given by

$$(10.40) \quad (HM^h)_{(i,j)}.$$

The autocovariance is computed by

$$(10.41) \quad \begin{aligned} \Gamma_i &= E(s_t s_{t-i}') \\ &= M^i \Gamma_0, \end{aligned}$$

where  $\Gamma_0 = E(s_t s_t')$  that is computed as follows. Let  $M = VDV^{-1}$  where  $D$  is a diagonal matrix that consists of eigen-values of  $M$ , and  $V$  is a matrix of corresponding eigen-vectors. By pre-multiplying  $V^{-1}$  on the both sides of (10.38), we get

$$(10.42) \quad V^{-1}s_{t+1} = DV^{-1}s_t + V^{-1}\hat{\epsilon}_{t+1}$$

or

$$(10.43) \quad \tilde{s}_{t+1} = D\tilde{s}_t + \tilde{\epsilon}_{t+1}.$$

Thus, we can compute the transformed autocovariance by

$$(10.44) \quad \begin{aligned} \tilde{\Gamma}_{0,ij} &= E(s_{it}s'_{jt}) \\ &= \frac{1}{1 - d_i d_j} \tilde{\Sigma}_{i,j} \end{aligned}$$

and

$$(10.45) \quad \Gamma_0 = V\tilde{\Gamma}_0V'.$$

We can also compute the autocovariance of other variables using

$$(10.46) \quad \begin{aligned} E(s_t w'_{t-i}) &= E(s_t (H s_{t-i})') = M^i \Gamma_0 H' \\ E(w_t w'_{t-i}) &= E(H s_t (H s_{t-i})') = H M^i \Gamma_0 H'. \end{aligned}$$

## 10.10 GMM and an ARCH Process

As explained in Chapter 2, an autoregressive conditional heteroskedastic (ARCH) process is frequently employed to model conditional heteroskedasticity. A typical estimation method for an ARCH model is the Maximum Likelihood (ML) estimator with the assumption that the conditional distribution of the error term follows normal

or t-distribution (see Bollerslev, Chou, and Kroner, 1992, for survey). However, ARCH models can also be estimated by GMM, which produces consistent estimates of the parameters without a specific distributional assumption (see, e.g., Mark, 1988; Simon, 1989). Further, as Rich, Raymond, and Butler (1991) point out, the GMM estimation directly allows for the specification test introduced by Hansen (1982).

An ARCH process is modeled as an innovation in the mean for some other stochastic process in most applications. Consider a regression model with ARCH( $q$ ) disturbances.

$$(10.47) \quad y_t = \mathbf{x}'_{2,t} \boldsymbol{\beta} + \epsilon_t$$

$$(10.48) \quad E(\epsilon_t | \mathbf{I}_{t-1}) = 0$$

$$(10.49) \quad E(\epsilon_t^2 | \mathbf{I}_{t-1}) = h_t$$

$$(10.50) \quad h_t = \alpha + \sum_{i=1}^q \gamma_i \epsilon_{t-i}^2; \quad \alpha > 0, \quad \sum_{i=1}^q \gamma_i < 1, \quad \gamma_i \geq 0$$

where  $y_t$  is the dependent variable,  $\mathbf{x}_{2,t}$  is a vector of explanatory variables in the information set  $\mathbf{I}_{t-1}$  which is assumed to be  $\mathbf{I}_{t-1} \subset \mathbf{I}_t$  for any  $t$  and  $\boldsymbol{\beta}, \alpha$  and  $\gamma$  are fixed parameters.

To apply GMM, Rich, Raymond, and Butler (1991) rewrite equations (10.47) and (10.49) as:

$$(10.51) \quad y_t = \mathbf{x}'_{2,t} \boldsymbol{\beta} + \epsilon_t$$

$$(10.52) \quad \epsilon_t^2 = \alpha + \sum_{i=1}^q \gamma_i \epsilon_{t-i}^2 + \eta_t$$

where

$$(10.53) \quad \eta_t = \epsilon_t^2 - h_t, \quad E(\eta_t | I_{t-1}) = 0$$

From these, we can obtain a system of two equations describing the innovations to the mean and variance of the ARCH( $q$ ) process, respectively,

$$(10.54) \quad \epsilon_t = y_t - \mathbf{x}'_{2,t} \boldsymbol{\beta}$$

$$(10.55) \quad \eta_t = (y_t - \mathbf{x}'_{2,t} \boldsymbol{\beta})^2 - \alpha - \sum_{i=1}^q \gamma_i (y_{t-i} - \mathbf{x}'_{2,t-i} \boldsymbol{\beta})^2$$

Let  $\tilde{\mathbf{b}}$  be the  $n$ -dimensional vector of parameters  $(\tilde{\boldsymbol{\beta}}', \tilde{\alpha}, \tilde{\boldsymbol{\gamma}})'$  of the ARCH model and  $\mathbf{x}_t = (y_t, \mathbf{x}'_{2,t})'$ . Let  $\mathbf{g}(\mathbf{x}_t, \tilde{\mathbf{b}})$  be a 2-dimensional vector of functions, then

$$(10.56) \quad \mathbf{g}(\mathbf{x}_t, \mathbf{b}_0) = \begin{bmatrix} \epsilon_t(\boldsymbol{\beta}) \\ \eta_t(\boldsymbol{\beta}, \alpha, \boldsymbol{\gamma}) \end{bmatrix}$$

$$(10.57) \quad E(\mathbf{g}(\mathbf{x}_t, \mathbf{b}_0) | I_{t-1}) = \mathbf{0}$$

where  $\mathbf{b}_0 = (\boldsymbol{\beta}', \alpha, \boldsymbol{\gamma}')'$  is the true parameter.

Suppose  $\mathbf{z}_{t-1}^1$  and  $\mathbf{z}_{t-1}^2$  are an  $(m_1 \times 1)$  and an  $(m_2 \times 1)$  vector of random variables in the information set  $I_{t-1}$ , uncorrelated with  $\epsilon_t$  and  $\eta_t$ , respectively, to serve as instrumental variables. Let  $\mathbf{z}_{t-1}$  be  $(m \times 2)$  block diagonal matrix where  $m = m_1 + m_2$ ,

$$(10.58) \quad \mathbf{z}_{t-1} = \begin{bmatrix} \mathbf{z}_{t-1}^1 & 0 \\ 0 & \mathbf{z}_{t-1}^2 \end{bmatrix}$$

By the law of iterative expectations, we obtain unconditional moment restrictions:

$$(10.59) \quad E(\mathbf{z}_{t-1} \mathbf{g}(\mathbf{x}_t, \mathbf{b}_0)) = \mathbf{0}$$

Equation (10.59) represents a set of  $m$  orthogonality conditions which are used to estimate  $\mathbf{b}_0$  with  $\mathbf{z}_{t-1}$  serving instruments in the ARCH model. Based on this procedure, Rich, Raymond, and Butler (1991) obtain results similar to ML estimates of Engle and Kraft's (1983) ARCH model of U.S. inflation.

This GMM framework can be extended to the generalized ARCH model, GARCH( $p, q$ ), where equation (10.50) allows for autoregressive components in the heteroskedastic variance:

$$(10.60) \quad h_t = \alpha + \sum_{i=1}^q \gamma_i \epsilon_{t-i}^2 + \sum_{j=1}^p \delta_j h_{t-j}$$

where  $\alpha > 0$ ,  $\sum_{i=1}^q \gamma_i < 1$ ,  $\gamma_i \geq 0$ ,  $\sum_{j=1}^p \delta_j < 1$ ,  $\delta_j \geq 0$ . In this case, we can still get the same moment conditions, equation (10.59), where  $\mathbf{b}_0 = (\boldsymbol{\beta}', \alpha, \boldsymbol{\gamma}', \boldsymbol{\delta}')$  is the true parameter.

## 10.11 Estimation and Testing of Linear Rational Expectations Models

In this section, econometric methods that impose and test the restrictions implied by linear rational expectations models are described. Many linear rational expectations models imply that an economic variable depends on a geometrically declining weighted sum of expected future values of another variable

$$(10.61) \quad y_t = aE\left(\sum_{i=1}^{\infty} \beta^i x_{t+i} | \mathbf{I}_t\right) + \mathbf{c}' \mathbf{z}_t,$$

where  $a$  and  $\beta$  are constants,  $\mathbf{c}$  is a vector of constants,  $y_t$  and  $x_t$  are random variables, and  $\mathbf{z}_t$  is a random vector. This implication imposes nonlinear restrictions on the VAR representation of  $y_t$ ,  $x_t$ , and  $\mathbf{z}_t$  as shown by Hansen and Sargent (1980). In

Section 10.11.1, these nonlinear restrictions are discussed. Section 10.11.2 describes econometric methods to utilize these restrictions.

### 10.11.1 The Nonlinear Restrictions

Consider West's (1987) model as an example of linear rational expectations model. Let  $p_t$  be the real stock price (after the dividend is paid) in period  $t$  and  $d_t$  be the real dividend paid to the owner of the stock at the beginning of period  $t$ . Then the arbitrage condition is

$$(10.62) \quad p_t = E[\beta(p_{t+1} + d_{t+1})|\mathbf{I}_t],$$

where  $\beta$  is the constant real discount rate,  $\mathbf{I}_t$  is the information set available to economic agents in period  $t$ . Solving (10.62) forward and imposing the no bubble condition, we obtain the present value formula:

$$(10.63) \quad p_t = E\left(\sum_{i=1}^{\infty} \beta^i d_{t+i}|\mathbf{I}_t\right).$$

We now derive restrictions for  $p_t$  and  $d_t$  implied by (10.63). Many linear rational expectations models imply that a variable is the expectation of a discounted infinite sum conditional on an information set. Hence similar restrictions can be derived for these rational expectations models. We consider two cases, depending on whether  $d_t$  is assumed to be covariance stationary or is unit root nonstationary.

Assume that  $d_t$  is covariance stationary with mean zero (imagine that data are demeaned), so that it has a Wold moving average representation

$$(10.64) \quad d_t = \alpha(L)\nu_t,$$

where  $\alpha(L) = 1 + \alpha_1 L + \alpha_2 L^2 + \dots$  and where

$$(10.65) \quad \nu_t = d_t - \hat{E}(d_t|\mathbf{H}_{t-1}).$$

Here,  $\hat{E}(\cdot|\mathbf{H}_t)$  is the linear projection operator onto the information set  $\mathbf{H}_t = \{d_t, d_{t-1}, d_{t-2}, \dots\}$ .

We assume that the econometrician uses the information set  $\mathbf{H}_t$ , which may be much smaller than the economic agents' information set,  $\mathbf{I}_t$ . Assuming that  $\alpha(L)$  is invertible,

$$(10.66) \quad \phi(L)d_t = \nu_t,$$

where  $\phi(L) = 1 - \phi_1 L - \phi_2 L^2 - \dots$ .

Using (10.63) and the law of iterated projections, we obtain

$$(10.67) \quad p_t = \hat{E}\left(\sum_{i=1}^{\infty} \beta^i d_{t+i} | \mathbf{H}_t\right) + w_t,$$

where

$$(10.68) \quad w_t = E\left(\sum_{i=1}^{\infty} \beta^i d_{t+i} | \mathbf{I}_t\right) - \hat{E}\left(\sum_{i=1}^{\infty} \beta^i d_{t+i} | \mathbf{H}_t\right),$$

and  $\hat{E}(w_t | \mathbf{H}_t) = 0$ . Since  $\hat{E}(\cdot | \mathbf{H}_t)$  is the linear projection operator onto  $\mathbf{H}_t$ ,

$$(10.69) \quad \hat{E}\left(\sum_{i=1}^{\infty} \beta^i d_{t+i} | \mathbf{H}_t\right) = \delta(L)d_t,$$

where  $\delta(L) = \delta_1 + \delta_2 L + \dots$ . Following Hansen and Sargent (1980, Appendix A), we obtain the restrictions imposed by (10.69) on  $\delta(L)$  and  $\phi(L)$ . The left-hand side of (10.69) can be written

$$(10.70) \quad \begin{aligned} \hat{E}\left(\sum_{i=1}^{\infty} \beta^i d_{t+i} | \mathbf{H}_t\right) &= \hat{E}\left(\frac{\beta L^{-1}}{1 - \beta L^{-1}} d_t | \mathbf{H}_t\right) \\ &= \left[\frac{\beta L^{-1} \alpha(L)}{1 - \beta L^{-1}}\right]_+ \nu_t \end{aligned}$$

where  $[B(L)]_+$  is an annihilator that removes negative power of the lag polynomial  $B(L)$ . The second equality holds because  $\nu_t$  is fundamental. Then by replacing  $L$  with  $z$  in (10.70), we have

$$(10.71) \quad \frac{\beta z^{-1} \alpha(z)}{1 - \beta z^{-1}} = \frac{\beta z^{-1} (\alpha(z) - \alpha(\beta))}{1 - \beta z^{-1}} + \frac{\beta z^{-1} \alpha(\beta)}{1 - \beta z^{-1}}.$$

Note that the first term in the right-hand side is removable singularity and the second term has only negative power of lag polynomial that is to be removed by the annihilator. Therefore we can write (10.70) as

$$(10.72) \quad \left[ \frac{\beta L^{-1} \alpha(L)}{1 - \beta L^{-1}} \right]_+ \nu_t = \left[ \frac{\beta z^{-1} (\alpha(z) - \alpha(\beta))}{1 - \beta L^{-1}} + \frac{\beta L^{-1} \alpha(\beta)}{1 - \beta L^{-1}} \right]_+ \nu_t \\ = \frac{\beta L^{-1} (\alpha(L) - \alpha(\beta))}{1 - \beta L^{-1}} \nu_t.$$

Since  $\nu_t = \phi(L)d_t$  as in (10.64), we have the following restriction

$$(10.73) \quad \delta(L) = \frac{\beta L^{-1} (\alpha(L) - \alpha(\beta))}{1 - \beta L^{-1}} \phi(L)$$

$$(10.74) \quad = \frac{\beta L^{-1} (1 - \phi^{-1}(\beta) \phi(L))}{1 - \beta L^{-1}}.$$

We now parameterize  $\phi(L)$  as a  $q$ -th order polynomial:

$$(10.75) \quad d_t = \phi_1 d_{t-1} + \cdots + \phi_q d_{t-q} + \nu_t.$$

Then, by using state space representation, (10.75) can be written as

$$(10.76) \quad D_t = AD_{t-1} + V_t$$

where  $D_t = (d_t, d_{t-1}, \dots, d_{t-q+1})'$  and

$$(10.77) \quad A = \begin{bmatrix} \phi_1 & \cdots & \cdots & \phi_q \\ 1 & & & 0 \\ & \ddots & & \vdots \\ & & 1 & 0 \end{bmatrix}$$

Then (10.67) can be written as

$$(10.78) \quad p_t = \hat{E} \left( \sum_{i=1}^{\infty} \beta^i d_{t+i} \mid \mathbf{H}_t \right) + w_t \\ = e_1 \beta A (I - \beta A)^{-1} D_t + w_t$$

where  $e_1 = (1, 0, \dots, 0)'$ .

Also (10.73) is used to show that  $\delta(L)$  is a finite order polynomial and to give a explicit formula for the coefficients for  $\delta(L)$ .<sup>13</sup> Thus

$$(10.79) \quad p_t = \delta_1 d_t + \dots + \delta_q d_{t-q+1} + w_t,$$

where  $\delta_i$ 's are functions of  $\beta$  and  $\phi_i$ 's. Comparing (10.78) and (10.79) yields the following nonlinear restriction

$$(10.80) \quad \delta_1 = \{1 - \phi(\beta)\}^{-1}$$

$$\delta_j = \delta\gamma(\beta)\{1 - \delta\phi(\beta)\}^{-1}(\phi_{j+1} + \beta\phi_{j+2} + \dots + \beta^p\phi_{j+p+1}) \quad \text{for } j = 2, \dots, p.$$

(?????γ(β)?) These are the nonlinear restrictions which (10.63) implies.

Masao needs to check this!

**Example 10.1** Consider the case where  $d_t$  is an AR(1) process, so that  $d_t = \phi_1 d_{t-1} + \nu_t$  where  $|\phi_1| < 1$ . Then  $\hat{E}(d_{t+i} | H_t) = \phi_1^i d_t$ , and hence  $\hat{E}(\sum_{i=1}^{\infty} \beta^i d_{t+i} | H_t) = \sum_{i=1}^{\infty} \beta^i \phi_1^i d_t = \frac{\beta\phi_1}{1-\beta\phi_1} d_t$ . Hence  $p_t = \delta_1 d_t + w_t$  where  $\delta_1 = \frac{\beta\phi_1}{1-\beta\phi_1}$ . ■

### 10.11.2 Econometric Methods

We focus on Hansen and Sargent's (1982) method which applies Hansen's (1982) Generalized Method of Moments (GMM) to linear rational expectations models.

Let  $\mathbf{z}_{1t}$  be a vector of random variables in  $H_t$ . For example,  $\mathbf{z}_{1t} = (d_t, \dots, d_{t-q+1})'$ . The unknown parameters  $\beta$  and  $\phi_i$ 's can be estimated by applying the GMM to orthogonality conditions  $E(\mathbf{z}_{1t}\nu_{t+1}) = \mathbf{0}$  and  $E(\mathbf{z}_{1t}w_t) = \mathbf{0}$  in the econometric system consisting of (10.75) and (10.79).

Let  $z_{2t}$  be a random variable in  $I_t$ , say  $d_t$ , and

$$(10.14) \quad p_t = \beta(p_{t+1} + d_{t+1}) + u_{t+1}.$$

<sup>13</sup>See West (1987), for the formula, which is based on Hansen and Sargent (1980), and on West (1988), for deterministic terms when  $d_t$  has a nonzero mean.

Then (10.62) implies another orthogonality condition  $E(z_{2t}u_{t+1}) = 0$ . This orthogonality condition can be used to estimate  $\beta$ . West (1987) forms a specification test *à la* Hausman (1978) by comparing the estimate of  $\beta$  from (10.14) with the estimate of  $\beta$  from (10.75) and (10.79). For this purpose, West forms a Wald test in the system consisting of (10.75), (10.79), and (10.14) without the restrictions (10.80) imposed. Another method to form West's specification test is to form a Lagrange Multiplier test or a likelihood ratio type test, which will require estimation constrained by the restrictions (10.80). This method may be preferable because of small sample problems with the Wald test for nonlinear restrictions (see Chapter 9 for discussions about these tests).

Some remarks are in order.

- (A) Hansen and Sargent's method described above does not require an assumption that  $d_t$  is exogenous. Relation (10.75) or (??) is obtained from the assumption that  $d_t$  is covariance stationary and that its Wold representation is invertible.
- (B) For the econometric system consisting of (10.75) and (10.79) (or (??) and (??)), random variables in  $H_t$  can be used as instruments, but the variables in  $I_t$  that are not in  $H_t$  are not valid instruments by construction.
- (C) Since  $u_{t+1}$  in (10.14) is in  $I_{t+1}$  and  $\nu_{t+1}$  in (10.75) is in  $H_{t+1}$ ,  $u_{t+1}$  and  $\nu_{t+1}$  are serially uncorrelated (see, e.g., Ogaki, 1993a, Section 6, for related discussions). However,  $w_t$  in (10.79) is not necessarily in  $H_{t+1}$ . Hence  $w_t$  has unknown order of serial correlation.

Masao  
needs to  
check this!

## 10.12 GMM for Consumption Euler Equations with Measurement Error

When data are contaminated by measurement error, the standard non-linear GMM yields inconsistent estimates (Garber and King, 1983; Amemiya, 1985). Such problem arises, for instance, in estimation of structural parameters in a non-linear consumption Euler equation when the consumption data contain measurement error.

To remedy this problem, Alan, Attanasio, and Browning (2005) propose two GMM estimators for consumption Euler equations in the presence of measurement error in data. Consider a simple life-cycle model with intertemporally additive and instantaneously iso-elastic utility. Under the assumption of rational expectations, a consumer's utility maximization yields the Euler equation,

$$(10.15) \quad E_t \left[ \beta \left( \frac{C_{t+1}^*}{C_t^*} \right)^{-\alpha} R_{t+1} \right] = 1,$$

where  $C_t^*$  is true consumption,  $R_{t+1}$  the gross real interest rate,  $\alpha$  the coefficient of relative risk aversion, and  $\beta < 1$  the discount factor. Call  $\beta (C_{t+1}^*/C_t^*)^{-\alpha} R_{t+1}$  an expectational error uncorrelated with the time  $t$  information. We wish to estimate the preference parameters  $\alpha$  and  $\beta$ . Suppose consumption data are observed with multiplicative error  $\epsilon_t$ :

$$C_t = C_t^* \epsilon_t,$$

where  $C_t$  is the observed consumption. Assume that the measurement error is stationary, serially uncorrelated, and uncorrelated with  $C_t^*$ ,  $R_t$ , and the expectational error for all  $t$ . Then, taking the expectations conditional on the time  $t$  information,

we can write

$$(10.16) \quad E_t \left[ \beta \left( \frac{C_{t+1}}{C_t} \right)^{-\alpha} R_{t+1} | \mathbf{I}_t \right] = E_t \left[ \beta \left( \frac{C_{t+1}^*}{C_t^*} \right)^{-\alpha} R_{t+1} | \mathbf{I}_t \right] E_t \left[ \left( \frac{\epsilon_{t+1}}{\epsilon_t} \right)^{-\alpha} | \mathbf{I}_t \right] = \kappa,$$

where  $\kappa$  is a constant. The first equality follows from the assumption that the measurement error is independent of the expectational error, and the second equality follows from the Euler equation (16.11) and the stationarity assumption of the measurement error. For  $\kappa \neq 1$ , equation (16.57) implies that the standard GMM without consideration for the measurement error would result in inconsistent estimates of  $\alpha$  and  $\beta$ . Similarly, consider the Euler equation representing the change in marginal utility between time  $t$  and  $t + 2$ :

$$E_t \left[ \beta^2 \left( \frac{C_{t+2}}{C_t} \right)^{-\alpha} R_{t+1} R_{t+2} | \mathbf{I}_t \right] = \kappa.$$

Now define

$$(10.17) \quad \begin{aligned} u_{t+1}^1 &\equiv \left[ \beta \left( \frac{C_{t+1}}{C_t} \right)^{-\alpha} R_{t+1} - \kappa \right], \\ u_{t+2}^2 &\equiv \left[ \beta^2 \left( \frac{C_{t+2}}{C_t} \right)^{-\alpha} R_{t+1} R_{t+2} - \kappa \right], \end{aligned}$$

where, by definition,  $u_{t+1}^1$  and  $u_{t+2}^2$  are uncorrelated with the time  $t$  information and  $E_t(u_{t+1}^1) = E_t(u_{t+2}^2) = 0$ .

The first estimator, the *GMM-LN estimator*, additionally assumes that the measurement error is log-normally distributed with mean  $\mu$  and variance  $\sigma^2$ . Let  $u_{t+2} = [ u_{t+1}^1 \quad u_{t+2}^2 ]'$  and  $z_t = [ c \quad z_{1t} ]'$  where  $c$  is a constant and  $z_{1t}$  is an instrument such as the lagged interest rate. Estimates of the parameters,  $\alpha$ ,  $\beta$ , and  $\kappa$  are obtained from four orthogonality conditions:

$$(10.18) \quad E\{u_{t+2} \otimes z_t\} = 0.$$

Under the assumption of log-normality,  $\kappa$  can be written as

$$(10.19) \quad \kappa = \exp(\alpha^2 \sigma^2).$$

Once  $\alpha$  and  $\kappa$  are estimated using the orthogonality conditions (16.61), the estimate of the variance of measurement error  $\sigma^2$  can be obtained from equation (16.58).

The second estimator, the *GMM-D estimator*, simply assumes stationarity and does not require any distributional assumption. Subtracting  $u_{t+2}^2$  from  $u_{t+1}^1$  in equations (16.60) yields

$$(10.20) \quad v_{t+2} = \left[ \beta \left( \frac{C_{t+1}}{C_t} \right)^{-\alpha} R_{t+1} \right] - \left[ \beta^2 \left( \frac{C_{t+2}}{C_t} \right)^{-\alpha} R_{t+1} R_{t+2} \right],$$

where  $v_{t+2}$  has zero mean and is independent of the time  $t - 1$  information. The orthogonality conditions for the GMM-D estimator are derived using equation (16.62) and a vector of instruments  $z_t$ . Note that because equation (16.62) takes the difference of the consumption growth (double-differencing), the GMM-D estimator is expected to be less precise than the GMM-LN estimator.<sup>14</sup>

Results from the Monte Carlo simulation in Alan, Attanasio, and Browning (2005) suggest that both proposed methods perform significantly better than conventional GMM estimators based on the log-linearized Euler equation or the exact Euler equation that ignores measurement error, especially when the panel length is short. In particular, both capture the true value of  $\beta$  remarkably well. They also report that when the measurement error is lognormally distributed, the distribution of  $\alpha$  is more dispersed under the GMM-D estimator than under the GMM-LN estimator.

---

<sup>14</sup>In the presence of measurement error, the lagged consumption growth rate - a common choice for an instrument in estimation of consumption Euler equations - would be invalid since it is correlated with  $u_{t+2}$ . Instead, one should use the consumption growth rate with two-period lags. On the other hand, a one-period lag is sufficient for the interest rates since they are unlikely to be correlated with the measurement error (Alan, Attanasio, and Browning, 2005).

## Exercises

**10.1** (Computer Exercise) In the text we considered four alternative measures of the intertemporal marginal rate of substitution,  $m_t$ :

$$(i) \quad m_t = \beta \left( \frac{c_{t+1}}{c_t} \right)^{-\gamma} \quad (\text{Hansen and Singleton})$$

$$(ii) \quad m_t = \beta^* (R_{t+1}^m)^{-\gamma} \quad (\text{Brown and Gibbons})$$

$$(iii) \quad m_t = \beta^+ (R_{t+1}^m)^\eta \left( \frac{c_{t+1}}{c_t} \right)^\theta \quad (\text{Epstein and Zin})$$

$$(iv) \quad m_t = \frac{\beta \{ S_{t+1}^{-\gamma} + \beta a_1 s_{t+2}^{-\gamma} \}}{E(S_t^{-\gamma} + \beta a_1 s_{t+1}^{-\gamma} | I_t)} \quad \text{where } s_t = c_t + a_1 c_{t-1} \quad (\text{Ferson and Constantinides}).$$

- (a) For each of the four alternative measures, estimate the unknown parameters and test the overidentifying restrictions implied by the asset pricing relation  $E(m_t R_{t+1}) = 1$ . Use quarterly data on nondurables and services for consumption  $c_t$ , real value-weighted returns from the New York Stock exchange for  $R_t^m$ , and ex post real returns on Treasury Bill returns for  $R_t$ . Use a constant, one-period and two-period lagged values of  $\frac{c_{t+1}}{c_t}$ , and one-period and two-period lagged values of  $R_{t+1}$  for instrumental variables. You can modify the `GMM.EXP` file for models (i), (ii), and (iii), and `GMMHF.EXP` for model (iv). Note that `GMM.EXP` uses monthly data and `GMMHF.EXP` uses quarterly data. You will need to modify `GMM.EXP` to use the quarterly data used by `GMMHF.EXP`. For Ferson and Constantinides's, report results for both the truncated kernel and the non-prewhitened QS kernel.

For each measure, state what value “`mas`” should take in the GMM program and explain why. Comment on the relative strengths and weaknesses of the four measures of  $m_t$  from both a theoretical and an empirical perspective.

Print out the “hu” procedure part of your program and the final GMM iteration output for each model and submit them.

- (b) Repeat the analysis in question (a) for the model (iii) using simultaneously the additional moment restrictions obtained by letting  $R_t = R_t^m$ . There is a difficulty in interpreting the empirical result for this case of multiple returns. What is the difficulty?

**10.2** Let  $p_t$  be the real stock price,  $d_t$  be the real dividend, and  $\beta$  be the constant ex ante discount rate. Assume that  $p_t$  and  $d_t$  are stationary with zero mean and finite second moments. The stock price satisfies

$$(10.E.1) \quad p_t = \beta E(p_{t+1} + d_{t+1} | I_t),$$

where  $I_t$  is the information set available at period  $t$ . We assume that  $I_t$  is generated from  $\mathbf{x}_t, \mathbf{x}_{t-1}, \dots$ , where  $\mathbf{x}_t$  is a random vector that includes  $p_t$  and  $d_t$  as its components. Solving (10.E.1) forward with the no bubble condition imposed, we obtain the present value formula:

$$(10.E.2) \quad p_t = \sum_{\tau=1}^{\infty} \beta^\tau E(d_{t+\tau} | I_t)$$

Suppose that  $d_t$  is stationary with zero mean and finite second moments and let  $H_t$  be the information set generated by the linear functions of  $\{d_t, d_{t-1}, d_{t-2}, \dots\}$ . Assume

$$(10.E.3) \quad \hat{E}(d_t | H_{t-1}) = \phi d_{t-1},$$

where  $|\phi| < 1$ , and  $\hat{E}(\cdot | H_{t-1})$  is the linear projection operator on  $H_t$ . Answer the following questions.

- (a) Suppose that you run a regression

$$(10.E.4) \quad p_t = \delta d_t + w_t.$$

Your estimator for  $\delta$  will converge to a number that can be expressed in terms of  $\phi$ , and  $\beta$ . Derive this expression for  $\delta$ . Show that  $\hat{E}(w_t|H_t) = 0$ . Is it possible to prove that  $E(w_t|I_t) = 0$ ? Explain.

- (b) Discuss whether or not  $w_t$  is serially correlated in general. If we make an additional assumption that  $p_t$  is in  $H_{t+1}$ , can you show that  $w_t$  is serially uncorrelated? Is this additional assumption realistic? Why?

- (c) Explain how to use (10.E.4),

$$(10.E.5) \quad d_{t+1} = \phi d_t + v_{t+1},$$

and

$$(10.E.6) \quad p_t = \beta(p_{t+1} + d_{t+1}) + u_t$$

to estimate  $\beta$  and  $\phi$  in the framework of the Generalized Method of Moments, imposing the restriction on  $\delta$  you derived. In particular, discuss the parameterized disturbances, valid instrumental variables, and appropriate methods to estimate the weighting matrix.

- (d) List three tests that can be used to test the restriction on  $\delta$  you derived. Discuss which tests may be better.

**10.3** Let  $p_t$  be the log price level and  $m_t$  be the log money supply. A version of the Cagan's hyperinflation model assume that the demand for real money balance is

$$(10.E.7) \quad m_t - p_t = \alpha(E(p_{t+1}|I_t) - p_t),$$

where  $I_t$  is the information set of the consumer,  $\alpha$  is a negative number, and  $-\alpha$  is the interest semi-elasticity of money demand when the real interest rate is assumed to be zero. Solving (10.E.7) as a difference equation for  $E(p_{t+i})$  for a fixed  $t$ , and imposing the stability condition that the solution for  $p_t$  is bounded for all bounded sequences of  $m_t$ , we obtain

$$(10.E.8) \quad p_t = \frac{1}{1-\alpha} E\left(\sum_{i=0}^{\infty} \left(\frac{\alpha}{\alpha-1}\right)^i m_{t+i} | I_t\right).$$

Suppose that  $m_t$  is stationary with zero mean and finite second moments (imagine that the data are already demeaned and detrended) and let  $H_t$  be the information set generated by the linear functions of  $\{m_t, m_{t-1}, m_{t-2}, \dots\}$ . Assume

$$(10.E.9) \quad \hat{E}(m_{t+1} | H_t) = \phi m_t,$$

where  $|\phi| < 1$ , and  $\hat{E}(\cdot | H_t)$  is the linear projection operator on  $H_t$ . Answer the following questions.

- (a) Suppose that you run a regression

$$(10.E.10) \quad p_t = \delta m_t + w_t$$

Your estimator for  $\delta$  will converge to a number that can be expressed in terms of  $\phi$ , and  $\alpha$ . Derive this expression for  $\delta$  (note that the summation in (10.E.8) starts from  $i = 0$  unlike West's present value model of the stock price in which the summation starts from  $i = 1$ ).

- (b) Discuss whether or not  $w_t$  is serially correlated in general. If we make an additional assumption that  $p_t$  is in  $H_{t+1}$ , can you show that  $w_t$  is serially uncorrelated? Is this additional assumption realistic? Why?

(c) Explain how to estimate  $\alpha$  from the equation (10.E.7) with a time series data set on  $m_t$  and  $p_t$ .

(d) Explain how to use (10.E.9), (10.E.10), and

$$(10.E.11) \quad m_{t+1} = \phi m_t + v_{t+1}$$

to estimate  $\alpha$  and  $\phi$  in the framework of the Generalized Method of Moments, imposing the restriction on  $\delta$  you derived in (i). In particular, discuss the parameterized disturbances, valid instrumental variables, and appropriate methods to estimate the weighting matrix.

(e) List three tests that can be used to test the restrictions on  $\delta$  you derived in (i). Discuss which tests may be better.

## References

- ABEL, A. B. (1990): "Asset Prices under Habit Formation and Catching Up with the Joneses," *American Economic Review*, 80(2), 38–42.
- AGUIAR, M., AND G. GOPINATH (2007): "Emerging Market Business Cycles: The Cycle Is the Trend," *Journal of Political Economy*, 115, 69–102.
- AI, C., AND X. CHEN (2003): "Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions," *Econometrica*, 71(6), 1795–1843.
- ALAN, S., O. ATTANASIO, AND M. BROWNING (2005): "Estimating Euler Equations with Noisy Data: Two Exact GMM Estimators," CAM Working Papers 2005-10, University of Copenhagen.
- ALEXOPOULOS, M. (2004): "Unemployment and the Business Cycle," *Journal of Monetary Economics*, 51(2), 277–298.
- AMBLER, S., E. CARDIA, AND C. ZIMMERMANN (2004): "International Business Cycles: What are the Facts?," *Journal of Monetary Economics*, 51(2), 257–276.
- AMEMIYA, T. (1985): *Advanced Econometrics*. Harvard University Press, Cambridge, Massachusetts.
- ATKESON, A., AND M. OGAKI (1996): "Wealth-Varying Intertemporal Elasticities of Substitution: Evidence from Panel and Aggregate Data," *Journal of Monetary Economics*, 38, 507–534.
- BACKUS, D. K., A. W. GREGORY, AND S. E. ZIN (1989): "Risk Premiums in the Term Structure - Evidence from Artificial Economies," *Journal of Monetary Economics*, 24(3), 371–399.

- BACKUS, D. K., AND P. J. KEHOE (1992): "International Evidence on the Historical Properties of Business Cycles," *American Economic Review*, 82(4), 864–888.
- BARSKY, R. B., AND J. A. MIRON (1989): "The Seasonal Cycle and the Business-Cycle," *Journal of Political Economy*, 97(3), 503–534.
- BEAULIEU, J. J., AND J. A. MIRON (1991): "The Seasonal Cycle in U.S. Manufacturing," *Economics Letters*, 37(2), 115–118.
- BOHN, H. (1991): "On Cash-in-Advance Models of Money Demand and Asset Pricing," *Journal of Money, Credit, and Banking*, 23, 224–242.
- BOLLERSLEV, T., R. Y. CHOU, AND K. F. KRONER (1992): "ARCH Modeling in Finance: A Review of the Theory and Empirical Evidence," *Journal of Econometrics*, 52, 5–59.
- BOSSAERTS, P. (1988): "Common Nonstationary Components of Asset Prices," *Journal of Economic Dynamics and Control*, 12, 347–364.
- BRAUN, R. A., AND C. L. EVANS (1998): "Seasonal Solow Residuals and Christmas: A Case for Labor Hoarding and Increasing Returns," *Journal of Money, Credit, and Banking*, 30(3), 306–330.
- BROWN, D. P., AND M. R. GIBBONS (1985): "A Simple Econometric Approach for Utility-Based Asset Pricing Models," *Journal of Finance*, 40, 359–381.
- BURNSIDE, C. (1999): "Real Business Cycle Models: Linear Approximation and GMM Estimation," Manuscript.
- BURNSIDE, C., M. EICHENBAUM, AND S. REBELO (1993): "Labor Hoarding and the Business Cycle," *Journal of Political Economy*, 101(2), 245–273.
- CAMPBELL, J. Y., AND J. H. COCHRANE (1999): "By Force of Habit: A Consumption-Based Explanation of Aggregate Stock Market Behavior," *Journal of Political Economy*, 107(2), 205–251.
- (2000): "Explaining the Poor Performance of Consumption-Based Asset Pricing Models," *Journal of Finance*, 55(6), 2863–2878.
- CHEN, X., AND S. C. LUDVIGSON (2004): "Land of Addicts? An Empirical Investigation of Habit-Based Asset Pricing Behavior," NBER Working Paper No. W10503.
- CHRISTIANO, L. J., AND M. EICHENBAUM (1992): "Current Real-Business-Cycle Theories and Aggregate Labor-Market Fluctuations," *American Economic Review*, 82(3), 430–450.
- CHRISTIANO, L. J., M. EICHENBAUM, AND D. MARSHALL (1991): "The Permanent Income Hypothesis Revisited," *Econometrica*, 59(2), 397–423.
- COCHRANE, J. H. (1989): "The Sensitivity of Tests of the Intertemporal Allocation of Consumption to Near-Rational Alternatives," *American Economic Review*, 79(3), 319–337.
- CONSTANTINIDES, G. M. (1990): "Habit Formation: A Resolution of the Equity Premium Puzzle," *Journal of Political Economy*, 98, 519–543.
- COOLEY, T. F., AND M. OGAKI (1996): "A Time Series Analysis of Real Wages, Consumption, and Asset Returns," *Journal of Applied Econometrics*, 11(2), 119–134.

- DUFFIE, D., AND K. J. SINGLETON (1993): "Simulated Moments Estimation of Markov-Models of Asset Prices," *Econometrica*, 61(4), 929–952.
- DUNN, K. B., AND K. J. SINGLETON (1986): "Modeling the Term Structure of Interest Rates under Non-Separable Utility and Durability of Goods," *Journal of Financial Economics*, 17, 27–55.
- DYNAN, K. E. (2000): "Habit Formation in Consumer Preferences: Evidence from Panel Data," *American Economic Review*, 90(3), 391–406.
- ECKSTEIN, Z., AND L. LEIDERMAN (1989): "Estimating an Intertemporal Model of Consumption, Money Demand, and Seignorage," Manuscript.
- EICHENBAUM, M., AND L. P. HANSEN (1990): "Estimating Models with Intertemporal Substitution Using Aggregate Time Series Data," *Journal of Business and Economic Statistics*, 8, 53–69.
- EICHENBAUM, M., L. P. HANSEN, AND K. J. SINGLETON (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., AND D. F. KRAFT (1983): "Multiperiod Forecast Error Variances of Inflation Estimated from ARCH Models," in *Applied Time Series Analysis of Economic Data*, ed. by A. Zellner, pp. 293–302. Bureau of the Census, Washington, DC.
- ENGLISH, W. B., J. A. MIRON, AND D. W. WILCOX (1989): "Seasonal Fluctuations and the Life-Cycle Permanent Income Model of Consumption: A Correction," *Journal of Political Economy*, 97(4), 988–991.
- EPSTEIN, L. G., AND S. E. ZIN (1991): "Substitution, Risk Aversion, and the Temporal Behavior of Consumption and Asset Returns: An Empirical Analysis," *Journal of Political Economy*, 99(2), 263–286.
- FERSON, W. E., AND G. M. CONSTANTINIDES (1991): "Habit Persistence and Durability in Aggregate Consumption: Empirical Tests," *Journal of Financial Economics*, 29(2), 199–240.
- FERSON, W. E., AND C. R. HARVEY (1992): "Seasonality and Consumption-Based Asset Pricing," *Journal of Finance*, 47, 511–552.
- FINN, M. G., D. L. HOFFMAN, AND D. E. SCHLAGENHAUF (1990): "Intertemporal Asset-Pricing Relationships in Barter and Monetary Economies: An Empirical Analysis," *Journal of Monetary Economics*, 25, 431–452.
- GARBER, P. M., AND R. G. KING (1983): "Deep Structural Excavation? A Critique of Euler Equation Methods," NBER Technical Working Paper No. 31.
- GHYSELS, E. (1990): "Unit-Root Tests and the Statistical Pitfalls of Seasonal Adjustment - The Case of United-States Postwar Real Gross-National-Product," *Journal of Business and Economic Statistics*, 8(2), 145–152.
- GROSSMAN, S. J., A. MELINO, AND R. SHILLER (1987): "Estimating the Continuous Time Consumption Based Asset Pricing Model," *Journal of Business and Economic Statistics*, 5, 315–327.
- HALL, R. E. (1978): "Stochastic Implications of the Life Cycle-Permanent Income Hypothesis: Theory and Evidence," *Journal of Political Economy*, 86, 971–987.

- (1988): “Intertemporal Substitution in Consumption,” *Journal of Political Economy*, 96, 339–357.
- HANSEN, L. P. (1982): “Large Sample Properties of Generalized Method of Moments Estimators,” *Econometrica*, 50(4), 1029–1054.
- HANSEN, L. P., AND R. JAGANNATHAN (1991): “Implications of Security Market Data for Models of Dynamic Economies,” *Journal of Political Economy*, 99, 225–262.
- HANSEN, L. P., AND S. F. RICHARD (1987): “The Role of Conditioning Information in Deducing Testable Restrictions Implied by Dynamic Asset Pricing Models,” *Econometrica*, 55, 587–613.
- HANSEN, L. P., AND T. J. SARGENT (1980): “Formulating and Estimating Dynamic Linear Rational Expectations Models,” *Journal of Economic Dynamics and Control*, 2(1), 7–46.
- (1982): “Instrumental Variables Procedures for Estimating Linear Rational Expectations Models,” *Journal of Monetary Economics*, 9(3), 263–296.
- (1983a): “Aggregation Over Time and the Inverse Optimal Predictor Problem for Adaptive Expectations in Continuous Time,” *International Economic Review*, 24, 1–20.
- (1983b): “The Dimensionality of the Aliasing Problem in Models with Rational Spectral Densities,” *Econometrica*, 51, 377–387.
- HANSEN, L. P., AND K. J. SINGLETON (1982): “Generalized Instrumental Variables Estimation of Nonlinear Rational Expectations Models,” *Econometrica*, 50(5), 1269–1286.
- (1984): “Generalized Instrumental Variables Estimation of Nonlinear Rational Expectations Models: Errata,” *Econometrica*, 52(1), 267–268.
- (1996): “Efficient Estimation of Linear Asset-Pricing Models with Moving Average Errors,” *Journal of Business and Economic Statistics*, 14, 53–68.
- HAUSMAN, J. A. (1978): “Specification Tests in Econometrics,” *Econometrica*, 46(6), 1251–1271.
- HAYASHI, F. (1982): “The Permanent Income Hypothesis: Estimation and Testing by Instrumental Variables,” *Journal of Political Economy*, 90, 895–916.
- HEATON, J. C. (1993): “The Interaction Between Time-Nonseparable Preferences and Time Aggregation,” *Econometrica*, 61, 353–385.
- (1995): “An Empirical Investigation of Asset Pricing with Temporally Dependent Preference Specifications,” *Econometrica*, 63, 681–717.
- HOFFMAN, D. L., AND A. R. PAGAN (1989): “Post-Sample Prediction Tests for Generalized-Method of Moments Estimators,” *Oxford Bulletin of Economics and Statistics*, 51(3), 333–343.
- HOUTHAKKER, H. S. (1960): “Additive Preferences,” *Econometrica*, 28(2), 244–257.
- IMROHOROGLU, S. (1991): “An Empirical Investigation of Currency Substitution,” Manuscript, University of Southern California.
- KING, R. G., C. I. PLOSSER, AND S. T. REBELO (1988a): “Production, Growth and Business Cycles: I. The Basic Neoclassical Models,” *Journal of Monetary Economics*, 21, 195–232.

- (1988b): “Production, Growth and Business Cycles: II. New Directions,” *Journal of Monetary Economics*, 21, 309–341.
- LEE, B. S., AND B. F. INGRAM (1991): “Simulation Estimation of Time-Series Models,” *Journal of Econometrics*, 47(2–3), 197–205.
- LI, Y. (2001): “Expected Returns and Habit Persistence,” *Review of Financial Studies*, 14(3), 861–899.
- LUCAS, JR., R. E., AND N. L. STOKEY (1987): “Money and Interest in a Cash-in-Advance Economy,” *Econometrica*, 55(3), 491–513.
- MANKIW, N. G. (1985): “Consumer Durables and the Real Interest Rate,” *Review of Economics and Statistics*, 67(3), 353–362.
- MANKIW, N. G., J. J. ROTEMBERG, AND L. H. SUMMERS (1985): “Intertemporal Substitution in Macroeconomics,” *Quarterly Journal of Economics*, 100(1), 225–251.
- MARK, N. (1988): “Time Varying Betas and Risk Premia in the Pricing of Forward Foreign Exchange Contracts,” *Journal of Financial Economics*, 22, 335–354.
- MARSHALL, D. A. (1992): “Inflation and Asset Returns in a Monetary Economy,” *Journal of Finance*, 47(4), 1315–1342.
- McFADDEN, D. (1989): “A Method of Simulated Moments for Estimation of Discrete Response Models without Numerical Integration,” *Econometrica*, 57(5), 995–1026.
- MEHRA, R., AND E. C. PRESCOTT (1985): “The Equity Premium: a Puzzle,” *Journal of Monetary Economics*, 15, 145–161.
- MENZLY, L., T. SANTOS, AND P. VERONESI (2004): “Understanding Predictability,” *Journal of Political Economy*, 112(1), 1–47.
- MIRON, J. A. (1986): “Seasonal Fluctuations and the Life-Cycle Permanent Income Model of Consumption,” *Journal of Political Economy*, 94(6), 1258–1279.
- NEWKEY, W. K., AND J. L. POWELL (2003): “Instrumental Variable Estimation of Nonparametric Models,” *Econometrica*, 71(5), 1565–1578.
- OGAKI, M. (1988): “Learning about Preferences from Time Trends,” Ph.D. thesis, University of Chicago.
- (1989): “Information in Deterministic Trends about Preferences,” Manuscript.
- (1990): “Demand for Foreign Bonds and the Term Structure of Interest Rates,” Manuscript.
- (1993a): “Generalized Method of Moments: Econometric Applications,” in *Handbook of Statistics: Econometrics*, ed. by G. S. Maddala, C. R. Rao, and H. D. Vinod, vol. 11, chap. 17, pp. 455–488. North-Holland, Amsterdam.
- (1993b): “GMM: A User Guide,” Rochester Center for Economic Research Working Paper No. 348.
- OGAKI, M., AND J. Y. PARK (1997): “A Cointegration Approach to Estimating Preference Parameters,” *Journal of Econometrics*, 82(1), 107–134.

- OGAKI, M., AND C. M. REINHART (1998a): "Intertemporal Substitution and Durable Goods: Long-Run Data," *Economics Letters*, 61, 85–90.
- (1998b): "Measuring Intertemporal Substitution: The Role of Durable Goods," *Journal of Political Economy*, 106, 1078–1098.
- OSANO, H., AND T. INOUE (1991): "Testing Between Competing Models Of Real Business Cycles," *International Economic Review*, 32(3), 669–688.
- PAKES, A., AND D. POLLARD (1989): "Simulation And The Asymptotics Of Optimization Estimators," *Econometrica*, 57(5), 1027–1057.
- PEARSON, N. (1991): "A Simulated Moments Estimator of Discrete time Asset Pricing Models," Manuscript, University of Rochester.
- POTERBA, J. M., AND J. J. ROTEMBERG (1987): "Money in the Utility Function: An Empirical Implementation," in *New Approaches to Monetary Economics: Proceedings of the Second International Symposium in Economic Theory and Econometrics*, ed. by W. Barnett, and K. Singleton, pp. 219–240. Cambridge University Press, Cambridge.
- RICH, R. W., J. RAYMOND, AND J. S. BUTLER (1991): "Generalized Instrumental Variables Estimation of Autoregressive Conditional Heteroskedastic Models," *Economics Letters*, 35(2), 179–185.
- ROLL, R. (1977): "A Critique of the Asset Pricing Theory's Tests, Part I: On Past and Potential Testability of the Theory," *Journal of Financial Economics*, 4(2), 129–176.
- SIL, K. (1992): "Money and Cash-In-Advance Models: An Empirical Implementation," Ph.D. thesis, University of Virginia.
- SIMON, D. P. (1989): "Expectations and Risk in the Treasury Bill Market: An Instrumental Variables Approach," *Journal of Financial and Quantitative Analysis*, 24(3), 357–365.
- SINGLETON, K. J. (1985): "Testing Specifications of Economic Agents' Intertemporal Optimum Problems in the Presence of Alternative Models," *Journal of Econometrics*, 30(1–2), 391–413.
- (1988): "Econometric Issues in the Analysis of Equilibrium Business Cycle Models," *Journal of Monetary Economics*, 21, 361–386.
- WEST, K. D. (1987): "A Specification Test for Speculative Bubbles," *Quarterly Journal of Economics*, 102, 553–580.
- (1988): "Asymptotic Normality, When Regressors Have a Unit Root," *Econometrica*, 56(6), 1397–1417.

# Chapter 11

## EXTREMUM ESTIMATORS

One of the common features across many estimators that are widely used in application such as ordinary least squares, instrumental variables, GMM, and maximum likelihood estimators, is that they are obtained by minimizing or maximizing an objective function. These estimators are called extremum estimators, or optimization estimators. This chapter explains a unified framework for this class of estimators.

### 11.1 Asymptotic Properties of Extremum Estimators

Let  $\{\mathbf{x}_t : t = 1, 2, \dots, T\}$  be a vector stochastic process,  $\mathbf{b}_0$  be a  $p$ -dimensional vector of parameters to be estimated, and  $J(\mathbf{b})$  be a real-valued objective function. For notational simplicity, the dependency of  $J(\mathbf{b})$  on  $\{\mathbf{x}_t : t = 1, 2, \dots, T\}$  is suppressed. An *extremum estimator* is a vector of parameters,  $\mathbf{b}_T$ , which minimizes the objective function,  $J_T(\mathbf{b})$ , with respect to  $\mathbf{b}$ . Under general regularity conditions, an extremum estimator is consistent and asymptotically normally distributed.<sup>1</sup>

There are two important assumptions that ensure the consistency and asymptotic normality of extremum estimators: convergence and identification.

---

<sup>1</sup>See the Appendix of Chapter 9 for a proof of consistency.

### 11.1.1 Convergence

The convergence assumption is that  $J_T(\mathbf{b})$  converges with probability one to some deterministic function  $J_0(\mathbf{b})$  as  $T \rightarrow \infty$  for all admissible values of  $\mathbf{b}$ . Convergence may take different forms such as uniform convergence and convergence in probability.

### 11.1.2 Identification

The identification assumption is that  $\mathbf{b}_0$  is the unique minimizer of  $J_0(\mathbf{b})$ .

## 11.2 Two Classes of Extremum Estimators

There are two classes of extremum estimators, *classical minimum distance estimators* and *M-estimators*.

### 11.2.1 Minimum Distance Estimators

An extremum estimator is a minimum distance estimator if the objective function is a quadratic function:

$$(11.1) \quad J_T(\mathbf{b}) = f_T(\mathbf{b})' \mathbf{W}_T f_T(\mathbf{b}),$$

where  $f(\cdot)$  is a  $q$ -dimensional vector of functions and  $\mathbf{W}_T$  is a sequence of matrix that satisfies

$$(11.2) \quad \lim_{T \rightarrow \infty} \mathbf{W}_T = \mathbf{W}_0$$

with probability one for a positive definite matrix  $\mathbf{W}_0$ . The matrices  $\mathbf{W}_T$  and  $\mathbf{W}_0$  are called the distance, or weighting, matrix.

A prominent example of the minimum distance estimator is the GMM estimator. In the GMM, the sample mean is used for  $f_T(\mathbf{b})$ , and the law of large number for the sample mean ensures convergence.

## 11.2.2 M-Estimators

An extremum estimator is an *M-estimator* if the objective function is a sample average:

$$(11.3) \quad Q_T(\mathbf{b}) = \frac{1}{T} \sum_{t=1}^T m(\mathbf{x}_t),$$

where  $m(\cdot)$  is a real-valued function. The maximum likelihood (ML) estimator is a leading example of the M-estimator. Suppose  $\{\mathbf{x}_t\}$  is an i.i.d. process with a known density function  $f(\mathbf{x}_t; \mathbf{b}_0)$  where  $\mathbf{b}_0$  is an unknown true parameter vector. The joint density of  $\{\mathbf{x}_t\}$  is given by

$$(11.4) \quad f(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T; \mathbf{b}_0) = \prod_{t=1}^T f(\mathbf{x}_t; \mathbf{b}_0).$$

If we replace  $\mathbf{b}_0$  with some arbitrary (random?) value  $\mathbf{b}$ , and interpret the density as a function of  $\mathbf{b}$ , it is called the *likelihood function*. The ML estimator for  $\mathbf{b}_0$  is a parameter vector  $\mathbf{b}$  that maximizes the likelihood function. Since the log transformation is a monotone transformation, maximizing the likelihood function is equivalent to minimizing the following:

$$(11.5) \quad -\log f(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T; \mathbf{b}_0) = -\sum_{t=1}^T \log f(\mathbf{x}_t; \mathbf{b}_0).$$

## 11.3 Examples of Minimum Distance Estimators

### 11.3.1 Two-Step Minimum Distance Estimators

Another example of the minimum distance estimator is a *two-step minimum distance estimator*. Suppose  $\mathbf{c}_0$  is the true values of some parameters of interest. In the first step, a consistent estimator for  $\mathbf{c}_0$ ,  $\mathbf{c}_T$ , is obtained. In the second step, the minimum

distance method is used to estimate another set of parameters based on  $\mathbf{c}_T$  from the first step.

One application of the two-step minimum distance estimation has an unrestricted estimator as  $\mathbf{c}_T$  and uses the minimum distance estimation to impose restrictions on  $\mathbf{c}_0$ . Suppose  $\mathbf{c}_T$  is an unrestricted estimator for a  $(p+s)$ -dimensional vector of parameters  $\mathbf{c}_0$ . Consider nonlinear restrictions

$$(11.6) \quad \phi(\mathbf{b}_0) = \mathbf{c}_0,$$

where  $\mathbf{b}_0$  is a  $p$ -dimensional vector of parameters. The minimum distance estimator,  $\mathbf{b}_T$ , minimizes

$$(11.7) \quad J_T(\mathbf{b}) = \{\phi(\mathbf{b}) - \mathbf{c}_T\}' \mathbf{W}_T \{\phi(\mathbf{b}) - \mathbf{c}_T\},$$

where  $\mathbf{W}_T$  is a positive definite distance matrix and converges to some positive definite matrix  $\mathbf{W}_0$  with probability one. As in the GMM, the optimal distance matrix is  $\mathbf{W} = \mathbf{\Omega}^{-1}$  and  $TJ_T(\mathbf{b}_T)$  has an (asymptotic) chi-square distribution with  $s$  degrees of freedom. The null hypothesis (11.6) is rejected when this statistic exceeds the critical value from a chi-square distribution. See Altug and Miller (1990) and Atkeson and Ogaki (1996) for empirical applications.

### 11.3.2 Two-Step Minimum Distance Estimation with Impulse Responses

Another application of the two-step minimum distance estimator is the estimation of parameters in a theoretical model by matching the model's theoretical impulse response functions with empirical impulse response functions estimated by vector autoregressions (VAR). Denoting a vector of model parameters by  $\boldsymbol{\beta}$ , the optimal

estimators are chosen so as to minimize the quadratic distance between empirical impulse responses, denoted by  $\hat{\Psi}$ , and the model-implied impulse responses:

$$(11.8) \quad \min_{\beta} \left[ \hat{\Psi} - \Psi(\beta) \right]' \Sigma^{-1} \left[ \hat{\Psi} - \Psi(\beta) \right],$$

where  $\Psi(\beta)$  denotes the mapping from  $\beta$  to the model impulse response functions, and  $\Sigma$  is a diagonal matrix whose diagonal elements are sample variances of the  $\hat{\Psi}$ 's.

Sbordone (2002) and Sbordone (2005) apply this method to estimate the degree of price stickiness from the NKPC. The so-called Calvo (1983) parameter measures the probability that a firm does not change its price in a given period. Letting  $\theta$  denote this probability, the average number of periods for which a price remains unchanged is  $(1 - \theta) \sum_{k=0}^{\infty} k\theta^{k-1} = 1/(1 - \theta)$ . Magnusson and Mavroeidis (2009) develop the identification robust minimum distance estimator with similar ideas as the identification robust GMM estimator. However, their confidence sets indicate that the minimum distance estimation applied to the NKPC is subject to the weak identification problem. For example, their 95% confidence interval for the average price duration has a lower bound of around 3.3 quarters and an upper bound of infinity.

A classic method to estimate  $\theta$  is the single-equation GMM using the NKPC (see, for example, Galí and Gertler (1999) and Eichenbaum and Fisher (2007)). In Galí and Gertler (1999),  $\theta$  is estimated to be around 0.8, implying the average price duration of 5 quarters. However, as surveyed by Kleibergen and Mavroeidis (2009), this estimation method is also subject to the weak identification problem. The 95% confidence interval for the average price duration using their recommended method has a lower bound of two quarters and an upper bound of infinity. Since the lower bound obtained from the minimum distance method is sharper than that from the

GMM, the minimum distance method outperforms the GMM when applied to a single equation using the NKPC.

Christiano, Eichenbaum, and Evans (2005) apply the two-step minimum distance method to a system of equations from their DSGE model to investigate the role of nominal rigidities in generating the observed persistent responses of inflation and output to a monetary policy shock. They first estimate the VAR impulse responses of 8 key macroeconomic variables using the post-war U.S. data. Let  $\mathbf{Y}_{1t}$  be a vector of observations on real GDP, real consumption, GDP deflator, real investment, and real wage,  $R_t$  denote the federal funds rate, and  $\mathbf{Y}_{2t}$  be a vector of real profits and the growth rate of M2. These variables are stacked as  $\mathbf{Y}_t = [ \mathbf{Y}'_{1t} \ R_t \ \mathbf{Y}'_{2t} ]'$ . This ordering ensures that the monetary policy shock is identified by two identifying assumptions. First, the variables in  $\mathbf{Y}_{1t}$  are assumed not to respond contemporaneously to the monetary policy shock, and second, the federal funds rate does not depend on the current values of the variables in  $\mathbf{Y}_{2t}$ . Using the first 25 estimated coefficients of each impulse response as elements of  $\hat{\Psi}$  in (16.11), model parameters are estimated as a solution to (16.11). Their estimate of  $\theta$  is 0.6 in the benchmark model, implying the average price duration of 2.5 quarters. Because they apply the method to a system of equations rather than a single equation, their system may be well identified. This is an important topic for further research.<sup>2</sup>

---

<sup>2</sup>Kim and Ogaki (2009) estimate the Calvo parameter in an exchange rate model with the Taylor rule without the NKPC. In their estimation for  $\theta$ , there is a substantial efficiency gain by applying the GMM to a system of equations rather than to a single equation. We expect an analogous substantial efficiency gain for the minimum distance estimation.

### 11.3.3 Minimum Distance to Estimate Data Statistics

Another application of the minimum distance method in the DSGE literature is to estimate various statistics of model variables such as mean, standard deviation, correlation, and autocorrelation. Although the GMM may be used, minimum distance may be more convenient.

Consider two stationary variables,  $x_t$  and  $y_t$ . Suppose we want to estimate their population moments,  $\mathbf{b}_0 = (E(x_t), E(x_t^2), E(y_t), E(y_t^2), E(x_t y_t), E(x_t x_{t-1}))$ . Let  $\mathbf{x}_t = (x_t, y_t)$  and  $f(\mathbf{x}_t, \mathbf{b}) = (x_t, x_t^2, y_t, y_t^2, x_t y_t, x_t x_{t-1})' - \mathbf{b}$ , where  $f(\mathbf{x}_t, \mathbf{b})$  is a disturbance defined at time  $t$ . The GMM minimizes a quadratic form of the sample average of  $f(\mathbf{x}_t, \mathbf{b})$ , to obtain an estimate of  $\mathbf{b}_0$ ,  $\mathbf{b}_T$ , and an estimate of covariance matrix of  $T^{\frac{1}{2}}(\mathbf{b}_T - \mathbf{b}_0)$ .

To obtain the standard errors of estimated statistics that are nonlinear functions of  $\mathbf{b}_0$  such as standard deviations, correlations, and autocorrelations, one can use the delta method explained in Proposition 5.8. For example, let  $a(\mathbf{b}_0)$  denote the standard deviation of  $x_t$ ,  $a(\mathbf{b}_0) = \sqrt{\text{var}(x_t)} = (E(x_t^2) - E(x_t)^2)^{\frac{1}{2}}$ , and  $a(\mathbf{b}_T)$  be a consistent estimator of  $a(\mathbf{b}_0)$ . By the delta method,  $\sqrt{T}(a(\mathbf{b}_T) - a(\mathbf{b}_0))$  has an approximate normal distribution with variance  $\mathbf{d}(\mathbf{b}_0) \text{Cov}(\boldsymbol{\Omega}^{-1}) \mathbf{d}(\mathbf{b}_0)'$  in a large sample where  $\mathbf{d}(\mathbf{b}_0)$  is the derivative of  $a(\cdot)$  evaluated at  $\mathbf{b}_0$ .

In the GMM, while parameters may enter moment conditions nonlinearly, sample moments may not because the moment conditions may not be equal to zero in that case. For example, in order to estimate the variance of  $x_t$  in the above example, the moment condition would be  $b - (x_t - \bar{x})^2$  where  $b$  is the variance to be estimated and  $\bar{x}$  is the sample mean. However, because the sample mean enters the moment condition in a nonlinear way,  $E(b - (x_t - \bar{x})^2)$  is not equal to zero, which prevents the

GMM estimation.

By contrast, in the minimum distance estimation, sample moments may enter moment conditions in nonlinear ways. For example, Ambler, Cardia, and Zimmermann (2004) (section 3) estimate a pair of correlations

$$(11.9) \quad \bar{\rho}_{1t} = \frac{(x_1 - \bar{x}_1)(x_2 - \bar{x}_2)}{\bar{\sigma}_1 \bar{\sigma}_2}$$

and

$$(11.10) \quad \bar{\rho}_{2t} = \frac{(x_3 - \bar{x}_3)(x_4 - \bar{x}_4)}{\bar{\sigma}_3 \bar{\sigma}_4},$$

where  $\bar{x}_i$  and  $\bar{\sigma}_i$  are the sample mean and variance of  $x_i$ . The optimal estimators are obtained by minimizing

$$(11.11) \quad \left\{ \frac{1}{T} \sum_{t=1}^T (\boldsymbol{\rho} - \bar{\boldsymbol{\rho}}_t) \right\}' \mathbf{W}_T \left\{ \frac{1}{T} \sum_{t=1}^T (\boldsymbol{\rho} - \bar{\boldsymbol{\rho}}_t) \right\}$$

where  $\boldsymbol{\rho}$  is a  $(2 \times 1)$  vector of population correlations of  $x_{it}$  for  $i = 1, 2, 3, 4$  and  $\bar{\boldsymbol{\rho}}_t = [ \bar{\rho}_{1t} \quad \bar{\rho}_{2t} ]'$ .

Although this setup resembles the GMM, it cannot be embedded in the standard GMM framework because the sample mean and variance enter the moment conditions nonlinearly. Instead, this is a minimum distance estimator.

The minimum distance estimator can be used to estimate a DSGE model by matching the model-implied moments with empirical moments in a similar way as GMM while allowing the sample mean to enter moment conditions nonlinearly. An application can be found in García-Cicco, Pancrazi, and Uribe (2009).

## 11.4 The Kalman Filter

We introduced the ML estimator for an i.i.d. process. However, this i.i.d. assumption rarely holds in time series data. In linear models with time dependence, the likelihood

function can be evaluated using a recursive linear algorithm called the Kalman filter (Kalman, 1960). The Kalman filter estimates an evolution of unobserved variable(s) of interest in a discrete-time dynamic system by sequentially updating a linear projection using current observations. Because this filtering process minimizes the mean squared prediction error, it yields an optimal estimator among the class of linear projections. Due to its accuracy and practicality, various extensions of the Kalman filter have been developed and applied in a broad area of study. In econometric, it is used to construct exact finite-sample forecasting, evaluate the exact likelihood function, and estimate parameters in ARMA models or time-varying parameters in linear regressions, just to name a few examples.

In order to formulate the Kalman filter algorithm, the process of interest is modeled in a set of linear equations called the *state-space representation*. This equation system characterizes the relationship between observed and unobserved variables. Let  $\mathbf{x}_t$  be an  $r$ -dimensional vector of unobserved variables,  $\mathbf{y}_t$  be an  $n$ -dimensional vector of observed variables, and  $\mathbf{z}_t$  be a  $k$ -dimensional vector of exogenous variables. Suppose  $\mathbf{y}_t$  depends linearly on  $\mathbf{x}_t$  and  $\mathbf{z}_t$ :

$$(11.12) \quad \mathbf{y}_t = \mathbf{A}' \cdot \mathbf{z}_t + \mathbf{H}' \cdot \mathbf{x}_t + \mathbf{e}_t,$$

where  $\mathbf{e}_t$  is  $(n \times 1)$  vector white noise with  $E(\mathbf{e}_t \mathbf{e}_j') = \mathbf{R}$  for  $t = j$  and  $\mathbf{0}$  otherwise, and  $\mathbf{A}'$  and  $\mathbf{H}'$  are  $(n \times k)$  and  $(n \times r)$  matrices of parameters, respectively.

The unobserved vector  $\mathbf{x}_t$ , called the *state vector*, is assumed to evolve according to a linear stochastic difference equation

$$(11.13) \quad \mathbf{x}_{t+1} = \mathbf{F} \cdot \mathbf{x}_t + \mathbf{u}_{t+1},$$

where  $\mathbf{u}_{t+1}$  is also  $(r \times 1)$  vector white noise with  $E(\mathbf{u}_t \mathbf{u}_j') = \mathbf{Q}$  for  $t = j$  and  $\mathbf{0}$

otherwise, and  $\mathbf{F}$  is an  $(r \times r)$  matrix of parameters. The disturbances  $\mathbf{e}_t$  and  $\mathbf{u}_t$  are assumed to be independent of each other at all lags,  $E(\mathbf{e}_t \mathbf{u}'_j) = 0$  for all  $t$  and  $j$ , and the initial state  $\mathbf{z}_1$  is uncorrelated with any realizations of  $\mathbf{e}_t$  and  $\mathbf{u}_t$ ,  $E(\mathbf{e}_t \mathbf{z}'_1) = 0$  and  $E(\mathbf{u}_t \mathbf{z}'_1) = 0$  for  $t = 1, \dots, T$ . Together with the state equation (11.13), the latter assumption implies that  $\mathbf{e}_t$  and  $\mathbf{u}_t$  are uncorrelated with all lagged values of  $\mathbf{x}_t$ :  $E(\mathbf{e}_t \mathbf{x}'_j) = 0$  and  $E(\mathbf{u}_t \mathbf{x}'_j) = 0$  for  $j = t - 1, t - 2, \dots, 1$ .

Equation (11.12) is called the *observation equation*, and equation (11.13) the *state equation*. Together, they comprise the state-space representation of the dynamics of  $\mathbf{y}$ .

The Kalman filter recursively generates least square forecasts of the unobserved state vector  $\mathbf{x}_t$  as a linear function of the observed data  $\mathbf{y}_t$  and  $\mathbf{z}_t$ . Let  $\hat{\mathbf{x}}_{t+1|t} \equiv \hat{E}(\mathbf{x}_{t+1} | \Omega_t)$  denote the best forecasts of  $\mathbf{x}_{t+1}$  based on the data available at time  $t$ ,  $\Omega_t \equiv (\mathbf{y}'_t, \mathbf{y}'_{t-1}, \dots, \mathbf{y}'_1, \mathbf{z}'_t, \mathbf{z}'_{t-1}, \dots, \mathbf{z}'_1)$ . The accuracy of each forecast is measured by an associated  $(r \times r)$  error covariance matrix,  $\mathbf{P}_{t+1|t} \equiv E[(\mathbf{x}_{t+1} - \hat{\mathbf{x}}_{t+1|t})(\mathbf{x}_{t+1} - \hat{\mathbf{x}}_{t+1|t})']$ .

In order to initiate the recursive process, the unconditional mean of the initial state  $\hat{\mathbf{x}}_{1|0}$  and its covariance  $\mathbf{P}_{1|0}$  must be chosen. If the eigenvalues of  $\mathbf{F}$  are inside the unit circle,  $\hat{\mathbf{x}}_{1|0}$  is simply set equal to  $\mathbf{0}$  with an associated covariance matrix whose column vectors are given by  $\text{vec}(\mathbf{P}_{1|0}) = [\mathbf{I}_{r^2} - (\mathbf{F} \times \mathbf{F})]^{-1} \cdot \text{vec}(\mathbf{Q})$ . Otherwise, the researcher's best guess of  $\mathbf{x}_{1|0}$  can be used as  $\hat{\mathbf{x}}_{1|0}$ , and a positive definite matrix that summarizes the confidence in this guess is used as  $\mathbf{P}_{1|0}$ .

Suppose we have data on  $(\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_T, \mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_T)$ . For simple illustration, assume that the matrices  $\mathbf{F}$ ,  $\mathbf{Q}$ ,  $\mathbf{A}$ ,  $\mathbf{H}$ , and  $\mathbf{R}$  are known and constant. Given  $\hat{\mathbf{x}}_{1|0}$  and  $\mathbf{P}_{1|0}$ , the linear projection of  $\hat{\mathbf{x}}_{t+1|t}$  and associated covariance of this forecast  $\mathbf{P}_{t+1|t}$

are iterated on

$$(11.14) \quad \begin{aligned} \hat{\mathbf{x}}_{t+1|t} &= \mathbf{F}\hat{\mathbf{x}}_{t|t-1} + \mathbf{K}_t(\mathbf{y}_t - \mathbf{A}'\mathbf{z}_t - \mathbf{H}'\hat{\mathbf{x}}_{t|t-1}), \\ \mathbf{P}_{t+1|t} &= \mathbf{F}[\mathbf{P}_{t|t-1} - \mathbf{P}_{t|t-1}\mathbf{H}(\mathbf{H}'\mathbf{P}_{t|t-1}\mathbf{H} + \mathbf{R})^{-1}\mathbf{H}'\mathbf{P}_{t|t-1}]\mathbf{F}' + \mathbf{Q}, \end{aligned}$$

for  $t = 1, 2, \dots, T$ , where  $\mathbf{K}_t \equiv \mathbf{F}\mathbf{P}_{t|t-1}\mathbf{H}(\mathbf{H}'\mathbf{P}_{t|t-1}\mathbf{H} + \mathbf{R})^{-1}$  is called the *Kalman gain*. That  $\mathbf{K}_t$  depends negatively on  $\mathbf{R}$  implies that, when computing the projection for next period, the Kalman filter attaches a smaller (larger) weight to the observation the larger (smaller) the noise in the observed data is (and hence the larger (smaller)  $\mathbf{R}$  is).

The previous period's projections are updated based on the current realization of the observable as follows:

$$(11.15) \quad \begin{aligned} \hat{\mathbf{x}}_{t|t} &= \hat{\mathbf{x}}_{t|t-1} + \mathbf{F}^{-1}\mathbf{K}(\mathbf{y}_t - \mathbf{A}'\mathbf{z}_t - \mathbf{H}'\hat{\mathbf{x}}_{t|t-1}), \\ \mathbf{P}_{t|t} &= \mathbf{P}_{t|t-1} - \mathbf{P}_{t|t-1}\mathbf{H}(\mathbf{H}'\mathbf{P}_{t|t-1}\mathbf{H} + \mathbf{R})^{-1}\mathbf{H}'\mathbf{P}_{t|t-1}. \end{aligned}$$

Notice that equations (16.61) and (16.62) are related by:

$$\begin{aligned} \hat{\mathbf{x}}_{t+1|t} &= \mathbf{F}\hat{\mathbf{x}}_{t|t}, \\ \mathbf{P}_{t+1|t} &= \mathbf{F}\mathbf{P}_{t|t}\mathbf{F}' + \mathbf{Q}. \end{aligned}$$

Thus, the Kalman filter repeats a project-and-update cycle in which it makes projections  $\hat{\mathbf{x}}_{t|t-1}$ , updates these projections based on the current observations to get  $\hat{\mathbf{x}}_{t|t}$ , and uses them to obtain next projections  $\hat{\mathbf{x}}_{t+1|t}$ . This recursive nature implies that all the necessary information is contained in previous forecasts and information sets, and hence the filtering does not require all the previous data to be stored and re-processed in each estimation step. This is one of the appealing features of the Kalman filter for practical implementations.

Finally, the forecast of  $\mathbf{y}_{t+1}$  is obtained as follows. The exogeneity assumption of  $\mathbf{z}_t$  implies that it contains no information about  $\mathbf{x}_t$  beyond what is contained in

the  $t - 1$  information set  $\mathbf{\Omega}_{t-1} \equiv (\mathbf{y}'_{t-1}, \mathbf{y}'_{t-2}, \dots, \mathbf{y}'_1, \mathbf{z}'_{t-1}, \mathbf{z}'_{t-2}, \dots, \mathbf{z}'_1)$ . Hence,

$$\hat{E}(\mathbf{x}_t | \mathbf{z}_t, \mathbf{\Omega}_{t-1}) = \hat{E}(\mathbf{x}_t | \mathbf{\Omega}_{t-1}) = \hat{\mathbf{x}}_{t|t-1}.$$

From the observation equation (11.12) and by the law of iterated projections, the forecast of  $\mathbf{y}_{t+1}$  is given by

$$\begin{aligned} \hat{\mathbf{y}}_{t+1|t} &\equiv \hat{E}(\mathbf{y}_{t+1} | \mathbf{z}_{t+1}, \mathbf{\Omega}_t) \\ &= \mathbf{A}'\mathbf{z}_{t+1} + \mathbf{H}'\hat{E}(\mathbf{x}_{t+1} | \mathbf{z}_{t+1}, \mathbf{\Omega}_t) \\ &= \mathbf{A}'\mathbf{z}_{t+1} + \mathbf{H}'\hat{\mathbf{x}}_{t+1|t}, \end{aligned}$$

with error covariance

$$E[(\mathbf{y}_{t+1} - \hat{\mathbf{y}}_{t+1|t})(\mathbf{y}_{t+1} - \hat{\mathbf{y}}_{t+1|t})'] = \mathbf{H}'\mathbf{P}_{t+1|t}\mathbf{H} + \mathbf{R}.$$

The Kalman filter minimizes the error covariance of the estimated objects; therefore, the forecasts  $\hat{\mathbf{x}}_{t+1|t}$  and  $\hat{\mathbf{y}}_{t+1|t}$  are best estimators within the class of linear filters (i.e. forecasts that are linear functions of  $(\mathbf{z}_t, \mathbf{\Omega}_{t-1})$ ). If we further assume that initial state  $\mathbf{x}_{1|0}$  and innovations  $\{\mathbf{e}_t, \mathbf{u}_t\}_{t=1}^T$  are multivariate Gaussian, then the forecasts are optimal among any functions of  $(\mathbf{z}_t, \mathbf{\Omega}_{t-1})$ .

### 11.4.1 Evaluation of the Likelihood Function using the Kalman Filter

One of the applications of the Kalman filter is the evaluation of unconditional likelihood for a DSGE model. Consider a state-space representation of the solution of the DSGE model:

$$\begin{aligned} \mathbf{x}_t &= \mathbf{F}(\mu)\mathbf{x}_{t-1} + \mathbf{u}_t \\ \mathbf{u}_t &= \mathbf{G}(\mu)\mathbf{v}_t, \end{aligned}$$

where  $\mathbf{x}_t$  is an  $(r \times 1)$  vector of model variables, and  $E(\mathbf{u}_t \mathbf{u}_t') = \mathbf{G}(\mu)E(\mathbf{v}_t \mathbf{v}_t')\mathbf{G}(\mu)' = \mathbf{Q}(\mu)$ . A measurement equation maps  $\mathbf{x}_t$  into the  $n \times 1$  vector of observable variables  $\mathbf{y}_t$ :

$$\mathbf{y}_t = \mathbf{H}(\mu)' \mathbf{x}_t + \mathbf{e}_t,$$

where  $\mathbf{e}_t$  is an  $n \times 1$  vector of measurement errors with  $E(\mathbf{e}_t \mathbf{e}_t') = \mathbf{R}$  for  $t = j$  and  $\mathbf{0}$  otherwise. Given time-series data and the model's parameter values  $\mu$  (so that  $F(\mu)$ ,  $G(\mu)$ ,  $Q(\mu)$ , and  $H(\mu)$  are known), the Kalman filter infers a sequence of conditional distribution for  $\mathbf{x}_t$  given  $\mathbf{x}_{t-1}$  and evaluate the likelihood.

In order to implement the Kalman filter, assume that  $\mathbf{e}_t$ ,  $\mathbf{u}_t$ , and  $\mathbf{v}_t$  are normally distributed. The initial unconditional values are given by

$$\hat{\mathbf{x}}_{1|0} = \mathbf{0}, \quad \mathbf{P}_{1|0} = \mathbf{F}\mathbf{P}_{1|0}\mathbf{F}' + \mathbf{Q}$$

where  $\text{vec}(\mathbf{P}_{1|0}) = (\mathbf{I} - \mathbf{F} \otimes \mathbf{F}')^{-1} \text{vec}(\mathbf{Q})$ .

Given the initial values, the projection  $\hat{\mathbf{x}}_{t|t-1}$  and its associated covariance matrix  $\mathbf{P}_{t|t-1}$  are iterated on:

$$\begin{aligned} \hat{\mathbf{x}}_{t|t-1} &= \mathbf{F}\hat{\mathbf{x}}_{t-1|t-1} \\ \mathbf{P}_{t|t-1} &= \mathbf{F}\mathbf{P}_{t-1|t-1}\mathbf{F}' + \mathbf{Q}, \end{aligned}$$

where  $\text{vec}(\mathbf{P}_{t|t-1}) = (\mathbf{I} - \mathbf{F} \otimes \mathbf{F}')^{-1} \text{vec}(\mathbf{Q})$ . These projections are then used to construct the conditional distribution of  $\mathbf{y}_t$ ,  $N(\hat{\mathbf{y}}_{t|t-1}, \Sigma_{t|t-1})$ , where the conditional mean  $\hat{\mathbf{y}}_{t|t-1}$  and conditional variance matrix  $\Sigma_{t|t-1}$  are given by

$$\begin{aligned} \hat{\mathbf{y}}_{t|t-1} &= \mathbf{H}'\hat{\mathbf{x}}_{t|t-1} \\ \Sigma_{t|t-1} &= E[(\mathbf{y}_t - \hat{\mathbf{y}}_{t|t-1})(\mathbf{y}_t - \hat{\mathbf{y}}_{t|t-1})'] \\ &= \mathbf{H}'\mathbf{P}_{t|t-1}\mathbf{H} + \mathbf{R}. \end{aligned}$$

The likelihood function for  $\mathbf{y}_t$  is thus given by:

$$L(\mathbf{y}_t|\boldsymbol{\mu}) = (2\pi)^{-m/2} |\boldsymbol{\Sigma}_{t|t-1}^{-1}|^{1/2} \exp \left[ -\frac{1}{2} (\mathbf{y}_t - \hat{\mathbf{y}}_{t|t-1})' \boldsymbol{\Sigma}_{t|t-1}^{-1} (\mathbf{y}_t - \hat{\mathbf{y}}_{t|t-1}) \right].$$

The next iteration is initiated by updating  $\hat{\mathbf{x}}_{t|t-1}$  and  $\mathbf{P}_{t|t-1}$ :

$$\begin{aligned} \mathbf{x}_{t|t} &= \mathbf{x}_{t|t-1} + \mathbf{P}_{t|t-1} \mathbf{H} \boldsymbol{\Sigma}_{t|t-1}^{-1} (\mathbf{y}_t - \mathbf{y}_{t|t-1}) \\ \mathbf{P}_{t|t} &= \mathbf{P}_{t|t-1} - \mathbf{P}_{t|t-1} \mathbf{H} \boldsymbol{\Sigma}_{t|t-1}^{-1} \mathbf{H}' \mathbf{P}_{t|t-1}. \end{aligned}$$

Finally, the likelihood from each iteration is multiplied to yield the sample likelihood:

$$L(\mathbf{y}|\boldsymbol{\mu}) = \prod_{t=1}^T L(\mathbf{y}_t|\boldsymbol{\mu}).$$

This likelihood function is maximized to yield the ML estimator of linearized DSGE models.

## Appendix

### 11.A Examples of State-Space Representations

This appendix contains examples of the state-space representation for AR( $p$ ) and MA( $p$ ) processes. There are several ways of representing a given process in state-space form. For more examples, see Hamilton (1994, Ch. 13).

Example 1: Univariate AR( $p$ ) Process Consider a univariate AR( $p$ ) process:

$$y_{t+1} - \mu = \phi_1(y_t - \mu) + \phi_2(y_{t-1} - \mu) + \cdots + \phi_p(y_{t-p+1} - \mu) + \varepsilon_{t+1},$$

where  $E(\varepsilon_t \varepsilon_j) = \sigma^2$  for  $j = t$  and 0 otherwise. One example of the state-space representation for this process is

$$\mathbf{x}_t = \begin{bmatrix} y_t - \mu \\ y_{t-1} - \mu \\ \vdots \\ y_{t-p+1} - \mu \end{bmatrix}, \mathbf{F} = \begin{bmatrix} \phi_1 & \phi_2 & \cdots & \phi_{p-1} & \phi_p \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix}, \mathbf{u}_{t+1} = \begin{bmatrix} \varepsilon_{t+1} \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \mathbf{Q} = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix},$$

$$\mathbf{y}_t = y_t, \mathbf{A}' = \mu, \mathbf{z}_t = 1, \mathbf{H}' = [1 \ 0 \ \cdots \ 0], \mathbf{e}_t = 0, \mathbf{R} = 0.$$

Example 2: Univariate MA(1) Process

For a univariate MA(1) process

$$y_t = \mu + \varepsilon_t + \theta\varepsilon_{t-1}$$

where  $E(\varepsilon_t\varepsilon_j) = \sigma^2$  for  $j = t$  and 0 otherwise, the state-space representation is given by

$$\mathbf{x}_t = \begin{bmatrix} \varepsilon_t \\ \varepsilon_{t-1} \end{bmatrix}, \mathbf{F} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}, \mathbf{u}_{t+1} = \begin{bmatrix} \varepsilon_{t+1} \\ 0 \end{bmatrix}, \mathbf{Q} = \begin{bmatrix} \sigma^2 & 0 \\ 0 & 0 \end{bmatrix},$$

$$\mathbf{y}_t = y_t, \mathbf{A}' = \mu, \mathbf{z}_t = 1, \mathbf{H}' = [1 \ \theta], \mathbf{e}_t = 0, \mathbf{R} = 0.$$

## References

- ALTUG, S., AND R. A. MILLER (1990): "Household Choices in Equilibrium," *Econometrica*, 58, 543–570.
- AMBLER, S., E. CARDIA, AND C. ZIMMERMANN (2004): "International Business Cycles: What are the Facts?," *Journal of Monetary Economics*, 51(2), 257–276.
- ATKESON, A., AND M. OGAKI (1996): "Wealth-Varying Intertemporal Elasticities of Substitution: Evidence from Panel and Aggregate Data," *Journal of Monetary Economics*, 38, 507–534.
- CALVO, G. A. (1983): "Staggered Prices and in a Utility-Maximizing Framework," *Journal of Monetary Economics*, 12(3), 383–398.
- CHRISTIANO, L. J., M. EICHENBAUM, AND C. L. EVANS (2005): "Nominal Rigidities and the Dynamic Effects of a Shock to Monetary Policy," *Journal of Political Economy*, 113(1), 1–45.
- EICHENBAUM, M., AND J. D. M. FISHER (2007): "Estimating the Frequency of Price Re-Optimization in Calvo-Style Models," *Journal of Monetary Economics*, 54(7), 2032–2047.

- GALÍ, J., AND M. GERTLER (1999): "Inflation Dynamics: A Structural Econometric Analysis," *Journal of Monetary Economics*, 44(2), 195–222.
- GARCÍA-CICCO, J., R. PANCRAZI, AND M. URIBE (2009): "Real Business Cycles in Emerging Countries? Expanded Version," Manuscript.
- HAMILTON, J. D. (1994): *Time Series Analysis*. Princeton University Press, Princeton.
- KALMAN, R. E. (1960): "A New Approach to Linear Filtering and Prediction Problems," *Journal of Basic Engineering*, 82(1), 35–45.
- KIM, H., AND M. OGAKI (2009): "Purchasing Power Parity and the Taylor Rule," Working Paper No. 09-03, Department of Economics, Ohio State University.
- KLEIBERGEN, F., AND S. MAVROEIDIS (2009): "Weak Instrument Robust Tests in GMM and the New Keynesian Phillips Curve," *Journal of Business and Economic Statistics*, 27(3), 293–311.
- MAGNUSSON, L. M., AND S. MAVROEIDIS (2009): "Identification-Robust Minimum Distance Estimation of the New Keynesian Phillips Curve," Working Paper 0904, Department of Economics, Tulane University.
- SBORDONE, A. M. (2002): "Prices and Unit Labor Costs: A New Test of Price Stickiness," *Journal of Monetary Economics*, 49(2), 265–292.
- (2005): "Do Expected Future Marginal Costs Drive Inflation Dynamics?," *Journal of Monetary Economics*, 52(6), 1183–1197.

# Chapter 12

## INTRODUCTION TO BAYESIAN APPROACH

Over the last decade, Bayesian analysis has become an increasingly popular method in economics. As you will see in this chapter, the Bayesian approach differs from the classical frequentist approach in various aspects. The fundamental difference lies in its probabilistic interpretation of the object of interest such as unknown parameters and random events. In the Bayesian framework, unknown parameters are treated as random variables while the observed data are treated as fixed. This interpretation allows us to assign a probability distribution associated with the parameters upon which Bayesian inferences are made.

This chapter introduces basic concepts and implementation of Bayesian analysis. Next section explains probability density functions in Bayesian statistics, followed by their application to generating point estimates and constructing Bayesian credible intervals. We then discuss posterior odds ratio tests for hypothesis testing and model comparison. Details of each topic can be found in DeJong and Dave (2007), Judge *et al*(1985), and Zellner (1996). The appendix to this chapter explains simulation methods that are widely used in the implementation of Bayesian analysis.

## 12.1 Bayes Theorem

Bayesian analysis centers around the representation of our uncertainty about the object of interest such as true values of unknown parameters. A prior distribution represents our initial knowledge or subjective beliefs about the unknown parameters held prior to observing data. After the data has been observed, sample information is incorporated into the prior to form a posterior distribution which assigns a probability to alternative parameter values based on the information from the prior and the data. Bayes' theorem is a mathematical formula in probability theory that relates the posterior distribution to the prior and the sample information represented by a likelihood function.

Suppose we are interested in a vector of unknown parameters  $\boldsymbol{\theta}$ . Let  $p(\boldsymbol{\theta})$  denote a prior density function for  $\boldsymbol{\theta}$ , and  $\mathbf{y}$  a vector of sample observations from a density  $f(\mathbf{y}|\boldsymbol{\theta})$ . A joint probability density for  $\boldsymbol{\theta}$  and  $\mathbf{y}$  is given by

$$(12.1) \quad P(\boldsymbol{\theta}, \mathbf{y}) = p(\boldsymbol{\theta})f(\mathbf{y}|\boldsymbol{\theta}) = f(\mathbf{y})p(\boldsymbol{\theta}|\mathbf{y}).$$

Rearranging the second equality in (12.1) yields a posterior density function for  $\boldsymbol{\theta}$ :

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\boldsymbol{\theta})f(\mathbf{y}|\boldsymbol{\theta})}{f(\mathbf{y})}.$$

This result is Bayes' theorem, showing how our prior knowledge  $p(\boldsymbol{\theta})$  is combined with sample information  $f(\mathbf{y}|\boldsymbol{\theta})$  to generate the posterior distribution. Since we are interested in the distribution of  $\boldsymbol{\theta}$ ,  $f(\mathbf{y})$  may be treated as a normalizing constant, and  $p(\boldsymbol{\theta}|\mathbf{y})$  is in general analyzed up to constant proportionality:

$$(12.2) \quad p(\boldsymbol{\theta}|\mathbf{y}) \propto p(\boldsymbol{\theta})f(\mathbf{y}|\boldsymbol{\theta}).$$

Here,  $f(\mathbf{y}|\boldsymbol{\theta})$  is algebraically identical to a likelihood function  $l(\boldsymbol{\theta}|\mathbf{y})$ , and (12.2) may be expressed as  $p(\boldsymbol{\theta}|\mathbf{y}) \propto p(\boldsymbol{\theta})l(\boldsymbol{\theta}|\mathbf{y})$ ; that is, the posterior distribution is proportional to the product of the prior and the likelihood function. The posterior distribution serves as an essential element of Bayesian inferences such as generating point estimates, constructing confidence intervals, and conducting hypothesis testing which we discuss in next sections.

## 12.2 Parameter Estimates

In general, Bayesian point estimates are obtained by specifying a loss function  $L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$  which quantifies the consequences of choosing  $\hat{\boldsymbol{\theta}}$  when the true value is  $\boldsymbol{\theta}$ . An optimal point estimate is the value  $\hat{\boldsymbol{\theta}}$  which minimizes the expected loss where the expectations are with respect to the posterior distribution of  $\boldsymbol{\theta}$ :

$$\min_{\hat{\boldsymbol{\theta}}} E \left( L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) \right) = \min_{\hat{\boldsymbol{\theta}}} \int L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta}.$$

In the case of a quadratic loss function  $L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})' \boldsymbol{\Phi} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})$  where  $\boldsymbol{\Phi}$  is a symmetric positive definite matrix, an optimal point estimate is given by the mean of the posterior distribution. Alternatively, if the loss is measured by an absolute error  $L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = |\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}|$ , then the median of the posterior distribution becomes an optimal point estimate.

## 12.3 Bayesian Intervals and Regions

A Bayesian counterpart of a classical confidence interval is called a posterior credible interval (or region if  $\boldsymbol{\theta}$  is a vector of parameters). For a scalar  $\theta$  in a parameter space

$\Omega$ , a  $100 \cdot (1 - \alpha)$  percent posterior credible interval is a subset  $S \subset \Omega$  such that

$$(12.3) \quad Pr(\theta \in S|\mathbf{y}) = \int_S p(\theta|\mathbf{y})d\theta = 1 - \alpha.$$

For any given  $\alpha$ , the interval  $S$  satisfying (12.3) may not be unique. Of those satisfying (12.3), a highest posterior density interval is obtained by imposing an additional condition that the value of  $p(\theta|\mathbf{y})$  at any  $\theta$  inside  $S$  is at least as large as that evaluated outside  $S$ ; that is,

$$p(\theta_i|\mathbf{y}) \geq p(\theta_j|\mathbf{y}) \text{ for all } \theta_i \in S \text{ and } \theta_j \notin S,$$

which implies that the end points of the interval, say  $\underline{\theta}$  and  $\bar{\theta}$ , satisfy  $p(\underline{\theta}|\mathbf{y}) = p(\bar{\theta}|\mathbf{y})$ .

If the posterior density is unimodal, a highest posterior density interval is an interval that satisfies (12.3) with a minimum distance between  $\underline{\theta}$  and  $\bar{\theta}$ .

While a highest posterior density interval is identical to a  $100 \cdot (1 - \alpha)$  percent confidence interval in the classical framework, their interpretations are different. A classical confidence interval is a random interval which would contain a fix value  $\theta$  with probability  $(1 - \alpha)$  if we repeatedly draw samples from population and construct an interval each time. On the other hand, a highest posterior density interval is a fixed interval within which a random variable  $\theta$  lies with probability  $(1 - \alpha)$ .

## 12.4 Posterior Odds Ratio and Hypothesis Testing

Posterior distributions are also employed to assess relative plausibility of competing hypotheses. We evaluate the relative plausibility with a ratio of posterior probabilities associated with the hypotheses, called a posterior odds ratio. Unlike the classical hypothesis testing, a posterior odds ratio test treats the competing hypotheses symmetrically, and its conclusion is not designed to necessarily accept or reject

the hypotheses. Instead, the test merely infers which hypothesis is more likely given the priors and sample information.

Suppose we are interested in comparing two hypotheses,  $H_0$  and  $H_1$ , with prior probabilities  $p(H_0)$  and  $p(H_1)$ . Let  $\boldsymbol{\theta}_i$  denote a parameter vector associated with hypothesis  $H_i$ ,  $i = 0, 1$ . For  $H_0$ , the joint density function for  $\mathbf{y}$ ,  $\boldsymbol{\theta}_0$ ,  $H_0$  is,

$$\begin{aligned} p(\mathbf{y}, \boldsymbol{\theta}_0, H_0) &= f(\mathbf{y})p(\boldsymbol{\theta}_0, H_0|\mathbf{y}) \\ &= p(\boldsymbol{\theta}_0, H_0)f(\mathbf{y}|\boldsymbol{\theta}_0, H_0) \end{aligned}$$

or

$$\begin{aligned} p(\boldsymbol{\theta}_0, H_0|\mathbf{y}) &= \frac{p(\boldsymbol{\theta}_0, H_0)f(\mathbf{y}|\boldsymbol{\theta}_0, H_0)}{f(\mathbf{y})} \\ (12.4) \qquad &= \frac{p(H_0)h(\boldsymbol{\theta}_0|H_0)f(\mathbf{y}|\boldsymbol{\theta}_0, H_0)}{f(\mathbf{y})}, \end{aligned}$$

where  $h(\boldsymbol{\theta}_0|H_0)$  is the conditional prior distribution for  $\boldsymbol{\theta}_0$  given  $H_0$ . The posterior distribution of  $H_0$  can be obtained by integrating (12.4) with respect to  $\boldsymbol{\theta}_0$ :

$$p(H_0|\mathbf{y}) = \frac{p(H_0) \int h(\boldsymbol{\theta}_0|H_0)f(\mathbf{y}|\boldsymbol{\theta}_0, H_0)d\boldsymbol{\theta}_0}{f(\mathbf{y})}.$$

Given that  $p(H_1|\mathbf{y})$  has been obtained in an analogous way, the posterior odds ratio is,

$$\begin{aligned} \frac{p(H_0|\mathbf{y})}{p(H_1|\mathbf{y})} &= \frac{p(H_0) \int h(\boldsymbol{\theta}_0|H_0)f(\mathbf{y}|\boldsymbol{\theta}_0, H_0)d\boldsymbol{\theta}_0}{p(H_1) \int h(\boldsymbol{\theta}_1|H_1)f(\mathbf{y}|\boldsymbol{\theta}_1, H_1)d\boldsymbol{\theta}_1} \\ (12.5) \qquad &= \frac{p(H_0) f(\mathbf{y}|H_0)}{p(H_1) f(\mathbf{y}|H_1)}. \end{aligned}$$

The larger the value of this ratio, the more the test is in favor of  $H_0$ .

The first term in (12.5),  $p(H_0)/p(H_1)$ , is called a prior odds ratio, and the second term  $f(\mathbf{y}|H_0)/f(\mathbf{y}|H_1)$  is the ratio of averaged likelihoods, called a Bayes factor. If

we assume, prior to observing the data, that the two hypotheses are equally likely, then the prior odds ratio is 1. In that case, the relative plausibility is determined by the Bayes factor, and we can conveniently interpret its value using the following scale developed by Jeffreys (1961):

Bayes factor	Evidence in favor of $H_0$
1:1 - 3:1	Very slight
3:1 - 10:1	Slight
10:1 - 100:1	Strong to very strong
100:1 -	Decisive

Although the posterior odds ratio itself does not make an explicit conclusion about accepting or rejecting one hypothesis with respect to the other, it is still possible to make an explicit choice between the two, if necessary. In such cases, a loss function is assumed to measure the consequences of choosing each hypothesis, and we accept one which yields the lowest expected loss, with the expectation with respect to the posterior probability of the hypothesis.

One useful application of a posterior odds ratio is the assessment of relative plausibility of competing models which may not be nested (for empirical applications, see Lubik and Schorfheide, 2007; Rabanal and Rubio-Ramirez, 2005). Its implementation follows the same procedure as simple hypothesis testing, but now the probabilities are conditional on the model specification, considering all possible parameter values rather than the parameters used by the model. Suppose we are interested in comparing two structural models  $\mathcal{M}_1$  and  $\mathcal{M}_2$  with an associated parameter vector  $\theta_i$  and prior probability  $p(\mathcal{M}_i)$ ,  $i = 1, 2$ . Let  $\mathbf{y}$  denote sample observations on variables in

the model. As in (12.5), the posterior odds ratio is given by

$$\begin{aligned} \frac{p(\mathcal{M}_1|\mathbf{y})}{p(\mathcal{M}_2|\mathbf{y})} &= \frac{p(\mathcal{M}_1) \int h(\boldsymbol{\theta}_1|\mathcal{M}_1)f(\mathbf{y}|\boldsymbol{\theta}_1, \mathcal{M}_1)d\boldsymbol{\theta}_1}{p(\mathcal{M}_2) \int h(\boldsymbol{\theta}_2|\mathcal{M}_2)f(\mathbf{y}|\boldsymbol{\theta}_2, \mathcal{M}_2)d\boldsymbol{\theta}_2} \\ &= \frac{p(\mathcal{M}_1) f(\mathbf{y}|\mathcal{M}_1)}{p(\mathcal{M}_2) f(\mathbf{y}|\mathcal{M}_2)}. \end{aligned}$$

Again, if the two models are equally likely a priori, the prior odds ratio is 1, and the Bayes factor can be interpreted according to Jeffreys' scale.

## Appendix

### 12.A Numerical Approximation Methods

As we have seen, calculating an explicit form of posterior distributions often involves evaluation of high-dimensional integrals. In practice, the integrals of high-order functions are increasingly difficult to solve analytically, and, as a result, the posterior distribution may be intractable. To overcome this difficulty, numerical approximation methods are prominently used in the Bayesian analysis. This section explains three leading simulation techniques popularly used in the literature: the Importance Sampling, the Gibbs sampler and the Metropolis-Hastings algorithm. The latter two are in the class of the Markov chain Monte Carlo methods.

#### 12.A.1 Importance Sampling

The idea behind the importance sampling is to obtain sample draws  $\{\theta_i\}$  from some known distribution and assign weights to each draw so that the limiting distribution of the weighted sample converges to the target distribution.

Suppose we are interested in evaluating

$$(12.A.1) \quad E[h(\theta)] = \int h(\theta)f(\theta)d\theta$$

but  $f(\theta)$  is not available as a sampling distribution. Let  $I(\theta|\mu)$  denote a known distribution from which  $\{\theta_i\}$  can be obtained. This distribution is called the importance sampler and  $\mu$  represents its parameterization. Equation (12.A.1) can be rewritten as

$$\begin{aligned} E[h(\theta)] &= \int h(\theta) \frac{f(\theta)}{I(\theta)} I(\theta) d\theta \\ (12.A.2) \qquad &= \int h(\theta) w(\theta) I(\theta) d\theta. \end{aligned}$$

where  $w(\theta) \equiv f(\theta)/I(\theta)$ . In (12.A.2),  $w(\theta)$  serves to mitigate the direct influence of  $I(\theta|\mu)$  on  $\theta_i$  by assigning the weight or “importance” of different points in the sample space.

After a sample  $\{\theta_i\}_{i=1}^N$  has been obtained from  $I(\theta)$  rather than  $f(\theta)$  for some large  $N$ ,  $E[h(\theta)]$  is approximated by the sample mean:

$$\hat{h} = \frac{1}{N} \sum_{i=1}^N h(\theta_i) w(\theta_i).$$

Geweke (1989) outlines criteria for choosing an importance sampler and formal diagnostics for the adequacy of a chosen sampler. Poor samplers tend to assign weights on only a small fraction of the sample rather than being approximately uniform, requiring a large number of draws to achieve convergence.

## 12.A.2 Markov Chain Monte Carlo

Markov chain Monte Carlo (MCMC) methods are iterative sampling schemes to generate sample draws  $\{x_i\}$  with the Markov property:

$$Pr(x_{i+1}|x_i, x_{i-1}, x_{i-2}, \dots) = Pr(x_{i+1}|x_i) \text{ for all } i$$

where  $i$  indexes the Monte Carlo draws. These computer-intensive algorithms are particularly powerful in approximating multi-dimensional integrals with high accu-

racy. This section explains two widely used methods to simulate Markov chains: the Gibbs sampler and the Metropolis-Hastings algorithm. Further details are provided by Casella and George (1992) for the Gibbs sampler and Chib and Greenberg (1995) for the Metropolis-Hastings algorithm.

### The Gibbs Sampler

Consider a  $q$ -dimensional vector of parameters  $\boldsymbol{\theta}$  that is partitioned into  $k$  blocks,  $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \dots, \boldsymbol{\theta}_k)$ ,  $k \leq q$ . Suppose we wish to obtain the marginal distribution of the  $i^{\text{th}}$  block:

$$P(\boldsymbol{\theta}_i|\mathbf{x}) = \int \cdots \int P(\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_k|\mathbf{y}) d\boldsymbol{\theta}_1, \dots, d\boldsymbol{\theta}_{i-1} d\boldsymbol{\theta}_{i+1} \cdots d\boldsymbol{\theta}_k$$

when the joint density  $P(\boldsymbol{\theta}|\mathbf{y})$  is intractable. We assume that, for all  $i$ , the conditional posterior probability density for  $\boldsymbol{\theta}_i$ ,  $P(\boldsymbol{\theta}_i|\mathbf{y}, \boldsymbol{\theta}_{-i})$ , is available as a sampling distribution where  $\boldsymbol{\theta}_{-i}$  denotes all components of  $\boldsymbol{\theta}$  excluding  $\boldsymbol{\theta}_i$ . The Gibbs sampler generates a Markov chain of random variables  $\boldsymbol{\theta}_i^{(1)}, \dots, \boldsymbol{\theta}_i^{(N)} \sim P(\boldsymbol{\theta}_i|\mathbf{y})$  by sampling from  $P(\boldsymbol{\theta}_i|\mathbf{y}, \boldsymbol{\theta}_{-i})$ .

The algorithm is initiated with some starting values  $\boldsymbol{\theta}^{(0)} = (\boldsymbol{\theta}_1^{(0)}, \boldsymbol{\theta}_2^{(0)}, \dots, \boldsymbol{\theta}_k^{(0)})$ , and the subsequent sampling proceeds as follows.

- (i) Draw a random observation  $\boldsymbol{\theta}_1^{(1)}$  from  $P(\boldsymbol{\theta}_1|\mathbf{y}, \boldsymbol{\theta}_2^{(0)}, \boldsymbol{\theta}_3^{(0)}, \dots, \boldsymbol{\theta}_k^{(0)})$ .
- (ii) Draw a random observation  $\boldsymbol{\theta}_2^{(1)}$  from  $P(\boldsymbol{\theta}_2|\mathbf{y}, \boldsymbol{\theta}_1^{(1)}, \boldsymbol{\theta}_3^{(0)}, \dots, \boldsymbol{\theta}_k^{(0)})$ .
- $\vdots$
- (iii) Draw a random observation  $\boldsymbol{\theta}_k^{(1)}$  from  $P(\boldsymbol{\theta}_k|\mathbf{y}, \boldsymbol{\theta}_1^{(1)}, \boldsymbol{\theta}_2^{(1)}, \dots, \boldsymbol{\theta}_{k-1}^{(1)})$ .
- (iv) Return to step 1 and draw  $\boldsymbol{\theta}_1^{(2)}$  from  $P(\boldsymbol{\theta}_1|\mathbf{y}, \boldsymbol{\theta}_2^{(1)}, \boldsymbol{\theta}_3^{(1)}, \dots, \boldsymbol{\theta}_k^{(1)})$ , and so on.

Repeating this process  $N$  times generates a Markov chain of length  $N$ ,  $\{\boldsymbol{\theta}^{(j)}\}_{j=1}^N$ .

The effect of the fixed starting values  $\boldsymbol{\theta}^{(0)}$  is eliminated by discarding some iterations at the beginning of the chain, a practice called a burn-in. With the remaining  $m$  observations,  $P(\boldsymbol{\theta}_i|\mathbf{y})$  is approximated by

$$\hat{P}(\boldsymbol{\theta}_i|\mathbf{x}) = \frac{1}{m} \sum_{j=1}^m P(\boldsymbol{\theta}_i|\mathbf{y}, \boldsymbol{\theta}_{-i}^{(j)}).$$

Alternatively, Gelfand and Smith (1990) suggest generating  $s$  independent Markov chains of length  $N$  and using the final value  $\boldsymbol{\theta}^{(N)}$  from each sequence. Other approaches to exploiting convergence are discussed in Casella and George (1992).

### Metropolis-Hastings Algorithm

The Gibbs sampler described above requires that the full conditional distribution is available in a tractable form as a sampling distribution for  $\boldsymbol{\theta}^{(i)}$ . There are also MCMC methods for the case in which it is unavailable. The best known of these is the Metropolis-Hastings algorithm.

Suppose the target density  $P(\boldsymbol{\theta}|\mathbf{x})$  is not available as a sampling distribution, but there is a known density  $g(\boldsymbol{\theta}|\boldsymbol{\theta}^{(i-1)}, \boldsymbol{\mu})$ , where  $\int g(\boldsymbol{\theta}|\boldsymbol{\theta}^{(i-1)}, \boldsymbol{\mu})d\boldsymbol{\theta} = 1$ , from which  $\boldsymbol{\theta}^{(i)}$  can be obtained. The Metropolis-Hastings algorithm is initialized with a starting value  $\boldsymbol{\theta}^{(0)}$  and, given  $\{\boldsymbol{\theta}^{(j)}\}_{j=1}^{i-1}$ ,  $\boldsymbol{\theta}^{(i)}$  is obtained as follows:

(i) Draw a random sample  $\tilde{\boldsymbol{\theta}}^{(i)}$  from  $g(\boldsymbol{\theta}|\boldsymbol{\theta}^{(i-1)}, \boldsymbol{\mu})$ . This serves as a candidate for  $\boldsymbol{\theta}^{(i)}$ .

(ii) Define the probability of accepting  $\tilde{\boldsymbol{\theta}}^{(i)}$  for  $\boldsymbol{\theta}^{(i)}$ :

$$(12.A.3) \quad \pi\left(\tilde{\boldsymbol{\theta}}^{(i)}|\boldsymbol{\theta}^{(i-1)}\right) = \min\left(1, \frac{P(\tilde{\boldsymbol{\theta}}^{(i)}|\mathbf{x})}{P(\boldsymbol{\theta}^{(i-1)}|\mathbf{x})} \frac{g(\boldsymbol{\theta}^{(i-1)}|\boldsymbol{\theta}^{(i-1)}, \boldsymbol{\mu})}{g(\tilde{\boldsymbol{\theta}}^{(i)}|\boldsymbol{\theta}^{(i-1)}, \boldsymbol{\mu})}\right).$$

(iii) Draw a value  $\delta$  from a uniform distribution on  $[0, 1]$ .

(iv) If  $\pi(\tilde{\boldsymbol{\theta}}^{(i)}|\boldsymbol{\theta}^{(i-1)}) > \delta$ , set  $\boldsymbol{\theta}^{(i)} = \tilde{\boldsymbol{\theta}}^{(i)}$ ; otherwise, discard  $\tilde{\boldsymbol{\theta}}^{(i)}$  and draw a new candidate.

A sequence of accepted draws  $\{\boldsymbol{\theta}^{(i)}\}_{i=1}^N$  is a Markov chain with transition probability  $\lambda(\tilde{\boldsymbol{\theta}}^{(i)}|\tilde{\boldsymbol{\theta}}^{(i-1)}) = \pi(\tilde{\boldsymbol{\theta}}^{(i)}|\boldsymbol{\theta}^{(i-1)})g(\tilde{\boldsymbol{\theta}}^{(i)}|\boldsymbol{\theta}^{(i-1)}, \boldsymbol{\mu})$  for  $i = 1, \dots, N$  and  $\tilde{\boldsymbol{\theta}}^{(i)} \neq \tilde{\boldsymbol{\theta}}^{(i-1)}$ . Under mild regularity conditions, this converges in distribution to  $P(\boldsymbol{\theta}|\mathbf{x})$  as  $N$  increases.

Note that the calculation of  $\pi(\tilde{\boldsymbol{\theta}}^{(i)}|\boldsymbol{\theta}^{(i-1)})$  does not require knowledge about a normalizing constant in  $P(\cdot)$  or  $g(\cdot)$  since they appear in both the numerator and the denominator of (12.A.3) and simply cancel out. This is one of the attractive features of this algorithm for approximating posterior distributions since they are often known up to constant proportionality as in (12.2).

In application, the candidate-generating density  $g(\boldsymbol{\theta}|\boldsymbol{\theta}^{(i-1)}, \boldsymbol{\mu})$  can be specified in various ways. A random walk chain utilizes  $g(\boldsymbol{\theta}|\boldsymbol{\theta}^{(i-1)}, \boldsymbol{\mu}) = g_1(\boldsymbol{\theta} - \boldsymbol{\theta}^{(i-1)}|\boldsymbol{\mu})$ , and  $\tilde{\boldsymbol{\theta}}^{(i)}$  follows the process  $\tilde{\boldsymbol{\theta}}^{(i)} = \boldsymbol{\theta}^{(i-1)} + \boldsymbol{\varepsilon}_i$  where  $\boldsymbol{\varepsilon}_i \sim g(\boldsymbol{\varepsilon})$  (Metropolis *et al.*, 1953). Choices for  $g_1$  include the multivariate normal and the multivariate- $t$  densities. Alternatively, an independent chain draws a candidate independently of the last accepted draw. This is implemented by choosing a density that is independent across all Monte Carlo replications:  $g(\boldsymbol{\theta}|\boldsymbol{\theta}^{(i-1)}, \boldsymbol{\mu}) = g_2(\boldsymbol{\theta}|\boldsymbol{\mu})$  (Hastings, 1970). Another possibility is an autoregressive chain. A vector autoregressive process of order 1 follows  $\tilde{\boldsymbol{\theta}}^{(i)} = \mathbf{a} + \mathbf{B}(\boldsymbol{\theta}^{(i-1)} - 1) + \mathbf{v}_i$  drawn from the density  $g(\boldsymbol{\theta}|\boldsymbol{\theta}^{(i-1)}, \boldsymbol{\mu}) = g(\tilde{\boldsymbol{\theta}}^{(i)} - \mathbf{a} - \mathbf{B}(\boldsymbol{\theta}^{(i-1)} - 1))$  where  $\mathbf{a}$  is a vector,  $\mathbf{B}$  is a matrix, and  $\mathbf{v}_i \sim g(\mathbf{v})$  (Tierney, 1994).

## 12.B Application of the MCMC methods

In this section, we describe an application of the MCMC methods by nan Chen, Watanabe, and Yabu (1990). They propose a new method of data augmentation based on the Gibbs sampler to account for an endogeneity problem arising from the

use of time-aggregated data. Their application considers the estimation of the effects of foreign exchange interventions by a central bank.

Suppose the exchange rate movements and the central bank's intervention in the foreign exchange market can be represented by the following two-equation system:

$$(12.B.4) \quad s_{t,h} - s_{t,h-1} = \alpha I_{t,h} + \varepsilon_{t,h}$$

$$(12.B.5) \quad I_{t,h} = \beta(s_{t,h-1} - s_{t-1,h-1}) + \eta_{t,h}$$

for  $t = 1, \dots, T$  and  $h = 1, \dots, 24$ , where  $s_{t,h}$  is the log price of the domestic currency per unit of the foreign currency at hour  $h$  of day  $t$ ,  $I_{t,h}$  is the central bank's purchase of the domestic currency between  $h - 1$  and  $h$  of day  $t$ ,  $\varepsilon_{t,h} \sim i.i.d.N(0, \sigma_\varepsilon^2)$ , and  $\eta_{t,h} \sim i.i.d.N(0, \sigma_\eta^2)$ . If  $s_{t,h}$  and  $I_{t,h}$  are both observable at the hourly frequency, we can obtain unbiased estimates of  $\alpha$  and  $\beta$  by estimating (12.B.4) and (12.B.5) by OLS.

Suppose instead that  $I_{t,h}$  is not observable and only the daily sum of hourly interventions  $I_t \equiv \sum_{h=1}^{24} I_{t,h}$  is available. The above model can be transformed into a daily-frequency model by summing up both sides of (12.B.4) and (12.B.5) over  $h$ :

$$(12.B.6) \quad s_{t,24} - s_{t-1,24} = \alpha I_t + \varepsilon_t$$

$$(12.B.7) \quad I_t = \beta \sum_{h=1}^{24} (s_{t,h-1} - s_{t-1,h-1}) + \eta_t$$

where  $s_{t,24} - s_{t-1,24} = \sum_{h=1}^{24} (s_{t,h} - s_{t,h-1})$  and  $x_t = \sum_{h=1}^{24} x_{t,h}$  for  $x = \{I, \varepsilon, \eta\}$ . This model, however, suffers from an endogeneity problem, and the OLS estimates from (16.56) and (16.57) may be biased. To see this, consider a rise in  $\varepsilon_{t,h}$ . It increases  $s_{t,h} - s_{t,h-1}$  in (12.B.4) and  $I_{t,h+1}$  in (12.B.5) for  $\beta > 0$ , and hence  $I_t$  and  $\varepsilon_t$  are positively correlated. Alternatively, a rise in  $\eta_{t,h}$  increases  $I_{t,h}$  in (12.B.5) and

appreciates the currency in (12.B.4) for  $\alpha < 0$ , implying that  $\sum(s_{t,h} - s_{t,h-1})$  and  $\eta_t$  are negatively correlated.

Recognizing this problem, nan Chen, Watanabe, and Yabu (1990) propose an algorithm to obtain a posterior distribution of the parameters using the Gibbs sampler. They first introduce an auxiliary variable  $I_{t,h}$  to substitute the unobserved hourly interventions, and assume a flat distribution as the priors of  $\alpha$  and  $\beta$ , and distributions  $IG\left(\frac{v_\varepsilon}{2}, \frac{\delta_\varepsilon}{2}\right)$  and  $IG\left(\frac{v_\eta}{2}, \frac{\delta_\eta}{2}\right)$  for  $\sigma_\varepsilon^2$  and  $\sigma_\eta^2$ . The algorithm proceeds as follows.<sup>1</sup>

(i) Generate  $\alpha$  conditional on  $s_{t,h}$ ,  $I_{t,h}$ , and  $\sigma_\varepsilon^2$ . The posterior distribution is  $\alpha \sim N(\phi_s, \omega_s)$  where  $\phi_s = \sum I_{t,h}(s_{t,h} - s_{t,h-1}) / \sum I_{t,h}^2$  and  $\omega_s = \sigma_\varepsilon^2 / \sum I_{t,h}^2$ .

(ii) Generate  $\beta$  conditional on  $s_{t,h}$ ,  $I_{t,h}$ , and  $\sigma_\eta^2$ . The posterior distribution is  $\beta \sim N(\phi_I, \omega_I)$  where  $\phi_I = \sum I_{t,h}(s_{t,h-1} - s_{t-1,h-1}) / \sum (s_{t,h-1} - s_{t-1,h-1})^2$ , and  $\omega_I = \sigma_\eta^2 / \sum (s_{t,h-1} - s_{t-1,h-1})^2$ .

(iii) Generate  $\sigma_\varepsilon^2$  conditional on  $s_{t,h}$ ,  $I_{t,h}$ , and  $\alpha$ . The posterior distribution is  $\sigma_\varepsilon^2 \sim IG\left(\frac{v_\varepsilon+T}{2}, \frac{\delta_\varepsilon+RRS_s}{2}\right)$  where  $RRS_s = \sum (s_{t,h} - s_{t,h-1} - \alpha I_{t,h})^2$ .

(iv) Generate  $\sigma_\eta^2$  conditional on  $s_{t,h}$ ,  $I_{t,h}$ , and  $\beta$ . The posterior distribution is  $\sigma_\eta^2 \sim IG\left(\frac{v_\eta+T}{2}, \frac{\delta_\eta+RRS_I}{2}\right)$  where  $RRS_I = \sum (I_{t,h} - \beta(s_{t,h-1} - s_{t-1,h-1}))^2$ .

(v) Generate  $I_{t,h}$  conditional on  $s_{t,h}$ ,  $I_t$ ,  $\alpha$ ,  $\beta$ ,  $\sigma_\varepsilon^2$ , and  $\sigma_\eta^2$ . The posterior distribution is derived as follows.

If  $I_t$  is unknown, the posterior distribution is given by  $(I_{t,1}, \dots, I_{t,24})' \sim N(\Xi_t, \Phi)$  where  $\Xi_t = (\xi_{t,1}, \dots, \xi_{t,24})'$  and  $\Phi = \text{diag}(\psi, \dots, \psi)$  with  $\psi = \left(\frac{1}{\sigma_\eta^2} + \frac{\alpha^2}{\sigma_\varepsilon^2}\right)^{-1}$  and  $\xi_{t,h} = \left(\psi \frac{1}{\sigma_\eta^2}\right) [\beta(s_{t,h-1} - s_{t-1,h-1})] + \left(\psi \frac{\alpha^2}{\sigma_\varepsilon^2}\right) [\alpha^{-1}(s_{t,h} - s_{t,h-1})]$ . Since  $I_t$  is known, consider the posterior distribution  $(I_{t,1}, \dots, I_{t,23}, I_t)' \sim N(\Xi_t^*, \Phi^*)$  where  $\Xi_t^* = \mathbf{B}\Xi_t$

<sup>1</sup>The summations indicate  $\sum \equiv \sum_{t=1}^T \sum_{h=1}^{24}$ .

and  $\Phi^* = \mathbf{B}\Phi\mathbf{B}'$  with

$$\mathbf{B}_{(24 \times 24)} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{bmatrix}.$$

Partition  $\Xi_t^*$  and  $\Phi^*$  as follows:

$$\Xi_t^* = \begin{bmatrix} \Xi_{t,1}^* \\ (23 \times 1) \\ \Xi_{t,2}^* \\ (1 \times 1) \end{bmatrix}, \Phi^* = \begin{bmatrix} \Phi_{11}^* & \Phi_{12}^* \\ (23 \times 23) & (23 \times 1) \\ \Phi_{21}^* & \Phi_{22}^* \\ (1 \times 23) & (1 \times 1) \end{bmatrix}.$$

The posterior distribution of  $(I_{t,1} \cdots, I_{t,23})$  conditional on  $I_t$  is given by

$$(I_{t,1} \cdots, I_{t,23} | I_t)' \sim N(\Xi_{t,1}^* + \Phi_{12}^* (\Phi_{22}^*)^{-1} (I_t - \Xi_{t,2}^*), \Phi_{11}^* - \Phi_{12}^* (\Phi_{22}^*)^{-1} \Phi_{21}^*).$$

After  $(I_{t,1} \cdots, I_{t,23})$  has been generated from this posterior distribution,  $I_{t,24}$  is obtained from  $I_{t,24} = I_t - \sum_{h=1}^{23} I_{t,h}$ .

Applying this method to the Japanese data, nan Chen, Watanabe, and Yabu (1990) generate three Markov chains of the length 2,000 after discarding the first 2,000 draws in each chain as a burn-in phase. They obtain the point estimate of each parameter using the mean of the generated posterior distribution, and find that the effect of intervention is more than twice as large as the magnitude estimated by OLS using daily observations, suggesting the quantitative significance of the endogeneity problem.

## References

- CASELLA, G., AND E. I. GEORGE (1992): "Explaining the Gibbs Sampler," *American Statistician*, 46(3), 167–174.
- CHIB, S., AND E. GREENBERG (1995): "Understanding the Metropolis-Hastings Algorithm," *American Statistician*, 49(4), 327–335.
- DEJONG, D. N., AND C. DAVE (2007): *Structural Macroeconometrics*. Princeton University Press.
- GELFAND, A. E., AND A. F. M. SMITH (1990): "Sampling-Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association*, 85(410), 398–409.

- GEWEKE, J. (1989): “Bayesian Inference in Econometric Models Using Monte Carlo Integration,” *Econometrica*, 57(6), 1317–1339.
- HASTINGS, W. K. (1970): “Monte Carlo Sampling Methods Using Markov Chains and Their Applications,” *Biometrika*, 57, 97–109.
- JEFFREYS, H. (1961): *Theory of Probability*. Oxford University Press, New York.
- JUDGE, G. G., W. E. GRIFFITHS, R. C. HILL, H. LÜTKEPOHL, AND T. LEE (1985): *The Theory and Practice of Econometrics*. Wiley, New York, 2nd edn.
- LUBIK, T., AND F. SCHORFHEIDE (2007): “Do Central Banks Respond to Exchange Rate Movements? A Structural Investigation,” *Journal of Monetary Economics*, 54(4), 1069–1087.
- METROPOLIS, N., A. W. ROSENBLUTH, M. N. ROSENBLUTH, A. H. TELLER, AND E. TELLER (1953): “Equation of State Calculations by Fast Computing Machines,” *Journal of Chemical Physics*, 21(6), 1087–1092.
- NAN CHEN, C., T. WATANABE, AND T. YABU (1990): “A New Method for Identifying the Effects of Foreign Exchange Interventions,” IMES Discussion Paper Series, No. 2009-E-6, Bank of Japan.
- RABANAL, P., AND J. F. RUBIO-RAMIREZ (2005): “Comparing New Keynesian Models of the Business Cycle: A Bayesian Approach,” *Journal of Monetary Economics*, 52(6), 1151–1166.
- TIERNEY, L. (1994): “Markov Chains for Exploring Posterior Distributions,” *Annals of Statistics*, 22(4), 1701–1728.
- ZELLNER, A. (1996): *An Introduction to Bayesian Inference in Econometrics*. Wiley, New York, Reprint of 1971 ed.

## Chapter 13

# UNIT ROOT NONSTATIONARY PROCESSES

This chapter concerns univariate stochastic processes. Since the seminal work of Nelson and Plosser (1982), much theoretical and empirical research has been done in the area of unit root nonstationarity. They found that the null hypothesis of unit root nonstationarity was not rejected for many macroeconomic series. When one or more variables of interest are unit root nonstationary, standard asymptotic distribution theory does not apply to the econometric system involving these variables. The spurious regression results discussed in Section 14.2 are concrete examples of this type of problem.

When a variable is unit root nonstationary, it has a stochastic trend. If linear combinations of two or more unit root nonstationary variables do not contain stochastic trends, then these variables are said to be cointegrated. Then the cointegrating vector, which eliminates the stochastic trends, can be estimated consistently by regressions without the use of instrumental variables, even when no variables are exogenous. If the cointegrating vector includes structural parameters, then the econometrician can estimate these structural parameters without making exogeneity

assumptions.<sup>1</sup>

The rest of this chapter is organized as follows. In Section 13.1, univariate unit root econometrics is discussed. It begins with definitions of basic concepts such as difference stationarity and trend stationarity. Then a decomposition of a difference stationary variable into a deterministic trend, a stochastic trend, and a stationary component is discussed. Spurious regression results, tests for the null of difference stationarity, and tests for the null of stationarity are reviewed.

## 13.1 Definitions

Consider a univariate stochastic process,  $\{x_t : t = \dots, -2, -1, 0, 1, 2, \dots\}$ , which is a sequence of random variables. Many macroeconomic variables tend to grow over time, so that their distributions shift upward over time. Hence they are not stationary. However, there are many possible forms of nonstationarity, and it is not clear which form of nonstationarity is appropriate in representing macroeconomic variables. It may be reasonable to assume that the growth rate or the first difference of the (natural) log of a variable is stationary for many macroeconomic variables. Let us now assume that the first difference of  $x_t$  ( $\Delta x_t = x_t - x_{t-1}$ ) is stationary. Then  $x_t$  is either difference stationary or trend stationary. If  $x_t$  is stationary after removing a deterministic time trend, then  $x_t$  is said to be trend stationary. Since  $\Delta x_t$  is assumed to be stationary,  $x_t$  has a linear time trend when  $x_t$  is trend stationary:

$$(13.1) \quad x_t = \theta + \mu t + \epsilon_t,$$

---

<sup>1</sup>Stock and Watson (1988), Diebold and Nerlove (1990), Campbell and Perron (1991), and Watson (1994) are examples of surveys for unit root econometrics.

where  $\epsilon_t$  is stationary with mean zero.<sup>2</sup> If  $\Delta x_t$  is stationary but  $x_t$  is not trend stationary, then  $x_t$  is said to be difference stationary. Alternatively, it is called unit root nonstationary or integrated of order one. The trend stationary and difference stationary processes have different properties on their long-run variances. The long-run variance of a stationary variable  $y_t$  is defined by

$$(13.2) \quad \omega^2 = \sum_{\tau=-\infty}^{\infty} E\{[y_t - E(y_t)][y_{t-\tau} - E(y_t)]\}.$$

After taking the first difference, a difference stationary process has a positive long-run variance, while trend stationary process has a long-run variance of zero.

A special case of a difference stationary process is a random walk. If  $E(x_{t+1} | x_t, x_{t-1}, x_{t-2}, \dots) = x_t$  and if  $E((\Delta x_{t+1})^2 | x_t, x_{t-1}, x_{t-2}, \dots)$  is constant over time, then  $x_t$  is a random walk. In general, if  $x_t$  is difference stationary, then  $\Delta x_t$  has nonzero serial correlation; however, if  $x_t$  is a random walk, then  $\Delta x_t$  does not have serial correlation.

## 13.2 Decompositions

It is often convenient to decompose a difference stationary process into components representing a deterministic trend, a stochastic trend, and a stationary component.

Let  $x_t$  be a difference stationary process:

$$(13.3) \quad x_t - x_{t-1} = \mu + \epsilon_t$$

for  $t \geq 1$  where  $\epsilon_t$  is stationary with mean zero. Here  $\mu$  is called a drift, which is the

---

<sup>2</sup>Note that  $\epsilon_t$  is not assumed to be *iid* because serial correlation is allowed in a stationary process.

mean of  $\Delta x_t$ . Then

$$\begin{aligned}
 (13.4) \quad x_t &= \mu + x_{t-1} + \epsilon_t \\
 &= 2\mu + x_{t-2} + \epsilon_{t-1} + \epsilon_t \\
 &= 3\mu + x_{t-3} + \epsilon_{t-2} + \epsilon_{t-1} + \epsilon_t \\
 &= \dots \\
 &= \mu t + x_0 + \sum_{\tau=1}^t \epsilon_\tau.
 \end{aligned}$$

Hence

$$(13.5) \quad x_t = \mu t + x_t^0,$$

where  $x_t^0$  is

$$(13.6) \quad x_t^0 = x_0 + \sum_{\tau=1}^t \epsilon_\tau,$$

where  $x_0$  is an initial value. Relation (13.5) decomposes the difference stationary process  $x_t$  into a deterministic trend arising from drift  $\mu$ , and the difference stationary process without drift,  $x_t^0$ .

Let us now consider Beveridge and Nelson (1981) decomposition, which further decompose  $x_t^0$  into a random walk component and a stationary component. Since  $\Delta x_t^0$  is covariance stationary, it has the Wold representation:

$$(13.7) \quad (1 - L)x_t^0 = A(L)\nu_t,$$

where  $L$  is the lag operator,  $A(L) = \sum_{\tau=0}^{\infty} A_\tau L^\tau$ ,  $A_0 = 1$ ,  $\nu_t = x_t^0 - \hat{E}(x_t^0 | x_{t-1}^0, x_{t-2}^0, \dots)$ , and  $\hat{E}(\cdot | x_{t-1}^0, x_{t-2}^0, \dots)$  is the linear projection operator. Then

$$(13.8) \quad x_t^0 = z_t + c_t,$$

where

$$(13.9) \quad z_t = z_{t-1} + A(1)\nu_t,$$

is the random walk component or a stochastic trend, and

$$(13.10) \quad c_t = -\left\{ \left( \sum_{\tau=1}^{\infty} A_{\tau} \right) \nu_t + \left( \sum_{\tau=2}^{\infty} A_{\tau} \right) \nu_{t-1} + \left( \sum_{\tau=3}^{\infty} A_{\tau} \right) \nu_{t-2} + \cdots \right\}$$

is the stationary component of  $x_t$ . Thus a difference stationary process  $x_t$  is decomposed into a deterministic trend, a stochastic trend, and a stationary component.

The variance of the random walk component,  $\text{Var}(\Delta z_t)$ , is equal to  $A(1)^2 \text{Var}(\nu_t)$ , which in turn is equal to the long-run variance of  $\Delta x_t$  and  $2\pi$  times the spectral density of  $\Delta x_t$  at frequency zero. If the long-run variance is zero, then  $x_t = \mu t + c_t$ , and  $x_t$  is trend stationary.

Cochrane (1988), among others, uses  $\frac{\text{Var}(\Delta z_t)}{\text{Var}(\Delta x_t)}$  as a measure of the persistence of  $x_t$ . This measure is zero for trend stationary  $x_t$  and is one for a random walk. He estimates  $\text{Var}(\Delta z_t)$  by  $\frac{1}{k}$  times the variance of  $k$ -differences of  $x_t$ ,  $\frac{1}{k} \text{Var}(\Delta^k x_t)$ , for a large enough  $k$ . His estimator is essentially the same as the Bartlett estimator, which was advocated by Newey and West (1987) in a different context. Any estimator of the long-run variance or the spectral density at frequency zero can be used for the purpose of estimating Cochrane's measure of persistence.

### 13.3 Tests for the Null of Difference Stationarity

This section explains Dickey-Fuller (1979), Said-Dickey (1984), Phillips-Perron (1988), and Park's (1990) tests for the null of difference stationarity. More recent work to improve small sample properties of tests includes Kahn and Ogaki (1990), Elliott, Rothenberg, and Stock (1996), and Hansen (1993).

### 13.3.1 Dickey-Fuller Tests

Dickey and Fuller (1979) propose to test for the null of a unit root in an AR(1) model:<sup>3</sup>

$$(13.11) \quad x_t = \theta + \mu t + \alpha x_{t-1} + \epsilon_t.$$

where  $\epsilon_t$  is NID. One of their tests is based on  $T(\hat{\alpha} - 1)$ , where  $T$  is the sample size and  $\hat{\alpha}$  is the OLS estimator for  $\alpha$  in (13.11). Another test is based on the  $t$ -ratio for the hypothesis  $\alpha = 1$ . These test statistics do not have standard distributions. Depending on whether or not a constant and a linear time trend are included, distributions of these tests under the null are different.<sup>4</sup> Fuller (1976, Tables 8.5.1 and 8.5.2) tabulates critical values for Dickey-Fuller tests.

Whether or not a constant and a linear time trend should be included in the regression depends on what type of alternative is appropriate. If the alternative hypothesis is that  $x_t$  is stationary with mean zero, then no deterministic terms should be included. This alternative is not appropriate for most macroeconomic time series. If the alternative hypothesis is that  $x_t$  is stationary with unknown mean, then a constant should be included. This alternative is appropriate for the time series that exhibit no consistent tendency to grow (or shrink) over time. On the other hand, if the alternative is that  $x_t$  is trend stationary, then a constant and a linear time trend should be included. This alternative is appropriate for the time series that exhibit a consistent tendency to grow (or shrink) over time. When these test statistics are

---

<sup>3</sup>It should be noted that Dickey and Fuller's (1981) joint tests with deterministic terms can have significantly lower power than Dickey and Fuller's (1979) one-tailed single unit root tests as explained by Park (1989).

<sup>4</sup>If the data are demeaned prior to the regression, then the test statistics have the same distributions as those from the regression with a constant in (13.11). If the data are detrended prior to the regression, then the test statistics have the same distributions as those from the regression with a constant and a linear time trend.

negative and greater than the appropriate critical value in absolute value, then the null of a unit root is rejected in favor of one of these alternatives.

Dickey-Fuller tests assume that the econometrician knows the order of autoregression. The following tests treat the case of unknown order of autoregression.

### 13.3.2 Said-Dickey Test

Said and Dickey (1984) extend the Dickey-Fuller's  $t$ -ratio test to the case where the order of autoregression is unknown. Consider an AR process of order  $p$ :

$$(13.12) \quad x_t = \theta + \mu t + a_1 x_{t-1} + a_2 x_{t-2} + \cdots + a_p x_{t-p} + \nu_t.$$

We assume that this process' autoregressive roots are less than or equal to one in absolute value, and that there is at most one root whose absolute value is equal to one. If there is a root with absolute value equal to one, then the root is assumed to be one, so that the process is unit root nonstationary. It should be noted that the null hypothesis that  $a_1 = 1$  in (13.12) does not have anything to do with the unit root hypothesis if  $p > 1$ . The unit root hypothesis is concerned with the autoregressive roots, and not with autoregressive coefficients. The first order autoregressive coefficient is equal to the autoregressive root only for an AR(1) process. For the purpose of testing for a unit root, it is convenient to reparameterize (13.12) as follows:<sup>5</sup>

$$(13.13) \quad \Delta x_t = \theta + \mu t + \rho x_{t-1} + \beta_1 \Delta x_{t-1} + \cdots + \beta_{p-1} \Delta x_{t-p+1} + \nu_t,$$

where

$$(13.14) \quad \rho = -(1 - a_1 - a_2 - \cdots - a_p),$$

---

<sup>5</sup>For example, consider an AR(2) process. Rearranging (13.12) yields  $x_t - x_{t-1} = \theta + \mu t - (1 - a_1 - a_2)x_{t-1} - a_2(x_{t-1} - x_{t-2}) + \nu_t$ . Therefore, we obtain  $\Delta x_t = \theta + \mu t + \rho x_{t-1} + \beta_1 \Delta x_{t-1} + \nu_t$ , where  $\rho = -(1 - a_1 - a_2)$  and  $\beta_1 = -a_2$ .

and

$$(13.15) \quad \beta_i = -[a_{i+1} + a_{i+2} + \cdots + a_p] \quad \text{for } i = 1, 2, \dots, p-1.$$

With this reparameterization,  $\Delta x_t$  has an invertible autoregressive representation when  $\rho = 0$ . Hence  $x_t$  is unit root nonstationary if and only if  $\rho = 0$ , and one can test the null hypothesis of unit root nonstationarity by testing  $\rho = 0$ . Said and Dickey show that the  $t$ -ratio for the hypothesis  $\rho = 0$  has the same asymptotic distribution as the Dickey-Fuller  $t$ -ratio test. Some authors call this test the augmented Dickey-Fuller (ADF) test while others reserve the word ADF for the corresponding cointegration test. A constant and a linear time trend are included or excluded according to the appropriate alternative hypothesis as before.

In many applications, the Said-Dickey test results are very sensitive to the choice of the order of autoregression,  $p$ . Ng and Perron (1995) analyze the choice of truncation lag, and categorize the existing methods into two rules: rules of thumb and data dependent rules. The former includes fixing  $p$  regardless of the sample size,  $T$ , or choosing  $p$  as a fixed function of  $T$  according to

$$(13.16) \quad p = \text{int}\left\{c\left(\frac{T}{100}\right)^{\frac{1}{d}}\right\},$$

where  $c = 4, 12$  and  $d = 4$  are used in Schwert (1989). The latter includes information-based rules such as Akaike information criterion (AIC) and Schwartz information criterion (SIC) according to

$$(13.17) \quad I_p = \log \hat{\sigma}_p^2 + p \frac{C_T}{T},$$

where  $\hat{\sigma}_p^2 = \frac{1}{T} \sum_{t=1}^T \hat{\nu}_t^2$ , and  $C_T = 2$  for AIC and  $C_T = \log T$  for SIC. Sequential tests for the significance of the coefficients on lags also fall into this category. Based on

Hall's (1994) work, Campbell and Perron (1991) recommend starting with a reasonably large value of  $p$  that is chosen a priori and decrease  $p$  until the coefficient on the last included lag is significant.<sup>6</sup> Ng and Perron (1995) show that rules of thumb are dominated by data dependent rules. They also show that general-to-specific sequential tests are better than information-based rules since the latter has severe size distortion.

### 13.3.3 Phillips-Perron Tests

Phillips (1987) and Phillips and Perron (1988) use a nonparametric method to correct for serial correlation of  $\epsilon_t$ . Their modification of the Dickey-Fuller  $T(\hat{\alpha} - 1)$  test is called  $Z(\alpha)$  test, while their modification of the Dickey-Fuller  $t$ -ratio test is called  $Z(t)$  test. These corrections are based on a nonparametric estimate of the long run variance of  $\epsilon_t$ . See Chapter 6 for a discussion of nonparametric estimation methods. Phillips-Perron tests are constructed so that they have the same asymptotic distributions as corresponding Dickey-Fuller tests.

An advantage of the Phillips-Perron tests over the Said-Dickey test is that they tend to be more powerful as shown in the Monte Carlo experiments of Phillips and Perron. A drawback of the Phillips-Perron tests is that they are subject to more severe size distortions than the Said-Dickey test (see Monte Carlo results of Phillips and Perron, 1988; Schwert, 1989). Size distortion exists when the actual size of a test in small samples is very different from the size of the test indicated by asymptotic theory. Such differences are due to approximations involved in the asymptotic theory.

---

<sup>6</sup>According to Hall (1994), compared to general-to-specific rules, specific-to-general rules are not generally asymptotically valid.

Table 13.1: Critical Values of Park's  $J(p, q)$  Tests for the Null of Difference Stationarity

Size	J(0,3)	J(1,5)	J(2,6)	J(3,8)	J(4,10)	J(5,11)
.010	.1118	.1228	.0886	.1093	.1348	.1157
.025	.2072	.1977	.1409	.1684	.1974	.1652
.050	.3385	.2950	.2050	.2394	.2660	.2210
.100	.5773	.4520	.3101	.3425	.3642	.3076
.150	.8042	.5959	.4034	.4299	.4516	.3800
.200	.9243	.7326	.4968	.5177	.5335	.4470

Source: Park and Choi's (1988) Table 1-B.

### 13.3.4 Park's $J$ Tests

Park's (1990)  $J$  tests based on a variable addition method were originally proposed by Park and Choi (1988). These tests are based on spurious regression results. Consider a regression

$$(13.18) \quad x_t = \sum_{\tau=0}^p \mu_{\tau} t^{\tau} + \sum_{\tau=p+1}^q \mu_{\tau} t^{\tau} + \eta_t.$$

Here the maintained hypothesis is that  $x_t$  possesses the deterministic time polynomials up to the order of  $p$  (typically,  $p$  is zero or one). The additional time polynomials are spurious time trends. Let  $F(p, q)$  be the standard Wald test statistic (without any correction for serial correlation of  $\eta_t$ ) for the null hypothesis  $\mu_{p+1} = \dots = \mu_q = 0$ . Under the null hypothesis that  $\eta_t$  is unit root nonstationary, spurious regression results imply that  $F(p, q)$  explodes, but  $\frac{1}{T}F(p, q)$  has an asymptotic distribution. The  $J(p, q)$  test is defined as  $\frac{1}{T}F(p, q)$ . The null hypothesis of difference stationarity is rejected against the alternative of trend stationarity when  $J(p, q)$  is *small* because  $J(p, q)$  converges to zero under the alternative hypothesis of trend stationarity. Part of Park and Choi's table of critical values for  $J$  tests are reproduced in Table 13.1 for convenience.

The  $J(p, q)$  tests do not require the estimation of the long-run variance of  $\eta_t$ ,

and thus have an advantage over the Said-Dickey and Phillips-Perron tests in that neither the order of autoregression nor the lag truncation number need to be chosen. Park and Choi's Monte Carlo experiments show that  $J$  tests have relatively stable sizes and are not dominated by Said-Dickey and Phillips-Perron tests in terms of size-adjusted power.

## 13.4 Testing the Null of Stationarity

In some cases, it is useful to test the null of stationarity (or trend stationarity) rather than the null of difference stationarity. For example, if an econometrician plans to apply econometric theory that assumes stationarity, a natural procedure is to test the null of stationarity rather than test the null of difference stationarity. Tests for the null of stationarity will also lead to tests for the null of cointegration as will be discussed in Chapter 14. However, most of the tests in the unit root literature take the null of a unit root rather than the null of stationarity. Only recently, Fukushige, Hatanaka, and Koto (1994), Kahn and Ogaki (1992), Kwiatkowski, Phillips, Schmidt, and Shin (1992), Bierens and Guo (1993), and Choi and Ahn (1999) among others have developed tests for the null of stationarity.

Park's (1990)  $G$  tests for the null of stationarity were first developed by Park and Choi's (1988). These tests, which have been used in empirical work by several researchers, are based on the same spurious regression results as Park's  $J$  tests. With the notations in Section 13.3.4,  $G(p, q) = F(p, q) \frac{\hat{\sigma}^2}{\hat{\omega}^2}$ , where  $\hat{\sigma}^2 = \frac{1}{T} \sum_{t=1}^T \hat{\eta}_t^2$ ,  $\hat{\omega}^2$  is an estimate of the long-run variance of  $\eta_t$ , and  $\hat{\eta}_t$  is the estimated residual in regression (13.18). Under the null that  $x_t$  is stationary after removing the maintained deterministic time terms of time polynomial of order  $p$ , the  $G(p, q)$  test statistic has

asymptotic chi-square distribution with the  $q - p$  degrees of freedom. Under the alternative hypothesis that  $x_t$  is difference stationary (after removing the maintained deterministic terms), the  $G(p, q)$  statistic diverges to infinity. This result is due to the spurious regression result that time polynomials tend to mimic a stochastic trend.

Unlike Park's  $J$  tests, Park's  $G$  tests require estimation of the long-run variance. Kahn and Ogaki's (1992) Monte Carlo experiments on Park's  $G$  tests suggest that it is advisable to use relatively small  $q$  when the sample size is small and not to use the prewhitening method discussed in Section 6.2.

## 13.5 Near Observational Equivalence

Most of the tests described in sections 13.3 and 13.4 seek to discriminate between difference stationary and trend stationary processes. In the finite samples that we observe, there is a conceptual difficulty with this task. In finite samples, any difference stationary process can be approximated arbitrarily well by a series of trend stationary processes. This evaluation can be done by driving the dominant autoregressive root of trend stationary processes to one from below. After all, it is very difficult to discriminate between the dominant autoregressive root of 0.999 and that of one. This type of problem exists for virtually any hypothesis testing. Hypothesis testing for unit root nonstationarity is special because the opposite is also true: any trend stationary process can be approximated arbitrarily well by a series of difference stationary processes. This approximation can be done by driving the long-run variance of the first difference of difference stationary processes to zero. Some authors call this problem the near observational equivalence problem (see, e.g., Cochrane, 1988; Campbell and Perron, 1991; Christiano and Eichenbaum, 1991; Blough, 1992; Faust,

1996).

## 13.6 Asymptotics for Unit Root Processes

This Appendix explains asymptotic theory for unit root processes. Many of the results depend on the Functional Central Limit Theorem (FCLT) explained in Appendix 5.B.

### 13.6.1 Continuous Mapping Theorem

**Theorem 13.1** *Let  $h : \mathcal{R} \rightarrow \mathcal{R}$  be a measurable function with discontinuity points confined to a set  $D$  where  $P(D) = 0$ . If  $X_n \Rightarrow X$ , then  $h(X_n) \Rightarrow h(X)$ .*

It is instructive to illustrate how the CMT can be used in the AR(1) model when  $\beta = 1$ :

$$y_t = \beta y_{t-1} + \varepsilon_t.$$

Consider the sampling error of the OLS estimator,

$$n(\hat{\beta} - 1) = \frac{\frac{1}{n} \sum_{t=2}^n y_{t-1} \varepsilon_t}{\frac{1}{n^2} \sum_{t=2}^n y_{t-1}^2}.$$

Asymptotic properties of the denominator can be established by the FCLT and the CMT. Let  $W_n(r) = \frac{y_{[nr]}}{\sqrt{n}}$ . Note that the denominator can be written

$$\frac{1}{n^2} \sum_{t=2}^n y_{t-1}^2 = \frac{1}{n} \sum_{t=2}^n \left( \frac{y_{t-1}}{\sqrt{n}} \right)^2 = \int_0^1 [W_n(r)]^2 dr.$$

Since  $W_n(r) \Rightarrow W(r)$  and the integral is the continuous function of  $W_n(r)$ , by the above theorem,

$$\int_0^1 [W_n(r)]^2 dr \Rightarrow \int_0^1 [W(r)]^2 dr.$$

### 13.6.2 Dickey-Fuller test with serially uncorrelated disturbances

We consider two cases for DF tests with the null: when the true process is a random walk with or without a drift, and when the equation is estimated with or without a trend. See Hamilton (1994) for details.

**The regression equation includes a constant term but no time trend when the true process is a random walk**

Suppose that the data are generated by a random walk without drift

$$y_t = y_{t-1} + \epsilon_t,$$

where  $\epsilon_t$  follows an i.i.d. sequence with mean zero, and variance  $\sigma^2$ . Consider a regression equation

$$\begin{aligned} \Delta y_t &= \alpha + \rho y_{t-1} + \epsilon_t \\ &= \mathbf{x}'_t \boldsymbol{\beta} + \epsilon_t, \end{aligned}$$

where  $\mathbf{x}_t = (1, y_{t-1})'$ , and  $\boldsymbol{\beta} = (\alpha, \rho)'$ . Define a scaling matrix

$$\mathbf{S}_T = \begin{bmatrix} \sqrt{T} & \mathbf{0} \\ \mathbf{0} & T \end{bmatrix}$$

and write the deviation of the OLS estimates using the scaling matrix

$$\begin{aligned} \mathbf{S}_T(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) &= \left\{ \mathbf{S}_T^{-1} \left[ \sum_{i=1}^T \mathbf{x}_i \mathbf{x}'_i \right] \mathbf{S}_T^{-1} \right\}^{-1} \left\{ \mathbf{S}_T^{-1} \left[ \sum_{i=1}^T \mathbf{x}_i \epsilon_i \right] \right\} \\ &= \begin{bmatrix} 1 & T^{-\frac{3}{2}} \sum_{i=1}^T y_{i-1} \\ \cdot & T^{-2} \sum_{i=1}^T y_{i-1}^2 \end{bmatrix}^{-1} \begin{bmatrix} T^{-\frac{1}{2}} \sum_{i=1}^T \epsilon_i \\ T^{-1} \sum_{i=1}^T y_{i-1} \epsilon_i \end{bmatrix}, \end{aligned}$$

where under the null of  $\Delta y_t = \epsilon_t$

$$\begin{aligned} \begin{bmatrix} 1 & T^{-\frac{3}{2}} \sum_{i=1}^T y_{i-1} \\ \cdot & T^{-2} \sum_{i=1}^T y_{i-1}^2 \end{bmatrix} &\xrightarrow{L} \begin{bmatrix} 1 & \sigma \int W(r) dr \\ \cdot & \sigma^2 \int W(r)^2 dr \end{bmatrix} \quad \text{and} \\ \begin{bmatrix} T^{-\frac{1}{2}} \sum_{i=1}^T \epsilon_i \\ T^{-1} \sum_{i=1}^T y_{i-1} \epsilon_i \end{bmatrix} &\xrightarrow{L} \begin{bmatrix} \sigma W(1) \\ \sigma^2 \int W dW \end{bmatrix}. \end{aligned}$$

Thus, we get

$$\begin{aligned} \begin{bmatrix} T^{\frac{1}{2}}\hat{\alpha} \\ T\hat{\rho} \end{bmatrix} &\xrightarrow{L} \begin{bmatrix} 1 & \sigma \int W(r)dr \\ \cdot & \sigma^2 \int W(r)^2 dr \end{bmatrix}^{-1} \begin{bmatrix} \sigma W(1) \\ \sigma^2 \int W dW \end{bmatrix} \\ &\xrightarrow{L} \begin{bmatrix} \sigma & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & \int W(r)dr \\ \cdot & \int W(r)^2 dr \end{bmatrix}^{-1} \begin{bmatrix} W(1) \\ \int W dW \end{bmatrix}. \end{aligned}$$

In particular,

$$\begin{aligned} T\hat{\rho} &\xrightarrow{L} [0 \quad 1] \begin{bmatrix} 1 & \int W(r)dr \\ \cdot & \int W(r)^2 dr \end{bmatrix}^{-1} \begin{bmatrix} W(1) \\ \int W dW \end{bmatrix} \\ &= \frac{\int W dW - W(1) \int W(r)dr}{\int W(r)^2 dr - (\int W(r)dr)^2}, \end{aligned}$$

which is the DF  $\rho$  test. Note that the coefficients on  $\Delta y_{t-i}$  follow a normal distribution asymptotically so that the usual test can be applied for restrictions on these variables.

Similarly, the variance of  $\hat{\beta}$  follows

$$\begin{aligned} \mathbf{S}_T \hat{\Sigma}_{\hat{\beta}} \mathbf{S}_T &= \hat{\sigma}^2 \left\{ \mathbf{S}_T^{-1} \left[ \sum_{i=1}^T \mathbf{x}_t \mathbf{x}_t' \right] \mathbf{S}_T^{-1} \right\}^{-1} \\ &= \hat{\sigma}^2 \begin{bmatrix} 1 & T^{-\frac{3}{2}} \sum_{i=1}^T y_{t-1} \\ \cdot & T^{-2} \sum_{i=1}^T y_{t-1}^2 \end{bmatrix}^{-1} \\ &\xrightarrow{L} \sigma^2 \begin{bmatrix} 1 & \sigma \int W(r)dr \\ \cdot & \sigma^2 \int W(r)^2 dr \end{bmatrix}^{-1}. \end{aligned}$$

In particular, the standard error of  $\hat{\rho}$  follows

$$T s_{\hat{\rho}} \xrightarrow{L} \frac{1}{[\int W(r)^2 dr - (\int W(r)dr)^2]^{\frac{1}{2}}}.$$

Therefore, we get the DF  $t$ -test

$$\begin{aligned} t_{\hat{\rho}} = \frac{T\hat{\rho}}{T s_{\hat{\rho}}} &\xrightarrow{L} \frac{[\int W dW - W(1) \int W(r)dr] / [\int W(r)^2 dr - (\int W(r)dr)^2]}{\{1 / [\int W(r)^2 dr - (\int W(r)dr)^2]\}^{\frac{1}{2}}} \\ &\xrightarrow{L} \frac{\int W dW - W(1) \int W(r)dr}{[\int W(r)^2 dr - (\int W(r)dr)^2]^{\frac{1}{2}}}. \end{aligned}$$

**The regression equation includes a constant term and a time trend when the true process is a unit root process with or without a drift**

Now, suppose that the data are generated by a random walk with or without a drift

$$y_t = \mu + y_{t-1} + \epsilon_t.$$

Consider a regression equation

$$\Delta y_t = \mu + \delta t + \rho y_{t-1} + \epsilon_t.$$

Note that the regression is subject to collinearity because  $y_{t-1}$  contains a deterministic time trend component if  $\mu \neq 0$ . To avoid the possible collinearity, rewrite the equation using a detrended series  $\xi_t = y_t - \mu t$

$$\begin{aligned} \Delta y_t &= \mu + \delta t + \rho(\xi_{t-1} + \mu(t-1)) + \epsilon_t \\ &= (1-\rho)\mu + (\delta + \rho\mu)t + \rho\xi_{t-1} + \epsilon_t \\ &= \alpha + \tau t + \rho\xi_{t-1} + \epsilon_t \\ &= \mathbf{x}'_t \boldsymbol{\beta} + \epsilon_t, \end{aligned}$$

where  $\mathbf{x}_t = (1, t, \xi_{t-1})'$ , and  $\boldsymbol{\beta} = (\alpha, \tau, \rho)'$ . Define a scaling matrix

$$\mathbf{S}_T = \begin{bmatrix} \sqrt{T} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \sqrt[3]{T} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & T \end{bmatrix}$$

and write the deviation of the OLS estimates using the scaling matrix

$$\begin{aligned} \mathbf{S}_T(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) &= \left\{ \mathbf{S}_T^{-1} \left[ \sum_{i=1}^T \mathbf{x}_i \mathbf{x}'_i \right] \mathbf{S}_T^{-1} \right\}^{-1} \left\{ \mathbf{S}_T^{-1} \left[ \sum_{i=1}^T \mathbf{x}_i \epsilon_i \right] \right\} \\ &= \begin{bmatrix} 1 & T^{-2} \sum_{i=1}^T t & T^{-\frac{3}{2}} \sum_{i=1}^T \xi_{t-1} \\ \cdot & T^{-3} \sum_{i=1}^T t^2 & T^{-\frac{5}{2}} \sum_{i=1}^T t \xi_{t-1} \\ \cdot & \cdot & T^{-2} \sum_{i=1}^T \xi_{t-1}^2 \end{bmatrix}^{-1} \begin{bmatrix} T^{-\frac{1}{2}} \sum_{i=1}^T \epsilon_t \\ T^{-\frac{3}{2}} \sum_{i=1}^T t \epsilon_t \\ T^{-1} \sum_{i=1}^T \xi_{t-1} \epsilon_t \end{bmatrix}, \end{aligned}$$

where under the null of  $\Delta\xi_t = \epsilon_t$

$$\begin{bmatrix} 1 & T^{-2} \sum_{i=1}^T t & T^{-\frac{3}{2}} \sum_{i=1}^T \xi_{t-1} \\ \cdot & T^{-3} \sum_{i=1}^T t^2 & T^{-\frac{5}{2}} \sum_{i=1}^T t\xi_{t-1} \\ \cdot & \cdot & T^{-2} \sum_{i=1}^T \xi_{t-1}^2 \end{bmatrix} \xrightarrow{L} \begin{bmatrix} 1 & \frac{1}{2} & \sigma \int W(r)dr \\ \cdot & \frac{1}{3} & \sigma \int rW(r)dr \\ \cdot & \cdot & \sigma^2 \int W(r)^2 dr \end{bmatrix} \text{ and}$$

$$\begin{bmatrix} T^{-\frac{1}{2}} \sum_{i=1}^T \epsilon_t \\ T^{-\frac{3}{2}} \sum_{i=1}^T t\epsilon_t \\ T^{-1} \sum_{i=1}^T \xi_{t-1}\epsilon_t \end{bmatrix} \xrightarrow{L} \begin{bmatrix} \sigma W(1) \\ \sigma \int rdW \\ \sigma^2 \int WdW \end{bmatrix}.$$

Due to the block diagonal property, we can write

$$\begin{bmatrix} T^{\frac{1}{2}} \hat{\alpha} \\ T^{\frac{3}{2}} \hat{\tau} \\ T \hat{\rho} \end{bmatrix} \xrightarrow{L} \begin{bmatrix} 1 & \frac{1}{2} & \sigma \int W(r)dr \\ \cdot & \frac{1}{3} & \sigma \int rW(r)dr \\ \cdot & \cdot & \sigma^2 \int W(r)^2 dr \end{bmatrix}^{-1} \begin{bmatrix} \sigma W(1) \\ \sigma \int rdW \\ \sigma^2 \int WdW \end{bmatrix}$$

$$\xrightarrow{L} \begin{bmatrix} \sigma & 0 & 0 \\ 0 & \sigma & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & \frac{1}{2} & \int W(r)dr \\ \cdot & \frac{1}{3} & \int rW(r)dr \\ \cdot & \cdot & \int W(r)^2 dr \end{bmatrix}^{-1} \begin{bmatrix} W(1) \\ \int rdW \\ \int WdW \end{bmatrix}.$$

In particular,

$$T \hat{\rho} \xrightarrow{L} [0 \ 0 \ 1] \begin{bmatrix} 1 & \frac{1}{2} & \int W(r)dr \\ \cdot & \frac{1}{3} & \int rW(r)dr \\ \cdot & \cdot & \int W(r)^2 dr \end{bmatrix}^{-1} \begin{bmatrix} W(1) \\ \int rdW \\ \int WdW \end{bmatrix},$$

which is the DF  $\rho$  test.

Similarly, the variance of  $\hat{\beta}$  follows

$$\begin{aligned} \mathbf{S}_T \hat{\Sigma}_{\hat{\beta}} \mathbf{S}_T &= \hat{\sigma}^2 \left\{ \mathbf{S}_T^{-1} \left[ \sum_{i=1}^T \mathbf{x}_t \mathbf{x}_t' \right] \mathbf{S}_T^{-1} \right\}^{-1} \\ &= \hat{\sigma}^2 \begin{bmatrix} 1 & T^{-2} \sum_{i=1}^T t & T^{-\frac{3}{2}} \sum_{i=1}^T \xi_{t-1} \\ \cdot & T^{-3} \sum_{i=1}^T t^2 & T^{-\frac{5}{2}} \sum_{i=1}^T t\xi_{t-1} \\ \cdot & \cdot & T^{-2} \sum_{i=1}^T \xi_{t-1}^2 \end{bmatrix}^{-1} \\ &\xrightarrow{L} \sigma^2 \begin{bmatrix} 1 & \frac{1}{2} & \sigma \int W(r)dr \\ \cdot & \frac{1}{3} & \sigma \int rW(r)dr \\ \cdot & \cdot & \sigma^2 \int W(r)^2 dr \end{bmatrix}^{-1}. \end{aligned}$$

In particular, the standard error of  $\hat{\rho}$  follows

$$T \mathbf{s}_{\hat{\rho}} \xrightarrow{L} \left\{ [0 \ 0 \ 1] \begin{bmatrix} 1 & \frac{1}{2} & \int W(r)dr \\ \cdot & \frac{1}{3} & \int rW(r)dr \\ \cdot & \cdot & \int W(r)^2 dr \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \right\}^{\frac{1}{2}}.$$

Therefore, we get the DF  $t$ -test

$$t_{\hat{\rho}} = \frac{T\hat{\rho}}{Ts_{\hat{\rho}}} \xrightarrow{L} \frac{[0 \ 0 \ 1] \begin{bmatrix} 1 & \frac{1}{2} & \int W(r)dr \\ \cdot & \frac{1}{3} & \int rW(r)dr \\ \cdot & \cdot & \int W(r)^2dr \end{bmatrix}^{-1} \begin{bmatrix} W(1) \\ \int rdW \\ \int WdW \end{bmatrix}}{\left\{ [0 \ 0 \ 1] \begin{bmatrix} 1 & \frac{1}{2} & \int W(r)dr \\ \cdot & \frac{1}{3} & \int rW(r)dr \\ \cdot & \cdot & \int W(r)^2dr \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \right\}^{\frac{1}{2}}}.$$

### 13.6.3 Said-Dickey test with serially correlated disturbances

We consider two cases for Said-Dickey tests with the null: when the true process is a random walk with or without a drift, and when the equation is estimated with or without a trend. See Hamilton (1994) for details.

**The regression equation includes a constant term but no time trend when the true process is a unit root process without a drift**

Consider a DGP:

$$a(L)y_t = \epsilon_t,$$

where  $\epsilon_t$  follows an i.i.d. sequence with mean zero, and variance  $\sigma^2$ . Let

$$a(L) = a(1)L + b(L)(1 - L),$$

where  $b(L) = 1 - \sum_{i=1}^{p-1} b_i L^i$  and  $b_i = -\sum_{j=i+1}^p a_j$ , and rearrange the equation

$$\begin{aligned} b(L)\Delta y_t &= -a(1)y_{t-1} + \epsilon_t \quad \text{or} \\ \Delta y_t &= \rho y_{t-1} + \sum_{i=1}^{p-1} b_i \Delta y_{t-i} + \epsilon_t, \end{aligned}$$

where  $\rho = -a(1) = -1 + \sum_{i=1}^p a_i$ . Note that the assumption of a single unit root in the DGP implies  $\rho = 0$ . Under the null, we get an MA representation

$$\begin{aligned} \Delta y_t &= c(L)\epsilon_t \\ &= u_t \end{aligned}$$

where  $c(L) = b(L)^{-1} = 1 + \sum_{i=1}^{\infty} c_i L^i$ .

Consider a regression equation

$$\begin{aligned}\Delta y_t &= \alpha + \rho y_{t-1} + \sum_{i=1}^{p-1} b_i \Delta y_{t-i} + \epsilon_t \\ &= \alpha + \rho y_{t-1} + \mathbf{z}'_t \mathbf{b} + \epsilon_t \\ &= \mathbf{x}'_t \boldsymbol{\beta} + \epsilon_t,\end{aligned}$$

where  $\mathbf{z}_t = (\Delta y_{t-1}, \dots, \Delta y_{t-p+1})'$ ,  $\mathbf{b} = (b_1, \dots, b_{p-1})'$ ,  $\mathbf{x}_t = (1, y_{t-1}, \mathbf{z}'_t)'$ , and  $\boldsymbol{\beta} = (\alpha, \rho, \mathbf{b}')'$ . Define a scaling matrix

$$\mathbf{S}_T = \begin{bmatrix} \sqrt{T} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & T & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \sqrt{T} \mathbf{I}_{p-1} \end{bmatrix}$$

and write the deviation of the OLS estimates using the scaling matrix

$$\begin{aligned}\mathbf{S}_T(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) &= \left\{ \mathbf{S}_T^{-1} \left[ \sum_{i=1}^T \mathbf{x}_t \mathbf{x}'_t \right] \mathbf{S}_T^{-1} \right\}^{-1} \left\{ \mathbf{S}_T^{-1} \left[ \sum_{i=1}^T \mathbf{x}_t \epsilon_t \right] \right\} \\ &= \begin{bmatrix} 1 & T^{-\frac{3}{2}} \sum_{i=1}^T y_{t-1} & T^{-1} \sum_{i=1}^T \mathbf{z}'_t \\ \cdot & T^{-2} \sum_{i=1}^T y_{t-1}^2 & T^{-\frac{3}{2}} \sum_{i=1}^T y_{t-1} \mathbf{z}'_t \\ \cdot & \cdot & T^{-1} \sum_{i=1}^T \mathbf{z}_t \mathbf{z}'_t \end{bmatrix}^{-1} \begin{bmatrix} T^{-\frac{1}{2}} \sum_{i=1}^T \epsilon_t \\ T^{-1} \sum_{i=1}^T y_{t-1} \epsilon_t \\ T^{-\frac{1}{2}} \sum_{i=1}^T \mathbf{z}_t \epsilon_t \end{bmatrix},\end{aligned}$$

where under the null of  $\Delta y_t = u_t$

$$\begin{aligned}\begin{bmatrix} 1 & T^{-\frac{3}{2}} \sum_{i=1}^T y_{t-1} & T^{-1} \sum_{i=1}^T \mathbf{z}'_t \\ \cdot & T^{-2} \sum_{i=1}^T y_{t-1}^2 & T^{-\frac{3}{2}} \sum_{i=1}^T y_{t-1} \mathbf{z}'_t \\ \cdot & \cdot & T^{-1} \sum_{i=1}^T \mathbf{z}_t \mathbf{z}'_t \end{bmatrix} &\xrightarrow{L} \begin{bmatrix} 1 & \lambda \int W(r) dr & \mathbf{0} \\ \cdot & \lambda^2 \int W(r)^2 dr & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{V} \end{bmatrix} \quad \text{and} \\ \begin{bmatrix} T^{-\frac{1}{2}} \sum_{i=1}^T \epsilon_t \\ T^{-1} \sum_{i=1}^T y_{t-1} \epsilon_t \\ T^{-\frac{1}{2}} \sum_{i=1}^T \mathbf{z}_t \epsilon_t \end{bmatrix} &\xrightarrow{L} \begin{bmatrix} \sigma W(1) \\ \sigma \lambda \int W dW \\ \mathbf{h} \end{bmatrix}.\end{aligned}$$

Due to the block diagonal property, we can write

$$\begin{aligned}\begin{bmatrix} T^{\frac{1}{2}} \hat{\alpha} \\ T \hat{\rho} \end{bmatrix} &\xrightarrow{L} \begin{bmatrix} 1 & \lambda \int W(r) dr \\ \cdot & \lambda^2 \int W(r)^2 dr \end{bmatrix}^{-1} \begin{bmatrix} \sigma W(1) \\ \sigma \lambda \int W dW \end{bmatrix} \\ &\xrightarrow{L} \begin{bmatrix} \sigma & 0 \\ 0 & \frac{\sigma}{\lambda} \end{bmatrix} \begin{bmatrix} 1 & \int W(r) dr \\ \cdot & \int W(r)^2 dr \end{bmatrix}^{-1} \begin{bmatrix} W(1) \\ \int W dW \end{bmatrix} \quad \text{and} \\ T^{-\frac{1}{2}}(\hat{\mathbf{b}} - \mathbf{b}) &\xrightarrow{L} \mathbf{V}^{-1} \mathbf{h} \sim N(\mathbf{0}, \sigma^2 \mathbf{V}^{-1}).\end{aligned}$$

In particular,

$$T\hat{\rho} \xrightarrow{L} \frac{\sigma}{\lambda} \begin{bmatrix} 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & \int W(r)dr \\ \cdot & \int W(r)^2 dr \end{bmatrix}^{-1} \begin{bmatrix} W(1) \\ \int W dW \end{bmatrix}.$$

From  $\frac{\lambda}{\sigma} = c(1) = b(1)^{-1}$ , we get the Said-Dickey  $\rho$  test

$$\frac{T\hat{\rho}}{1 - \sum_{i=1}^{p-1} \hat{b}_i} \xrightarrow{L} \frac{\int W dW - W(1) \int W(r)dr}{\int W(r)^2 dr - (\int W(r)dr)^2},$$

which follows the same asymptotic distribution as the DF  $\rho$  test. Note that the coefficients on  $\Delta y_{t-i}$  follow a normal distribution asymptotically so that the usual test can be applied for restrictions on these variables.

Similarly, the variance of  $\hat{\beta}$  follows

$$\begin{aligned} \mathbf{S}_T \hat{\Sigma}_{\hat{\beta}} \mathbf{S}_T &= \hat{\sigma}^2 \left\{ \mathbf{S}_T^{-1} \left[ \sum_{i=1}^T \mathbf{x}_t \mathbf{x}_t' \right] \mathbf{S}_T^{-1} \right\}^{-1} \\ &= \hat{\sigma}^2 \begin{bmatrix} 1 & T^{-\frac{3}{2}} \sum_{i=1}^T y_{t-1} & T^{-1} \sum_{i=1}^T \mathbf{z}_t' \\ \cdot & T^{-2} \sum_{i=1}^T y_{t-1}^2 & T^{-\frac{3}{2}} \sum_{i=1}^T y_{t-1} \mathbf{z}_t' \\ \cdot & \cdot & T^{-1} \sum_{i=1}^T \mathbf{z}_t \mathbf{z}_t' \end{bmatrix}^{-1} \\ &\xrightarrow{L} \sigma^2 \begin{bmatrix} 1 & \lambda \int W(r)dr & \mathbf{0} \\ \cdot & \lambda^2 \int W(r)^2 dr & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{V} \end{bmatrix}^{-1}. \end{aligned}$$

In particular, the standard error of  $\hat{\rho}$  follows

$$T s_{\hat{\rho}} \xrightarrow{L} \frac{\sigma/\lambda}{[\int W(r)^2 dr - (\int W(r)dr)^2]^{\frac{1}{2}}}.$$

Therefore, we get the Said-Dickey  $t$ -test

$$\begin{aligned} t_{\hat{\rho}} = \frac{T\hat{\rho}}{T s_{\hat{\rho}}} &\xrightarrow{L} \frac{(\sigma/\lambda) [\int W dW - W(1) \int W(r)dr] / [\int W(r)^2 dr - (\int W(r)dr)^2]}{(\sigma/\lambda) \{1/ [\int W(r)^2 dr - (\int W(r)dr)^2]\}^{\frac{1}{2}}} \\ &\xrightarrow{L} \frac{\int W dW - W(1) \int W(r)dr}{[\int W(r)^2 dr - (\int W(r)dr)^2]^{\frac{1}{2}}}, \end{aligned}$$

which follows the same asymptotic distribution as the DF  $t$  test.

The regression equation includes a constant term and a time trend when the true process is a unit root process with or without a drift

Now, consider a DGP:

$$a(L)y_t = \mu + \epsilon_t$$

and rearrange the equation

$$\begin{aligned} b(L)\Delta y_t &= \mu - a(1)y_{t-1} + \epsilon_t \quad \text{or} \\ \Delta y_t &= \mu + \rho y_{t-1} + \sum_{i=1}^{p-1} b_i \Delta y_{t-i} + \epsilon_t. \end{aligned}$$

Under the null, we get an MA representation

$$\begin{aligned} \Delta y_t &= \theta + c(L)\epsilon_t \\ &= \theta + u_t \end{aligned}$$

where  $\theta = c(1)\mu$ .

Consider a regression equation

$$\Delta y_t = \mu + \delta t + \rho y_{t-1} + \sum_{i=1}^{p-1} b_i \Delta y_{t-i} + \epsilon_t.$$

Note that the regression is subject to collinearity because  $y_{t-1}$  contains a deterministic time trend component if  $\mu \neq 0$ . To avoid the possible collinearity, rewrite the equation using a detrended series  $\xi_t = y_t - \mu t$

$$\begin{aligned} \Delta y_t &= \mu + \delta t + \rho(\xi_{t-1} + \mu(t-1)) + \sum_{i=1}^{p-1} b_i(\Delta \xi_{t-i} + \mu) + \epsilon_t \\ &= (1 - \rho + \sum_{i=1}^{p-1} b_i)\mu + (\delta + \rho\mu)t + \rho\xi_{t-1} + \mathbf{z}'_t \mathbf{b} + \epsilon_t \\ &= \alpha + \tau t + \rho\xi_{t-1} + \mathbf{z}'_t \mathbf{b} + \epsilon_t \\ &= \mathbf{x}'_t \boldsymbol{\beta} + \epsilon_t, \end{aligned}$$

where  $\mathbf{z}_t = (\Delta\xi_{t-1}, \dots, \Delta\xi_{t-p+1})'$ ,  $\mathbf{b} = (b_1, \dots, b_{p-1})'$ ,  $\mathbf{x}_t = (1, t, \xi_{t-1}, \mathbf{z}_t)'$ , and  $\boldsymbol{\beta} = (\alpha, \tau, \rho, \mathbf{b}')'$ . Define a scaling matrix

$$\mathbf{S}_T = \begin{bmatrix} \sqrt{T} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \sqrt[3]{T} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & T & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \sqrt{T} \mathbf{I}_{p-1} \end{bmatrix}$$

and write the deviation of the OLS estimates using the scaling matrix

$$\begin{aligned} \mathbf{S}_T(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) &= \left\{ \mathbf{S}_T^{-1} \left[ \sum_{i=1}^T \mathbf{x}_i \mathbf{x}_i' \right] \mathbf{S}_T^{-1} \right\}^{-1} \left\{ \mathbf{S}_T^{-1} \left[ \sum_{i=1}^T \mathbf{x}_i \epsilon_i \right] \right\} \\ &= \begin{bmatrix} 1 & T^{-2} \sum_{i=1}^T t & T^{-\frac{3}{2}} \sum_{i=1}^T \xi_{t-1} & T^{-1} \sum_{i=1}^T \mathbf{z}_t' \\ \cdot & T^{-3} \sum_{i=1}^T t^2 & T^{-\frac{5}{2}} \sum_{i=1}^T t \xi_{t-1} & T^{-2} \sum_{i=1}^T t \mathbf{z}_t' \\ \cdot & \cdot & T^{-2} \sum_{i=1}^T \xi_{t-1}^2 & T^{-\frac{3}{2}} \sum_{i=1}^T \xi_{t-1} \mathbf{z}_t' \\ \cdot & \cdot & \cdot & T^{-1} \sum_{i=1}^T \mathbf{z}_t \mathbf{z}_t' \end{bmatrix}^{-1} \begin{bmatrix} T^{-\frac{1}{2}} \sum_{i=1}^T \epsilon_t \\ T^{-\frac{3}{2}} \sum_{i=1}^T t \epsilon_t \\ T^{-1} \sum_{i=1}^T \xi_{t-1} \epsilon_t \\ T^{-\frac{1}{2}} \sum_{i=1}^T \mathbf{z}_t \epsilon_t \end{bmatrix}, \end{aligned}$$

where under the null of  $\Delta\xi_t = u_t$

$$\begin{aligned} \begin{bmatrix} 1 & T^{-2} \sum_{i=1}^T t & T^{-\frac{3}{2}} \sum_{i=1}^T \xi_{t-1} & T^{-1} \sum_{i=1}^T \mathbf{z}_t' \\ \cdot & T^{-3} \sum_{i=1}^T t^2 & T^{-\frac{5}{2}} \sum_{i=1}^T t \xi_{t-1} & T^{-2} \sum_{i=1}^T t \mathbf{z}_t' \\ \cdot & \cdot & T^{-2} \sum_{i=1}^T \xi_{t-1}^2 & T^{-\frac{3}{2}} \sum_{i=1}^T \xi_{t-1} \mathbf{z}_t' \\ \cdot & \cdot & \cdot & T^{-1} \sum_{i=1}^T \mathbf{z}_t \mathbf{z}_t' \end{bmatrix} \xrightarrow{L} \begin{bmatrix} 1 & \frac{1}{2} & \lambda \int W(r) dr & \mathbf{0} \\ \cdot & \frac{1}{3} & \lambda \int r W(r) dr & \mathbf{0} \\ \cdot & \cdot & \lambda^2 \int W(r)^2 dr & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{V} \end{bmatrix} \text{ and} \\ \begin{bmatrix} T^{-\frac{1}{2}} \sum_{i=1}^T \epsilon_t \\ T^{-\frac{3}{2}} \sum_{i=1}^T t \epsilon_t \\ T^{-1} \sum_{i=1}^T \xi_{t-1} \epsilon_t \\ T^{-\frac{1}{2}} \sum_{i=1}^T \mathbf{z}_t \epsilon_t \end{bmatrix} \xrightarrow{L} \begin{bmatrix} \sigma W(1) \\ \sigma \int r dW \\ \sigma \lambda \int W dW \\ \mathbf{h} \end{bmatrix}. \end{aligned}$$

Due to the block diagonal property, we can write

$$\begin{aligned} \begin{bmatrix} T^{\frac{1}{2}} \hat{\alpha} \\ T^{\frac{3}{2}} \hat{\tau} \\ T \hat{\rho} \end{bmatrix} &\xrightarrow{L} \begin{bmatrix} 1 & \frac{1}{2} & \lambda \int W(r) dr \\ \cdot & \frac{1}{3} & \lambda \int r W(r) dr \\ \cdot & \cdot & \lambda^2 \int W(r)^2 dr \end{bmatrix}^{-1} \begin{bmatrix} \sigma W(1) \\ \sigma \int r dW \\ \sigma \lambda \int W dW \end{bmatrix} \\ &\xrightarrow{L} \begin{bmatrix} \sigma & 0 & 0 \\ 0 & \sigma & 0 \\ 0 & 0 & \frac{\sigma}{\lambda} \end{bmatrix} \begin{bmatrix} 1 & \frac{1}{2} & \int W(r) dr \\ \cdot & \frac{1}{3} & \int r W(r) dr \\ \cdot & \cdot & \int W(r)^2 dr \end{bmatrix}^{-1} \begin{bmatrix} W(1) \\ \int r dW \\ \int W dW \end{bmatrix} \text{ and} \\ T^{-\frac{1}{2}}(\hat{\mathbf{b}} - \mathbf{b}) &\xrightarrow{L} \mathbf{V}^{-1} \mathbf{h} \sim N(\mathbf{0}, \sigma^2 \mathbf{V}^{-1}). \end{aligned}$$

In particular,

$$T \hat{\rho} \xrightarrow{L} \frac{\sigma}{\lambda} \begin{bmatrix} 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & \frac{1}{2} & \int W(r) dr \\ \cdot & \frac{1}{3} & \int r W(r) dr \\ \cdot & \cdot & \int W(r)^2 dr \end{bmatrix}^{-1} \begin{bmatrix} W(1) \\ \int r dW \\ \int W dW \end{bmatrix}.$$

From  $\frac{\lambda}{\sigma} = c(1) = b(1)^{-1}$ , we get the Said-Dickey  $\rho$  test

$$\frac{T\hat{\rho}}{1 - \sum_{i=1}^{p-1} \hat{b}_i} \xrightarrow{L} [0 \ 0 \ 1] \begin{bmatrix} 1 & \frac{1}{2} & \int W(r)dr \\ \cdot & \frac{1}{3} & \int rW(r)dr \\ \cdot & \cdot & \int W(r)^2dr \end{bmatrix}^{-1} \begin{bmatrix} W(1) \\ \int rdW \\ \int WdW \end{bmatrix} \text{ and}$$

which follows the same asymptotic distribution as the DF  $\rho$  test. Note that the coefficients on  $\Delta y_{t-i}$  follow a normal distribution asymptotically so that the usual test can be applied for restrictions on these variables.

Similarly, the variance of  $\hat{\beta}$  follows

$$\begin{aligned} \mathbf{S}_T \hat{\Sigma}_{\hat{\beta}} \mathbf{S}_T &= \hat{\sigma}^2 \left\{ \mathbf{S}_T^{-1} \left[ \sum_{i=1}^T \mathbf{x}_t \mathbf{x}_t' \right] \mathbf{S}_T^{-1} \right\}^{-1} \\ &= \hat{\sigma}^2 \begin{bmatrix} 1 & T^{-2} \sum_{i=1}^T t & T^{-\frac{3}{2}} \sum_{i=1}^T \xi_{t-1} & T^{-1} \sum_{i=1}^T \mathbf{z}_t' \\ \cdot & T^{-3} \sum_{i=1}^T t^2 & T^{-\frac{5}{2}} \sum_{i=1}^T t \xi_{t-1} & T^{-2} \sum_{i=1}^T t \mathbf{z}_t' \\ \cdot & \cdot & T^{-2} \sum_{i=1}^T \xi_{t-1}^2 & T^{-\frac{3}{2}} \sum_{i=1}^T \xi_{t-1} \mathbf{z}_t' \\ \cdot & \cdot & \cdot & T^{-1} \sum_{i=1}^T \mathbf{z}_t \mathbf{z}_t' \end{bmatrix}^{-1} \\ &\xrightarrow{L} \sigma^2 \begin{bmatrix} 1 & \frac{1}{2} & \lambda \int W(r)dr & \mathbf{0} \\ \cdot & \frac{1}{3} & \lambda \int rW(r)dr & \mathbf{0} \\ \cdot & \cdot & \lambda^2 \int W(r)^2dr & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{V} \end{bmatrix}^{-1}. \end{aligned}$$

In particular, the standard error of  $\hat{\rho}$  follows

$$T \mathbf{s}_{\hat{\rho}} \xrightarrow{L} \frac{\sigma}{\lambda} \left\{ [0 \ 0 \ 1] \begin{bmatrix} 1 & \frac{1}{2} & \int W(r)dr \\ \cdot & \frac{1}{3} & \int rW(r)dr \\ \cdot & \cdot & \int W(r)^2dr \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \right\}^{\frac{1}{2}}.$$

Therefore, we get the Said-Dickey  $t$ -test

$$t_{\hat{\rho}} = \frac{T\hat{\rho}}{T \mathbf{s}_{\hat{\rho}}} \xrightarrow{L} \frac{[0 \ 0 \ 1] \begin{bmatrix} 1 & \frac{1}{2} & \int W(r)dr \\ \cdot & \frac{1}{3} & \int rW(r)dr \\ \cdot & \cdot & \int W(r)^2dr \end{bmatrix}^{-1} \begin{bmatrix} W(1) \\ \int rdW \\ \int WdW \end{bmatrix}}{\left\{ [0 \ 0 \ 1] \begin{bmatrix} 1 & \frac{1}{2} & \int W(r)dr \\ \cdot & \frac{1}{3} & \int rW(r)dr \\ \cdot & \cdot & \int W(r)^2dr \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \right\}^{\frac{1}{2}}},$$

which follows the same asymptotic distribution as the DF  $t$  test.

### 13.6.4 Phillips-Perron test

We consider two cases for PP tests with the null: when the true process is a random walk with or without a drift, and when the equation is estimated with or without a trend. See Hamilton (1994) for details.

**The regression equation includes a constant term but no time trend when the true process is a unit root process without a drift**

Consider a DGP:

$$a(L)y_t = \epsilon_t,$$

where  $\epsilon_t$  follows an i.i.d. sequence with mean zero, and variance  $\sigma^2$ . Let

$$a(L) = a(1)L + b(L)(1 - L),$$

where  $b(L) = 1 - \sum_{i=1}^{p-1} b_i L^i$  and  $b_i = -\sum_{j=i+1}^p a_j$ , and rearrange the equation

$$\begin{aligned} b(L)\Delta y_t &= -a(1)y_{t-1} + \epsilon_t \quad \text{or} \\ \Delta y_t &= \rho y_{t-1} + \sum_{i=1}^{p-1} b_i \Delta y_{t-i} + \epsilon_t, \end{aligned}$$

where  $\rho = -a(1) = -1 + \sum_{i=1}^p a_i$ . Note that the assumption of a single unit root in the DGP implies  $\rho = 0$ . Under the null, we get an MA representation

$$\begin{aligned} \Delta y_t &= c(L)\epsilon_t \\ &= u_t \end{aligned}$$

where  $c(L) = b(L)^{-1} = 1 + \sum_{i=1}^{\infty} c_i L^i$ .

Consider a regression equation

$$\begin{aligned} \Delta y_t &= \alpha + \rho y_{t-1} + u_t \\ &= \mathbf{x}'_t \boldsymbol{\beta} + u_t, \end{aligned}$$

where  $\mathbf{x}_t = (1, y_{t-1})'$ ,  $\boldsymbol{\beta} = (\alpha, \rho)'$ , and  $u_t$  is a regression error with mean zero and variance  $\sigma_u^2$ . Define a scaling matrix

$$\mathbf{S}_T = \begin{bmatrix} \sqrt{T} & \mathbf{0} \\ \mathbf{0} & T \end{bmatrix}$$

and write the deviation of the OLS estimates using the scaling matrix

$$\begin{aligned} \mathbf{S}_T(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) &= \left\{ \mathbf{S}_T^{-1} \left[ \sum_{i=1}^T \mathbf{x}_t \mathbf{x}_t' \right] \mathbf{S}_T^{-1} \right\}^{-1} \left\{ \mathbf{S}_T^{-1} \left[ \sum_{i=1}^T \mathbf{x}_t u_t \right] \right\} \\ &= \begin{bmatrix} 1 & T^{-\frac{3}{2}} \sum_{i=1}^T y_{t-1} \\ \cdot & T^{-2} \sum_{i=1}^T y_{t-1}^2 \end{bmatrix}^{-1} \begin{bmatrix} T^{-\frac{1}{2}} \sum_{i=1}^T u_t \\ T^{-1} \sum_{i=1}^T y_{t-1} u_t \end{bmatrix}, \end{aligned}$$

where under the null of  $\Delta y_t = u_t$

$$\begin{aligned} \begin{bmatrix} 1 & T^{-\frac{3}{2}} \sum_{i=1}^T y_{t-1} \\ \cdot & T^{-2} \sum_{i=1}^T y_{t-1}^2 \end{bmatrix} &\xrightarrow{L} \begin{bmatrix} 1 & \lambda \int W(r) dr \\ \cdot & \lambda^2 \int W(r)^2 dr \end{bmatrix} \quad \text{and} \\ \begin{bmatrix} T^{-\frac{1}{2}} \sum_{i=1}^T u_t \\ T^{-1} \sum_{i=1}^T y_{t-1} u_t \end{bmatrix} &\xrightarrow{L} \begin{bmatrix} \lambda W(1) \\ \lambda^2 \int W dW \end{bmatrix} + \begin{bmatrix} 0 \\ \frac{1}{2}(\lambda^2 - \gamma_0) \end{bmatrix}. \end{aligned}$$

Thus,

$$\begin{bmatrix} T^{\frac{1}{2}} \hat{\alpha} \\ T \hat{\rho} \end{bmatrix} \xrightarrow{L} \begin{bmatrix} 1 & \lambda \int W(r) dr \\ \cdot & \lambda^2 \int W(r)^2 dr \end{bmatrix}^{-1} \left\{ \begin{bmatrix} \lambda W(1) \\ \lambda^2 \int W dW \end{bmatrix} + \begin{bmatrix} 0 \\ \frac{1}{2}(\lambda^2 - \gamma_0) \end{bmatrix} \right\}.$$

In particular,

$$\begin{aligned} T \hat{\rho} &\xrightarrow{L} [0 \quad 1] \begin{bmatrix} 1 & \int W(r) dr \\ \cdot & \int W(r)^2 dr \end{bmatrix}^{-1} \left\{ \begin{bmatrix} \lambda W(1) \\ \lambda^2 \int W dW \end{bmatrix} + \begin{bmatrix} 0 \\ \frac{\lambda^2 - \gamma_0}{2\lambda^2} \end{bmatrix} \right\} \\ &= \frac{\int W dW - W(1) \int W(r) dr}{\int W(r)^2 dr - (\int W(r) dr)^2} + \frac{(\lambda^2 - \gamma_0)/2\lambda^2}{\int W(r)^2 dr - (\int W(r) dr)^2} \end{aligned}$$

Note that the second component can be consistently estimated by

$$\frac{T^2 s_{\hat{\rho}}^2}{\hat{\sigma}_u^2} \frac{\hat{\lambda}^2 - \hat{\gamma}_0}{2}$$

because

$$T^2 \mathbf{S}_{\hat{\rho}}^2 \xrightarrow{L} \frac{\sigma_u^2 / \lambda^2}{\int W(r)^2 dr - (\int W(r) dr)^2}.$$

Accordingly, we get the PP  $\rho$  test

$$T\hat{\rho} - \frac{T^2\mathbf{s}_{\hat{\rho}}^2}{\hat{\sigma}_u^2} \frac{\hat{\lambda}^2 - \hat{\gamma}_0}{2} \xrightarrow{L} \frac{\int W dW - W(1) \int W(r) dr}{\int W(r)^2 dr - (\int W(r) dr)^2},$$

which follows the same asymptotic distribution as the DF  $\rho$  test.

Similarly, the variance of  $\hat{\beta}$  follows

$$\begin{aligned} \mathbf{S}_T \hat{\Sigma}_{\hat{\beta}} \mathbf{S}_T &= \hat{\sigma}_u^2 \left\{ \mathbf{S}_T^{-1} \left[ \sum_{i=1}^T \mathbf{x}_i \mathbf{x}_i' \right] \mathbf{S}_T^{-1} \right\}^{-1} \\ &= \hat{\sigma}_u^2 \begin{bmatrix} 1 & T^{-\frac{3}{2}} \sum_{i=1}^T y_{t-1} \\ \cdot & T^{-2} \sum_{i=1}^T y_{t-1}^2 \end{bmatrix}^{-1} \\ &\xrightarrow{L} \sigma_u^2 \begin{bmatrix} 1 & \lambda \int W(r) dr \\ \cdot & \lambda^2 \int W(r)^2 dr \end{bmatrix}^{-1}. \end{aligned}$$

In particular, the standard error of  $\hat{\rho}$  follows

$$T\mathbf{s}_{\hat{\rho}} \xrightarrow{L} \frac{\sigma_u/\lambda}{[\int W(r)^2 dr - (\int W(r) dr)^2]^{\frac{1}{2}}}$$

and the  $t$ -test follows

$$\begin{aligned} t_{\hat{\rho}} = \frac{T\hat{\rho}}{T\mathbf{s}_{\hat{\rho}}} &\xrightarrow{L} \frac{\left\{ [\int W dW - W(1) \int W(r) dr] + \frac{\lambda^2 - \gamma_0}{2\lambda^2} \right\} / [\int W(r)^2 dr - (\int W(r) dr)^2]}{(\sigma_u/\lambda) \left\{ 1 / [\int W(r)^2 dr - (\int W(r) dr)^2] \right\}^{\frac{1}{2}}} \\ &\xrightarrow{L} \frac{\lambda}{\sigma_u} \left\{ \frac{\int W dW - W(1) \int W(r) dr}{[\int W(r)^2 dr - (\int W(r) dr)^2]^{\frac{1}{2}}} + \frac{\frac{\lambda^2 - \gamma_0}{2\lambda^2}}{[\int W(r)^2 dr - (\int W(r) dr)^2]^{\frac{1}{2}}} \right\}. \end{aligned}$$

Note that the second component can be consistently estimated by

$$\frac{T\mathbf{s}_{\hat{\rho}}}{\hat{\sigma}_u} \frac{\hat{\lambda}^2 - \hat{\gamma}_0}{2\hat{\lambda}}$$

because

$$T^2\mathbf{s}_{\hat{\rho}}^2 \xrightarrow{L} \frac{\sigma_u^2/\lambda^2}{\int W(r)^2 dr - (\int W(r) dr)^2}.$$

Accordingly, we get the PP  $t$  test

$$\frac{\hat{\sigma}_u t_{\hat{\rho}}}{\hat{\lambda}} - \frac{T\mathbf{s}_{\hat{\rho}}}{\hat{\sigma}_u} \frac{\hat{\lambda}^2 - \hat{\gamma}_0}{2\hat{\lambda}} \xrightarrow{L} \frac{\int W dW - W(1) \int W(r) dr}{[\int W(r)^2 dr - (\int W(r) dr)^2]^{\frac{1}{2}}},$$

which follows the same asymptotic distribution as the DF  $t$  test. Note that  $\gamma_0 (= E(u_t^2))$  can be consistently estimated by  $\hat{\sigma}_u^2 (= \frac{1}{T-2} \sum_{t=1}^T \hat{u}_t^2)$  and that  $\lambda$  can be consistently estimated by the Newey-West estimator

$$\hat{\lambda}^2 = \hat{\gamma}_0 + 2 \sum_{j=1}^q \left(1 - \frac{j}{q+1}\right) \hat{\gamma}_j,$$

where  $\hat{\gamma}_j = \frac{1}{T} \sum_{t=j+1}^T \hat{u}_t \hat{u}_{t-j}$ .

**The regression equation includes a constant term and a time trend when the true process is a unit root process with or without a drift**

Now, consider a DGP:

$$a(L)y_t = \mu + \epsilon_t$$

and rearrange the equation

$$\begin{aligned} b(L)\Delta y_t &= \mu - a(1)y_{t-1} + \epsilon_t \quad \text{or} \\ \Delta y_t &= \mu + \rho y_{t-1} + \sum_{i=1}^{p-1} b_i \Delta y_{t-i} + \epsilon_t. \end{aligned}$$

Under the null, we get an MA representation

$$\begin{aligned} \Delta y_t &= \theta + c(L)\epsilon_t \\ &= \theta + u_t \end{aligned}$$

where  $\theta = c(1)\mu$ .

Consider a regression equation

$$\Delta y_t = \mu + \delta t + \rho y_{t-1} + u_t.$$

Note that the regression is subject to collinearity because  $y_{t-1}$  contains a deterministic time trend component if  $\mu \neq 0$ . To avoid the possible collinearity, rewrite the equation

using a detrended series  $\xi_t = y_t - \mu t$

$$\begin{aligned}\Delta y_t &= \mu + \delta t + \rho(\xi_{t-1} + \mu(t-1))u_t \\ &= (1-\rho)\mu + (\delta + \rho\mu)t + \rho\xi_{t-1}u_t \\ &= \alpha + \tau t + \rho\xi_{t-1} + u_t \\ &= \mathbf{x}'_t \boldsymbol{\beta} + \epsilon_t,\end{aligned}$$

where  $\mathbf{x}_t = (1, t, \xi_{t-1})'$ , and  $\boldsymbol{\beta} = (\alpha, \tau, \rho)'$ . Define a scaling matrix

$$\mathbf{S}_T = \begin{bmatrix} \sqrt{T} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \sqrt[3]{T} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & T \end{bmatrix}$$

and write the deviation of the OLS estimates using the scaling matrix

$$\begin{aligned}\mathbf{S}_T(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) &= \left\{ \mathbf{S}_T^{-1} \left[ \sum_{i=1}^T \mathbf{x}_i \mathbf{x}'_i \right] \mathbf{S}_T^{-1} \right\}^{-1} \left\{ \mathbf{S}_T^{-1} \left[ \sum_{i=1}^T \mathbf{x}_i u_i \right] \right\} \\ &= \begin{bmatrix} 1 & T^{-2} \sum_{i=1}^T t & T^{-\frac{3}{2}} \sum_{i=1}^T \xi_{t-1} \\ \cdot & T^{-3} \sum_{i=1}^T t^2 & T^{-\frac{5}{2}} \sum_{i=1}^T t \xi_{t-1} \\ \cdot & \cdot & T^{-2} \sum_{i=1}^T \xi_{t-1}^2 \end{bmatrix}^{-1} \begin{bmatrix} T^{-\frac{1}{2}} \sum_{i=1}^T u_t \\ T^{-\frac{3}{2}} \sum_{i=1}^T t u_t \\ T^{-1} \sum_{i=1}^T \xi_{t-1} u_t \end{bmatrix},\end{aligned}$$

where under the null of  $\Delta \xi_t = u_t$

$$\begin{aligned}\begin{bmatrix} 1 & T^{-2} \sum_{i=1}^T t & T^{-\frac{3}{2}} \sum_{i=1}^T \xi_{t-1} \\ \cdot & T^{-3} \sum_{i=1}^T t^2 & T^{-\frac{5}{2}} \sum_{i=1}^T t \xi_{t-1} \\ \cdot & \cdot & T^{-2} \sum_{i=1}^T \xi_{t-1}^2 \end{bmatrix} &\xrightarrow{L} \begin{bmatrix} 1 & \frac{1}{2} & \lambda \int W(r) dr \\ \cdot & \frac{1}{3} & \lambda \int r W(r) dr \\ \cdot & \cdot & \lambda^2 \int W(r)^2 dr \end{bmatrix} \text{ and} \\ \begin{bmatrix} T^{-\frac{1}{2}} \sum_{i=1}^T u_t \\ T^{-\frac{3}{2}} \sum_{i=1}^T t u_t \\ T^{-1} \sum_{i=1}^T \xi_{t-1} u_t \end{bmatrix} &\xrightarrow{L} \begin{bmatrix} \lambda W(1) \\ \lambda \int r dW \\ \lambda^2 \int W dW \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ \frac{1}{2}(\lambda^2 - \gamma_0) \end{bmatrix}.\end{aligned}$$

Thus, we get

$$\begin{aligned}\begin{bmatrix} T^{\frac{1}{2}} \hat{\alpha} \\ T^{\frac{3}{2}} \hat{\tau} \\ T \hat{\rho} \end{bmatrix} &\xrightarrow{L} \begin{bmatrix} 1 & \frac{1}{2} & \lambda \int W(r) dr \\ \cdot & \frac{1}{3} & \lambda \int r W(r) dr \\ \cdot & \cdot & \lambda^2 \int W(r)^2 dr \end{bmatrix}^{-1} \left\{ \begin{bmatrix} \lambda W(1) \\ \lambda \int r dW \\ \lambda^2 \int W dW \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ \frac{1}{2}(\lambda^2 - \gamma_0) \end{bmatrix} \right\} \\ &\xrightarrow{L} \begin{bmatrix} \lambda & 0 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & \frac{1}{2} & \int W(r) dr \\ \cdot & \frac{1}{3} & \int r W(r) dr \\ \cdot & \cdot & \int W(r)^2 dr \end{bmatrix}^{-1} \left\{ \begin{bmatrix} W(1) \\ \int r dW \\ \int W dW \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ \frac{1}{2}(\lambda^2 - \gamma_0)/\lambda^2 \end{bmatrix} \right\}.\end{aligned}$$

In particular,

$$T\hat{\rho} \xrightarrow{L} [0 \ 0 \ 1] \begin{bmatrix} 1 & \frac{1}{2} & \int W(r)dr \\ \cdot & \frac{1}{3} & \int rW(r)dr \\ \cdot & \cdot & \int W(r)^2dr \end{bmatrix}^{-1} \begin{bmatrix} W(1) \\ \int r dW \\ \int W dW \end{bmatrix} \\ + \frac{\lambda^2 - \gamma_0}{2\lambda^2} [0 \ 0 \ 1] \begin{bmatrix} 1 & \frac{1}{2} & \int W(r)dr \\ \cdot & \frac{1}{3} & \int rW(r)dr \\ \cdot & \cdot & \int W(r)^2dr \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}.$$

Note that the second component can be consistently estimated by

$$\frac{T^2 s_{\hat{\rho}}^2}{\hat{\sigma}_u^2} \frac{\hat{\lambda}^2 - \hat{\gamma}_0}{2}$$

because

$$T^2 \mathbf{s}_{\hat{\rho}}^2 \xrightarrow{L} \frac{\sigma_u^2}{\lambda^2} [0 \ 0 \ 1] \begin{bmatrix} 1 & \frac{1}{2} & \int W(r)dr \\ \cdot & \frac{1}{3} & \int rW(r)dr \\ \cdot & \cdot & \int W(r)^2dr \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}.$$

Accordingly, we get the PP  $\rho$  test

$$T\hat{\rho} - \frac{T^2 \mathbf{s}_{\hat{\rho}}^2}{\hat{\sigma}_u^2} \frac{\hat{\lambda}^2 - \hat{\gamma}_0}{2} \xrightarrow{L} [0 \ 0 \ 1] \begin{bmatrix} 1 & \frac{1}{2} & \int W(r)dr \\ \cdot & \frac{1}{3} & \int rW(r)dr \\ \cdot & \cdot & \int W(r)^2dr \end{bmatrix}^{-1} \begin{bmatrix} W(1) \\ \int r dW \\ \int W dW \end{bmatrix}$$

which follows the same asymptotic distribution as the DF  $\rho$  test.

Similarly, the variance of  $\hat{\beta}$  follows

$$\mathbf{S}_T \hat{\Sigma}_{\hat{\beta}} \mathbf{S}_T = \hat{\sigma}_u^2 \left\{ \mathbf{S}_T^{-1} \left[ \sum_{i=1}^T \mathbf{x}_t \mathbf{x}_t' \right] \mathbf{S}_T^{-1} \right\}^{-1} \\ = \hat{\sigma}_u^2 \begin{bmatrix} 1 & T^{-2} \sum_{i=1}^T t & T^{-\frac{3}{2}} \sum_{i=1}^T \xi_{t-1} \\ \cdot & T^{-3} \sum_{i=1}^T t^2 & T^{-\frac{5}{2}} \sum_{i=1}^T t \xi_{t-1} \\ \cdot & \cdot & T^{-2} \sum_{i=1}^T \xi_{t-1}^2 \end{bmatrix}^{-1} \\ \xrightarrow{L} \sigma_u^2 \begin{bmatrix} 1 & \frac{1}{2} & \lambda \int W(r)dr \\ \cdot & \frac{1}{3} & \lambda \int rW(r)dr \\ \cdot & \cdot & \lambda^2 \int W(r)^2dr \end{bmatrix}^{-1}.$$

In particular, the standard error of  $\hat{\rho}$  follows

$$T \mathbf{s}_{\hat{\rho}} \xrightarrow{L} \frac{\sigma_u}{\lambda} \left\{ [0 \ 0 \ 1] \begin{bmatrix} 1 & \frac{1}{2} & \int W(r)dr \\ \cdot & \frac{1}{3} & \int rW(r)dr \\ \cdot & \cdot & \int W(r)^2dr \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \right\}^{\frac{1}{2}}$$

and the  $t$ -test follows

$$t_{\hat{\rho}} = \frac{T\hat{\rho}}{Ts_{\hat{\rho}}} \xrightarrow{L} \left(\frac{\lambda}{\sigma_u}\right) \frac{[0 \ 0 \ 1] \begin{bmatrix} 1 & \frac{1}{2} & \int W(r)dr \\ \cdot & \frac{1}{3} & \int rW(r)dr \\ \cdot & \cdot & \int W(r)^2dr \end{bmatrix}^{-1} \begin{bmatrix} W(1) \\ \int rdW \\ \int WdW \end{bmatrix}}{\left\{ [0 \ 0 \ 1] \begin{bmatrix} 1 & \frac{1}{2} & \int W(r)dr \\ \cdot & \frac{1}{3} & \int rW(r)dr \\ \cdot & \cdot & \int W(r)^2dr \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \right\}^{\frac{1}{2}}} \\ + \left(\frac{\lambda}{\sigma_u}\right) \frac{\lambda^2 - \gamma_0}{2\lambda^2} \left\{ [0 \ 0 \ 1] \begin{bmatrix} 1 & \frac{1}{2} & \int W(r)dr \\ \cdot & \frac{1}{3} & \int rW(r)dr \\ \cdot & \cdot & \int W(r)^2dr \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \right\}^{\frac{1}{2}}.$$

Thus, we get

$$\left(\frac{\sigma_u}{\lambda}\right)t_{\hat{\rho}} \xrightarrow{L} \frac{[0 \ 0 \ 1] \begin{bmatrix} 1 & \frac{1}{2} & \int W(r)dr \\ \cdot & \frac{1}{3} & \int rW(r)dr \\ \cdot & \cdot & \int W(r)^2dr \end{bmatrix}^{-1} \begin{bmatrix} W(1) \\ \int rdW \\ \int WdW \end{bmatrix}}{\left\{ [0 \ 0 \ 1] \begin{bmatrix} 1 & \frac{1}{2} & \int W(r)dr \\ \cdot & \frac{1}{3} & \int rW(r)dr \\ \cdot & \cdot & \int W(r)^2dr \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \right\}^{\frac{1}{2}}} \\ + \frac{\lambda^2 - \gamma_0}{2\lambda^2} \left\{ [0 \ 0 \ 1] \begin{bmatrix} 1 & \frac{1}{2} & \int W(r)dr \\ \cdot & \frac{1}{3} & \int rW(r)dr \\ \cdot & \cdot & \int W(r)^2dr \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \right\}^{\frac{1}{2}}.$$

Note that the second component can be consistently estimated by

$$\frac{Ts_{\hat{\rho}}}{\hat{\sigma}_u} \frac{\hat{\lambda}^2 - \hat{\gamma}_0}{2\hat{\lambda}}$$

because

$$T^2 \mathbf{S}_{\hat{\rho}}^2 \xrightarrow{L} \frac{\sigma_u^2}{\lambda^2} [0 \ 0 \ 1] \begin{bmatrix} 1 & \frac{1}{2} & \int W(r)dr \\ \cdot & \frac{1}{3} & \int rW(r)dr \\ \cdot & \cdot & \int W(r)^2dr \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}.$$

Accordingly, we get the PP  $t$  test

$$\frac{\hat{\sigma}_u t_{\hat{\rho}} - \frac{Ts_{\hat{\rho}}}{\hat{\sigma}_u} \frac{\hat{\lambda}^2 - \hat{\gamma}_0}{2\hat{\lambda}}}{\hat{\lambda}} \xrightarrow{L} \frac{[0 \ 0 \ 1] \begin{bmatrix} 1 & \frac{1}{2} & \int W(r)dr \\ \cdot & \frac{1}{3} & \int rW(r)dr \\ \cdot & \cdot & \int W(r)^2dr \end{bmatrix}^{-1} \begin{bmatrix} W(1) \\ \int rdW \\ \int WdW \end{bmatrix}}{\left\{ [0 \ 0 \ 1] \begin{bmatrix} 1 & \frac{1}{2} & \int W(r)dr \\ \cdot & \frac{1}{3} & \int rW(r)dr \\ \cdot & \cdot & \int W(r)^2dr \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \right\}^{\frac{1}{2}}},$$

which follows the same asymptotic distribution as the DF  $t$  test.

## Appendix

### 13.A Asymptotic Theory

#### 13.A.1 Functional Central Limit Theorem

For the purpose of deriving asymptotic distributions for unit root tests, it is convenient to generalize the concept of convergence in distribution. Instead of considering a sequence of random variables or random vectors, we will consider a sequence of random functions. This consideration leads to a generalized version of the central limit theorem.

Let  $(S, \mathcal{F}, Pr)$  be a probability space and  $S$  be a metric space with a metric  $d$ . The class  $\mathcal{B}$  of *Borel sets* in  $M$  is the  $\sigma$ -field generated by the open sets of  $M$ . If a function  $x$  which maps  $S$  into  $M$  is measurable  $\mathcal{F}/\mathcal{B}$ , then  $x$  is a *random element*. A random element  $x$  induces a probability measure  $Pr^*$  on  $(M, \mathcal{B})$  when we define  $Pr^*(B) = Pr(x \in B)$  for any  $B$  in  $\mathcal{B}$ . A sequence  $\{x_j : j \geq 1\}$  of random elements is said to *converge in distribution* to a random element  $x_0$  if

$$(13.A.1) \text{ ???}$$

Masao  
needs to  
check this!

### 13.B Procedures for Unit Root Tests

#### 13.B.1 Said-Dickey Test (ADF.EXP)

Said-Dickey test with the general-to-specific rules proceeds as follows:

- (i) Choose whether or not a constant and a time trend should be included in the regression by selecting an appropriate alternative hypothesis. If the variable of interest does not exhibit any secular trend, an appropriate alternative hypothesis should be that the variable is stationary with non-zero mean and without a

time trend. In this case, the regression should include a constant but no time trend. On the other hand, if the variable of interest exhibits a secular trend, an appropriate alternative hypothesis is that the variable is trend stationary. Therefore, the regression should include both a constant and a linear time trend.

- (ii) Select the maximum order of lagged polynomials (the corresponding variable to be determined is  $P$ ).
- (iii) Determine the order of autoregressive process by following Campbell and Perron (1991)'s recommendation.
- (iv) If the  $t$  ratio consistent with the specification of the regression form is negative and greater than the appropriate critical value in absolute value, then reject the null of a unit root.

### 13.B.2 Park's $J$ Test (JPQ.EXP)

Park's  $J(p, q)$  test proceeds as follows:

- (i) Choose the order of the maintained trend in the regression (the corresponding variable in the program is  $P$ ). If the variable of interest does not exhibit a secular time trend, the maintained hypothesis is that it includes only a constant (set  $P=0$ ). However, if it shows a secular time trend, the maintained hypothesis is that it possesses a linear time trend (set  $P=1$ ).
- (ii) Select the largest order of additional time polynomials (the corresponding variable in the program is  $Q$ ) and its range (the corresponding variable in the program is  $DQ$ ) in the regression. If the variable of interest does not exhibit a secular time trend, the maintained hypothesis is that it includes only a constant

(set  $Q=1$ ). However, if it shows a secular time trend, the maintained hypothesis is that it possesses a linear time trend (set  $Q=2$ ). Choose an appropriate  $DQ$  depending on how many test results you want. We recommend either  $DQ=2$  or  $DQ=3$ .

- (iii) If  $J(p, q)$  is smaller than the appropriate critical value, then reject the null of difference stationarity.

### 13.B.3 Park's $G$ Test (GPQ.EXP)

Park's  $G(p, q)$  test proceeds as follows:

- (i) Choose the order of the maintained trend in the regression (the corresponding variable in the program is  $P$ ). If the variable of interest does not exhibit a secular time trend, the maintained hypothesis is that it includes only a constant (set  $P=0$ ). However, if it shows a secular time trend, the maintained hypothesis is that it possesses a linear time trend (set  $P=1$ ).
- (ii) Select the largest order of additional time polynomials (the corresponding variable in the program is  $Q$ ) and its range (the corresponding variable in the program is  $DQ$ ) in the regression. If the variable of interest does not exhibit a secular time trend, the maintained hypothesis is that it includes only a constant (set  $Q=1$ ). However, if it shows a secular time trend, the maintained hypothesis is that it possesses a linear time trend (set  $Q=2$ ). Choose an appropriate  $DQ$  depending on how many test results you want. We recommend either  $DQ=2$  or  $DQ=3$ .
- (iii) Specify an appropriate method to estimate the long-run covariance matrix,  $\Omega_T$ .

See chapter 6 for more details (the corresponding variables to be specified are MAXD, ST, BST, and MSERHO).

- (iv) If  $G(p, q)$  is greater than the appropriate critical value, then reject the null of stationarity.

## Exercises

**13.1** Imagine that you are applying the Said-Dickey (augmented Dickey-Fuller) test to the log real GDP for the United States. Explain the Said-Dickey test (the definition, the null and alternative hypotheses that are appropriate in this context, and the small sample properties compared with the Phillips and Perron test). If the test statistic takes the value of -3.33, do you reject the null hypothesis at the 5 percent level? What if the value is -1.47? What if the value is +3.99? The critical values for the Said-Dickey test are given in Table 13.2, in which  $p$  is the order of time polynomial included in the regression.

Table 13.2: Probability of smaller values

0.01	0.025	0.05	0.10	0.90	0.95	0.975	0.99
$p = 0$ (a constant)							
-3.43	-3.12	-2.86	-2.57	-0.44	-0.07	0.23	0.60
$p = 1$ (a constant and a time trend)							
-3.96	-3.66	-3.41	-3.12	-1.25	-0.94	-0.66	-0.33

**13.2** Imagine that you are applying the Said-Dickey (augmented Dickey-Fuller) test to the log real exchange rate for the United States and United Kingdom for the purpose of testing Purchasing Power Parity. Explain the Said-Dickey test (the definition, the null and alternative hypotheses which are appropriate in this context, and the

small sample properties compared with the Phillips and Perron test.) If the test statistic takes the value of  $-2.93$ , do you reject the null hypothesis at the 5 percent level? What if the value is  $-2.67$ ? What if the value is  $+3.99$ . The critical values for the Said-Dickey test are given in 13.2, in which  $p$  is the order of time polynomial included in the regression.

## References

- BEVERIDGE, S., AND C. R. NELSON (1981): "A New Approach to Decomposition of Economic Time Series into Permanent and Transitory Components with Particular Attention to Measurement of the 'Business Cycle'," *Journal of Monetary Economics*, 7, 151–174.
- BIERENS, H. J., AND S. GUO (1993): "Testing Stationarity and Trend Stationarity Against the Unit Root Hypothesis," *Econometric Reviews*, 12, 1–32.
- BLOUGH, S. R. (1992): "The Relationship between Power and Level for Generic Unit Root Tests in Finite Samples," *Journal of Applied Econometrics*, 7, 295–308.
- CAMPBELL, J. Y., AND P. PERRON (1991): "Pitfalls and Opportunities: What Macroeconomists Should Know about Unit Roots," in *NBER Macroeconomics Annual*, ed. by O. J. Blanchard, and S. Fischer, pp. 141–201. MIT Press, Cambridge, MA.
- CHOI, I., AND B. C. AHN (1999): "Testing the Null of Stationarity for Multiple Time Series," *Journal of Econometrics*, 88(1), 41–77.
- CHRISTIANO, L. J., AND M. EICHENBAUM (1991): "Unit Roots in Real GNP: Do We Know, and Do We Care?," *Carnegie-Rochester Conference Series on Public Policy*, 32, 7–61.
- COCHRANE, J. H. (1988): "How Big is the Random Walk in GNP?," *Journal of Political Economy*, 96(5), 893–920.
- DICKEY, D. A., AND W. A. FULLER (1979): "Distribution of the Estimators for Autoregressive Time Series With a Unit Root," *Journal of the American Statistical Association*, 74, 427–431.
- (1981): "Likelihood Ratio Statistics for Autoregressive Time Series With a Unit Root," *Econometrica*, 49(4), 1057–1072.
- DIEBOLD, F. X., AND M. NERLOVE (1990): "Unit Roots in Economic Time Series: A Selective Survey," in *Advances in Econometrics: Cointegration, Spurious Regressions, and Unit Roots*, ed. by T. B. Fomby, and G. F. Rhodes, pp. 3–69. JAI Press, Greenwich, Connecticut.
- ELLIOTT, G., T. J. ROTHENBERG, AND J. H. STOCK (1996): "Efficient Tests for an Autoregressive Unit Root," *Econometrica*, 64(4), 813–836.
- FAUST, J. (1996): "Near Observational Equivalence and Theoretical Size Problems with Unit Root Tests," *Econometric Theory*, 12, 724–731.
- FUKUSHIGE, M., M. HATANAKA, AND Y. KOTO (1994): "Testing for the Stationarity and the Stability of Equilibrium," in *Advances in Econometrics, Sixth World Congress*, ed. by C. A. Sims, vol. I, pp. 3–45. Cambridge University Press.

- FULLER, W. A. (1976): *Introduction to Statistical Time Series*. John Wiley and Sons, New York.
- HALL, A. R. (1994): "Testing for a Unit Root in Time Series with Pretest Data-Based Model Selection," *Journal of Business and Economic Statistics*, 12, 461–470.
- HAMILTON, J. D. (1994): *Time Series Analysis*. Princeton University Press, Princeton.
- HANSEN, L. P. (1993): "Semiparametric Efficiency Bound for Linear Time-Series Models," in *Models, Methods, and Applications of Econometrics: Essays in Honor of A.R. Bergstrom*, ed. by P. C. B. Phillips. Blackwell, Oxford.
- KAHN, J. A., AND M. OGAKI (1990): "A Chi-Square Test for a Unit Root," *Economics Letters*, 34, 37–42.
- (1992): "A Consistent Test for the Null of Stationarity Against the Alternative of a Unit Root," *Economics Letters*, 39, 7–11.
- KWIATKOWSKI, D., P. C. B. PHILLIPS, P. SCHMIDT, AND Y. C. SHIN (1992): "Testing the Null Hypothesis of Stationarity against the Alternative of a Unit-Root - How Sure are We that Economic Time-Series Have a Unit-Root," *Journal of Econometrics*, 54(1–3), 159–178.
- NELSON, C. R., AND C. I. PLOSSER (1982): "Trends and Random Walks in Macroeconomic Time Series: Some Evidence and Implications," *Journal of Monetary Economics*, 10, 139–162.
- NEWBY, W. K., AND K. D. WEST (1987): "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, 55(3), 703–708.
- NG, S., AND P. PERRON (1995): "Unit Root Tests in ARMA Models with Data-Dependent Methods for the Selection of the Truncation Lag," *Journal of the American Statistical Association*, 90(429), 268–281.
- PARK, J. Y. (1989): "On the Joint Test of A Unit Root and Time Trend," Manuscript, Cornell University.
- (1990): "Testing for Unit Roots and Cointegration by Variable Addition," in *Advances in Econometrics: Cointegration, Spurious Regressions and Unit Roots*, ed. by T. Fomby, and G. Rhodes, vol. 8, pp. 107–133. JAI Press, Greenwich, CT.
- PARK, J. Y., AND B. CHOI (1988): "A New Approach to Testing for a Unit Root," CAE Working Paper No. 88-23, Cornell University.
- PHILLIPS, P. C. B. (1987): "Time Series Regression with a Unit Root," *Econometrica*, 55(2), 277–301.
- PHILLIPS, P. C. B., AND P. PERRON (1988): "Testing for a Unit Root in Time Series Regression," *Biometrika*, 75(2), 335–346.
- SAID, E. S., AND D. A. DICKEY (1984): "Testing for Unit Roots in Autoregressive-Moving Average Models of Unknown Order," *Biometrika*, 71(3), 599–607.
- SCHWERT, G. W. (1989): "Tests for Unit Roots: A Monte Carlo Investigation," *Journal of Business and Economic Statistics*, 7, 147–159.
- STOCK, J. H., AND M. W. WATSON (1988): "Variable Trends in Economic Time Series," *Journal of Economic Perspectives*, 2(3), 147–74.

- WATSON, M. W. (1994): "Vector Autoregressions and Cointegration," in *Handbook of Econometrics*, ed. by R. F. Engle, and D. L. McFadden, vol. IV, chap. 47, pp. 2843–2915. Elsevier Science Publishers.

## Chapter 14

# COINTEGRATING AND SPURIOUS REGRESSIONS

This chapter reviews properties of regression estimators and test statistics based on the estimators when the regressors and regressant are difference stationary. When the stochastic trends of two or more difference stationary variables are eliminated by forming a linear combination of these variables, the variables are said to be cointegrated in the terminology of Engle and Granger (1987). Let  $\mathbf{z}_t$  be a  $n \times 1$  vector of difference stationary random variables with  $\Delta\mathbf{z}_t$  being stationary. If there exists a nonzero vector of real numbers  $\boldsymbol{\beta}$  such that  $\boldsymbol{\beta}'\mathbf{z}_t$  is stationary, then  $\mathbf{z}_t$  is said to be *cointegrated* with a *cointegrating vector*  $\boldsymbol{\beta}$ . If  $\boldsymbol{\beta}$  is a cointegrating vector,  $b\boldsymbol{\beta}$  is also a cointegrating vector for any real number  $b$ . There may exist more than one linearly independent cointegrating vector. This chapter covers the case in which there is only one linearly independent cointegrating vector, and the case in which there exists no cointegrating vector. Chapter 16 concerns the case when there are more than one linearly independent cointegrating vectors.

When there is one cointegrating vector, a regression of one variable in  $\mathbf{z}_t$  on the others is called a *cointegrating regression*. What is striking about cointegration

is that a cointegrating vector that eliminates the stochastic trends can be estimated consistently by a cointegrating regression without using instrumental variables, even when no variables are exogenous.

When there is no cointegrating vector, a regression of one variable in  $\mathbf{z}_t$  on the others is called a *spurious regression*. One reason why macroeconomists need to be careful about unit root nonstationary variables is that the standard asymptotic theory for regressions in Chapter 5 can be very misleading when variables in a regression are difference stationary.

In the first section, cointegration, stochastic cointegration, and the deterministic cointegration restriction are defined. Then some estimators for cointegrating vectors are described. Tests for the null of no cointegration and the null of cointegration as well as tests for the number of cointegrating vectors are presented. Section 14.6 discusses how cointegration may be combined with standard econometric methods that assume stationarity.

## 14.1 Definitions

If  $\boldsymbol{\beta}$  is a cointegrating vector,  $b\boldsymbol{\beta}$  is also a cointegrating vector for any real number  $b$ . It is often convenient to normalize one of the elements of  $\boldsymbol{\beta}$  by one. Suppose that the first element of  $\boldsymbol{\beta}$  is nonzero, then partition  $\mathbf{z}_t$  by  $\mathbf{z}_t = (y_t, \mathbf{x}_t)'$  and normalize  $\boldsymbol{\beta}$  by  $\boldsymbol{\beta} = (1, -\mathbf{c})'$ . Here  $y_t$  is a difference stationary process,  $\mathbf{x}_t$  is a vector difference stationary process, and  $\mathbf{c}$  is a normalized cointegrating vector.

For most macroeconomic time series such as aggregate income, consumption, and investment, we observe secular upward trends. A secular upward trend of a time series implies that the expected value of the first difference of the series is positive,

which implies that the drift term of the series is positive if the series is difference stationary.

Nonzero drift terms in a system of difference stationary series introduce the deterministic trends in addition to the stochastic trends. Hence the cointegrating vector, which eliminates the stochastic trends, may or may not eliminate the deterministic trends from the system. In order to distinguish these cases, we now introduce the notions of stochastic cointegration and the deterministic cointegration restriction, as defined by Ogaki and Park (1997).<sup>1</sup> Consider a vector difference stationary process  $\mathbf{x}_t$  with drift:

$$(14.1) \quad \mathbf{x}_t - \mathbf{x}_{t-1} = \boldsymbol{\mu}_x + \mathbf{v}_t$$

for  $t \geq 1$  where  $\boldsymbol{\mu}_x$  is an  $(n-1)$ -dimensional vector of real numbers and  $\mathbf{v}_t$  is stationary with mean zero. Recursive substitution in (14.1) yields

$$(14.2) \quad \mathbf{x}_t = \boldsymbol{\mu}_x t + \mathbf{x}_t^0$$

where  $\mathbf{x}_t^0$  is difference stationary without drift. Relation (14.2) decomposes the difference stationary process  $\mathbf{x}_t$  into deterministic trends arising from drift  $\boldsymbol{\mu}_x$  and the difference stationary process without drift,  $\mathbf{x}_t^0$ . Suppose that  $y_t$  is a scalar difference stationary process with drift  $\mu_y$ . Similarly, decompose  $y_t$  into a deterministic trend  $\mu_y t$  and a difference stationary process without drift,  $y_t^0$ , as in (14.2):

$$(14.3) \quad y_t = \mu_y t + y_t^0.$$

Difference stationary processes  $y_t$  and  $\mathbf{x}_t$  are said to be *stochastically cointegrated*

---

<sup>1</sup>Ogaki (1988) introduces these notions and calls them the stochastic and deterministic parts of cointegration. West (1988) considers estimation under the deterministic cointegration restriction for the special case of one stochastic trend in the system. Hansen (1992a) and Park (1992) consider the deterministic cointegration restriction under more general cases.

with a *normalized cointegrating vector*  $\mathbf{c}$  when there exists an  $(n-1)$ -dimensional vector  $\mathbf{c}$  such that  $y_t - \mathbf{c}'\mathbf{x}_t$  is trend stationary.<sup>2</sup> This property means that stochastic cointegration only requires that stochastic trend components of the series are cointegrated. We may then write  $y_t^0 - \mathbf{c}'\mathbf{x}_t^0 = \theta_c + \epsilon_t$ , where  $\epsilon_t$  is stationary with mean zero. Then by (14.2) and (14.3),

$$(14.4) \quad y_t = \theta_c + m_c t + \mathbf{c}'\mathbf{x}_t + \epsilon_t$$

where

$$(14.5) \quad m_c = \mu_y - \mathbf{c}'\boldsymbol{\mu}_x.$$

Suppose that

$$(14.6) \quad \mu_y = \mathbf{c}'\boldsymbol{\mu}_x$$

holds. Then the deterministic cointegration restriction is said to hold. This means that the cointegrating vector that eliminates the stochastic trends also eliminates the deterministic trends. If this restriction is satisfied, then

$$(14.7) \quad y_t = \theta_c + \mathbf{c}'\mathbf{x}_t + \epsilon_t.$$

and  $(y_t, \mathbf{x}_t)'$  is cointegrated.

Another way to explain the deterministic cointegration is to use an idea of cotrending. Suppose that a vector  $\mathbf{c}^*$  satisfies

$$(14.8) \quad \mu_y = \mathbf{c}^{*'}\boldsymbol{\mu}_x.$$

Then  $y_t - \mathbf{c}^{*'}\mathbf{x}_t$  does not possess any deterministic trend, and  $y_t$  and  $\mathbf{x}_t$  are *cotrended* with a *normalized cotrending vector*  $\mathbf{c}^*$ . If  $n > 2$  and if one of the components

---

<sup>2</sup>If  $y_t^0 - \mathbf{c}'\mathbf{x}_t^0$  is stationary rather than trend stationary,  $y_t$  and  $\mathbf{x}_t$  are said to be cointegrated.

of  $\boldsymbol{\mu}_x$  is nonzero, there are infinitely many cotrending vectors. Consider an extra restriction that the normalized cointegrating vector  $\mathbf{c}$  is a cotrending vector. This restriction, which we call the *deterministic cointegration restriction*, requires that the cointegrating vector eliminates both the stochastic and deterministic trends. In this case, Equation (14.7) holds and  $(y_t, \mathbf{x}_t)'$  is cointegrated.

## 14.2 Exact Finite Sample Properties of Regression Estimators

This section studies exact finite sample properties of cointegrating and spurious regression estimators. In the literature on unit root econometrics, asymptotic theory and the method of Monte Carlo studies have been typically used. However, the conditional Gauss-Markov theorem in Chapter 5 can be applied to study exact finite sample properties as in Ogaki and Choi (2001).

Consider a regression of the form

$$(14.9) \quad y_t = \mathbf{h}'\mathbf{d}_t + \mathbf{c}'\mathbf{x}_t + \epsilon_t.$$

where  $\mathbf{d}_t$  is a function of time,  $t$ . For example,  $\mathbf{d}_t = (1, t)'$  as in (14.4) or  $\mathbf{d}_t = 1$  as in (14.7). If  $\epsilon_t$  is stationary for some  $\mathbf{c}$ , (14.9) is a cointegrating regression. If  $\epsilon_t$  is difference stationary for any  $\mathbf{c}$ , then (14.9) is a spurious regression.

### 14.2.1 Spurious Regressions

Suppose that  $y_t$  is a random walk and  $x_t$  is a random walk that is independent of  $y_t$ . Granger and Newbold (1974) find that the standard Wald test statistic for the hypothesis that the coefficient on  $x_t$  is zero tends to be large (compared with standard critical values) in ordinary least squares (OLS) regressions of  $y_t$  onto  $x_t$  in their Monte

Carlo experiments. Later, Phillips (1986) show that the Wald test statistic diverges to infinity as the sample size is increased. In a regression with two independent difference stationary variables without drift, the random walk components will dominate the stationary components at least asymptotically. Hence these spurious regression results imply that the absolute value of the  $t$ -ratio of the regressor tends to be larger than the critical value implied by the standard statistical theory that assumes stationarity. An econometrician who ignores unit root nonstationarity issues tends to spuriously conclude that two independent difference stationary variables are related.

Another example of the spurious regression results is in Durlauf and Phillips (1988). When a difference stationary variable without drift,  $y_t$ , is regressed onto a constant and a linear time trend, the Wald test statistic for the hypothesis that a coefficient for the linear trend is zero diverges to infinity as the sample size increases.

The Gauss Markov theorem provides us with a tool to understand exact small sample properties of estimators and test statistics of spurious regressions. The asymptotic theories of Phillips (1987, 1998) have been used to understand the spurious regression problem, but have not been used to provide a solution to the problem. The Gauss Markov theorem indicates a simple solution to the problem.

Let  $y_t$  be a random walk that is generated from

$$(14.10) \quad \Delta y_t = \epsilon_t$$

with an initial random variable  $y_0$  and a white noise  $\epsilon_t$  that is conditionally homoskedastic. Let  $x_t$  be another random walk that is generated from

$$(14.11) \quad \Delta x_t = v_t$$

with an initial random variable  $x_0$  and a white noise  $v_t$  that is conditionally ho-

moskedastic. We assume that  $\{\epsilon_t\}_{t=1}^T$  and  $y_0$  are independent, and that they are independent from  $\{v_t\}_{t=1}^T$  and  $x_0$ , so that  $x_t$  and  $y_t$  are independent random walks. Let  $\mathbf{y} = \{y_t\}_{t=1}^T$ ,  $\mathbf{X} = \{x_t\}_{t=1}^T$  and  $\mathbf{e} = \{e_t\}_{t=1}^T$  where  $\Delta e_t = \epsilon_t$ , and consider the OLS estimator for  $\mathbf{y} = \mathbf{X}b_0 + \mathbf{e}$ . Then the true value of the regression coefficient is zero:  $b_0 = 0$ .

Let  $I_x$  be the information set generated from  $y_0$  and  $\mathbf{X}$ . Assumptions 5.1, 5.2, and 5.4 of the strict version of the theorem in Chapter 5 hold for the spurious regression. However, Assumption 5.3 is violated because

$$(14.12) \quad E(\mathbf{e}\mathbf{e}'|I_x) = \sigma^2\mathbf{\Phi}$$

where  $\sigma^2 = E(\epsilon_t^2)$ , and

$$(14.13) \quad \mathbf{\Phi} = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 & 1 \\ 1 & 2 & 2 & \cdots & 2 & 2 \\ 1 & 2 & 3 & \cdots & 3 & 3 \\ \cdot & \cdot & \cdot & \cdots & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot & \cdot \\ 1 & 2 & 3 & \cdots & T-1 & T-1 \\ 1 & 2 & 3 & \cdots & T-1 & T \end{bmatrix}.$$

Thus the spurious regression violates Assumption 5.3, but not the other assumptions. The OLS estimator is still unbiased. One can apply a GLS correction and obtain a more efficient estimator.

When Assumption 5.5 is made, by applying GLS to the spurious regression, we can solve the spurious regression problem: we can obtain the exact (unconditional)  $t$  distribution for the usual  $t$  statistic.

We now consider spurious regressions of the form (14.9) which do not satisfy the strict exogeneity assumption. For this purpose, we consider a particular data generating process that leads to a spurious regression.

Let  $e_t$  and  $\mathbf{z}_t$  be two time series of dimensions 1 and  $k$ , respectively, that are generated from

$$(14.14) \quad \Delta e_t = \epsilon_t, \quad t = 1, 2, 3, \dots$$

$$(14.15) \quad \Delta \mathbf{z}_t = \boldsymbol{\mu} + \mathbf{v}_t, \quad t = -q, \dots, -1, 0, 1, \dots$$

where  $(\epsilon_t, \mathbf{v}_t)'$  is a covariance stationary series,  $\boldsymbol{\mu}$  is a  $k$ -dimensional vector of real numbers,  $e_0 = 0$ , and  $\mathbf{z}_{-q}$  is a given random vector. We assume that the long-run covariance matrix of  $\mathbf{v}_t$ ,

$$(14.16) \quad \boldsymbol{\Omega} = \lim_{j \rightarrow \infty} \sum_{-j}^j E(\mathbf{v}_t \mathbf{v}'_{t-j})$$

is nonsingular. We assume that  $\mathbf{z}_t$  is strictly exogenous with respect to  $\epsilon_t$ .

An implication of the strict exogeneity assumption is that  $e_t$  and  $\mathbf{z}_t$  are not cointegrated: that is, there is no nonzero vector  $\boldsymbol{\beta}$  such that  $e_t - \boldsymbol{\beta}' \mathbf{z}_t$  is stationary. This property results because the assumption implies that

$$(14.17) \quad \lim_{j \rightarrow \infty} \sum_{-j}^j E(\epsilon_t \mathbf{v}'_{t-j}) = 0.$$

Consider a series  $y_t$  that is generated from

$$(14.18) \quad y_t = \mathbf{h}' \mathbf{d}_t + \mathbf{c}' \mathbf{z}_t + \gamma(L^{-1}) \Delta \mathbf{z}_t + \boldsymbol{\eta}(L) \Delta \mathbf{z}_t + e_t, \quad t = 1, 2, 3, \dots,$$

where  $\gamma(L^{-1}) = \gamma_1 L^{-1} + \dots + \gamma_p L^{-p}$ ,  $\boldsymbol{\eta}(L) = \boldsymbol{\eta}_0 + \boldsymbol{\eta}_1 L + \dots + \boldsymbol{\eta}_q L^q$  and  $\mathbf{d}_t$  is a vector of deterministic variables that is  $(1, t)'$  or 1 for example. Here  $\gamma_1, \dots, \gamma_p$ , and  $\boldsymbol{\eta}_0, \dots, \boldsymbol{\eta}_q$  are  $1 \times k$  vectors, and we assume that at least one of them is nonzero.

Under these assumptions, consider a regression of  $y_t$  onto  $\mathbf{d}_t$  and  $\mathbf{z}_t$ :

$$(14.19) \quad y_t = \mathbf{h}^* \mathbf{d}_t + \mathbf{c}^* \mathbf{z}_t + e_t^*$$

This regression is a spurious regression: that is, for any vector  $\mathbf{c}^*$ ,  $e_t^*$  is unit root nonstationary. To see this property, assume that  $e_t^*$  is stationary for a vector  $\mathbf{c}^*$ . Then (14.18) implies that  $e_t - (\mathbf{c}^* - \mathbf{c})'\mathbf{z}_t$  is stationary. It follows that a cointegrating relationship exists between  $e_t$  and  $\mathbf{z}_t$  contradicting the strict exogeneity assumption.

Given that  $e_t$  in (14.18) satisfies the exogeneity condition,  $\mathbf{c}$  can be considered the true value of the spurious regression coefficient  $\mathbf{c}^*$  in (14.19). With this interpretation, one problem with (14.19) is that the strict exogeneity assumption is violated.

Let  $\mathbf{X}$  be a matrix whose  $t$ -th row is given by  $(\mathbf{d}'_t, \mathbf{z}'_t, \Delta\mathbf{z}'_{t+p}, \Delta\mathbf{z}'_{t+p-1}, \dots, \Delta\mathbf{z}'_t, \Delta\mathbf{z}'_{t-1}, \dots, \Delta\mathbf{z}'_{t-q})$ ,  $\mathbf{y} = \{y_t\}_{t=1}^T$ , and  $\mathbf{e}^* = \{e_t^*\}_{t=1}^T$ . When

$$(14.20) \quad E(\mathbf{e}^* \mathbf{e}^{*\prime} | \mathbf{X}) = \sigma^2 \mathbf{\Psi}$$

with a known matrix  $\mathbf{\Psi}$  and a possibly unknown number  $\sigma$ , then the GLS can be applied to (14.18). If  $e_t$  is a random walk, then with  $\mathbf{\Phi}$  given by (14.13) and  $\sigma^2 = E(\epsilon_t^2)$ . Just as in the strict exogenous case, the finite sample properties of the GLS estimators and test statistics based on GLS can be analyzed.

The GLS correction is basically the same as taking first differences for the case of strictly exogenous regressors. The GLS correction, however, can be useful in applications for which the strict exogeneity assumption is violated.

## 14.2.2 Cointegrating Regressions

Let  $e_t$  and  $\mathbf{z}_t$  be two time series of dimensions 1 and  $k$ , respectively. We assume that  $\mathbf{z}_t$  is generated from (14.15), where  $(e_t, \mathbf{v}'_t)'$  is a covariance stationary series,  $\boldsymbol{\mu}$  is a  $k$ -dimensional vector of real numbers,  $e_0 = 0$ , and  $\mathbf{z}_{-q}$  is a given random vector. We assume that the long-run covariance matrix of  $\mathbf{v}_t$ ,  $\boldsymbol{\Omega} = \lim_{j \rightarrow \infty} \sum_{-j}^j E(\mathbf{v}_t \mathbf{v}'_{t-j})$  is nonsingular. We assume that  $\mathbf{z}_t$  is strictly exogenous with respect to  $e_t$ .

Consider a series  $y_t$  that is generated from

$$(14.21) \quad y_t = \mathbf{h}'\mathbf{d}_t + \mathbf{c}'\mathbf{z}_t + \gamma(L^{-1})\Delta\mathbf{z}_t + \boldsymbol{\eta}(L)\Delta\mathbf{z}_t + e_t, \quad t = 1, 2, 3, \dots,$$

where  $\gamma(L^{-1})$ ,  $\boldsymbol{\eta}(L)$  and  $\mathbf{d}_t$  are defined in (14.18).

Under these assumptions, consider a regression of  $y_t$  onto  $\mathbf{d}_t$  and  $\mathbf{z}_t$ :

$$(14.22) \quad y_t = \mathbf{h}^{*'}\mathbf{d}_t + \mathbf{c}^{*'}\mathbf{z}_t + e_t^*$$

This regression is a cointegrating regression. With an appropriate choice of  $\mathbf{h}^*$  and  $\mathbf{c}^* = \mathbf{c}$ ,  $e_t^*$  is stationary. However, since the strict exogeneity assumption is not satisfied, the OLS estimator for (14.22) is biased.

In contrast, the OLS estimator for (14.21) is unbiased. It is the BLUE if  $e_t$  is serially uncorrelated. This is because Assumptions 5.1, 5.2, 5.3 and 5.4 are satisfied, and the conditional Gauss-Markov theorem applies. The OLS estimator for (14.21) is called the dynamic OLS estimator. The GLS estimator for (14.21) is called the dynamic GLS estimator.

### 14.3 Large Sample Properties

An important feature of the cointegration regression is that the OLS estimator is consistent without any exogeneity assumption (see Phillips and Durlauf, 1986; Stock, 1987). Along with the spurious regression results discussed in the last section, it is another example of the fact that the standard asymptotic theory in Chapter 5 does not apply to regressions in the presence of unit root nonstationary variables. This fact is well known in the literature. On the other hand, the fact that the conditional probability version of the Gauss Markov theorem applies to cointegrating regressions under the assumptions of the theorem has not been emphasized in the literature. In

the context of cointegration, an assumption of the theorem requires that  $\mathbf{x}_t$  is strictly exogenous.

In most applications, the strict exogeneity assumption is too restrictive. This section discusses econometric methods for when the assumption is violated. The OLS estimator is consistent (see Phillips and Durlauf, 1986; Stock, 1987), but is asymptotically biased. It also has a nonstandard distribution, which makes statistical inference very difficult. For example, the OLS standard errors calculated in the standard econometric packages for OLS are not very meaningful for cointegrating regressions. Many efficient estimation methods that solve all or some of these problems have been developed. Dynamic OLS and GLS estimators introduced in the last section were proposed by Stock and Watson (1993). Phillips and Loretan (1991) and Saikkonen (1991) have proposed similar estimators.

Dynamic OLS and GLS estimators correct the endogeneity problem parametrically. Estimators proposed by Phillips and Hansen (1990) and Park's (1992) Canonical Cointegrating Regressions correct the endogeneity problem nonparametrically. In Chapter 16, we will explain Johansen (1988, 1991) Maximum Likelihood Estimation method.

### 14.3.1 Canonical Cointegrating Regression

Johansen's maximum likelihood estimation makes a parametric correction for long-run correlation of  $\Delta\mathbf{x}_t$  and  $\mathbf{e}_t$ . Another way to obtain an efficient estimator is to utilize a nonparametric estimate of the long-run covariance parameters. Both Phillips and Hansen (1990) and Park (1992) employ such covariance estimates. Here, attention is confined to Park's Canonical Cointegration Regressions (CCR).

Consider a cointegrated system

$$(14.23) \quad y_t = \mathbf{h}'\mathbf{d}_t + \mathbf{c}'\mathbf{x}_t + \epsilon_t$$

$$(14.24) \quad \Delta\mathbf{x}_t = \mathbf{v}_t,$$

where  $\mathbf{d}_t$  is a deterministic term that are usually constants, time trends, or both,  $y_t$  and  $\mathbf{x}_t$  are difference stationary, and  $\epsilon_t$  and  $\mathbf{v}_t$  are stationary with zero mean. Here  $y_t$  is a scalar and  $\mathbf{x}_t$  is a  $(n-1) \times 1$  random vector. Let

$$(14.25) \quad \mathbf{w}_t = (\epsilon_t, \mathbf{v}_t)'$$

Define  $\Phi(i) = E(\mathbf{w}_t \mathbf{w}'_{t-i})$ ,  $\Sigma = \Phi(0)$ ,  $\Gamma = \sum_{i=0}^{\infty} \Phi(i)$ , and  $\Omega = \sum_{i=-\infty}^{\infty} \Phi(i)$ . Here  $\Omega$  is the matrix version of (14.16) and is the long run variance (or covariance) matrix of  $\mathbf{w}_t$ . Partition  $\Omega$  as

$$(14.26) \quad \Omega = \begin{bmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{bmatrix}$$

where  $\Omega_{11}$  is a scalar, and  $\Omega_{22}$  is a  $(n-1) \times (n-1)$  matrix, and partition  $\Gamma$  conformably.

Define

$$(14.27) \quad \Omega_{11.2} = \Omega_{11} - \Omega_{12}\Omega_{22}^{-1}\Omega_{21}$$

and  $\Gamma_2 = (\Gamma'_{12}, \Gamma'_{22})'$ . The CCR procedure assumes that  $\Omega_{22}$  is positive definite, implying that  $\mathbf{x}_t$  is not itself cointegrated (see, e.g., Phillips, 1986; Engle and Granger, 1987). This assumption assures that  $(1, -\mathbf{c})$  is the unique cointegrating vector (up to a scale factor).<sup>3</sup>

---

<sup>3</sup>For many applications, it is natural to assume that  $\Delta^{-1}\epsilon_t$  is not cointegrated with  $\mathbf{x}_t$ . This assumption implies that  $\Omega_{11.2}$  is positive. Park (1992) calls cointegration between  $y_t$  and  $\mathbf{x}_t$  singular when  $\Omega_{11.2}$  is zero. For the singular models, either a different CCR procedure described by Park is necessary (the removable singularity case) or the CCR procedure is not applicable (the essential singularity case).

The OLS estimator in (14.23) is super-consistent in that the estimator converges to  $\mathbf{c}$  at the rate of  $T$  (sample size) even when  $\Delta\mathbf{x}_t$  and  $\epsilon_t$  are correlated. The OLS estimator, however, is not asymptotically efficient. Consider transformations

$$(14.28) \quad y_t^* = y_t + \boldsymbol{\pi}_y' \mathbf{w}_t$$

$$(14.29) \quad \mathbf{x}_t^* = \mathbf{x}_t + \boldsymbol{\pi}_x' \mathbf{w}_t.$$

Since  $\mathbf{w}_t$  is stationary,  $y_t^*$  and  $\mathbf{x}_t^*$  are cointegrated with the same cointegrating vector  $(1, -\mathbf{c})$  as  $y_t$  and  $\mathbf{x}_t$  for any  $\boldsymbol{\pi}_y$  and  $\boldsymbol{\pi}_x$ . The idea of the CCR is to choose  $\boldsymbol{\pi}_y$  and  $\boldsymbol{\pi}_x$ , so that the OLS estimator is asymptotically efficient when  $y_t^*$  is regressed on

$\mathbf{x}_t^*$ .<sup>4</sup> This requires

$$(14.30) \quad \boldsymbol{\pi}_y = \boldsymbol{\Sigma}^{-1} \boldsymbol{\Gamma}_2 \mathbf{c} + (0, \boldsymbol{\Omega}_{12} \boldsymbol{\Omega}_{22}^{-1})'$$

$$(14.31) \quad \boldsymbol{\pi}_x = \boldsymbol{\Sigma}^{-1} \boldsymbol{\Gamma}_2.$$

In practice, long-run covariance parameters in these formulas are estimated, and estimated  $\boldsymbol{\pi}_y$  and  $\boldsymbol{\pi}_x$  are used to transform  $y_t$  and  $\mathbf{x}_t$ . As long as these parameters are estimated consistently, the resultant CCR estimator is asymptotically efficient.

Here we have considered a single regression. If there are many cointegrating regressions with disturbances with nonzero long-run covariances in an econometric system of interest, then asymptotically it is more efficient to apply seemingly unrelated regressions. Park and Ogaki (1991a) develop a method of Seemingly Unrelated Canonical Cointegrating Regressions (SUCCR) for this case. In the SUCCR, transformations of  $y_t$  and  $\mathbf{x}_t$  that are slightly different from (14.28) and (14.29) are applied

---

<sup>4</sup>Under general conditions, a sequence of functions  $\frac{1}{\sqrt{T}} \sum_{t=1}^T \mathbf{w}_t$  converges in distribution to a vector Brownian motion  $B$  with covariance matrix  $\boldsymbol{\Omega}$ . The OLS estimator converges in distribution to ????

Masao  
needs to  
check this!

in each regression. After transforming the variables, the standard seemingly unrelated regression method is applied to the transformed variables.

### 14.3.2 Estimation of Long-Run Covariance Parameters

In order to use efficient estimators for cointegrating vectors based on nonparametric correction such as CCR estimators, it is necessary to estimate long-run covariance parameters  $\mathbf{\Omega}$  and  $\mathbf{\Gamma}$ .

In many applications of cointegration, the order of serial correlation is unknown.

Let  $\mathbf{\Phi}(\tau) = E(\mathbf{w}_t \mathbf{w}'_{t-\tau})$ ,

$$(14.32) \quad \mathbf{\Phi}_T(\tau) = \frac{1}{T} \sum_{t=\tau+1}^T \hat{\mathbf{w}}_t \hat{\mathbf{w}}'_{t-\tau} \quad \text{for } \tau \geq 0,$$

and  $\mathbf{\Phi}_T(\tau) = \mathbf{\Phi}_T(-\tau)'$  for  $\tau < 0$ , where  $\hat{\mathbf{w}}_t$  is constructed from a consistent estimate of the cointegrating vector. Many estimators for  $\mathbf{\Omega}$  in the literature have the form

$$(14.33) \quad \mathbf{\Omega}_T = \frac{T}{T-p} \sum_{\tau=-T+1}^{T-1} k\left(\frac{\tau}{S_T}\right) \mathbf{\Phi}_T(\tau),$$

where  $k(\cdot)$  is a real-valued kernel, and  $S_T$  is a band-width parameter. The factor  $\frac{T}{T-p}$  is a small sample degrees of freedom adjustment. See Andrews (1991) for examples of kernels. Similarly,  $\mathbf{\Gamma}$  is estimated by

$$(14.34) \quad \mathbf{\Gamma}_T = \frac{T}{T-p} \sum_{\tau=0}^{T-1} k\left(\frac{\tau}{S_T}\right) \mathbf{\Phi}_T(\tau),$$

Park and Ogaki (1991b) extend Andrews and Monahan's (1992) VAR prewhitening method to the estimation of  $\mathbf{\Gamma}$  so that it can be applied to cointegrating regressions. The first step in the VAR prewhitening method is to run a VAR:

$$(14.35) \quad \mathbf{w}_t = \mathbf{A}_1 \mathbf{w}_{t-1} + \mathbf{A}_2 \mathbf{w}_{t-2} + \cdots + \mathbf{A}_k \mathbf{w}_{t-k} + \mathbf{f}_t.$$

Note that the model (14.35) need not be a true model in any sense. Then the estimated VAR is used to form an estimate  $\mathbf{f}_t$  and estimators of the form (14.33) and (14.34) are applied to the estimated  $\mathbf{f}_t$  to estimate the long-run variance of  $\mathbf{f}_t$ ,  $\mathbf{\Omega}^*$  and the parameter  $\mathbf{\Gamma}$  for  $\mathbf{f}_t$ ,  $\mathbf{\Gamma}^*$ . The estimator based on the QS kernel with the automatic bandwidth parameter can be used for  $\mathbf{f}_t$  for example. Then the sample counterpart of the formulas

$$(14.36) \quad \mathbf{\Omega} = [\mathbf{I} - \sum_{i=1}^k \mathbf{A}_i]^{-1} \mathbf{\Omega}^* [\mathbf{I} - \sum_{i=1}^k \mathbf{A}'_i]^{-1}$$

$$(14.37) \quad \mathbf{\Gamma} = \mathbf{\Phi}(0) + [\mathbf{I} - \sum_{i=1}^k \mathbf{A}_i]^{-1} (\mathbf{\Gamma}^* - E(\mathbf{f}_t \mathbf{f}'_t)) [\mathbf{I} - \sum_{i=1}^k \mathbf{A}'_i]^{-1} \\ + [\mathbf{I} - \sum_{i=1}^k \mathbf{A}_i]^{-1} \sum_{j=0}^{k-1} \sum_{i=j+1}^k \mathbf{A}_i \mathbf{\Phi}(-i)$$

are used to form estimates of  $\mathbf{\Omega}$  and  $\mathbf{\Gamma}$ .<sup>5</sup>

Monte Carlo experiments in Park and Ogaki (1991b) show that the VAR prewhitening improves small sample properties of CCR estimators substantially.

## 14.4 Tests for the Null Hypothesis of No Cointegration

Many tests for cointegration apply unit root tests to the residuals of a cointegrating regression. When tests for the null hypothesis of unit root nonstationarity are applied to residuals, the null of no cointegration is tested against the alternative of cointegration. It should be noted that the asymptotic distributions of these tests generally depend on the number of the variables in the cointegrating regression.

Engle and Granger's (1987) augmented Dickey-Fuller (ADF) test applies the Said-Dickey test to the residual from cointegrating regressions. The asymptotic prop-

---

<sup>5</sup>See Park and Ogaki (1991a) for a derivation of (14.36) and (14.37).

erties of the ADF test are studied in Phillips and Ouliaris (1990). These authors and MacKinnon (1990) tabulate critical values from Monte Carlo simulations. Note that these critical values assume the OLS is used for the cointegrating regression, so that the efficient estimation methods discussed in Section 14.3 above should not be used for this test. Just as the Said-Dickey test, the ADF test may be sensitive to the choice of the order of AR.

Phillips and Ouliaris (1990) also study asymptotic properties of tests for cointegration obtained by applying the Phillips-Perron test to OLS cointegrating regression residuals. Asymptotic critical values are reported by Phillips and Ouliaris. This test requires an estimate of the long run variance of the residual.

Park's (1990)  $I(p,q)$  test basically applies his  $J(p,q)$  test to OLS cointegrating regression residuals. This test was originally developed by Park, Ouliaris, and Choi (1988). The  $I(p,q)$  test is computed by adding spurious time trends as additional regressors in the cointegrating regression:

$$(14.38) \quad y_t = \sum_{\tau=0}^p \mu_{\tau} t^{\tau} + \sum_{\tau=p+1}^q \mu_{\tau} t^{\tau} + \mathbf{c}' \mathbf{x}_t + \epsilon_t.$$

Here, time polynomials up to the order of  $p$  represent maintained trends, while higher order time polynomials are spurious trends. Part of Park, Ouliaris, and Choi's (1988) table of critical values for  $I(p,q)$  tests are reproduced here in Table 14.1. This test has an advantage over ADF and Phillips-Ouliaris tests in that neither the order of AR nor the bandwidth parameter needs to be chosen.

Table 14.1: Critical Values of Park's  $I(p, q)$  Tests for Null of No Cointegration

Number of Regressors	Size	$I(0, 3)$	$I(1, 5)$
1	0.01	0.06864	0.10269
	0.05	0.23286	0.25064
	0.10	0.39897	0.49845
2	0.01	0.05520	0.00819
	0.05	0.17539	0.21040
	0.10	0.29622	0.32251

Note: These critical values are from Park, Ouliaris, and Choi (1988).

## 14.5 Tests for the Null Hypothesis of Cointegration

When an economic model implies cointegration, it is often more appealing to test for the null of cointegration, so that an econometrician can control the probability of rejecting a valid economic model. Phillips and Ouliaris (1990) discussed why it was hard to develop tests for the null of cointegration. More recently, Fukushige, Hatanaka, and Koto (1994), Hansen (1992b), and Kwiatkowski, Phillips, Schmidt, and Shin (1992), among others, have developed tests for the null of cointegration.

Park's (1990)  $H(p, q)$  test is computed by applying the CCR to (14.38). Thus, this test essentially applies Park's  $G(p, q)$  test to CCR residuals. A similar test was originally developed by Park, Ouliaris, and Choi (1988), where  $G(p, q)$  tests were applied to OLS residuals, and their tests have nonstandard distributions. In contrast, Park's  $H(p, q)$  tests have asymptotic chi-square distributions with  $q - p$  degrees of freedom. Under the alternative of no cointegration, the  $H(p, q)$  statistic diverges to infinity because spurious trends try to mimic the stochastic trend left in the residual. Therefore, this test is consistent.

In many applications, it is appropriate to model each variable in the econometric system as first difference stationary with drift. Each variable possess a linear deterministic trend as well as a stochastic trend in Section 14.1 because of drift. In this case,  $H(1, q)$  statistics test the null hypothesis of stochastic cointegration. The  $H(0, 1)$  test can be considered as a test for the deterministic cointegration restriction because the restriction implies that the cointegrating vector that eliminates the stochastic trends also eliminates the linear deterministic trends.

## 14.6 Generalized Method of Moments and Unit Roots

When difference stationary variables are involved in the econometric system, standard econometric methods that assume stationarity are not applicable because of spurious regression problems. Hence econometricians detrend data by taking growth rates of variables, for example. However, by detrending data, the econometrician loses the information contained in stochastic and deterministic trends. It is thus natural to seek a method to combine standard econometric methods and cointegrating regressions. Estimating an error correction representation explained in Section 16.4 is an example of such a method in vector autoregressions. Let us now consider this problem in the context of Hansen's (1982) Generalized Method of Moments (GMM) estimation. This case is particularly useful because many estimators can be considered special cases of GMM.

The asymptotic theory of GMM does not make strong distributional assumptions, such as that the variables are normally distributed. However, Hansen assumes that  $\mathbf{x}_t$  is stationary. Hence if variables are difference stationary, the econometrician needs to transform the variables to induce stationarity. One such transformation is

to take the first difference of a variable, or to take the growth rate of the variables if the log of the variable is difference stationary. But it may not be possible to take growth rates of all variables for some functions in  $f(\mathbf{x}_t, \mathbf{b}_0)$  while retaining moment conditions. In such cases, it may be possible to use cointegrating relationships to induce stationarity by taking linear combinations of variables. In empirical applications of Eichenbaum and Hansen (1990) and Eichenbaum, Hansen, and Singleton (1988), their economic models imply some variables are cointegrated with a known cointegrating vector. They use this cointegration relationship to induce stationarity for the equations involving the first order condition that equate the relative price and the marginal rate of substitution.

In Cooley and Ogaki (1996) and Ogaki and Reinhart (1998a,b) explained in the next chapter, their economic model implies a cointegration relationship, but the cointegrating vector is not known. They employ a two-step procedure. In the first step, they estimate the cointegrating vector, using a cointegrating regression. In the second step, they plug in estimates from the first step into GMM functions,  $f(\mathbf{x}_t, \mathbf{b}_0)$ . This two step procedure is similar to Engle and Granger's two step procedure for the error correction model discussed in Section 16.4. Asymptotic distributions of GMM estimators in the second step are not affected by the first step estimation because cointegrating regression estimators converge at a faster rate than  $\sqrt{T}$ .

## Appendix

### 14.A Procedures for Cointegration Tests

#### 14.A.1 Park's CCR and $H$ Test (CCR.EXP)

Park's canonical cointegrating regression (CCR) and  $H(p, q)$  test proceed as follows:

- (i) Define a regressand (the corresponding variable to be specified is  $Y$ ) and regressors. The latter includes both a vector of deterministic regressors<sup>6</sup> (the corresponding variable to be specified is  $X1$ ) and difference stationary regressors (the corresponding variable to be specified is  $X2$ ).
- (ii) Choose the order of the maintained trend (the corresponding variable to be specified is  $P$ ) in the residuals to test for the null hypothesis of cointegration ( $H(p, q)$  test). If either each variable exhibits no secular trend or some variables show a secular trend with the deterministic cointegration restriction<sup>7</sup>, set  $P=0$ . On the other hand, when some variables exhibit a time trend without the deterministic cointegration, set  $P=1$ .
- (iii) Select the largest order of additional time polynomials (the corresponding variable to be specified is  $Q$ ). If either each variable exhibits no secular trend or some variables show a secular trend with the deterministic cointegration restriction, set  $Q=1$ . But when some variables exhibit a time trend without the deterministic cointegration, set  $Q=2$ . Choose an appropriate  $DQ$  depending on how many test results you want. We recommend either  $DQ=2$  or  $DQ=3$ .
- (iv) Determine an appropriate method to estimate the long-run covariance matrix,  $\Omega_T$ . See chapter 6 for details (the corresponding variables to be specified are  $MAXD$ ,  $ST$ ,  $BST$ , and  $MSERHO$ . The default of the program is the prewhitened  $QS$  kernel with automatic bandwidth selection).
- (v) Impose restrictions on the cointegrating vector (the corresponding variable in

---

<sup>6</sup>It is typically either a constant or a constant and a linear time trend

<sup>7</sup>Typically, the economic model for the application tells us whether or not this restriction is satisfied

the program is  $B$ ), if any (the corresponding variables to be specified are  $R$  and  $RV$  matrices).

- (vi) Check the statistical evidence about estimates and tests. For CCR estimates, report the third stage result. For the  $H(p, q)$  test, report the fourth stage result. For a linear restriction  $RB=RV$ , report the third stage result when the alternative hypothesis is cointegration with  $RB \neq RV$ , and report the fourth stage result when the alternative hypothesis is no cointegration.

### 14.A.2 Park's $I$ Test (IPQ.EXP)

Park's  $I(p, q)$  test proceeds as follows:

- (i) Define a regressand (the corresponding variable to be specified is  $Y$ ) and regressors. The latter includes both a vector of deterministic regressors<sup>8</sup> (the corresponding variable to be specified is  $X1$ ) and difference stationary regressors (the corresponding variable to be specified is  $X2$ ).
- (ii) Choose the order of the maintained trend in the regression (the corresponding variable in the program is  $P$ ). If the variable of interest does not exhibit a secular time trend, the maintained hypothesis is that it includes only a constant (set  $P=0$ ). However, if it shows a secular time trend, the maintained hypothesis is that it possesses a linear time trend (set  $P=1$ ).
- (iii) Select the largest order of additional time polynomials (the corresponding variable in the program is  $Q$ ) and its range (the corresponding variable in the program is  $DQ$ ) in the regression. If the variable of interest does not exhibit a

---

<sup>8</sup>It is typically either a constant or a constant and a linear time trend

secular time trend, the maintained hypothesis is that it includes only a constant (set  $Q=1$ ). However, if it shows a secular time trend, the maintained hypothesis is that it possesses a linear time trend (set  $Q=2$ ). Choose an appropriate  $DQ$  depending on how many test results you want. We recommend either  $DQ=2$  or  $DQ=3$ .

- (iv) Impose restrictions on the cointegrating vector (the corresponding variable in the program is  $B$ ), if any (the corresponding variables to be specified are  $R$  and  $RV$  matrices).
- (v) If  $I(p, q)$  is smaller than the appropriate critical value, then reject the null of no cointegration.

## 14.B Weak Convergence to Stochastic Integral

Before we provide formal theorems of weak convergence to stochastic integral, by using a cointegrating regression, we show why the FCLT alone (even with the CMT) is not enough to establish the asymptotic properties of the OLS estimator.

Consider the following cointegrating regression:

$$y_t = \beta x_t + u_t,$$

where  $x_t$  is an  $I(1)$  process and  $u_t$  is an  $I(0)$  process. The OLS estimator is given by

$$\hat{\beta} = \frac{\sum_{t=1}^T y_t u_t}{\sum_{t=1}^T x_t^2},$$

and its sample error can be written as

$$T(\hat{\beta} - \beta) = \frac{\frac{1}{T} \sum_{t=1}^T x_t u_t}{\frac{1}{T^2} \sum_{t=1}^T x_t^2}.$$

For the denominator, by the FCLT (along with the CMT) it can be shown that:

$$\frac{1}{T^2} \sum_{t=1}^T x_t^2 \xrightarrow{d} \int_0^1 W^2(r) dr.$$

However, the numerator cannot be analyzed by the FCLT alone. It is evident that the asymptotic distribution of the numerator cannot be established by the FCLT alone, because the numerator is a mixture of I(1) and I(0) random variables. Therefore we need a different tool, so-called “weak convergence to the stochastic integral.” In below, we present the most general version of the theorem.

**Theorem 14.1** *Let  $\{U_{nt}, W_{nt}\}$  be a  $(2 \times 1)$  stochastic array, let  $X_n(r) = \sum_{t=1}^{[nr]} U_{nt}$  and  $Y_n(r) = \sum_{t=1}^{[nr]} W_{nt}$ , and suppose that  $(X_n(r), Y_n(r)) \xrightarrow{d} (B_X(r), B_Y(r))$ . Assume  $\{U_{nt}\}$  is  $L_r$ -bounded and  $L_2$ -NED of size -1 on  $\{\mathbf{V}_{nt}\}$  with respect to constants  $\{c_{nt}^U\}$ . If the one of the following assumptions hold:*

1.  $\{W_{nt}, \mathcal{H}_{nt}\}$  is a martingale difference array, where  $\mathcal{H}_{nt} = \sigma((W_{nk}, U_{n,k-1}, k \leq t)$ , and  $E(W_{n,t+1}^2 | \mathcal{H}_{n,t}) \ll (c_{n,t}^W)^2 < \infty$ , a.s.
2.  $W_{nt} = \sum_{k=0}^{\infty} \theta_k V_{1n,t-k}$  where  $V_{1nt} \in \mathbf{V}_{nt}$  is a  $L_r$ -bounded zero-mean random variable, independent of  $V_{n,t'}$  for all  $t \neq t'$ , and  $\sum_{t=0}^{\infty} \sum_{k=t}^{\infty} |\theta_k| < \infty$

Then,

$$G_n = \sum_{j=1}^{n-1} \left( \sum_{i=1}^j U_{n,i} \right) W_{n,j+1} \xrightarrow{d} \int_0^1 B_X(r) dB_Y(r) + \Lambda_{XY}$$

where  $\Lambda_{XY} = \lim_{n \rightarrow \infty} \sum_{i=1}^{n-1} \sum_{m=0}^{i-1} E(U_{n,i-m} W_{n,i+1})$

## References

ANDREWS, D. W. K. (1991): “Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation,” *Econometrica*, 59(3), 817–858.

- ANDREWS, D. W. K., AND J. C. MONAHAN (1992): "An Improved Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimator," *Econometrica*, 60(4), 953–966.
- COOLEY, T. F., AND M. OGAKI (1996): "A Time Series Analysis of Real Wages, Consumption, and Asset Returns," *Journal of Applied Econometrics*, 11(2), 119–134.
- DURLAUF, S. N., AND P. C. B. PHILLIPS (1988): "Trends versus Random Walks in Time Series Analysis," *Econometrica*, 56, 1333–1354.
- EICHENBAUM, M., AND L. P. HANSEN (1990): "Estimating Models with Intertemporal Substitution Using Aggregate Time Series Data," *Journal of Business and Economic Statistics*, 8, 53–69.
- EICHENBAUM, M., L. P. HANSEN, AND K. J. SINGLETON (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., AND C. GRANGER (1987): "Co-Integration and Error Correction: Representation, Estimation, and Testing," *Econometrica*, 55, 251–276.
- FUKUSHIGE, M., M. HATANAKA, AND Y. KOTO (1994): "Testing for the Stationarity and the Stability of Equilibrium," in *Advances in Econometrics, Sixth World Congress*, ed. by C. A. Sims, vol. I, pp. 3–45. Cambridge University Press.
- GRANGER, C. W. J., AND P. NEWBOLD (1974): "Spurious Regressions in Econometrics," *Journal of Econometrics*, 2, 111–120.
- HANSEN, B. E. (1992a): "Efficient Estimation and Testing of Cointegrating Vectors in the Presence of Deterministic Trends," *Journal of Econometrics*, 53, 87–121.
- (1992b): "Tests for Parameter Instability in Regression with I(1) Processes," *Journal of Business and Economic Statistics*, 10, 321–335.
- HANSEN, L. P. (1982): "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50(4), 1029–1054.
- JOHANSEN, S. (1988): "Statistical Analysis of Cointegration Vectors," *Journal of Economic Dynamics and Control*, 12, 231–254.
- (1991): "Estimation and Hypothesis Testing of Cointegration Vectors in Gaussian Vector Autoregressive Models," *Econometrica*, 59(6), 1551–1580.
- KWIATKOWSKI, D., P. C. B. PHILLIPS, P. SCHMIDT, AND Y. C. SHIN (1992): "Testing the Null Hypothesis of Stationarity against the Alternative of a Unit-Root - How Sure are We that Economic Time-Series Have a Unit-Root," *Journal of Econometrics*, 54(1–3), 159–178.
- MACKINNON, J. (1990): "Critical Values for Cointegration Tests," in *Long-Run Economic Relationships: Readings in Cointegration*, ed. by R. F. Engle, and C. W. J. Granger, pp. 267–276. Oxford University Press, Oxford.
- OGAKI, M. (1988): "Learning about Preferences from Time Trends," Ph.D. thesis, University of Chicago.
- OGAKI, M., AND C.-Y. CHOI (2001): "The Gauss-Markov Theorem and Spurious Regressions," Working Paper No. 01-13, Department of Economics, Ohio State University.

- OGAKI, M., AND J. Y. PARK (1997): "A Cointegration Approach to Estimating Preference Parameters," *Journal of Econometrics*, 82(1), 107–134.
- OGAKI, M., AND C. M. REINHART (1998a): "Intertemporal Substitution and Durable Goods: Long-Run Data," *Economics Letters*, 61, 85–90.
- (1998b): "Measuring Intertemporal Substitution: The Role of Durable Goods," *Journal of Political Economy*, 106, 1078–1098.
- PARK, J. Y. (1990): "Testing for Unit Roots and Cointegration by Variable Addition," in *Advances in Econometrics: Cointegration, Spurious Regressions and Unit Roots*, ed. by T. Fomby, and G. Rhodes, vol. 8, pp. 107–133. JAI Press, Greenwich, CT.
- (1992): "Canonical Cointegrating Regressions," *Econometrica*, 60(1), 119–143.
- PARK, J. Y., AND M. OGAKI (1991a): "Seemingly Unrelated Canonical Cointegrating Regressions," RCER Working Paper No. 280.
- (1991b): "VAR Prewhitening to Estimate Short-Run Dynamics: On Improved Method of Inference in Cointegrated Models," RCER Working Paper No. 281.
- PARK, J. Y., S. OULIARIS, AND B. CHOI (1988): "Spurious Regressions and Tests for Cointegration," CAE Working Paper No. 88-07, Cornell University.
- PHILLIPS, P. C. B. (1986): "Understanding Spurious Regressions in Econometrics," *Journal of Econometrics*, 33, 311–340.
- (1987): "Time Series Regression with a Unit Root," *Econometrica*, 55(2), 277–301.
- (1998): "New Tools for Understanding Spurious Regressions," *Econometrica*, 66, 1299–1325.
- PHILLIPS, P. C. B., AND S. N. DURLAUF (1986): "Multiple Time Series Regression with Integrated Processes," *Review of Economic Studies*, 53, 473–495.
- PHILLIPS, P. C. B., AND B. E. HANSEN (1990): "Statistical Inference in Instrumental Variables Regression with I(1) Processes," *Review of Economic Studies*, 57, 99–125.
- PHILLIPS, P. C. B., AND M. LORETAN (1991): "Estimating Long-Run Economic Equilibria," *Review of Economic Studies*, 58, 407–436.
- PHILLIPS, P. C. B., AND S. OULIARIS (1990): "Asymptotic Properties of Residual Based Tests for Cointegration," *Econometrica*, 58(1), 165–193.
- SAIKKONEN, P. (1991): "Asymptotically Efficient Estimation of Cointegrating Regressions," *Econometric Theory*, 7, 1–21.
- STOCK, J. H. (1987): "Asymptotic Properties of Least Squares Estimators of Cointegrating Vectors," *Econometrica*, 55, 1035–1056.
- STOCK, J. H., AND M. W. WATSON (1993): "A Simple Estimator of Cointegrating Vectors in Higher Order Integrated Systems," *Econometrica*, 61(4), 783–820.
- WEST, K. D. (1988): "Asymptotic Normality, When Regressors Have a Unit Root," *Econometrica*, 56(6), 1397–1417.

# Chapter 15

## ECONOMIC MODELS AND COINTEGRATING REGRESSIONS

Economic models often imply cointegration. This chapter illustrates how to derive cointegration restrictions from economic models.

In order to derive cointegration restrictions, one must show that a linear combination of difference stationary random variables is stationary. This is often done by showing that a linear combination of difference stationary variables is a time independent function of a finite number of stationary random variables (see Proposition 2.2).

For many economic models, Proposition 2.2 can be directly used to show cointegration. In some other models, one more step is necessary. There are cases in which economic models imply that a linear combination of difference stationary variables is a conditional expectation of a variable. For example, suppose that an economic model implies

$$\mathbf{b}'\mathbf{y}_t = E(z_t|\mathbf{I}_t),$$

where  $z_t$  can be shown to be a stationary random variable because of Proposition

2.2. Here,  $I_t$  is typically the information set available to the economic agents, which includes both stationary and difference stationary random variables. Since  $z_t$  is stationary,  $E(z_t|I_t)$  is likely to be stationary. However, in order to formally show that  $E(z_t|I_t)$  is stationary, we need an additional assumption that  $E(z_t|I_t)$  is equal to  $E(z_t|J_t)$ , where  $J_t$  is a subset of  $I_t$  and includes only a finite number of stationary variables. Then  $E(z_t|J_t)$  is a time independent function of a finite number of stationary variables, and we can use Proposition 2.2.

This additional assumption is not stringent as long as  $z_t$  is stationary. In order to see this property, suppose that  $I_t$  is generated by the current and past values of a difference stationary vector process  $\mathbf{x}_t$ :

$$\Delta \mathbf{x}_t = \mathbf{A} \Delta \mathbf{x}_{t-1} + \mathbf{u}_t.$$

where  $\mathbf{u}_t$  is a vector i.i.d. white noise process and all roots of the characteristic equation lie outside the unit circle. Here  $\mathbf{x}_t$  can include current and lagged values of many economic variables. If  $z_t = \mathbf{c}' \Delta \mathbf{x}_{t+1}$ , then

$$E(z_t|I_t) = \mathbf{c}' \mathbf{A} \Delta \mathbf{x}_t = E(z_t|\Delta \mathbf{x}_t).$$

## 15.1 The Permanent Income Hypothesis of Consumption

The standard version of the permanent income hypothesis of consumption implies cointegration. The exact form of cointegration depends on the assumption on the form of the difference stationarity of labor income.

Consider a representative consumer who maximizes a quadratic utility function

$$(15.1) \quad E_t \left[ \sum_{i=0}^{\infty} \beta^i (C_{t+i} - \gamma)^2 \right],$$

subject to the budget constraint

$$(15.2) \quad A_{t+1} = (1+r)A_t + Y_t^l - C_t$$

and a no-Ponzi-game condition

$$(15.3) \quad \lim_{i \rightarrow \infty} (1+r)^i A_{t+i} = 0 \quad \text{almost surely.}$$

Here  $Y_t^l$  denotes labor income, and  $A_t$  is wealth at date  $t$ . Assuming that  $\beta = (1+r)^{-1}$ , the optimal consumption is

$$(15.4) \quad C_t = rA_t + (1-\beta)E_t\left[\sum_{i=0}^{\infty} \beta^i Y_{t+i}^l\right].$$

Substituting (15.4) back in the budget constraint, we obtain

$$(15.5) \quad A_{t+1} - A_t = Y_t^l - (1-\beta)E_t\left[\sum_{i=0}^{\infty} \beta^i Y_{t+i}^l\right].$$

Let  $Y_t = rA_t + Y_t^l$  be total income, which includes both labor income and property income. Then (15.4) implies

$$(15.6) \quad C_t - Y_t = (1-\beta)E_t\left[\sum_{i=0}^{\infty} \beta^i Y_{t+i}^l\right] - Y_t^l.$$

The cointegration implication of the permanent income hypothesis is different depending on whether we assume that the level of labor income is difference stationary or we assume that the log of labor income is difference stationary as pointed out by Cochrane and Sbordone (1988).

First, assume that the level of labor income is difference stationary so that  $Y_t^l - Y_{t-1}^l$  is stationary. Then rewrite (15.6) as

$$(15.6') \quad C_t - Y_t = (1-\beta)E_t\left[\sum_{i=0}^{\infty} \beta^i \{Y_{t+i}^l - Y_t^l\}\right].$$

Since the right hand side of (15.6') is stationary, the permanent income hypothesis implies that  $C_t - Y_t$  is stationary, which can be called the stationarity restriction. It remains to show that  $C_t$  and  $Y_t$  are difference stationary. From (15.5),

$$(15.7) \quad \begin{aligned} Y_{t+1} - Y_t &= -r(1 - \beta)E_t\left[\sum_{i=0}^{\infty} \beta^i \{Y_{t+i}^l - Y_t^l\}\right] \\ &+ Y_{t+1}^l - Y_t^l \end{aligned}$$

because the right hand side is stationary, the left hand side is stationary, and  $Y_t$  is difference stationary. The stationarity restriction implies that  $C_t$  is a sum of difference stationary  $Y_t$  and a stationary variable. Hence  $C_t$  is difference stationary. Therefore in this case,  $C_t$  and  $Y_t$  are cointegrated with a  $(1, -1)$  cointegrating vector.

Second, assume that the log of labor income is difference stationary, so that  $\ln(Y_t^l) - \ln(Y_{t-1}^l)$  is stationary. Divide the both sides of (15.6) by  $Y_t$  to obtain

$$(15.8) \quad \frac{C_t}{Y_t} = 1 + (1 - \beta)E_t\left[\sum_{i=0}^{\infty} \beta^i \frac{Y_{t+i}^l}{Y_t}\right] - \frac{Y_t^l}{Y_t}.$$

With an additional assumption that  $\frac{Y_t^l}{A_t}$  is stationary, the right hand side of (15.8) is stationary, and  $\ln(Y_t)$  and  $\ln(C_t)$  are difference stationary. This additional assumption holds in standard general equilibrium models (see, e.g., King, Plosser, and Rebelo, 1988).

Thus, depending on whether the level of labor income or the log of labor income is assumed to be difference stationary, the permanent income hypothesis predicts different forms of cointegration. Which assumption is more appropriate? We observe from many economic data that the growth rate of an economic variable is stable over time. From this observation, it is more appropriate to assume that the log labor income is difference stationary rather than the level of labor income is difference stationary.

Another observation is that the assumption that the level of labor income is difference stationary implies that the level of saving,  $Y_t - C_t$ , is stationary. Since  $Y_t$  is nonstationary, the saving rate  $\frac{Y_t - C_t}{Y_t}$  is nonstationary under this assumption. In contrast, the assumption that the log labor income is difference stationary implies that the saving rate,  $\frac{Y_t - C_t}{Y_t} = 1 - \frac{C_t}{Y_t}$ , is stationary. Hence this assumption is consistent with Kuznets' (1946) stylized fact that the saving rate is stable in the U.S. in the long-run.

## 15.2 Present Value Models of Asset Prices

The standard present value model implies that the stock price and the dividend are cointegrated. The exact form of cointegration implied by the model depends on whether the level or the log of the dividend is assumed to be difference stationary as pointed out by Cochrane and Sbordone (1988).

Let  $p_t$  be the real stock price (after the dividend is paid) in period  $t$ , and  $d_t$  be the real dividend paid to the owner of the stock in period  $t$ . Then the arbitrage condition is

$$(15.9) \quad p_t = E[b(p_{t+1} + d_{t+1}) | \mathbf{I}_t],$$

where  $b$  is the constant real discount rate, and  $E(\cdot | \mathbf{I}_t)$  is the mathematical expectation operator conditioned on the information set  $\mathbf{I}_t$  in period  $t$ . Solving (15.9) forward and imposing the no bubble condition, we obtain the present value formula:

$$(15.10) \quad p_t = E\left(\sum_{i=1}^{\infty} b^i d_{t+i} | \mathbf{I}_t\right).$$

First, assume that  $d_t$  is difference stationary, following Campbell and Shiller

(1987). Then

$$(15.11) \quad p_t - \frac{b}{1-b}d_t = E\left[\sum_{i=1}^{\infty} b^i(d_{t+i} - d_t) | I_t\right].$$

Since  $d_{t+i} - d_t$  is stationary for any  $i$ , the right hand side of (15.11) is stationary. Hence we obtain a stationarity restriction that  $p_t - \frac{b}{1-b}d_t$  is stationary. This restriction implies that  $p_t$  is a sum of a difference stationary random variable and a stationary random variable. Hence  $p_t$  is difference stationary. This restriction also implies that  $p_t$  and  $d_t$  are cointegrated with a cointegrating vector  $[1, -\frac{b}{1-b}]'$ .

Second, assume that  $\ln(d_t)$  is difference stationary. Then dividing both sides of (15.10) by  $d_t$  yields

$$(15.12) \quad \frac{p_t}{d_t} = E\left[\sum_{i=1}^{\infty} b^i \frac{d_{t+i}}{d_t} | I_t\right].$$

The right hand side of this equation is stationary. Hence, taking the log of both sides of (15.12), we obtain a stationarity restriction that  $\ln(p_t) - \ln(d_t)$  is stationary. This restriction implies that  $\ln(p_t)$  is a sum of a difference stationary random variable and a stationary random variable. Hence  $\ln(p_t)$  is difference stationary. This restriction also implies that  $\ln(p_t)$  and  $\ln(d_t)$  are cointegrated with a cointegrating vector  $(1, -1)'$ .

When  $d_t$  is difference stationary, the cointegrating vector involves the discount factor,  $b$ . Hence a cointegrating regression can be used to estimate this structural parameter without making exogeneity assumptions. In addition to testing for cointegration, one can test the model by obtaining another estimate of  $b$  and compare it with a cointegrating regression estimate of  $b$  as explained in the next chapter. In contrast, when  $\ln(d_t)$  is assumed to be difference stationary, the cointegrating vector is known, and no structural parameter can be estimated by a cointegrating regression. Even in this case, it is possible to test the model by testing for cointegration.

The near observational equivalence problem, however, tells us that cointegration test results are not reliable. Hence it is more interesting to assume that  $d_t$  is difference stationary than to assume  $\ln(d_t)$  is difference stationary. Unfortunately, it is more reasonable to assume that  $\ln(d_t)$  is difference stationary because the growth rates of the stock price and the dividends are relatively stable over time.

### 15.3 Applications to Money Demand Functions

Another application of cointegration is to assume directly that a demand or supply function is stable in the long run. The stable function can be estimated by a cointegrating regression, and the model can be tested by testing for cointegration. The most important application of this type is estimating a money demand function (see, e.g., Hoffman and Rasche, 1991; Stock and Watson, 1993).

Let  $M_t$  be the real money balance,  $Y_t$  be real income, and  $i_t$  be the nominal interest rate. Let the money demand function be

$$(15.13) \quad \ln(M_t) = a_0 + a_1 \ln(Y_t) + a_2 i_t + u_t.$$

If the money demand function is stable in the long run,  $u_t$  is stationary. If we assume that  $\ln(Y_t)$  and  $i_t$  are difference stationary, and that they are not stochastically cointegrated, then the stable money demand function implies that  $\ln(M_t)$  is difference stationary, and that  $\ln(M_t)$ ,  $\ln(Y_t)$ , and  $i_t$  are cointegrated with  $(1, -a_1, -a_2)'$  as a cointegrating vector.

### 15.4 The Cointegration Approach to Estimating Preference Parameters

Ogaki and Park (1997) develop a cointegration approach to estimating preference

parameters by utilizing the information in stochastic and deterministic time trends.

The first order condition that equates the relative price and the contemporaneous

Masao  
needs to  
check this!

marginal rate of substitution of two goods is used to derive the restriction that the relative price and consumption of the two goods are cointegrated.<sup>1</sup> The cointegrating vector involves preference parameters that are estimated with a cointegrating regression. In their application, they estimate the (long-run) intertemporal elasticity of substitution (IES) of nondurable consumption, which is a key parameter in a Consumption-Based Asset Pricing Model (C-CAPM). The parameter was also estimated by Hansen's (1982) GMM in a C-CAPM. The C-CAPM is rejected strongly by Hansen and Singleton (1982) when stock returns and Treasury Bill rates are used together. Possible reasons for the rejection of the C-CAPM have been pointed out. These include liquidity constraints (see, e.g., Hayashi, 1985; Zeldes, 1989), unknown preference shocks (e.g., Garber and King, 1983), time-nonseparable preferences (e.g., Eichenbaum, Hansen, and Singleton, 1988; Constantinides, 1990; Eichenbaum and Hansen, 1990; Ferson and Constantinides, 1991; Ferson and Harvey, 1992; Cooley and Ogaki, 1996; Heaton, 1995), and small information cost (Cochrane, 1998). GMM estimation of nonlinear Euler equations also assumes that there are no measurement errors.

---

<sup>1</sup>Ogaki and Park use Houthakker's (1960) addilog utility function. The cointegration approach can also be used to estimate the curvature parameters of the extended addilog utility function as in Atkeson and Ogaki (1996), and the CES utility function as in Ogaki and Reinhart (1998). Deaton and Wigley (1971), Deaton (1974), Miron (1986), and Ball (1990), among others, have estimated addilog utility functions. Ogaki (1988) introduces the cointegration approach to estimate preference parameters of the addilog utility function. Ogaki (1992) uses the cointegration approach to estimate income elasticities for food and other goods; Braun (1994), to estimate a utility function for cash and credit goods; Cooley and Ogaki (1996), to estimate a utility function for consumption and leisure; Amano and Wirjanto (1996) and Amano, Ho, and Wirjanto (1998) to estimate models of import demand; and Amano and Wirjanto (1997) to estimate a model of government spending. Working independently, Clarida (1994, 1996) estimates addilog utility functions to estimate price and income elasticities for imported goods with cointegrating regressions.

The cointegration approach provides an estimator that is consistent even in the presence of factors such as liquidity constraints, aggregation over heterogeneous consumers, unknown preference shocks, a general form of time-nonseparability, measurement errors, and the possibility that consumers do not know the true stochastic law of motion of the economy. The GMM estimator is not consistent in the presence of these factors, but the cointegrating regression estimator is consistent under certain assumptions. It is important to develop such an estimator because a great amount of recent research simulates economies with features that accounts for GMM's rejections of the C-CAPM such as liquidity constraints in recent works (see, e.g., Deaton, 1991; Marcet and Singleton, 1991; Heaton and Lucas, 1992). An estimator that is consistent in the presence of liquidity constraints can be used to guide the choice of parameters for these simulations.

In Section 15.5, we will discuss Cooley and Ogaki's (1996) test that compares the estimates obtained using cointegration techniques with those obtained using GMM in the spirit of Hausman's (1978) specification test. Since the GMM estimator is not consistent but the cointegrating regression estimator is consistent in the presence of factors such as liquidity constraints, this test can be interpreted as a test for the C-CAPM against an alternative hypothesis that such factors are present.

### 15.4.1 The Time Separable Addilog Utility Function

Suppose that a representative consumer maximizes the lifetime utility function<sup>2</sup>

$$(15.14) \quad U = E_0 \left[ \sum_{t=0}^{\infty} \beta^t u_t(C_t) \right]$$

---

<sup>2</sup>The existence of a representative consumer under complete markets was discussed by Ogaki (1997) for the general concave utility functions and by Atkeson and Ogaki (1996) for the extended addilog utility function. Ogaki and Park (1997) discussed a sufficient condition for aggregation under incomplete markets for the cointegration approach.

subject to a life time budget constraint in complete markets at period 0, where  $\beta$  is a discount factor and  $C_t = (C_{1t}, C_{2t})$ . Here  $C_{it}$  is real consumption of the  $i$ -th good, and  $E_t(\cdot)$  denotes expectations conditional on the information available at period  $t$ . The intraperiod utility function is assumed to be of a monotone transformation of the addilog utility function:

$$(15.15) \quad u_t(C_t) = f_t\left(\sum_{i=1}^2 \sigma_{it} \frac{C_{it}^{1-\alpha_i} - 1}{1 - \alpha_i}\right).$$

where  $\alpha_i > 0$  for  $i = 1, 2$ . The stochastic process  $\{\sigma_{1t}, \sigma_{2t}\}$ , which is assumed to be (strictly) stationary, represents preference shocks. We refer to parameters  $\alpha_1$  and  $\alpha_2$  as curvature parameters. Nonseparability across goods is allowed by an arbitrary monotone transformation  $f_t$  with  $f'_t > 0$ .<sup>3</sup> This utility function includes Houthakker's (1960) addilog utility function and the Cobb-Douglas utility function ( $\alpha_1 = \alpha_2 = 1$ ) as special cases. When  $\alpha_1 \neq \alpha_2$ , preferences are not homothetic.

Since time separability is assumed, a two-stage budgeting scheme can be applied to show that the consumer maximizes his/her intraperiod utility (15.15) subject to the intraperiod budget constraint

$$(15.16) \quad P_{1t}C_{1t} + P_{2t}C_{2t} = E_t,$$

where  $E_t$  is the total consumption expenditure at period  $t$  and  $P_{it}$  is the price of the  $i$ -th good. Let the first good be the numeraire for each period ( $P_{1t} \equiv 1$ ).

The first order necessary conditions for the intraperiod optimization problem include

$$(15.17) \quad P_{2t} = \frac{\sigma_{2t} C_{2t}^{-\alpha_2}}{\sigma_{1t} C_{1t}^{-\alpha_1}}.$$

---

<sup>3</sup>Ogaki and Park (1992) showed that the cointegration approach allows for measurement errors, liquidity constraints, aggregation over heterogeneous consumers, and a general form of time-nonseparability in preferences. The present paper shows that the cointegration approach also allows for nonseparability across goods as long as time separability is assumed.

Since the first good is the numeraire,  $P_{2t} = \frac{P_{2t}}{P_{1t}}$  is the relative price between the second good and the first good. Taking the natural logarithm of both side of (15.17) yields

$$(15.18) \quad p_{2t} - \alpha_1 c_{1t} + \alpha_2 c_{2t} = \ln\left(\frac{\sigma_{2t}}{\sigma_{1t}}\right)$$

where  $p_{it} = \ln(P_{it})$ ,  $c_{it} = \ln(C_{it})$ . Thus the first order condition (15.17) implies a restriction that  $p_2(t) - \alpha_1 c_{1t} + \alpha_2 c_{2t}$  be stationary. We call this restriction the stationarity restriction.

The stationarity restriction summarizes the long-run implication from the demand side. In order to model the supply side in the simplest way, let us consider an endowment economy without production. Let  $C_{it}^*$  be the endowment of the  $i$ -th good and  $c_{it}^* = \ln(C_{it}^*)$ , so that  $c_{it} = c_{it}^*$  in an equilibrium. In a production economy, we require that equilibrium consumption satisfies the trend properties of  $c_i^*$  that we assume. The trend properties of equilibrium consumption of the  $i$ -th good is likely to be closely related to those of the technology shock to the  $i$ -th good industry in a production economy. The stationarity restriction comes from an assumption of stable preference shocks in the long-run. Preference parameters can be identified from the stationarity restriction if the supply side exhibits much more volatility in the long-run than the demand side. This can be done by assuming that at least one of  $c_{1t}^*$  and  $c_{2t}^*$  has a stochastic trend. Stable preferences and technological shocks with stochastic trends seem to be plausible assumptions for identification.<sup>4</sup>

First, let us consider the case where both  $c_{1t}^*$  and  $c_{2t}^*$  are difference stationary:

**Assumption 15.1a** The process  $\{c_{it}^* : t \geq 0\}$  is difference stationary for  $i = 1, 2$ .

---

<sup>4</sup>Ogaki (1988, 1989) showed that the Hansen and Singleton's (1982) GMM approach cannot be applied to the intraperiod first order condition of the addilog utility function when either  $c_{1t}^*$  or  $c_{2t}^*$  is difference stationary. For this reason, Ogaki (1988, 1989) assumed the trend stationarity for  $c_{1t}^*$  and  $c_{2t}^*$  to apply the GMM.

**Assumption 15.1b** The processes  $\{c_{1t}^* : t \geq 0\}$  and  $\{c_{2t}^* : t \geq 0\}$  are not stochastically cointegrated.

Assumption 15.1b will be satisfied for equilibrium consumption in a production economy if the technological shock in the second good industry has a different stochastic trend component from the technological shock in the food industry. Under assumption 15.1a and 15.1b, the stationarity restriction implies that the stochastic trends in  $(p_{2t}, c_{1t}, c_{2t})'$  are eliminated by a cointegrating vector  $(1, -\alpha_1, \alpha_2)$ . The stationarity restriction also implies that the cointegrating vector eliminates the deterministic trends in  $(p_{2t}, c_{1t}, c_{2t})'$ . Thus the deterministic cointegration restriction will be satisfied under assumption 15.1a and 15.1b.

Second, consider the case where the log of the endowment of one good is difference stationary and that of the other good is trend stationary. There are two cases depending on which good is assumed to be trend stationary.

**Assumption 15.2** The process  $\{c_{1t}^* : t \geq 0\}$  is difference stationary, and the process  $\{c_{2t}^* : t \geq 0\}$  is trend stationary with a nonzero trend.

**Assumption 15.2'** The process  $\{c_{1t}^* : t \geq 0\}$  is trend stationary with a nonzero trend, and the process  $\{c_{2t}^* : t \geq 0\}$  is difference stationary.

Assumption 15.2 or Assumption 15.2' will be satisfied for equilibrium consumption in a production economy if the technological shock in one good is trend stationary and the technological shock in the other good is difference stationary. For example, Costello (1990, chapter III) analyzed trend properties of Solow residuals of several industries and found some evidence that the Solow residual of the food industry is trend stationary and that of other industries is difference stationary.

Under assumption 15.2', the stationarity restriction implies that  $p_{2t}$  and  $c_{2t}$  are stochastically cointegrated with a cointegrating vector  $(1, \alpha_2)'$  and  $(p_{2t}, c_{1t}, c_{2t})'$  is cotrended with a cotrending vector  $(1, -\alpha_1, \alpha_2)'$ . The curvature parameters can be identified from these conditions.

### 15.4.2 The Time Nonseparable Addilog Utility Function

The intra-period utility function is assumed to be of the addilog form

$$(15.19) \quad u_t = \sum_{i=1}^n \sigma_{it} \frac{S_{it}^{1-\alpha_i} - 1}{1 - \alpha_i},$$

where  $\alpha_i > 0$  for  $i = 1, \dots, n$  and  $\sigma_i$ 's represent preference shocks. Here the stochastic process  $\{(\sigma_{1t}, \dots, \sigma_{nt})' : -\infty < t < \infty\}$  is assumed to be (strictly) stationary. This includes the case where some or all of  $\sigma_i$ 's are constant. When  $\alpha_i = 1$ , we interpret  $\frac{S_{it}^{1-\alpha_i} - 1}{1 - \alpha_i}$  to be  $\ln(S_{it})$ . Here  $S_{it}$  is the service flow from consumption purchases of good  $i$ . Purchases of consumption goods and service flows are related by

$$(15.20) \quad S_{it} = \{a_0^i C_{it} + a_1^i C_{i,t-1} + \dots + a_k^i C_{i,t-k}\} \exp(\theta_i^s t)$$

for  $i = 1, \dots, n$ , where  $C_{it}$  is real consumption expenditure for good  $i$  in period  $t$ . Following Eichenbaum and Hansen (1990), we allow for the possibility of technological progress in the transformation of purchases of good  $i$  into  $S_{it}$  in (15.20) via the exponential deterministic trend  $\exp(\theta_i^s t)$ . Below, we will consider the case in which the  $\theta_i^s$ 's are known to be zero as well as the case in which the  $\theta_i^s$ 's are unknown. Note that the purchase of one unit of good  $i$  at period  $t$  increases  $S_{i,t+\tau}$  by  $a_\tau^i \exp(\theta_i^s t)$  units for nonnegative  $\tau \leq k$ . This type of method of specifying time-nonseparability is used by Hayashi (1982), Eichenbaum, Hansen, and Singleton (1988), Eichenbaum and Hansen (1990), and Heaton (1995), among others.

In our empirical work, we take a measure of nondurable consumption as one good (say good 1) and interpret the curvature parameter for nondurable consumption ( $\alpha_1$ ) as the long-run intertemporal elasticity of substitution (IES) for the consumption of nondurables.<sup>5</sup> As we will discuss in Section 2.4????????????, this interpretation relies on the assumption of additive separability across the goods. It should be noted that this separability assumption is already made in Hansen and Singleton (1982) and Ferson and Constantinides (1991), both of which use the GMM approach and are closely related to this paper?????.

Masao  
needs to  
check this!

Masao  
needs to  
check this!

Let  $P_{it}$  be the purchase price of consumption good  $i$ . We take good 1 as a numeraire for each period:  $P_{1t} \equiv 1$ . The first order condition that equates the relative price between good  $i$  and good 1 ( $P_{it} = \frac{P_{it}}{P_{1t}}$ ) with the marginal rate of substitution of these goods is

$$\begin{aligned}
 (15.21) \quad P_{it} &= \frac{\partial U / \partial C_{it}}{\partial U / \partial C_{1t}} \\
 &= \frac{E_t[\sum_{\tau=0}^k \beta^\tau \frac{\partial u_{t+\tau}}{\partial C_{it}}]}{E_t[\sum_{\tau=0}^k \beta^\tau \frac{\partial u_{t+\tau}}{\partial C_{1t}}]} \\
 &= \frac{E_t[\sum_{\tau=0}^k \beta^\tau \sigma_{i,t+\tau} a_\tau^i \exp(\theta_{i,t+\tau}^s) \{S_{i,t+\tau}\}^{-\alpha_i}]}{E_t[\sum_{\tau=0}^k \beta^\tau \sigma_{1,t+\tau} a_\tau^1 \exp(\theta_{1,t+\tau}^s) \{S_{1,t+\tau}\}^{-\alpha_1}]}
 \end{aligned}$$

This first order condition forms the basis of the cointegration approach and summarizes the information needed from the demand side. In order to model the supply side in the simplest way, let us consider an endowment economy without production. Let  $C_{it}^*$  be the endowment of good  $i$  and  $c_{it}^* = \ln(C_{it}^*)$ . In equilibrium,  $c_{it} = \ln(C_{it}) = c_{it}^*$ . In a production economy, we require that equilibrium consumption satisfies the trend properties we assume for  $c_{it}^*$ . The trend properties of equilibrium consumption are

<sup>5</sup>This parameter is the long-run IES for nondurable consumption when we allow current and past consumption to adjust. When preferences are time nonseparable, the short-run IES is different from the long-run IES because we take past consumption to be fixed in the short-run.

likely to be closely related to those of the technology shock to the good  $i$  industry in a production economy.

We consider three alternative assumptions about the trend properties of  $C_{it}^*$ . In each of the three assumptions,  $\frac{C_{it}^*}{C_{i,t-1}^*}$  is stationary for all  $i$ . This assumption ensures that  $\frac{S_{it}}{C_{it} \exp(\theta_i^s t)}$  is stationary in equilibrium. To see this property, let  $S_{it}^*$  be the  $S_{it}$  implied by  $C_{it}^*$  and note that  $\frac{C_{i,t+\tau}^*}{C_{it}^*}$  is stationary for any fixed integer  $\tau$  because  $\frac{C_{i,t+\tau}^*}{C_{it}^*} = \frac{C_{i,t+\tau}^*}{C_{i,t+\tau-1}^*} \frac{C_{i,t+\tau-1}^*}{C_{i,t+\tau-2}^*} \dots \frac{C_{i,t+1}^*}{C_{it}^*}$ . It follows that the process  $\left\{ \frac{S_{i,t+\tau}^*}{C_{it}^* \exp(\theta_i^s t)} : -\infty < t < \infty \right\}$  is also stationary for any  $\tau$  because the right hand side of

$$(15.22) \quad \frac{S_{i,t+\tau}^*}{C_{it}^* \exp(\theta_i^s t)} = \left\{ a_0^i \frac{C_{i,t+\tau}^*}{C_{it}^*} + a_1^i \frac{C_{i,t+\tau-1}^*}{C_{it}^*} + \dots + a_k^i \frac{C_{i,t+\tau-k}^*}{C_{it}^*} \right\} \exp(\theta_i^s \tau)$$

is stationary. We also make an extra assumption that the expectation of a stationary variable conditional on the consumer's information set is equal to the expectation conditional on the stationary variables included in his information set. Then

$$\frac{P_{it} \exp(\theta_1^s t) [C_{1t}^* \exp(\theta_1^s t)]^{-\alpha_1}}{\exp(\theta_i^s t) [C_{it}^* \exp(\theta_i^s t)]^{-\alpha_i}}$$

is stationary because the right hand side of

$$(15.23) \quad \frac{P_{it} \exp(\theta_1^s t) [C_{1t}^* \exp(\theta_1^s t)]^{-\alpha_1}}{\exp(\theta_i^s t) [C_{it}^* \exp(\theta_i^s t)]^{-\alpha_i}} = \frac{E_t \left[ \sum_{\tau=0}^k \beta^\tau \sigma_{i,t+\tau} a_\tau^i \exp(\theta_i^s \tau) \left\{ \frac{S_{i,t+\tau}^*}{C_{it}^* \exp(\theta_i^s t)} \right\}^{-\alpha_i} \right]}{E_t \left[ \sum_{\tau=0}^k \beta^\tau \sigma_{1,t+\tau} a_\tau^1 \exp(\theta_1^s \tau) \left\{ \frac{S_{1,t+\tau}^*}{C_{1t}^* \exp(\theta_1^s t)} \right\}^{-\alpha_1} \right]}$$

is stationary. The right hand side of (15.23) is the ratio of conditional expectations of the functions of stationary variables.

Taking the natural log of the left hand side, we define  $z_t$  by

$$(15.24) \quad z_t = p_{it} - \alpha_1 c_{1t}^* + \alpha_i c_{it}^* + (1 - \alpha_1) \theta_1^s t - (1 - \alpha_i) \theta_i^s t$$

where  $p_{it} = \ln(P_{it})$ ,  $c_{it}^* = \ln(C_{it}^*)$  for  $i = 1, \dots, n$  and conclude that  $z_t$  is stationary. We shall call this restriction as the stationary restriction. This restriction implies that  $p_{it} - \alpha_1 c_{1t}^* + \alpha_i c_{it}^*$  ( $= z_t - (1 - \alpha_1)\theta_1^s t + (1 - \alpha_i)\theta_i^s t$ ) is trend stationary in general, and is stationary if and only if  $(1 - \alpha_1)\theta_1^s - (1 - \alpha_i)\theta_i^s = 0$ .

In this section, we study the implications of the stationarity restriction. We consider only the pair of good 1 and good 2 since our results generalize to any pair of goods. The stationarity restriction is a result of the assumption of the long-run stability of preferences. Preference parameters can be identified from the stationarity restriction if the supply side is substantially more volatile than the demand side in the long-run. This condition requires the assumption that at least one of  $c_{1t}^*$  and  $c_{2t}^*$  has a stochastic trend.<sup>6</sup> Stable preferences and technological shocks with stochastic trends seem to be plausible assumptions for identification.

First, consider the case in which both  $c_{1t}^*$  and  $c_{2t}^*$  are difference stationary:<sup>7</sup>

**Assumption 15.3a** The process  $\{c_{it}^* : t \geq 0\}$  is difference stationary for  $i = 1, 2$ .

**Assumption 15.3b** The processes  $\{c_{1t}^* : t \geq 0\}$  and  $\{c_{2t}^* : t \geq 0\}$  are not stochastically cointegrated.

Assumption 15.3b will be satisfied for equilibrium consumption in a production economy if the technological shock in the good 1 industry has a different stochastic trend component from the technological shock in the good 2 industry. Under assumption 15.3a and 15.3b, the stationarity restriction implies that  $p_{2t} - \alpha_1 c_{1t}^* + \alpha_2 c_{2t}^*$  is trend stationary. Thus  $(p_{2t}, c_{1t}^*, c_{2t}^*)'$  is stochastically cointegrated with a cointegrating vector

<sup>6</sup>Ogaki (1988) develops an econometric method based on GMM which uses the information in deterministic trends to estimate the preference parameters of the addilog utility function when both of  $c_{1t}^*$  and  $c_{2t}^*$  are trend stationary.

<sup>7</sup>A special case is that  $c_{1t}^*$  and  $c_{2t}^*$  are martingale when the real interest rate is constant and  $C_{it}^*$  is lognormally distributed.

$(1, -\alpha_1, \alpha_2)'$ . However, the deterministic cointegration restriction is not necessarily satisfied under assumption 15.3a and 15.3b. The stationarity restriction implies that  $p_{2t} - \alpha_1 c_{1t}^* + \alpha_2 c_{2t}^*$  is stationary under the condition that there is no technological progress in the transformation technology from consumption purchases to service flows (namely,  $\theta_i^s = 0$  for  $i = 1, 2$ ). Hence, consider the following assumption:

**Assumption 15.4** Assumption 15.3a and 15.3b are satisfied and  $\theta_i^s = 0$  for  $i = 1, 2$ .

Under assumption 15.4,  $(p_{2t}, c_{1t}^*, c_{2t}^*)'$  is stochastically cointegrated with a cointegrating vector  $(1, -\alpha_1, \alpha_2)'$  and satisfies the deterministic cointegration restriction.

Second, consider the case where the log of the endowment of good 1 is difference stationary and that of good 2 is trend stationary:

**Assumption 15.5a** The process  $\{c_{1t}^* : t \geq 0\}$  is difference stationary and the process  $\{c_{2t}^* : t \geq 0\}$  is trend stationary with a nonzero linear trend.

**Assumption 15.5b**  $\theta_i^s = 0$  for  $i = 1, 2$ .

Assumption 15.5a will be satisfied for equilibrium consumption in a production economy if the technological shock in the good 1 industry is difference stationary and the technological shock in the good 2 industry is trend stationary. Under assumption 15.5a, the stationarity restriction implies that  $p_{2t}$  and  $c_{1t}^*$  are stochastically cointegrated with a cointegrating vector  $(1, -\alpha_1)'$ . Assumption 15.5a is enough to identify  $\alpha_1$ . In order to identify  $\alpha_2$  as well as  $\alpha_1$ , we need assumption 15.5b. Under assumption 15.5a and 15.5b, the stationarity restriction implies that  $(p_{2t}, c_{1t}^*, c_{2t}^*)'$  is cotrended with a cotrending vector  $(1, -\alpha_1, \alpha_2)'$ .

### 15.4.3 Engel's Law and Cointegration

The key assumption for the cointegration approach to estimating preference parameters is that preferences are stable over time. Ogaki (1992) tests this assumption by comparing total expenditure elasticities (income elasticities in the context of the static models) estimated from time series data obtained by the cointegration approach with those estimated from household level cross-sectional data. The nonhomotheticity of preferences studied by Ogaki (1992) also important implications on intertemporal consumption decisions as in Atkeson and Ogaki (1996).

In cross sectional data, it is widely observed that a higher share of total expenditure goes to food for poorer households than is the case for richer households. A time series counterpart of this observation, Engel's law, is that the expenditure share on food declines as the economy grows. Ogaki (1992) tests if Houthakker's (1960) addilog utility function can explain both of these cross sectional and time series observations simultaneously. The cointegration approach is used to estimate parameters of the addilog utility function governing total expenditure elasticities of demand from time series data. Information in stochastic and deterministic trends is exploited in this approach.

Define  $\mu = \frac{\partial \ln(C_{1t})}{\partial \ln(E(t))}$  as the total expenditure elasticity of demand for the first good, using the intraperiod optimization problem. It can be shown that the addilog utility function implies that the expenditure elasticity of demand for the first good is

$$(15.25) \quad \mu = \left\{ \frac{\alpha_1}{\alpha_2} + \omega_{1t} \left( 1 - \frac{\alpha_1}{\alpha_2} \right) \right\}^{-1},$$

where  $\omega_{it} = \frac{P_{it}C_{it}}{E_t}$  is the budget share of the  $i$ -th good. Thus the expenditure elasticity for given levels of  $E_t$ ,  $C_{1t}$ , and  $P_{2t}$  can be estimated once  $\frac{\alpha_1}{\alpha_2}$  is estimated.

Comparing the expenditure elasticities implied by the addilog utility function estimated from the cointegration approach and the estimates of the elasticities estimated from cross-sectional household data provides a tests for the cointegration approach. The crucial assumption in the cointegration approach is that preferences are stable relative to the trends in equilibrium consumption expenditures. The most important factor that could cause problems with this assumption would probably be trending demographic changes. If this factor causes important problems, then the cointegration approach estimates from aggregate time series data will differ from the estimates from cross-sectional data.

Ogaki (1992) shows that the cointegration approach estimates of the expenditure elasticities from U.S. aggregate time series data are consistent with those from cross-sectional household data for food, clothing, household operation, and transportation. These empirical results support the assumption of stable preferences.

It should be noted that the expenditure elasticity is not constant. Suppose that  $\alpha_1 > \alpha_2$ , so that the first good is a necessary good. For very poor consumers,  $\omega_{it}$  is close to one, and the elasticity is equal to one. For very rich consumers,  $\omega_{it}$  is close to zero, and the elasticity is equal to  $\frac{\alpha_2}{\alpha_1}$ . When the relative price is constant,  $\omega_{it}$  falls from one to zero as a consumer becomes richer, and the expenditure elasticity falls from one to  $\frac{\alpha_2}{\alpha_1}$ .

International comparison of elasticities is also of interest. Houthakker (1957) finds some tendency for the expenditure elasticity of the demand for food to be higher in low income countries than in high income countries. However, it seems important to allow for subsistence levels for low income countries. Atkeson and Ogaki (1996) estimate the extended addilog utility function, which generalizes the addilog utility

function by allowing for subsistence levels:

$$(15.26) \quad u(C) = \sum_{i=1}^n \frac{\theta_i}{1 - \alpha_i} [(C_i - \gamma_i)^{1 - \alpha_i} - 1]$$

where  $\alpha_i > 0$  and  $\theta_i > 0$  for  $i = 1, \dots, n$ . We refer to the parameters  $\gamma_i$  as subsistence parameters and the parameters  $\alpha_i$  as curvature parameters. This utility function contains as special cases two utility functions commonly used in demand studies. If  $\alpha_i = 1$  for  $i = 1, \dots, n$ , then this utility function yields the linear expenditure system in that the intratemporal demand functions for consumption of each good in excess of subsistence consumption are linear in expenditure in excess of subsistence expenditure. More generally, if  $\alpha_1 = \alpha_2 = \dots = \alpha_n$ , then these preferences are quasi-homothetic. If  $\gamma_i = 0$  for  $i = 1, \dots, n$ , then this utility function is Houthakker's (1960) addilog utility function.

Atkeson and Ogaki (1996) discuss technical difficulties in estimating fixed subsistence levels from nonstationary time series data, and estimate them from Indian household panel data. The cointegration approach is applied to estimate the curvature parameters in Indian and U.S. aggregate time series data after factoring the estimated subsistence levels. They find little evidence against the hypothesis that preferences are identical for Indian and U.S. households when we maintain the hypothesis that the subsistence levels are the same for the two countries.

Houthakker (1957) finds that the expenditure elasticity of the demand for food is much lower for the typical Indian household than for the typical U.S. households in cross-sectional data. This finding can be consistent with identical preferences for Indian and U.S. households because the extended addilog utility function implies that the total expenditure elasticity of the demand for food will be different for rich and poorer households. Ogaki (1992) reports that the extended addilog utility function

estimated by Atkeson and Ogaki (1996) explains the ratio of Houthakker's estimates of the elasticities for India and United States.

## 15.5 The Cointegration-Euler Equation Approach

This section explains Cooley and Ogaki's (1996) cointegration-Euler Equation approach, which combines the cointegration approach to estimating preference parameters with Hansen and Singleton's (1982) Euler equation approach based on GMM. In the first step of this approach, a cointegrating regression is applied to an intratemporal first order condition for the household's maximization problem to estimate some preference parameters. In the second step, GMM is applied to an Euler equation after plugging in point estimates from the cointegrating regression in the first step. Since the first step estimators are super consistent, asymptotic properties of the GMM estimators in the second step are not affected by the first step estimation.

This section explains Cooley and Ogaki's application of the approach on the consumption-leisure choice model for time nonseparable preferences that are additively separable for consumption and leisure. The next section explains Ogaki and Reinhart's (1998) application to estimate the intertemporal elasticity of substitution when preferences are nonseparable over nondurable and durable goods.

Cooley and Ogaki reexamine whether the time series properties of aggregate consumption, real wages, and asset returns are consistent with a simple neoclassical representative agent economy. Previous empirical explorations of this issue have rejected the neoclassical model in large part because the marginal rate of substitution between consumption and leisure does not equal the real wage as is implied by the first order conditions of the model. They argue that an optimal labor contract

model is more appropriate for understanding the time series behavior of real wages and consumption. They show that a version of the optimal contract model restricts the long-run relation between real wages and consumption. They exploit this long-run restriction (cointegration restriction) to estimate preference parameters and test the model. First, they employ the cointegration approach to estimate the long-run intertemporal elasticity of substitution for nondurable consumption from a cointegrating regression. They test the model by testing for the cointegration restriction.

As further analysis, they use this estimated preference parameter in the asset pricing equation implied by this economy to estimate the discount factor and a coefficient of time-nonseparability using Hansen's (1982) Generalized Method of Moments (GMM). From this they are able to construct another specification test of the model.

Mankiw, Rotemberg, and Summers (1985, hereafter Mankiw *et al.*) subjected the Euler equations of an intertemporal labor supply model to a battery of tests and found no evidence to support it. Not only did their formal tests reject the model, but their point estimates of preference parameters implied a convex utility function. They concluded that the observed "... economic fluctuations do not easily admit of a neoclassical interpretation."

Eichenbaum, Hansen, and Singleton (1988, hereafter Eichenbaum *et al.*) also used the Euler equation approach, but their point estimates of preference parameters were more reasonable. They attributed their different finding to two factors. First, they removed trends by taking growth rates of variables and taking ratios of variables while Mankiw *et al.* did not. Second, Eichenbaum *et al.* allowed time-nonseparability of preferences. Though their point estimates were reasonable, their formal test statistics typically rejected the model at the one percent level when they tested both asset

pricing equations and the first order condition that equates the real wage with the marginal rate of substitution between leisure and consumption. When they removed the first order condition and tested the asset pricing equations, their tests did not reject the model. However, the loss of precision of their estimates was substantial when the first order condition was removed. Eichenbaum *et al.* interpreted their results as suggesting that the optimal labor contract model might be appropriate for understanding real wages.<sup>8</sup>

A given Pareto optimal allocation can be consistent with a wide variety of institutional arrangements. In optimal labor contract models (see, e.g. Azariadis, 1975; Rosen, 1985; Wright, 1988), labor income contains a component that provides workers with some degree of protection against business cycle fluctuations (also see Hall, 1980). This insurance component of labor income inserts a wedge between the marginal rate of substitution between leisure and consumption and wages. In their empirical work, Gomme and Greenwood (1995) showed that accounting for this component could help explain the observed pattern of fluctuations in income. These arguments combined with the findings of Eichenbaum *et al.* suggest that the imposition of the requirement that wages equal the marginal rate of substitution between consumption and leisure is too confining.

Cooley and Ogaki use a restriction on the time series properties of real wages and consumption that is implied by optimal labor contract to estimate preference parameters and test the model. In the optimal contract model, the first order condition for real wages and consumption does not hold on a period-by-period basis.

---

<sup>8</sup>Osano and Inoue (1991) used an approach similar to Eichenbaum *et al.* to test the overidentifying restrictions of Euler equations, using aggregate Japanese data. They also noted that there was much less evidence against the model when they removed the Euler equation associated with the equation of real wages and the marginal product of labor.

They show, however, that a version of the optimal contract model implies that the real wage rate is equated with the marginal rate of substitution between consumption and labor in the long-run. They exploit this long-run restriction for estimating and testing the model.

In contrast to the research cited above, the cointegration approach yields results that are supportive of the representative agent model. In the first step of our econometric procedure, we test the null hypothesis of cointegration and estimate the long-run IES for three measures of nondurable consumption. Cooley and Ogaki do not reject the null of cointegration and obtain reasonable estimates. The long-run IES appears in the asset pricing equation derived from the representative consumer model. Cooley and Ogaki use the estimated IES parameter from the cointegrating regression in the first step in the asset pricing equation and apply GMM to estimate the discount parameter and a coefficient of time-nonseparability. They use both stock and nominal risk free returns. They form a specification test *a la* Hausman (1978) through these steps. This specification test does not reject the model.

### 15.5.1 The Economy

We consider an economy populated by  $N$  households who have preferences defined over consumption and the flow of services from their leisure time. Each household maximizes

$$(15.27) \quad U = E_0 \left[ \sum_{t=0}^{\infty} \beta^t u_t \right]$$

where  $E_t$  denotes the expectation conditioned on the information available at  $t$ . In order to develop intuition, let us first consider a simple intraperiod utility function that is assumed to be time- and state-separable and separable in nondurable consumption,

durable consumption, and leisure

$$(15.28) \quad u_t = \frac{C_t^{1-\alpha} - 1}{1-\alpha} + v(l_t)$$

where  $v(\cdot)$  is a continuously differentiable concave function,  $C_t$  is nondurable consumption, and  $l_t$  is leisure.

For now, assume that real wages do not contain any insurance component. Then the usual first order condition for a household that equates the real wage rate with the marginal rate of substitution between leisure and consumption is:

$$(15.29) \quad W_t = \frac{v'(l_t)}{C_t^{-\alpha}}$$

where  $W_t$  is the real wage rate. We assume that the stochastic process of leisure is (strictly) stationary in the equilibrium as in Eichenbaum, Hansen, and Singleton (1988) and that the random variables used to form the conditional expectations for stationary variables are stationary. Then an implication of the first order condition is that  $\ln(W_t) - \alpha \ln(C_t) = \ln(v'(l_t))$  is stationary. When we assume that the log of consumption is difference stationary, this assumption implies that the log of the real wage rate and the log of consumption are cointegrated with a cointegrating vector  $(1, -\alpha)'$ . We exploit this cointegration restriction to identify the curvature parameter  $\alpha$  from cointegrating regressions.

Given that the saving rate is stable in the long-run in the U.S. (as Kuznets, 1946, found), it is natural to impose a restriction that the ratio of total consumption expenditure and labor income is stable at least when a consumer is rich enough. Since we assume that consumption and leisure are additively separable in intertemporal preferences, this restriction implies that  $\alpha$  is equal to one when total consumption expenditure is used as  $C_t$  in our model. In our empirical work, we use a measure

of nondurable consumption as  $C_t$ , assuming that the other consumption goods (say, durable consumption goods) are additively separable from the measure of nondurable consumption good used in our analysis. For this reason,  $\alpha$  can be different from one even when the saving rate is stationary for rich enough consumers.<sup>9</sup>

Masao  
needs to  
check this!

We now introduce time-nonseparability of preferences. The intraperiod utility function is assumed to be

$$(15.30) \quad u_t = \frac{S_t^{1-\alpha} - 1}{1 - \alpha} + v(l_t, l_{t-1}, \dots, l_{t-k}),$$

where  $S_t$  is the service flow from nondurable consumption:

$$(15.31) \quad S_t = C_t + \lambda C_{t-1}.$$

This type of time nonseparable specification of leisure has been used by many authors and is useful because it can capture the fact that households may use leisure time in a household production technology to augment a stock of household capital (Kydland, 1984; Greenwood and Hercowitz, 1991; Benhabib, Rogerson, and Wright, 1991).

The time-nonseparable specification for nondurable consumption is similar to that considered by Eichenbaum, Hansen, and Singleton (1988), Eichenbaum and Hansen (1990), Constantinides (1990), Heaton (1993, 1995), Allen (1992), and Braun, Constantinedes, and Ferson (1993) among others, except that some of these authors considered a more general form of time-nonseparability for nondurable consumption than (15.31). We have habit formation for nondurable consumption when  $\lambda$  is neg-

---

<sup>9</sup>Since many economic models imply known cointegrating vectors when the log of the variables are taken and because an attractive feature of cointegration is that unknown parameters can be estimated without exogeneity assumptions, the fact that  $\alpha$  is unknown in the model is important. For this reason, this point that  $\alpha$  can be different from one is explained in details in the Appendix.????????????????

ative and local substitutability or durability when  $\lambda$  is positive.<sup>10</sup> Note that the time-nonseparability does not affect the IES in the long-run when  $C_t$  and  $C_{t-1}$  are equal.<sup>11</sup> We will refer to  $\frac{1}{\alpha}$  as the long-run IES for nondurable consumption.

The usual first order condition for a household that equates real wage rate with the marginal rate of substitution between leisure and consumption is now:

$$\begin{aligned}
 (15.32) \quad W_t &= \frac{\partial U / \partial l_t}{\partial U / \partial C_t} \\
 &= \frac{E_t[\sum_{\tau=0}^K \beta^\tau \frac{\partial u_{t+\tau}}{\partial l_t}]}{E_t[\frac{\partial u_t}{\partial C_t} + \frac{\partial u_{t+1}}{\partial C_t}]} \\
 &= \frac{E_t[\sum_{\tau=0}^K \beta^\tau \frac{\partial v_{t+\tau}}{\partial l_t}]}{E_t[S_t^{-\alpha} + \beta \lambda S_{t+1}^{-\alpha}]}.
 \end{aligned}$$

We assume that  $\ln(C_t)$  is difference stationary in the equilibrium. Then

$$(15.33) \quad \frac{S_{t+\tau}}{C_t} = \frac{C_{t+\tau}}{C_t} + \lambda \frac{C_{t+\tau-1}}{C_t}$$

is stationary for any  $\tau$ . Combined with the first order condition (15.32), it follows that

$$(15.34) \quad W_t C_t^{-\alpha} = \frac{E_t[\sum_{\tau=0}^K \beta^\tau \frac{\partial v_{t+\tau}}{\partial l_t}]}{E_t[\{\frac{S_t}{C_t}\}^{-\alpha} + \beta \lambda \{\frac{S_{t+1}}{C_t}\}^{-\alpha}]}$$

is stationary. Taking logs,  $\ln(W_t) - \alpha \ln(C_t)$  is stationary as in the time-separable case we discussed.

In Cooley and Ogaki's empirical work, they estimate and test the first order condition (15.32) through the cointegration restriction for aggregated real wages and consumption. They also estimate and test the standard asset pricing equation for the

<sup>10</sup>The time-nonseparability for nondurable consumption allows us to separate the IES in the short-run and the reciprocal of the RRA coefficient as Constantinides (1990) described, which could help explain the equity premium puzzle of Mehra and Prescott (1985). Ferson and Constantinides (1991) found evidence in favor of the asset pricing model with habit formation, using GMM.

<sup>11</sup>Alternatively,  $C_t$  grows at a constant rate in the long-run.

time-nonseparable utility function

$$(15.35) \quad \frac{E_t[\beta\{S_{t+1}^{-\alpha} + \beta\lambda S_{t+2}^{-\alpha}\}R_{t+1}]}{E_t[S_t^{-\alpha} + \beta\lambda S_{t+1}^{-\alpha}]} = 1$$

for any gross asset return  $R_t$ .

In optimal labor contract models, labor income contains a component that provides workers with some degree of protection against business cycle fluctuations. This insurance component of labor income inserts a wedge between the marginal rate of substitution between leisure and consumption and wages. To utilize information in the first order condition (15.32) for estimation and testing, we start from the observation that the cointegration restriction is robust as long as the measured wage rate has the same trend as the marginal rate of substitution. Even when there is a wedge between the real wage rate and the marginal rate of substitution, the stationary restriction holds as long as the insurance component does not have (stochastic or deterministic) trends. Intuition suggests that the fraction of the insurance component in the wage rate is likely to be stationary rather than trending. Cooley and Ogaki formalize this intuition by considering a version of an optimal contract model.

### 15.5.2 The 2-Step Estimation Method

In the first step, a cointegration regression is used to estimate  $\alpha$  from the stationarity restriction. Since the log real wage rate and log consumption are cointegrated, either variable can be used as a regressand. In finite samples, the empirical results will be different depending on the choice of the regressand. However, the results should be approximately the same as long as cointegration holds and the sample size is large enough.

The econometric model for our GMM procedure is based on the asset pricing

equation (15.35), which implies  $E_t(\epsilon_{gt}^0) = 0$ , where

$$(15.36) \quad \epsilon_{gt}^0 = \beta[(C_{a,t+1} + \lambda C_{a,t})^{-\alpha} + \lambda\beta(C_{a,t+2} + \lambda C_{a,t+1})^{-\alpha}]R_{t+1} \\ - [(C_{a,t} + \lambda C_{a,t-1})^{-\alpha} + \lambda\beta(C_{a,t+1} + \lambda C_{a,t})^{-\alpha}]$$

where  $C_a$  indicates aggregate nondurable consumption. We define

$$\epsilon_{gt} = \frac{\epsilon_{gt}^0 g(\lambda)}{(1 + \beta\lambda)\{C_{a,t} + \lambda C_{a,t-1}\}^{-\alpha}}$$

where  $g(\lambda) = 1$  if  $\lambda \leq 1$  and  $g(\lambda) = 1 + (\lambda - 1)^2$  if  $\lambda > 1$ . We use  $\epsilon_{gt}$  as the disturbance for the GMM estimation. Since the scale factor  $\frac{g(\lambda)}{(1 + \beta\lambda)\{C_{a,t} + \lambda C_{a,t-1}\}^{-\alpha}}$  is in the information available at  $t$ ,  $E_t(\epsilon_{gt}) = 0$ . We scale the disturbance to achieve stationarity required for the GMM,<sup>12</sup> to avoid the trivial solutions that cause an identification problem, and to incorporate the prior information that  $\lambda$  is likely to be smaller than one in absolute value.<sup>13</sup> Even though the asymptotic theory justifies this type of scaling, small sample properties of the GMM estimator are affected by the choice of the scaling factor. For this reason, the  $g(\lambda)$  function is designed not to affect the disturbance when  $\lambda \leq 1$ : we have little prior information about which admissible value of  $\lambda$  is more plausible when the absolute value of  $\lambda$  is less than one. The disturbance term is MA of order one because of the time-nonseparable specification. The weighting matrix for the GMM estimation must take account of the serial correlation.

A formal test statistic can be formed by using the estimate of  $\alpha$  from the cointegrating regression in the GMM procedure to obtain restricted estimates. In this

<sup>12</sup>The stationarity assumption of the GMM can be relaxed to some extent, but unit-root nonstationarity is not allowed. Hence the stationary inducing transformation is necessary for our model.

<sup>13</sup>Certain values of  $\lambda$  are not admissible because  $C_{a,t} + \lambda C_{a,t-1}$  cannot be negative. In order to exclude these values in the GMM nonlinear search, a very large positive number was returned as  $\epsilon_{gt}$  when they are tried. The numerical derivative program was modified accordingly. See Ogaki (1993b) for details.

restricted GMM estimation, we estimate only  $\beta$  and  $\lambda$ . We use the same weighting matrix to form unrestricted estimates. We then take the difference of Hansen's (1982) chi-square test (Hansen's  $J_T$  test) statistic for the overidentifying restrictions from the restricted estimation and that from the unrestricted estimation, in which  $\beta, \lambda$ , and  $\alpha$  are estimated. The difference is the likelihood ratio type test (denoted by  $C_T$ ), which has an asymptotic chi-square distribution with one degree of freedom.<sup>14</sup> This two step procedure does not alter the asymptotic distribution of GMM estimators and test statistics because our cointegrating regression estimator is super consistent and converges at a rate faster than  $\sqrt{T}$ .

### 15.5.3 Measuring Intertemporal Substitution: The Role of Durable Goods

Masao  
needs to  
check this!

[To Be added??????]

## 15.6 Purchasing Power Parity

Masao  
needs to  
check this!

[?????? This section is incomplete]

Assume that there is only one good in the world economy, and that the law of one price for the world economy (called Purchasing Power Parity (PPP)) holds at each point in time. Let  $P_t$  be the domestic price of the good, and  $P_t^F$  be the foreign price of the good at  $t$ . Define the *real exchange rate* as

$$(15.37) \quad S_t^r = \frac{S_t P_t^F}{P_t}$$

When the good is measured with the same unit in the two countries, PPP implies that the real exchange rate is equal to one. This version of PPP is called *absolute* PPP.

<sup>14</sup>See, e.g., Ogaki (1993a) for an explanation of the likelihood ratio type test for GMM.

When the good is measured with different units in the two countries, PPP implies that the real exchange rate is constant. This version of PPP is called *relative* PPP.

Two cases are worth noting. First, if infinitely many stationary random variables are involved in an economic model, it is often possible to show that an infinite sum of a series of random variables (or vectors) converges to a stationary random variable (or vector). Then it is possible to use Proposition ???.

Masao  
needs to  
check this!

## Exercises

**15.1** Suppose that a representative consumer maximizes the life time utility function

$$(15.E.1) \quad U = E_0 \sum_{t=0}^{\infty} \beta^t u_t$$

at time 0, where  $E_t(\cdot)$  denotes expectations conditional on the information available at time  $t$ ,  $I_t$ , subject to a life time budget constraint in an Arrow-Debreu economy with two goods. The intra-period utility function is assumed to be

$$(15.E.2) \quad u_t = \frac{C_{1t}^{1-\alpha_1} - 1}{1 - \alpha_1} + \sigma_2 \frac{S_{2t}^{1-\alpha_2} - 1}{1 - \alpha_2}$$

where  $\alpha_i > 0$  for  $i = 1, 2$  and

$$(15.E.3) \quad S_{2t} = e^{\theta t} (C_{2t} + \delta C_{2,t-1})$$

is service flow from purchases of the second consumption good. Let  $P_{2t}$  be the purchasing price of the second good in terms of the first good and  $R_t$  be the ex post gross rate of return of an asset in terms of the second good. Assume that  $\{\frac{C_{i,t+1}}{C_{it}}\}$  is stationary for  $i = 1, 2$ .

- (a) Write down the parametric form of the first order condition that  $p_{2t}, C_{1t}, C_{2,t-1}, C_{2t}$  and  $C_{2,t+1}$  should satisfy in an equilibrium. Explain your answer.

- (b) Show that  $\ln \frac{S_{2t}}{C_{2t}}$  is trend stationary.
- (c) Give the definitions of stochastic cointegration and the deterministic cointegration restriction. In each of the following cases, which variables are stochastically cointegrated? Give a cointegrating vector and explain whether or not the deterministic cointegration restriction is satisfied for these variables that are stochastically cointegrated. Explain your answers.

Case 1:  $\theta = 0$  and  $\ln C_{it}$  is difference stationary for  $i = 1, 2$ .

Case 2:  $\theta \neq 0$  and  $\ln C_{it}$  is difference stationary for  $i = 1, 2$ .

Case 3:  $\theta = 0$  and  $\ln C_{1t}$  is difference stationary, and  $\ln C_{2t}$  is stationary.

Case 4:  $\theta \neq 0$  and  $\ln C_{1t}$  is difference stationary, and  $\ln C_{2t}$  is stationary.

Case 5:  $\theta = 0$  and  $\ln C_{1t}$  is difference stationary, and  $\ln C_{2t}$  is trend stationary with a nonzero linear trend.

Case 6:  $\theta \neq 0$  and  $\ln C_{1t}$  is difference stationary, and  $\ln C_{2t}$  is trend stationary with a nonzero linear trend.

**15.2** Take nondurables as the first good and durables in the national account as the second good in Ogaki's (1992) model. To obtain per equivalent adult consumption, place an weight of 1 for the civilian noninstitutional population with ages 16 and over, and 0.55 on the rest of the total population. The consumption data are in QNRND91.DAT(nondurables) and QNRD91.DAT(durables). These files also include data descriptions in detail. These quarterly data files contain the current dollar consumption in the first column and the 1987 dollar consumption in the second column over the period 1947:1-1993:4. The population data are in MPOP92.DAT which contains the total population in the first column and the total civilian noninstitutional

population with ages 16 and over in the second column. This monthly data covers 1947:1-1992:1. Take the quarterly average of the equivalent adult population. Use the sample period 1947:2-1989:4. All the files necessary for this exercise are in <http://economics.sbs.ohio-state.edu/ogaki>. You can modify and rename \*.EXP file for each of the \*.SET file mentioned in the problems. Imagine that you were reporting your empirical results in a section of a paper to be published in a professional journal. Report results in tables and explain purposes of tests and your results.

- (a) Report  $G(1, q)$  test statistics with  $q = 2, 3$  for  $\ln C_{1t}$  and  $\ln C_{2t}$  with non-prewhitened QS kernel. Use GPQ.SET.
- (b) Report augmented Dickey-Fuller (Said-Dickey) test statistics for  $\ln C_{1t}$  and  $\ln C_{2t}$ .
- (c) Report the third stage CCR estimators for preference parameters with  $\ln C_{1t}$  as the regressand. Also report  $H(0, 1)$ ,  $H(1, q)$  test statistics with  $q = 2, 3$ , and Wald test statistics for the null hypothesis  $\alpha_1 = \alpha_2 = 1$  from the fourth stage CCR with the singular values for the prewhitening VAR matrix bounded by 0.99 and the automatic bandwidth parameter bounded by  $\sqrt{T}$  where  $T$  is the sample size. Use CCR.SET.

### References

- ALLEN, E. R. (1992): "Cross Sectional and Time Series Measures of Asset Pricing Model Fit," Manuscript, University of Huston.
- AMANO, R. A., W.-M. HO, AND T. S. WIRJANTO (1998): "Intraperiod and Intertemporal Substitution in Import Demand," Bank of Canada, CREFE Working paper No. 84.
- AMANO, R. A., AND T. S. WIRJANTO (1996): "Intertemporal Substitution, Imports, and the Permanent-Income Model," *Journal of International Economics*, 40, 439–457.
- (1997): "Intratemporal Substitution and Government Spending," *Review of Economics and Statistics*, 40, 605–609.

- ATKESON, A., AND M. OGAKI (1996): "Wealth-Varying Intertemporal Elasticities of Substitution: Evidence from Panel and Aggregate Data," *Journal of Monetary Economics*, 38, 507–534.
- AZARIADIS, C. (1975): "Implicit Contracts and Unemployment Equilibria," *Journal of Political Economy*, 83, 1183–1202.
- BALL, L. (1990): "Intertemporal Substitution and Constraints on Labor Supply - Evidence from Panel Data," *Economic Inquiry*, 28(4), 706–724.
- BENHABIB, J., R. ROGERSON, AND R. WRIGHT (1991): "Homework in Macroeconomics: Household Production and Aggregate Fluctuations," *Journal of Political Economy*, 99, 1166–1187.
- BRAUN, P. A., G. M. CONSTANTINEDES, AND W. E. FERSON (1993): "Time Nonseparability in Aggregate Consumption: International Evidence," *European Economic Review*, 37, 897–920.
- BRAUN, R. A. (1994): "How Large is the Optimal Inflation Tax?," *Journal of Monetary Economics*, 34, 201–214.
- CAMPBELL, J. Y., AND R. J. SHILLER (1987): "Cointegration and Tests of Present Value Models," *Journal of Political Economy*, 95(5), 1062–1088.
- CLARIDA, R. H. (1994): "Cointegration, Aggregate Consumption, and the Demand for Imports: A Structural Econometric Investigation," *American Economic Review*, 84(1), 298–308.
- (1996): "Consumption, Import Prices, and the Demand for Imported Consumer Durables: A Structural Econometric Investigation," *Review of Economics and Statistics*, 78, 369–374.
- COCHRANE, J. H. (1998): "What Do the VARs Mean? - Measuring the Output Effects of Monetary Policy," *Journal of Monetary Economics*, 41, 277–300.
- COCHRANE, J. H., AND A. M. SBORDONE (1988): "Multivariate Estimates of the Permanent Components of GNP and Stock Prices," *Journal of Economic Dynamics and Control*, 12, 255–296.
- CONSTANTINIDES, G. M. (1990): "Habit Formation: A Resolution of the Equity Premium Puzzle," *Journal of Political Economy*, 98, 519–543.
- COOLEY, T. F., AND M. OGAKI (1996): "A Time Series Analysis of Real Wages, Consumption, and Asset Returns," *Journal of Applied Econometrics*, 11(2), 119–134.
- COSTELLO, D. M. (1990): "Productivity Growth, the Transfer of Technology and International Business Cycles," Ph.D dissertation, University of Rochester.
- DEATON, A. (1991): "Saving and Liquidity Constraints," *Econometrica*, 59(5), 1221–1248.
- DEATON, A. S. (1974): "The Analysis of Consumer Demand in the United Kingdom, 1900–1970," *Econometrica*, 42, 341–367.
- DEATON, A. S., AND K. K. WIGLEY (1971): "Econometric Models for the Personal Sector," *Bulletin of the Oxford Institute of Statistics*, 33.
- EICHENBAUM, M., AND L. P. HANSEN (1990): "Estimating Models with Intertemporal Substitution Using Aggregate Time Series Data," *Journal of Business and Economic Statistics*, 8, 53–69.

- EICHENBAUM, M., L. P. HANSEN, AND K. J. SINGLETON (1988): "A Time Series Analysis of Representative Agent Models of Consumption and Leisure Choice under Uncertainty," *Quarterly Journal of Economics*, 103, 51–78.
- FERSON, W. E., AND G. M. CONSTANTINIDES (1991): "Habit Persistence and Durability in Aggregate Consumption: Empirical Tests," *Journal of Financial Economics*, 29(2), 199–240.
- FERSON, W. E., AND C. R. HARVEY (1992): "Seasonality and Consumption-Based Asset Pricing," *Journal of Finance*, 47, 511–552.
- GARBER, P. M., AND R. G. KING (1983): "Deep Structural Excavation? A Critique of Euler Equation Methods," NBER Technical Working Paper No. 31.
- GOMME, P., AND J. GREENWOOD (1995): "On the Cyclical Allocation of Risk," *Journal of Economic Dynamics and Control*, 19(1–2), 91–124.
- GREENWOOD, J., AND Z. HERCOWITZ (1991): "The Allocation of Goods and Time over the Business Cycle," *Journal of Political Economy*, 99, 1188–1214.
- HALL, R. E. (1980): "Labor Supply and Aggregate Fluctuations," *Carnegie-Rochester Conference Series on Public Policy*, 12, 7–33.
- HANSEN, L. P. (1982): "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50(4), 1029–1054.
- HANSEN, L. P., AND K. J. SINGLETON (1982): "Generalized Instrumental Variables Estimation of Nonlinear Rational Expectations Models," *Econometrica*, 50(5), 1269–1286.
- HAUSMAN, J. A. (1978): "Specification Tests in Econometrics," *Econometrica*, 46(6), 1251–1271.
- HAYASHI, F. (1982): "The Permanent Income Hypothesis: Estimation and Testing by Instrumental Variables," *Journal of Political Economy*, 90, 895–916.
- (1985): "The Effect of Liquidity Constraints on Consumption: A Cross-Sectional Analysis," *Quarterly Journal of Economics*, 100, 183–206.
- HEATON, J. C. (1993): "The Interaction Between Time-Nonseparable Preferences and Time Aggregation," *Econometrica*, 61, 353–385.
- (1995): "An Empirical Investigation of Asset Pricing with Temporally Dependent Preference Specifications," *Econometrica*, 63, 681–717.
- HEATON, J. C., AND D. LUCAS (1992): "The Effects of Incomplete Insurance Markets and Trading Costs in a Consumption-Based Asset Pricing Model," *Journal of Economic Dynamics and Control*, 16(3–4), 601–620.
- HOFFMAN, D. L., AND R. H. RASCHE (1991): "Long-Run Income and Interest Elasticities of Money Demand in the United States," *Review of Economics and Statistics*, 73(4), 665–674.
- HOUTHAKKER, H. S. (1957): "An International Comparison of Household Expenditure Patterns, Commemorating the Centenary of Engel's Law," *Econometrica*, 25(4), 532–551.
- (1960): "Additive Preferences," *Econometrica*, 28(2), 244–257.

- KING, R. G., C. I. PLOSSER, AND S. T. REBELO (1988): "Production, Growth and Business Cycles: II. New Directions," *Journal of Monetary Economics*, 21, 309–341.
- KUZNETS, S. (1946): *National Income: A Summary of Findings*. National Bureau of Economic Research, New York.
- KYDLAND, F. E. (1984): "Labor-Force Heterogeneity and the Business Cycle," *Carnegie-Rochester Conference Series on Public Policy*, 21, 173–208.
- MANKIW, N. G., J. J. ROTEMBERG, AND L. H. SUMMERS (1985): "Intertemporal Substitution in Macroeconomics," *Quarterly Journal of Economics*, 100(1), 225–251.
- MAR CET, A., AND K. J. SINGLETON (1991): "Optimal Consumption-Savings Decisions and Equilibrium Asset Prices in A Model with Heterogeneous Agents Subject to Portfolio Constraints," Manuscript, Stanford University.
- MEHRA, R., AND E. C. PRESCOTT (1985): "The Equity Premium: a Puzzle," *Journal of Monetary Economics*, 15, 145–161.
- MIRON, J. A. (1986): "Seasonal Fluctuations and the Life-Cycle Permanent Income Model of Consumption," *Journal of Political Economy*, 94(6), 1258–1279.
- OGAKI, M. (1988): "Learning about Preferences from Time Trends," Ph.D. thesis, University of Chicago.
- (1989): "Information in Deterministic Trends about Preferences," Manuscript.
- (1992): "Engel's Law and Cointegration," *Journal of Political Economy*, 100(5), 1027–1046.
- (1993a): "Generalized Method of Moments: Econometric Applications," in *Handbook of Statistics: Econometrics*, ed. by G. S. Maddala, C. R. Rao, and H. D. Vinod, vol. 11, chap. 17, pp. 455–488. North-Holland, Amsterdam.
- (1993b): "Unit Roots in Macroeconometrics: A Survey," *Monetary and Economic Studies*, 11(2), 131–154.
- (1997): "Aggregation under Complete Markets," Working Paper No. 97-05, Department of Economics, Ohio State University.
- OGAKI, M., AND J. Y. PARK (1992): "A Cointegration Approach to Estimating Preference Parameters," Manuscript.
- (1997): "A Cointegration Approach to Estimating Preference Parameters," *Journal of Econometrics*, 82(1), 107–134.
- OGAKI, M., AND C. M. REINHART (1998): "Measuring Intertemporal Substitution: The Role of Durable Goods," *Journal of Political Economy*, 106, 1078–1098.
- OSANO, H., AND T. INOUE (1991): "Testing Between Competing Models Of Real Business Cycles," *International Economic Review*, 32(3), 669–688.
- ROSEN, S. (1985): "Implicit Contracts: A Survey," *Journal of Economic Literature*, 23, 1144–1175.
- STOCK, J. H., AND M. W. WATSON (1993): "A Simple Estimator of Cointegrating Vectors in Higher Order Integrated Systems," *Econometrica*, 61(4), 783–820.

- WRIGHT, R. (1988): "The Observational Implications of Labor Contracts in A Dynamic General Equilibrium Model," *Journal of Labor Economics*, 6, 530–551.
- ZELDES, S. (1989): "Consumption And Liquidity Constraints: An Empirical Investigation," *Journal of Political Economy*, 97(2), 305–346.

## Chapter 16

# VECTOR AUTOREGRESSIONS WITH UNIT ROOT NONSTATIONARY PROCESSES

This chapter explains econometric methods related to VARs and cointegration. We first introduce a broader concept of cointegration that allows us to treat the case in which a vector time series includes both stationary and nonstationary variables. In the previous chapters, cointegration is only defined for a vector time series that does not include stationary variables. Then we discuss a method to impose long-run restrictions for VARs with stationary variables for which the nonstationary variables in the vector time series are not cointegrated. We will explain various representations of a cointegrated system such as Vector Error Correction Model (VECM) and Phillips' triangular representation. Then we will present methods to impose long-run restrictions imposed on Phillips' triangular representation and VECM representation. We will introduce a structural Error Correction Model (ECM) by considering a foreign exchange rate model in which prices and the exchange rate adjusts toward a long-run equilibrium level. A method to estimate the structural speed of the adjustment coefficient toward the long-run equilibrium level will be discussed. In the Appendix, we

will discuss long-run and short-run restrictions imposed on VECM.

## 16.1 Identification on Structural VAR Models

### 16.1.1 Long-Run Restrictions for Structural VAR Models

Blanchard and Quah (1989) propose using long-run restrictions to identify the underlying shocks in a VAR system. Let  $y_t$  be the logarithm of GDP and  $u_t$  be the level of the unemployment rate. Here  $y_t$  is assumed to be difference stationary and  $u_t$  is assumed to be stationary. Let  $\mathbf{y}_t = (\Delta y_t, u_t)'$ , and let  $\mathbf{e}_t = (e_t^s, e_t^d)'$  be the underlying shocks of the economy, where  $e_t^d$  is the demand shock, and  $e_t^s$  is the supply shock. It is assumed that the demand and supply shocks are uncorrelated, and that  $\mathbf{y}_t$  has an MA representation in terms of  $\mathbf{e}_t$ :

$$\begin{aligned} (16.1) \quad \mathbf{y}_t &= \boldsymbol{\mu} + \boldsymbol{\Phi}(L)\mathbf{e}_t \\ &= \boldsymbol{\mu} + \boldsymbol{\Phi}_0\mathbf{e}_t + \boldsymbol{\Phi}_1\mathbf{e}_{t-1} + \boldsymbol{\Phi}_2\mathbf{e}_{t-2} + \cdots, \end{aligned}$$

where  $\boldsymbol{\Phi}(1)$  is normalized so that its principal diagonal components are 1's, and  $E(\mathbf{e}_t\mathbf{e}_t') = \boldsymbol{\Lambda}$ .

The long-run restrictions are that the demand shock does not have any long-run effect, and the supply shock does not have any long-run effect on unemployment, but may have a long-run effect on the level of output. These restrictions imply that the matrix  $\boldsymbol{\Phi}(1)$  is lower triangular.

Let  $\mathbf{y}_t = \boldsymbol{\mu} + \boldsymbol{\Psi}(L)\boldsymbol{\epsilon}_t$  be the Wold representation, which can be estimated by inverting the VAR representation for  $\mathbf{y}_t$ . Then  $\boldsymbol{\epsilon}_t = \boldsymbol{\Phi}_0\mathbf{e}_t$ ,  $\boldsymbol{\Sigma}_\epsilon = E(\boldsymbol{\epsilon}_t\boldsymbol{\epsilon}_t') = \boldsymbol{\Phi}_0\boldsymbol{\Lambda}\boldsymbol{\Phi}_0'$ , and  $\boldsymbol{\Phi}_j = \boldsymbol{\Psi}_j\boldsymbol{\Phi}_0$  for all  $j$ . Once  $\boldsymbol{\Phi}_0$  is known, we can obtain  $\mathbf{e}_t$  from  $\boldsymbol{\epsilon}_t$ , and  $\boldsymbol{\Phi}_j$  from  $\boldsymbol{\Psi}_j$ . Is  $\boldsymbol{\Phi}_0$  identified? An informal argument by Blanchard and Quah suggest that

it is. Given  $\Sigma_\epsilon$ , the equation  $\Phi_0 \Lambda \Phi_0' = \Sigma_\epsilon$  gives three restrictions because  $\Sigma_\epsilon$  is symmetric. Given  $\Psi(1)$ , the equation that the upper right-hand entry of  $\Phi(1)$  is zero gives one more restriction. There exist four restrictions for four unknown parameters in  $\Phi_0$ .

The assumption that  $\Phi(1)$  is lower triangular gives  $\frac{n(n-1)}{2}$  necessary conditions. From  $\Phi(1)\mathbf{e}_t = \Psi(1)\boldsymbol{\epsilon}_t$  it follows

$$(16.2) \quad \Phi(1)\Lambda\Phi(1)' = \Psi(1)\Sigma_\epsilon\Psi(1)'$$

Let  $\mathbf{P}$  be a lower triangular matrix of the Cholesky decomposition of  $\Psi(1)\Sigma_\epsilon\Psi(1)'$  so that  $\mathbf{P}\mathbf{P}' = \Psi(1)\Sigma_\epsilon\Psi(1)'$ . Then,

$$(16.3) \quad \Phi(1) = \mathbf{P}\Lambda^{-\frac{1}{2}}$$

and

$$(16.4) \quad \Phi_0 = \Psi(1)^{-1}\Phi(1),$$

where  $\Lambda = [\text{diag}(\mathbf{P})]^2$ . Lastrapes and Selgin (1995) apply this model to study liquidity effects using  $\mathbf{y}_t = [r_t, y_t, (m_t - p_t), m_t]'$ .

Galí (1999) uses similar long-run restrictions to identify shocks. The main methodological difference from Blanchard and Quah is that Galí uses different variables, log productivity and log hours. Log productivity replaces log GDP. Log hours (or the first difference of log hours) replaces the unemployment rate. The log GDP and unemployment rate used by Blanchard and Quah can lead shocks such as government purchases and permanent labor-supply shocks to be mislabeled as the technological shock. Galí defines correlation of two variables when all shocks but one are shut down as conditional correlation. The estimated conditional correlations of hours and

productivity are negative for nontechnology shocks. Hours show a persistent decline in response to a positive technology shock. These findings are hard to reconcile with a RBC model, but are consistent with a model with monopolistic competition and sticky price.

### 16.1.2 Short-run and Long-Run Restrictions for Structural VAR Models

Galí (1992) uses both short-run and long-run restrictions to identify a structural VAR. He considers an IS-LM model that consists of output ( $y_t$ ), money supply ( $m_t$ ), the nominal interest rate ( $r_t$ ), and the price level ( $p_t$ )<sup>1</sup>:

$$(16.5) \quad \mathbf{B}(L)\mathbf{y}_t = \boldsymbol{\delta} + \mathbf{e}_t$$

where  $\mathbf{B}(L) = \mathbf{B}_0 - \sum_{i=1}^p \mathbf{B}_i L^i$ ,  $\mathbf{B}_0$  has ones on its diagonal,  $\mathbf{y}_t = (\Delta y_t, \Delta r_t, r_t - \Delta p_t, \Delta m_t - \Delta p_t)'$ ,  $p$  is the lag order of VAR,  $L$  is the lag operator, and  $\mathbf{e}_t = (e_t^s, e_t^{ms}, e_t^{md}, e_t^{is})'$  is the vector stochastic process describing supply, money supply, money demand, and spending (IS) disturbances that are assumed to be serially uncorrelated. Let  $n$  denote the dimension of  $\mathbf{y}_t$ , that is,  $n = 4$  in this model.

The model (16.5) can be estimated by the reduced form VAR:

$$(16.6) \quad \mathbf{A}(L)\mathbf{y}_t = \boldsymbol{\delta}_\epsilon + \boldsymbol{\epsilon}_t$$

where  $\mathbf{A}(L) = \mathbf{I} - \sum_{i=1}^p \mathbf{A}_i L^i$ ,  $\mathbf{A}_0 = \mathbf{I}$ , and  $\boldsymbol{\epsilon}_t$  is the vector of innovations in the elements of  $\mathbf{y}_t$ . Let  $\boldsymbol{\Sigma}_\epsilon$  denote the variance-covariance matrix of  $\boldsymbol{\epsilon}_t$ . Provided that  $\mathbf{B}_0$  is identified, all the structural parameters in (16.5) are computed from the estimates of (16.6) using  $\boldsymbol{\delta} = \mathbf{B}_0 \boldsymbol{\delta}_\epsilon$  and  $\mathbf{B}_i = \mathbf{B}_0 \mathbf{A}_i$  for  $i = 1, 2, \dots, p$ . Structural shocks are also constructed by  $\mathbf{e}_t = \mathbf{B}_0 \boldsymbol{\epsilon}_t$ .

---

<sup>1</sup> $y_t$ ,  $m_t$ , and  $p_t$  are in logarithms.

In order to identify  $\mathbf{B}_0$ , Galí (1992) imposes an orthogonality condition ( $\mathcal{R}0$ ) that the variance-covariance matrix of structural shocks,  $\mathbf{\Lambda}$ , is diagonal. From  $\mathbf{B}_0 \mathbf{\Sigma}_\epsilon \mathbf{B}'_0 = \mathbf{\Lambda}$  we have  $\frac{n(n+1)}{2} = 10$  independent restrictions, and leave  $\frac{n(n-1)}{2} = 6$  free parameters in  $\mathbf{B}_0$ .

A second set of restrictions, building on Blanchard and Quah (1989), specifies that the supply shock has long-run effects on the level of output but the three aggregate demand shocks ( $e_t^{ms}$ ,  $e_t^{md}$ , and  $e_t^{is}$ ) have no long-run effects on the level of output ( $\mathcal{R}1$ ,  $\mathcal{R}2$ , and  $\mathcal{R}3$ ). These restrictions identify the supply shock ( $e_t^s$ ) from the other shocks. These restrictions are denoted by  $\Phi(1)_{1j} = 0$  for  $j = 2, 3$ , and 4.

A third set of restrictions is that the money supply and the money demand shocks have no contemporaneous effects on output ( $\mathcal{R}4$  and  $\mathcal{R}5$ ). These restrictions identify the IS shock from the two types of monetary shocks. Let  $\Phi(L) = \mathbf{B}(L)^{-1}$ , in particular,  $\Phi_0 = \mathbf{B}_0^{-1}$ . These two restrictions are denoted by  $\Phi_{0,1j} = 0$  for  $j = 2$  and 3.

The final restriction identifies the money supply shock from the money demand shock. Galí (1992) assumes that the contemporaneous price does not enter the money supply rule that is denoted by  $\mathbf{B}_{0,23} + \mathbf{B}_{0,24} = 0$  ( $\mathcal{R}6$ ).<sup>2</sup>

The estimation of Galí (1992) is dramatic, and is well described by Pagan and Robertson (1995, 1998). From the long-run restrictions ( $\mathcal{R}1 \sim \mathcal{R}3$ ),  $\Phi(1)$  becomes a block lower triangular matrix, where  $\Phi(L) = \mathbf{B}(L)^{-1}$  in (16.5). Inverting  $\Phi(1)$ , we also have a block lower triangular matrix  $\mathbf{B}(1)$  so that  $\mathbf{B}_{12}(1) = \mathbf{B}_{13}(1) = \mathbf{B}_{14}(1) = 0$ . We can impose this set of restrictions directly on the coefficients of the structural

---

<sup>2</sup>Galí (1992) suggests two more alternative assumptions; contemporaneous output does not enter the money supply rule ( $\mathcal{R}7$ ) and contemporaneous homogeneity in money demand ( $\mathcal{R}8$ ). In this section, we focus on ( $\mathcal{R}6$ ).

VAR. For notational convention, let  $b_{ij}$  and  $b_{s,ij}$  be the  $(i, j)$  components of  $\mathbf{B}_0$  and  $\mathbf{B}_s$ , respectively. By imposing these long-run restrictions ( $\mathcal{R}1 \sim \mathcal{R}3$ ), we can reparameterize the first equation of (16.5) as

$$(16.7) \quad y_{1t} = -b_{12}\Delta^p y_{2t} - b_{13}\Delta^p y_{3t} - b_{14}\Delta^p y_{4t} + \sum_{i=1}^p b_{i,11}y_{1,t-i} \\ + \sum_{i=1}^{p-1} b_{i,12}\Delta^{p-i}y_{2,t-i} + \sum_{i=1}^{p-1} b_{i,13}\Delta^{p-i}y_{3,t-i} + \sum_{i=1}^{p-1} b_{i,14}\Delta^{p-i}y_{4,t-i} + e_{1t},$$

where  $\Delta^p y_{2t}$  is, for example,  $y_{2t} - y_{2,t-p}$ , and estimate the coefficients by instrumental variables using  $y_{it-1}$  for  $\Delta^p y_{it}$  for  $i = 2, 3, 4$ . Similarly, with the short-run restriction ( $\mathcal{R}6$ ), we can reparameterize the second equation of (16.5) as

$$(16.8) \quad y_{2t} = -b_{21}y_{1t} - b_{23}(y_{3t} - y_{4t}) \\ + \sum_{i=1}^p b_{i,21}y_{1,t-i} + \sum_{i=1}^p b_{i,22}y_{2,t-i} + \sum_{i=1}^p b_{i,23}y_{3,t-i} + \sum_{i=1}^p b_{i,24}y_{4,t-i} + e_{2t},$$

where we use  $\hat{\epsilon}_{1t}$ , a sample counterpart of the first error in (16.6) from a reduced form VAR, and  $\hat{e}_{1t}$ , a sample counterpart of the first shock in (16.7) from a structural VAR, for  $y_{1t}$  and  $y_{3t} - y_{4t}$  as an instrument, respectively. This result follows because  $\epsilon_{1t}$  is orthogonal to  $e_{2t}$  by the short-run restriction ( $\mathcal{R}4$ ) and  $e_{1t}$  is orthogonal to  $e_{2t}$  by the orthogonality conditions. The third equation is given by

$$(16.9) \quad y_{3t} = -b_{31}y_{1t} - b_{32}y_{2t} - b_{34}y_{4t} \\ + \sum_{i=1}^p b_{i,31}y_{1,t-i} + \sum_{i=1}^p b_{i,32}y_{2,t-i} + \sum_{i=1}^p b_{i,33}y_{3,t-i} + \sum_{i=1}^p b_{i,34}y_{4,t-i} + e_{3t},$$

where  $\hat{\epsilon}_{1t}$ ,  $\hat{e}_{1t}$ , and  $\hat{e}_{2t}$  are used as the instrumental variables for  $y_{1t}$ ,  $y_{2t}$ , and  $y_{4t}$ , respectively. The short-run restriction ( $\mathcal{R}5$ ) ensures that  $\epsilon_{1t}$  is orthogonal to  $e_{3t}$ , while the orthogonality conditions are used for  $e_{1t}$  and  $e_{2t}$ . Finally, the fourth equation is

given by

$$(16.10) \quad y_{4t} = -b_{41}y_{1t} - b_{42}y_{2t} - b_{43}y_{3t} \\ + \sum_{i=1}^p b_{i,41}y_{1,t-i} + \sum_{i=1}^p b_{i,42}y_{2,t-i} + \sum_{i=1}^p b_{i,43}y_{3,t-i} + \sum_{i=1}^p b_{i,44}y_{4,t-i} + e_{4t}$$

and estimated by instrumental variables using  $\hat{e}_{1t}$ ,  $\hat{e}_{2t}$ , and  $\hat{e}_{3t}$  for  $y_{1t}$ ,  $y_{2t}$ , and  $y_{3t}$ , respectively from the orthogonality conditions.

The estimation method described above is a two-step instrumental variables method because the reduced form VAR is estimated in the first step and some of the residuals estimated in the first step are used for instrumental variables in the second step.

## 16.2 Representations for the Cointegrated System

This section introduces four useful representations of a cointegrating system: the *vector moving average representation* and *Phillips' triangular representation*. For example, these representations are useful in developing different methods to impose long-run restrictions.<sup>3</sup> For the illustration below, consider a vector of difference stationary processes  $\mathbf{z}_t = (\mathbf{y}_t, \mathbf{x}_t)'$  with a cointegrating vector  $\boldsymbol{\beta} = (\mathbf{I}, -\mathbf{c})'$ .

### 16.2.1 Vector Moving Average Representation

The cointegrating relationship between  $\mathbf{y}_t$  and  $\mathbf{x}_t$ , and the difference stationarity of  $\mathbf{x}_t$  can be written as

$$(16.11) \quad \mathbf{y}_t = \mathbf{c}'\mathbf{x}_t + \mathbf{u}_t$$

$$(16.12) \quad \mathbf{x}_t = \mathbf{x}_{t-1} + \mathbf{v}_t,$$

---

<sup>3</sup>Details of these representations are discussed in Section 19.1 of Hamilton (1994).

where  $\mathbf{u}_t$  and  $\mathbf{v}_t$  are stationary with zero mean.

Differencing (16.11) yields

$$(16.13) \quad \Delta \mathbf{y}_t = \mathbf{c}' \Delta \mathbf{x}_t + \Delta \mathbf{u}_t = \mathbf{c}' \mathbf{v}_t + \mathbf{u}_t - \mathbf{u}_{t-1}.$$

Let  $\mathbf{e}_{1,t} \equiv \mathbf{c}' \mathbf{v}_t + \mathbf{u}_t$  and  $\mathbf{e}_{2,t} \equiv \mathbf{v}_t$ . Then, (16.56) can be written as

$$\Delta \mathbf{y}_t = \mathbf{e}_{1,t} - (\mathbf{e}_{1,t-1} - \mathbf{c}' \mathbf{e}_{2,t-1}) = (\mathbf{I} - L) \mathbf{e}_{1,t} + \mathbf{c}' L \mathbf{e}_{2,t}.$$

Stacking this along with (16.12) in a vector system yields the vector moving average representation for  $(\Delta \mathbf{y}_t, \Delta \mathbf{x}_t)'$ ,

$$\begin{bmatrix} \Delta \mathbf{y}_t \\ \Delta \mathbf{x}_t \end{bmatrix} = \Phi(L) \begin{bmatrix} \mathbf{e}_{1,t} \\ \mathbf{e}_{2,t} \end{bmatrix},$$

where

$$\Phi(L) \equiv \begin{bmatrix} \mathbf{I} - L & \mathbf{c}' L \\ \mathbf{0} & \mathbf{I} \end{bmatrix}.$$

Note that the polynomial  $\Phi(z)$  has a root at unity,  $|\Phi(1)| = \mathbf{0}$ , and hence is non-invertible. This suggests that  $\Delta \mathbf{z}_t$  cannot be represented by any finite-order vector autoregression since  $[\Phi(L)]^{-1} \Delta \mathbf{z}_t = \mathbf{e}_t$  does not exist.

Stationarity of  $\beta' \mathbf{z}_t$  requires that the vector moving average representation satisfies two necessary conditions. First, the matrix polynomial associated with the moving average must satisfy

$$\beta' \Phi(1) = \mathbf{0}.$$

Further, if some of the series in  $\mathbf{z}_t$  exhibit nonzero drift and thus include the deterministic trend component  $\mu_z t$ ,

$$\mathbf{z}_t = \mu_z t + \mathbf{z}_t^0,$$

where  $\mu_z \neq \mathbf{0}$ , and  $\mathbf{z}_t^0$  is difference stationary without drift, then the stationarity requires that the deterministic cointegration restriction holds (Engle and Yoo, 1987;

Ogaki and Park, 1997). That is, the cointegrating vector must eliminate the deterministic trend from the system:

$$\beta' \mu_z = \mathbf{0}.$$

Otherwise, the linear combination  $\beta' \mathbf{z}_t$  will grow deterministically at the rate  $\beta' \mu_z$ .

### 16.2.2 Phillips' Triangular Representation

Phillips's (1991) triangular representation takes the form:

$$(16.14) \quad \mathbf{y}_t - \mathbf{c}' \mathbf{x}_t = \mathbf{u}_t,$$

$$(16.15) \quad \Delta \mathbf{x}_t = \mathbf{v}_t.$$

To derive this, suppose an  $n \times 1$  vector  $\mathbf{z}_t = (\mathbf{y}_t, \mathbf{x}_t)'$  is characterized by  $h$  cointegrating relations. The matrix of  $h$  cointegrating vectors can be written as

$$\beta' = \begin{bmatrix} \mathbf{b}'_1 \\ \mathbf{b}'_2 \\ \vdots \\ \mathbf{b}'_h \end{bmatrix} = \begin{bmatrix} 1 & b_{12} & b_{13} & \cdots & b_{1n} \\ b_{21} & b_{22} & b_{23} & \cdots & b_{2n} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ b_{h1} & b_{h2} & b_{h3} & \cdots & b_{hn} \end{bmatrix},$$

where the (1,1)-th element has been normalized to unity. After appropriate row operations, it can be transformed as

$$\beta' = \begin{bmatrix} 1 & 0 & \cdots & 0 & b_{1,h+1}^* & b_{1,h+2}^* & \cdots & b_{1,n}^* \\ 0 & 1 & \cdots & 0 & b_{2,h+1}^* & b_{2,h+2}^* & \cdots & b_{2,n}^* \\ \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & 1 & b_{h,h+1}^* & b_{h,h+2}^* & \cdots & b_{h,n}^* \end{bmatrix} = [\mathbf{I}_h \quad -\mathbf{c}'].$$

Therefore, with  $\mathbf{z}_t$  correspondingly partitioned into an  $h \times 1$  vector  $\mathbf{y}_t$  and a  $(n-h) \times 1$  vector  $\mathbf{x}_t$ ,

$$\beta' \mathbf{z}_t = [\mathbf{I}_h \quad -\mathbf{c}'] \begin{bmatrix} \mathbf{y}_t \\ \mathbf{x}_t \end{bmatrix} = \mathbf{y}_t - \mathbf{c}' \mathbf{x}_t$$

is stationary in equation (16.57). Equation (16.58) comes from the assumption that  $\mathbf{z}_t$  is difference stationary. Thus, in Phillips' triangular representation, variables on

the left hand side are all stationary, and are expressed in the form of the moving average.

The triangular representation has been widely used for estimating cointegrating vectors. One of the reasons is that when presented in this way, the model's (unknown) coefficients appear only in equation (16.57). Therefore, we can estimate the cointegrating relationship using standard estimation methods for a system of simultaneous equations.

As an example of Phillips' representation, consider the 4-variable system of Shapiro and Watson (1988). The model consists of four variables: labor input  $h_t$ , output  $y_t$ , the inflation rate  $\pi_t$ , and the long-run real interest rate  $i_t - \pi_t$ . In the short-run, these variables deviate from their long-run steady state values due to four types of serially uncorrelated shocks: labor supply shocks  $v_t$ , technological shocks  $e_t$ , and two aggregate demand shocks  $\nu_t^1$  and  $\nu_t^2$ . Labor supply shocks and technology shocks are uncorrelated with each other and with the aggregate demand shocks. In this model, all shocks are assumed to have only short-term effects on the real interest rate. That is, the nominal interest rate and the inflation rate are cointegrated so the real interest rate is stationary. Let

$$\mathbf{z}_t = [ i_t \quad \pi_t \quad h_t \quad y_t ]'$$

with a cointegrating vector

$$\boldsymbol{\beta}' = [ 1 \quad -1 \quad 0 \quad 0 ].$$

We can partition  $\mathbf{z}_t$  into  $z_{1,t} = i_t$ , and  $\mathbf{z}_{2,t} = ( \pi_t \quad h_t \quad y_t )'$ . With the model's long-

run restrictions, Phillips' triangular representation for this cointegrating system is

$$\begin{aligned} i_t - \pi_t &= c_1 + \Phi_i(L) [ v_t \ e_t \ \nu_t^1 \ \nu_t^2 ]', \\ \Delta\pi_t &= c_2 + \Phi_\pi(L) [ v_t \ e_t \ \nu_t^1 \ \nu_t^2 ]', \\ \Delta h_t &= c_3 + \Sigma_h(L)v_t + (1-L)\Phi_h(L) [ v_t \ e_t \ \nu_t^1 \ \nu_t^2 ]', \\ \Delta y_t &= c_4 + \Sigma_h(L)v_t + \alpha^{-1}\Sigma_\epsilon(L)e_t + (1-L)\Phi_y(L) [ v_t \ e_t \ \nu_t^1 \ \nu_t^2 ]', \end{aligned}$$

where  $c_i$  for  $i = 1, \dots, 4$ , are constant, and the lag polynomials  $\Sigma_h(L)$  and  $\Sigma_\epsilon(L)$  are assumed to have absolutely summable coefficients and roots outside the unit circle.

### 16.2.3 Vector Error Correction Model Representation

Vector autoregressive models originating with Sims (1980) have the following reduced form:

$$(16.16) \quad \mathbf{A}(L)\mathbf{y}_t = \boldsymbol{\delta}_\epsilon + \boldsymbol{\epsilon}_t,$$

where  $\mathbf{A}(L) = \mathbf{I}_n - \sum_{i=1}^p \mathbf{A}_i L^i$ ,  $\mathbf{A}(0) = \mathbf{I}_n$ , and  $\boldsymbol{\epsilon}_t$  is *white noise* with mean zero and variance  $\Sigma_\epsilon$ . From the reduced form of the VAR model,  $\mathbf{A}(L)$  can be re-parameterized as  $\mathbf{A}(1)L + \mathbf{A}^*(L)(1-L)$ , where  $\mathbf{A}(1)$  has a reduced rank,  $r < n$ . Engle and Granger (1987) showed that there exists an error correction representation:

$$(16.17) \quad \mathbf{A}^*(L)\Delta\mathbf{y}_t = \boldsymbol{\delta}_\epsilon - \mathbf{A}(1)\mathbf{y}_{t-1} + \boldsymbol{\epsilon}_t,$$

where  $\mathbf{A}^*(L) = \mathbf{I}_n - \sum_{i=1}^{p-1} \mathbf{A}_i^* L^i$ , and  $\mathbf{A}_i^* = -\sum_{j=i+1}^p \mathbf{A}_j$ . Since  $\mathbf{y}_t$  is assumed to be cointegrated  $I(1)$ ,  $\Delta\mathbf{y}_t$  is  $I(0)$ , and  $-\mathbf{A}(1)$  can be decomposed as  $\boldsymbol{\alpha}\boldsymbol{\beta}'$ , where  $\boldsymbol{\alpha}$  and  $\boldsymbol{\beta}$  are  $n \times r$  matrices with full column rank,  $r$ .

Monte Carlo experiments of Qureshi (2008) show that for OLS estimates of level VAR very often exhibit explosive autoregressive roots for typical macro data. In

contrast, the frequency of encountering explosive roots in OLS estimates of VECM is much fewer. Because there is a general consensus among macroeconomists that the absolute value of autoregressive roots is at most one, this is an important advantage for VECM over level VAR.

### 16.2.4 Common Trend Representation

Another representation of a cointegrated VAR system is Stock and Watson (1988b) common trend representation, which is a generalization of Beveridge-Nelson decomposition. Since  $\Delta y_t$  is stationary, we have

$$(16.18) \quad (1 - L)y_t = \Phi(L)\epsilon_t.$$

Then

$$(16.19) \quad \begin{aligned} y_t &= \frac{\Phi(L)}{1 - L} \\ &= \frac{\Phi(1)}{1 - L}\epsilon_t + \frac{\Phi(L) - \Phi(1)}{1 - L}\epsilon_t \\ &= A \begin{bmatrix} z_{1,t} \\ \vdots \\ z_{n-r,t} \end{bmatrix} + B(L)\epsilon_t \end{aligned}$$

where  $z_{i,t}$  is a random walk and is called a stochastic trend. In a  $n$ -variable system, there exist  $r$  cointegration relationship if and only if there exist  $(n - r)$  common stochastic trend.

**Example 16.1** If we have income and consumption,  $y_t$  and  $c_t$ , such that

$$(16.20) \quad y_t = z_t + e_t^y$$

$$(16.21) \quad c_t = z_t + e_t^c$$

where  $z_t$  is a random walk, and  $e_t^y$  and  $e_t^c$  are transitory income and consumption shock, respectively. Then,

$$(16.22) \quad \begin{pmatrix} y_t \\ c_t \end{pmatrix} = z_t + \begin{pmatrix} e_t^y \\ e_t^c \end{pmatrix}.$$

where  $z_t$  is a common stochastic trend. In this case, there is one cointegrating relationship so that  $y_t - c_t = e_t^y - e_t^c$  is stationary. ■

### 16.3 Long-Run Restrictions on Phillips' Triangular Representation

Long-run restrictions can be imposed on Phillips' Triangular representation. As an illustration, consider the model of Shapiro and Watson (1988). In this model,  $\mathbf{y}_t = (\Delta h_t, \Delta y_t, \Delta \pi_t, i_t - \pi_t)'$ , where  $h_t$  denotes labor supply,  $y_t$  output,  $\pi_t$  inflation, and  $i_t$  the nominal interest rate. Since  $h_t$ ,  $y_t$ , and  $\pi_t$  are assumed to be  $I(1)$ ,  $\Delta h_t$ ,  $\Delta y_t$ , and  $\Delta \pi_t$  are stationary  $I(0)$ . There are three sources of disturbances: labor supply  $v_t$ , technology  $e_t$ , and aggregate demand disturbances  $\nu_t^1$  and  $\nu_t^2$ , and thus  $\mathbf{e}_t = (v_t, e_t, \nu_t^1, \nu_t^2)'$ . The first two disturbances may be referred as supply shocks, and are assumed to be orthogonal and serially uncorrelated, and uncorrelated with the demand shocks. Since  $\mathbf{y}_t$  has been assumed to be stationary, none of the shocks has a long-run effect on  $\Delta h_t$ ,  $\Delta y_t$ ,  $\Delta \pi_t$ , or  $i_t - \pi_t$ .

Shapiro and Watson (1988) make two identifying restrictions: first, the aggregate demand shocks have no permanent effect on the level of output; and second, the long-run level of labor supply is exogenous. To impose these restrictions, consider, for example, the long-run effect of  $\nu_t^1$  on  $y_t$ . In their setup,  $\phi_{23k}$  is the effect of  $\nu_t^1$  on  $\Delta y_t$  after  $k$  periods, and therefore  $\sum_{k=1}^l \phi_{23k}$  is the effect of  $\nu_t^1$  on  $y_t$  itself after  $l$  periods. For  $\nu_t^1$  to have no effect on  $y_t$  in the long run, then we must have that  $\sum_{k=0}^{\infty} \phi_{23k} = 0$ .

Thus, the two assumptions impose restrictions that the long-run multipliers from  $\nu_t^1$  and  $\nu_t^2$  to  $h_t$  and  $y_t$ , and from  $e_t$  to  $h_t$  are zero. The resulting matrix of long-run multipliers,  $\Phi(1)$ , is block lower triangular:

$$\Phi(1) = \begin{bmatrix} \phi_{11} & 0 & 0 & 0 \\ \phi_{21} & \phi_{22} & 0 & 0 \\ \phi_{31} & \phi_{32} & \phi_{33} & \phi_{34} \\ \phi_{41} & \phi_{42} & \phi_{43} & \phi_{44} \end{bmatrix}.$$

Because there are no restrictions on  $\phi_{34}$ , this identification scheme cannot be used to disentangle the two aggregate demand shocks  $\nu_t^1$  and  $\nu_t^2$ , and only their joint impact can be estimated.

In order to estimate  $e_t$  and  $\Phi(L)$  using the observed data, Shapiro and Watson (1988) follow Blanchard and Quah (1989), and use the block lower triangular structure of  $\Phi(1)$  and the assumption that the shocks are serially and mutually uncorrelated. The Wold representation  $\mathbf{y}_t = \delta + \Psi(L)\epsilon_t$  can be obtained by first estimating and then inverting the VAR representation of  $\mathbf{y}_t$  in the usual way.

The equation for  $\Delta h_t$  can be written as

$$\Delta h_t = \sum_{j=1}^p \beta_{hh,j} \Delta h_{t-j} + \sum_{j=0}^p \beta_{hy,j} \Delta y_{t-j} + \sum_{j=0}^p \beta_{h\pi,j} \Delta \pi_{t-j} + \sum_{j=0}^p \beta_{hi,j} (i_{t-j} - \pi_{t-j}) + v_t.$$

Because the long-run multipliers from  $e_t$ ,  $\nu_t^1$ , and  $\nu_t^2$  to  $h_t$  are zero,  $\sum_{j=0}^p \beta_{hn,j} = 0$  for  $n = y, \pi, i$ . Imposing these constraints yields second differences. For example, consider the long-run restriction of  $e_t$  on  $h_t$ :

$$\begin{aligned} \sum_{j=0}^p \beta_{hy,j} \Delta y_{t-j} &= \beta_{hy,0} \Delta y_t + \cdots + \beta_{hy,p-1} \Delta y_{t-(p-1)} + \beta_{hy,p} \Delta y_{t-p} \\ &= \beta_{hy,0} (\Delta y_t - \Delta y_{t-1}) + (\beta_{hy,0} + \beta_{hy,1}) (\Delta y_{t-1} - \Delta y_{t-2}) + \cdots \\ &\quad + (\beta_{hy,0} + \beta_{hy,1} + \cdots + \beta_{hy,p-1}) (\Delta y_{t-(p-1)} - \Delta y_{t-p}) \\ &\quad + (\beta_{hy,0} + \beta_{hy,1} + \cdots + \beta_{hy,p-1} + \beta_{hy,p}) (\Delta y_{t-p}) \end{aligned}$$

The long-run restriction requires that  $\beta_{hy,0} + \beta_{hy,1} + \cdots + \beta_{hy,p-1} + \beta_{hy,p} = 0$ , and hence the coefficient on  $\Delta y_{t-p}$  is zero. Thus we have

$$\begin{aligned} \sum_{j=0}^p \beta_{hy,j} \Delta y_{t-j} &= \beta_{hy,0} \Delta^2 y_t + (\beta_{hy,0} + \beta_{hy,1}) \Delta^2 y_{t-1} + \cdots + (\beta_{hy,0} + \beta_{hy,1} + \cdots + \beta_{hy,p-1}) \Delta^2 y_{t-(p-1)} \\ &= \gamma_{hy,0} \Delta^2 y_t + \gamma_{hy,1} \Delta^2 y_{t-1} + \cdots + \gamma_{hy,p-1} \Delta^2 y_{t-(p-1)} \\ &= \sum_{j=0}^{p-1} \gamma_{hy,s} \Delta^2 y_{t-j}. \end{aligned}$$

The same operations can be done for  $\sum_{j=0}^p \beta_{h\pi,j}$  and  $\sum_{j=0}^p \beta_{hi,j}$  as well. The resulting equation to be estimated is

$$\Delta h_t = \sum_{j=1}^p \beta_{hh,j} \Delta h_{t-j} + \sum_{j=0}^{p-1} \gamma_{hy,j} \Delta^2 y_{t-j} + \sum_{j=0}^{p-1} \gamma_{h,\pi} \Delta^2 \pi_{t-j} + \sum_{j=0}^{p-1} \gamma_{hi,j} (\Delta i_{t-j} - \Delta \pi_{t-j}) + v_t.$$

This equation cannot be consistently estimated by OLS because it includes contemporaneous values of some of the regressors which are correlated with  $v_t$ . Therefore, the IV estimation is used with  $\{\Delta h_{t-s}, \Delta y_{t-s}, \Delta \pi_{t-s}, i_{t-s} - \pi_{t-s}\}_{s=1}^p$  as instruments.

Similarly, the equation for  $\Delta y_t$  is

$$\Delta y_t = \sum_{j=1}^p \beta_{yh,j} \Delta h_{t-j} + \sum_{j=1}^p \beta_{yy,j} \Delta y_{t-j} + \sum_{j=0}^{p-1} \Delta^2 \pi_{t-j} + \sum_{j=0}^{p-1} \gamma_{yi,j} (\Delta i_{t-j} - \Delta \pi_{t-j}) + \beta_{yv} v_t + e_t.$$

Note that the contemporaneous value of  $\Delta h_t$  do not enter this equation since  $v_t$  enters directly. Again, the correlations between  $e_t$  and contemporaneous values of some of the regressors require that it is estimated by the IV estimation using the same set of instruments plus  $\{v_{t-s}\}_{s=1}^p$  as instruments.

The equations estimated for  $\Delta \pi_t$  and  $\pi_t - i_t$  are reduced forms. They are

$$\Delta \pi_t = \sum_{j=1}^p \beta_{\pi h,j} \Delta h_{t-j} + \sum_{j=0}^p \beta_{\pi y,j} \Delta y_{t-j} + \sum_{j=1}^p \beta_{\pi \pi,j} \Delta \pi_{t-j} + \sum_{j=1}^p \beta_{\pi i,j} (i_{t-j} - \pi_{t-j}) + \beta_{\pi v} v_t + \beta_{\pi e} e_t + a_t^1,$$

and

$$i_t - \pi_t = \sum_{j=1}^p \beta_{ih,j} \Delta h_{t-j} + \sum_{j=0}^p \beta_{iy,j} \Delta y_{t-j} + \sum_{j=1}^p \beta_{i\pi,j} \Delta \pi_{t-j} + \sum_{j=1}^p \beta_{ii,j} (i_{t-j} - \pi_{t-j}) + \beta_{iv} v_t + \beta_{ie} e_t + a_t^2.$$

The error terms  $a_t^1$  and  $a_t^2$  are linear combinations of the structural aggregate shocks  $\nu_t^1$  and  $\nu_t^2$ . Since these disturbances are uncorrelated with the regressions, these two equations can be estimated by OLS.

### 16.3.1 Long-run Restrictions and VECM

An alternative method to impose long-run restrictions is to use VECM. As  $\Delta \mathbf{y}_t$  is assumed to be stationary, it has a unique Wold representation:

$$(16.23) \quad \Delta \mathbf{y}_t = \boldsymbol{\mu} + \boldsymbol{\Psi}(L)\boldsymbol{\epsilon}_t,$$

where  $\boldsymbol{\mu} = \boldsymbol{\Psi}(1)\boldsymbol{\delta}_\epsilon$  and  $\boldsymbol{\Psi}(L) = \mathbf{I}_n + \sum_{i=1}^{\infty} \boldsymbol{\Psi}_i L^i$ . The above, which is in reduced form, can be represented in structural form as:

$$(16.24) \quad \begin{aligned} \Delta \mathbf{y}_t &= \boldsymbol{\mu} + \boldsymbol{\Phi}(L)\mathbf{e}_t \\ \boldsymbol{\Phi}(L) &= \boldsymbol{\Psi}(L)\boldsymbol{\Phi}_0 \\ \mathbf{e}_t &= \boldsymbol{\Phi}_0^{-1}\boldsymbol{\epsilon}_t, \end{aligned}$$

where  $\boldsymbol{\Phi}(L) = \boldsymbol{\Phi}_0 + \sum_{i=1}^{\infty} \boldsymbol{\Phi}_i L^i$ , and  $\mathbf{e}_t$  is a vector of structural innovations with mean zero and variance  $\boldsymbol{\Lambda}$ .

Long-run restrictions are imposed on the structural form, as in Blanchard and Quah (1989). Stock and Watson (1988a) developed a common trend representation that was shown to be equivalent to a VECM representation. When cointegrated variables have a reduced rank,  $r$ , there exist  $k = n - r$  common trends. These common trends can be considered to be generated by permanent shocks, so that  $\mathbf{e}_t$  can be decomposed into  $(\mathbf{e}_t^{k'}, \mathbf{e}_t^{r'})'$ , in which  $\mathbf{e}_t^{k'}$  is a  $k$ -dimensional vector of permanent shocks and  $\mathbf{e}_t^{r'}$  is an  $r$ -dimensional vector of transitory shocks. As developed in King, Plosser, Stock, and Watson (1989, 1991, KPSW for short), this decomposition ensures

that

$$(16.25) \quad \Phi(1) = \begin{bmatrix} \mathbf{A} & \mathbf{0} \end{bmatrix},$$

where  $\mathbf{A}$  is an  $n \times k$  matrix and  $\mathbf{0}$  is an  $n \times r$  matrix with zeros, representing long-run effects of permanent shocks and transitory shocks, respectively. In order to identify permanent shocks, in general, causal chains, in the sense of Sims (1980), are imposed on permanent shocks:

$$(16.26) \quad \mathbf{A} = \hat{\mathbf{A}}\mathbf{\Pi},$$

where  $\hat{\mathbf{A}}$  is an  $n \times k$  matrix, and  $\mathbf{\Pi}$  is a  $k \times k$  lower triangular matrix with ones in the diagonal. As Jang (2001a) shows,  $\hat{\mathbf{A}}$  is constructed using the cointegrating vectors:

$$(16.27) \quad \hat{\mathbf{A}} = \hat{\beta}_{\perp}.$$

See Appendix 16.A for detail.

### 16.3.2 Identification of Permanent Shocks

The main interest lies in the identification of structural permanent shocks, not in structural transitory shocks.<sup>4</sup> Following KPSW, we decompose  $\Phi_0$  and  $\Phi_0^{-1}$  as:

$$(16.28) \quad \Phi_0 = \begin{bmatrix} \mathbf{H} & \mathbf{J} \end{bmatrix}, \quad \Phi_0^{-1} = \begin{bmatrix} \mathbf{G} \\ \mathbf{E} \end{bmatrix}$$

where  $\mathbf{H}$ ,  $\mathbf{J}$ ,  $\mathbf{G}$  and  $\mathbf{E}$  are  $n \times k$ ,  $n \times r$ ,  $k \times n$ , and  $r \times n$  matrices, respectively. Note that the permanent shocks are identified once  $\mathbf{H}$  (or  $\mathbf{G}$ ) is identified, and that these two matrices have a one-to-one relation,  $\mathbf{G} = \mathbf{\Lambda}^k \mathbf{H}' \Sigma_{\epsilon}^{-1}$ , where  $\mathbf{\Lambda}^k$  is the variance-covariance matrix of permanent shocks,  $\mathbf{e}_t^k$ .<sup>5</sup> Therefore, the above decomposition of  $\Phi_0$  does not generate additional free parameters.

<sup>4</sup>Fisher, Fackler, and Orden (1995) consider the identification of transitory shocks imposing causal chains on transitory shocks.

<sup>5</sup>One can easily derive this relation from the relation  $\Phi_0^{-1} \Sigma_{\epsilon} = \mathbf{\Lambda} \Phi_0'$ .

The identifying scheme below basically follows that of KPSW, but enables one to generalize their model as described below. See Jang (2001a) for details. Following KPSW, let  $\mathbf{D} = (\hat{\boldsymbol{\beta}}'_{\perp} \hat{\boldsymbol{\beta}}_{\perp})^{-1} \hat{\boldsymbol{\beta}}'_{\perp} \boldsymbol{\Psi}(1)$  and  $\mathbf{P}$  be a lower triangular matrix chosen from the Cholesky decomposition of  $\mathbf{D}\boldsymbol{\Sigma}_{\epsilon}\mathbf{D}'$ . Then  $\boldsymbol{\Pi}$  and  $\boldsymbol{\Lambda}^k$  are uniquely determined by

$$(16.29) \quad \boldsymbol{\Pi} = \mathbf{P}(\boldsymbol{\Lambda}^k)^{-\frac{1}{2}},$$

where  $\boldsymbol{\Lambda}^k = [\text{diag}(\mathbf{P})]^2$ , and  $\mathbf{H}$  and  $\mathbf{G}$  are identified by

$$(16.30) \quad \mathbf{H} = \begin{bmatrix} \mathbf{D} \\ \boldsymbol{\alpha}'\boldsymbol{\Sigma}_{\epsilon}^{-1} \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{\Pi} \\ \mathbf{0} \end{bmatrix}$$

and

$$(16.31) \quad \mathbf{G} = \boldsymbol{\Lambda}^k \mathbf{H}' \boldsymbol{\Sigma}_{\epsilon}^{-1}.$$

Accordingly, the permanent shocks and the short run dynamics are identified by

$$(16.32) \quad \mathbf{e}_t^k = \mathbf{G}\epsilon_t$$

and

$$(16.33) \quad \boldsymbol{\Phi}(L)^k = \boldsymbol{\Psi}(L)\mathbf{H},$$

where  $\boldsymbol{\Phi}(L)^k$  denotes the first  $k$  columns of  $\boldsymbol{\Phi}(L)$ .

The specific solutions for  $\mathbf{H}$  and  $\mathbf{G}$  in the form of matrices enable one to generalize the model. Jang (2001b) considered a structural VECM in which structural shocks are partially identified using long-run restrictions and are fully identified by means of additional short-run restrictions (See Jang, 2001b, for the method of identification in structural VECMs with short-run and long-run restrictions). Jang and Ogaki (2001) consider a special case, where impulse response analysis is used to examine the effects

of only one permanent shock, and the recursive assumption on the permanent shocks in (16.26) can be relaxed, which implies  $\mathbf{\Pi}$  is lower block triangular. Note that we can compute the impulse responses to the  $k_{th}$  shock as long as the  $k_{th}$  column of  $\mathbf{H}$ ,  $\mathbf{H}_k$ , is identified. Note also that the third column of  $\mathbf{\Pi}$  does not contain any unknown parameters. Analogous to (16.30),  $\mathbf{H}_k$  is identified by

$$(16.34) \quad \mathbf{H}_k = \begin{bmatrix} \mathbf{D} \\ \boldsymbol{\alpha}' \boldsymbol{\Sigma}_\epsilon^{-1} \end{bmatrix}^{-1} \mathbf{S}_k$$

where  $\mathbf{S}_k$  is an  $n$ -dimensional selection vector with one at the  $k_{th}$  row and zeros at other rows. Similarly,  $\mathbf{G}_k$  is identified by:

$$(16.35) \quad \mathbf{G}_k = \boldsymbol{\Lambda}_{k,k}^k \mathbf{H}_k' \boldsymbol{\Sigma}_\epsilon^{-1}$$

and it follows from the identity relation of  $\mathbf{G}\mathbf{H} = \mathbf{I}_k$  that

$$(16.36) \quad \boldsymbol{\Lambda}_{k,k}^k = (\mathbf{H}_k' \boldsymbol{\Sigma}_\epsilon^{-1} \mathbf{H}_k)^{-1},$$

where  $\boldsymbol{\Lambda}_{k,k}^k$  is the variance of the  $k_{th}$  permanent shock. Thus, the  $k_{th}$  permanent shock is identified by

$$(16.37) \quad e_{t,k}^k = \mathbf{G}_k \boldsymbol{\epsilon}_t.$$

### 16.3.3 Impulse Response Functions

Impulse response analysis has been widely used in the applied VAR literature. It is, however, not straightforward to compute the impulse response from VECMs. The reduced-form VECM is usually converted to a *levels* VAR model for impulse response analysis.<sup>6</sup> Noting that the presence of unit roots prevents the inversion of a *levels*

<sup>6</sup>Mellander, Vredin, and Warne (1992) provide an algorithm to compute impulse response without converting VECM to *levels* VAR following the scheme in Campbell and Shiller (1988) and Warne (1991).

VAR model to a moving average (MA) representation, Lütkepohl and Reimers (1992) suggested the following algorithm to get impulse responses recursively in a cointegrated system. First, estimate the reduced-form VECM in (16.17), then convert the VECM to a *levels* VAR representation in (16.16) using the following relations:<sup>7</sup>

$$(16.38) \quad \mathbf{A}_i = \begin{cases} \mathbf{I}_n - \mathbf{A}(1) + \mathbf{A}_1^* & i = 1 \\ \mathbf{A}_i^* - \mathbf{A}_{i-1}^* & \text{for } 2 \leq i \leq p-1 \\ -\mathbf{A}_{p-1}^* & i = p. \end{cases}$$

Though a Wold representation does not exist in the presence of unit roots, Lütkepohl and Reimers (1992) showed that impulse responses can be recursively computed by

$$(16.39) \quad \Psi_m = \sum_{l=1}^p \Psi_{m-l} \mathbf{A}_l, \quad m = 1, 2, 3, \dots$$

$$(16.40) \quad \Phi_m = \Psi_m \Phi_0,$$

where  $\Psi_0 = \mathbf{I}_n$ ,  $\Phi_m = (\phi_{m,ij})$ , and  $\phi_{m,ij}$  is an  $m$ -step response of the  $i_{th}$  variable to the  $j_{th}$  innovation.<sup>8</sup> In particular, the impulse response function of permanent shocks in this paper is calculated by<sup>9</sup>

$$(16.41) \quad \Phi_m^k = \Psi_m \mathbf{H}, \quad m = 1, 2, \dots$$

As a special case, discussed in Section 16.3.2, the impulse response function of the  $k_{th}$  permanent shock is uniquely calculated from

$$(16.42) \quad \Phi_{m,k}^k = \Psi_m \mathbf{H}_k, \quad m = 1, 2, \dots$$

where  $\Phi_{m,k}^k$  is equivalent to the  $k_{th}$  column of  $\Phi_m^k$  in (16.41).

<sup>7</sup>We assume that  $n > p$  without any loss of generality.

<sup>8</sup>This algorithm can be simplified by rewriting VAR in (16.16) as a companion VAR(1) form. Then,  $\Psi_m$  is the first  $n$  row and  $n$  column submatrix of  $\mathbf{A}_c^m$ , in which  $\mathbf{A}_c$  is a companion form coefficient matrix.

<sup>9</sup>One may calculate the impulse response to a one standard deviation permanent shock by  $\Psi_m \mathbf{H}(\boldsymbol{\Lambda}^k)^{\frac{1}{2}}$ .

### 16.3.4 Forecast-Error Variance Decomposition

Denoting the  $h$ -step forecast error by

$$(16.43) \quad \begin{aligned} \mathbf{y}_{t+h} - E_t \mathbf{y}_{t+h} &= \sum_{i=0}^{\infty} \Psi_i (\boldsymbol{\epsilon}_{t+h-i} - E_t \boldsymbol{\epsilon}_{t+h-i}) \\ &= \sum_{i=0}^{h-1} \Psi_i \boldsymbol{\epsilon}_{t+h-i}, \end{aligned}$$

the forecast error variance is computed by the diagonal components of

$$(16.44) \quad E(\mathbf{y}_{t+h} - E_t \mathbf{y}_{t+h})^2 = \sum_{i=0}^{h-1} \Psi_i \Sigma_{\epsilon} \Psi_i'.$$

In particular, the forecast error variance of the  $l_{th}$  variable,  $y_{l,t+h}$ , is computed by

$$(16.45) \quad \sum_{i=0}^{h-1} \Psi_{i,l} \Sigma_{\epsilon} \Psi_{i,l}'.$$

where  $\Psi_{i,l}$  denotes the  $l_{th}$  row of  $\Psi_i$ .

To isolate the fraction of the forecast error variance attributed to permanent shocks, it is convenient and necessary to decompose the contribution of permanent shocks and transitory shocks as follows:

$$(16.46) \quad \begin{aligned} \mathbf{y}_{t+h} - E_t \mathbf{y}_{t+h} &= \sum_{i=0}^{\infty} \Psi_i \Phi_0 (\mathbf{e}_{t+h-i} - E_t \mathbf{e}_{t+h-i}) \\ &= \sum_{i=0}^{h-1} \Psi_i \begin{bmatrix} \mathbf{H} & \mathbf{J} \end{bmatrix} \begin{bmatrix} \mathbf{e}_{t+h-i}^k \\ \mathbf{e}_{t+h-i}^r \end{bmatrix}, \end{aligned}$$

where  $\Psi_i$  is defined in (16.39). Since  $\mathbf{e}_t$  is serially uncorrelated,

$$(16.47) \quad \begin{aligned} E(\mathbf{y}_{t+h} - E_t \mathbf{y}_{t+h})^2 &= \sum_{i=0}^{h-1} \Psi_i \begin{bmatrix} \mathbf{H} & \mathbf{J} \end{bmatrix} \begin{bmatrix} \Lambda^k & \mathbf{0} \\ \mathbf{0} & \Lambda^r \end{bmatrix} \begin{bmatrix} \mathbf{H}' \\ \mathbf{J}' \end{bmatrix} \Psi_i' \\ &= \sum_{i=0}^{h-1} \Psi_i (\mathbf{H} \Lambda^k \mathbf{H}' + \mathbf{J} \Lambda^r \mathbf{J}') \Psi_i'. \end{aligned}$$

Therefore, the contribution of permanent shocks to forecast error variance of the  $h$ -step forecast is estimated by the diagonal components of

$$(16.48) \quad \sum_{i=0}^{h-1} \Phi_i^k \Lambda^k \Phi_i^{k'}$$

In particular, the contribution of the  $m_{th}$  permanent shock,  $e_m^k$ , to the forecast error variance of the  $l_{th}$  variable,  $y_{l,t+h}$ , is<sup>10</sup>

$$(16.49) \quad \sum_{i=0}^{h-1} (\Phi_{i,lm}^k)^2 \Lambda_{m,m}^k$$

where  $\Lambda_{m,m}^k$  is the variance of the  $m_{th}$  permanent shock.

Finally, dividing (16.49) by (16.45) yields the fraction of the  $h$ -step forecast error variance of the  $l_{th}$  variable attributed to the  $m_{th}$  structural shock.

Section 16.3.2 discusses the special case of the contribution of the  $k_{th}$  permanent shock,  $e_k^k$ , to the forecast error variance of the  $l_{th}$  variable,  $y_{l,t+h}$ , which is computed by

$$(16.50) \quad \sum_{i=0}^{h-1} (\Phi_{i,lk}^k)^2 \Lambda_{k,k}^k$$

where  $\Lambda_{k,k}^k$  is the variance of the  $k_{th}$  permanent shock. Dividing (16.50) by (16.45) gives the portion of the contribution of the  $k_{th}$  structural shock to the  $h$ -step forecast error variance of the  $l_{th}$  variable.

### 16.3.5 Summary

In summary, the estimation and identification of VECM with long-run restrictions are executed by the following procedure:

1. Select the lag length of VECM using some criteria such as AIC and BIC.

---

<sup>10</sup>By the virtue of the assumption that permanent shocks are uncorrelated mutually, we can separate the contribution of each permanent shock.

2. Estimate cointegrating vectors and determine the rank of cointegrating vectors in (16.17).
3. Convert VECM to levels VAR using (16.38).
4. Impose long-run restrictions implied by economic theory<sup>11</sup>, and identify structural parameters using (16.30) and (16.31).
5. Compute impulse responses to a structural shock using (16.41).
6. Compute forecast-error variance decompositions using (16.45) and (16.49).
7. Compute confidence intervals of impulse responses and standard errors of forecast-error variance decompositions using Monte Carlo integration as described in Appendix 16.B.

## 16.4 Structural Vector Error Correction Models

In this section, we introduce ECM. Let  $\mathbf{y}_t$  be an  $n$ -dimensional vector of first difference stationary and stationary random variables. Let  $\boldsymbol{\ell}_i = (0, \dots, 0, 1, 0, \dots, 0)'$  with 1 on the  $i_{th}$  element. If the  $i_{th}$  element of  $\mathbf{y}_t$  is stationary, then  $\boldsymbol{\ell}_i \mathbf{y}_t$  is stationary. When a time series includes stationary variables, we extend the definition of cointegration, and say that  $\mathbf{y}_t$  is cointegrated with  $\boldsymbol{\ell}_i$  as a cointegrating vector. Suppose that  $\mathbf{y}_t$  has a VAR representation

$$(16.51) \quad \mathbf{y}_t = \boldsymbol{\delta}_\epsilon + \mathbf{A}_1 \mathbf{y}_{t-1} + \mathbf{A}_2 \mathbf{y}_{t-2} + \cdots + \mathbf{A}_p \mathbf{y}_{t-p} + \boldsymbol{\epsilon}_t.$$

---

<sup>11</sup>For example, one may adopt a long-run restriction that a monetary shock does not affect the level of real output.

where  $\delta_\epsilon$  is an  $n \times 1$  vector. Just as in Said-Dickey's reparameterization for the univariate case, it is convenient to reparameterize Equation (16.51) as

$$(16.52) \quad \Delta \mathbf{y}_t = \delta_\epsilon - \mathbf{A}(1)\mathbf{y}_{t-1} + \mathbf{A}_1^* \Delta \mathbf{y}_{t-1} + \cdots + \mathbf{A}_{p-1}^* \Delta \mathbf{y}_{t-p+1} + \epsilon_t,$$

where

$$(16.53) \quad \mathbf{A}(1) = \mathbf{I}_n - \sum_{j=1}^p \mathbf{A}_j \quad \text{and} \quad \mathbf{A}_i^* = - \sum_{j=i+1}^p \mathbf{A}_j \quad \text{for } i = 1, 2, \dots, p-1.$$

This reparameterization is convenient because  $-\mathbf{A}(1)$  summarizes the long-run properties of the series. We assume that there exist  $r$  linearly independent cointegrating vectors, so that  $\beta' \mathbf{y}_{t-1}$  is stationary, where  $\beta'$  is a  $r \times n$  matrix of real numbers whose rows are linearly independent cointegrating vectors. Then  $-\mathbf{A}(1) = \alpha \beta'$  for an  $n \times r$  matrix of real numbers,  $\alpha$ . Hence Equation (16.52) can be written as

$$(16.54) \quad \Delta \mathbf{y}_t = \delta_\epsilon + \alpha \beta' \mathbf{y}_{t-1} + \mathbf{A}_1^* \Delta \mathbf{y}_{t-1} + \cdots + \mathbf{A}_{p-1}^* \Delta \mathbf{y}_{t-p+1} + \epsilon_t,$$

This representation is called an ECM.

In many applications of standard ECMs, elements in  $\alpha$  are given structural interpretations as parameters of the speed of adjustment toward the long-run equilibrium represented by  $\beta' \mathbf{y}_{t-1}$ . It is of interest to study conditions under which the elements in  $\alpha$  can be given such a structural interpretation. In the model of the next section, the domestic price level gradually adjusts to its PPP level with a speed of adjustment parameter  $b$ . We will investigate conditions under which  $b$  can be estimated as an element in  $\alpha$  from (16.54).

The standard ECM, (16.54), is a reduced form model. A class of structural models can be written in the following form of a structural ECM:

$$(16.55) \quad \mathbf{B}_0 \Delta \mathbf{y}_t = \boldsymbol{\mu}^* + \boldsymbol{\alpha}^* \beta' \mathbf{y}_{t-1} + \mathbf{B}_1 \Delta \mathbf{y}_{t-1} + \cdots + \mathbf{B}_{p-1} \Delta \mathbf{y}_{t-p+1} + \mathbf{e}_t,$$

where  $\mathbf{B}_i$  is an  $n \times n$  matrix,  $\boldsymbol{\mu}^*$  is an  $n \times 1$  vector, and  $\boldsymbol{\alpha}^*$  is an  $n \times r$  matrix of real numbers. Here  $\mathbf{B}_0$  is a nonsingular matrix of real numbers with ones along its principal diagonal, and  $\mathbf{e}_t$  is a stationary  $n$ -dimensional vector of random variables with  $\hat{E}[\mathbf{e}_t | \mathbf{H}_{t-\tau}] = 0$ , where  $\tau > 0$ . Even though cointegrating vectors are not unique, we assume that there is a normalization that uniquely determines  $\boldsymbol{\beta}$ , so that parameters in  $\boldsymbol{\alpha}^*$  have structural meanings.

In order to see the relationship between the standard ECM and the structural ECM, we premultiply both sides of (16.55) by  $\mathbf{B}_0^{-1}$  to obtain the standard ECM (16.54), where  $\boldsymbol{\delta}_\epsilon = \mathbf{B}_0^{-1}\boldsymbol{\mu}^*$ ,  $\boldsymbol{\alpha} = \mathbf{B}_0^{-1}\boldsymbol{\alpha}^*$ ,  $\mathbf{A}_i^* = \mathbf{B}_0^{-1}\mathbf{B}_i$ , and  $\boldsymbol{\epsilon}_t = \mathbf{B}_0^{-1}\mathbf{e}_t$ . Thus the standard ECM estimated by Engle and Granger's two step method or Johansen's (1988) Maximum Likelihood method is a reduced form model. Hence it cannot be used to recover structural parameters in  $\boldsymbol{\alpha}^*$ , nor can the impulse-response functions based on  $\boldsymbol{\epsilon}_t$  be interpreted in a structural way unless some restrictions are imposed on  $\mathbf{B}_0$ .

As in a VAR, various restrictions are possible for  $\mathbf{B}_0$ . One example is to assume that  $\mathbf{B}_0$  is lower triangular. If  $\mathbf{B}_0$  is lower triangular, then the first row of  $\boldsymbol{\alpha}$  is equal to the first row of  $\boldsymbol{\alpha}^*$ , and structural parameters in the first row of  $\boldsymbol{\alpha}^*$  are estimated by the standard methods to estimate an ECM.

## 16.5 An Exchange Rate Model with Sticky Prices

This section presents a simple exchange rate model in which the domestic price adjusts slowly toward the long-run equilibrium level implied by Purchasing Power Parity (PPP). Kim, Ogaki, and Yang (2007) use this model to motivate a particular form of a structural ECM in the previous section. This model's two main components

are a slow adjustment equation and a rational expectations equation for the exchange rate. The single equation method is only based on the slow adjustment equation. The system method utilizes both the slow adjustment and rational expectations equations. A similar method was applied to an exchange rate model with the Taylor rule by Kim and Ogaki (2009).

Let  $p_t$  ( $p_t^*$ ) be the log domestic (foreign) price level, and  $e_t$  be the log nominal exchange rate. We assume that these variables are first difference stationary and PPP holds in the long-run, so that the real exchange rate,  $p_t - p_t^* - e_t$ , is stationary, or  $\mathbf{y}_t = (p_t, e_t, p_t^*)'$  is cointegrated with a cointegrating vector  $(1, -1, -1)$ . Let  $\mu = E[p_t - p_t^* - e_t]$ , then  $\mu$  can be nonzero when different units are used to measure prices in the two countries.

Using Mussa's (1982) model, the domestic price is assumed to adjust slowly to the PPP level

$$(16.56) \quad \Delta p_{t+1} = b(\mu + p_t^* + e_t - p_t) + E_t[p_{t+1}^* + e_{t+1}] - (p_t^* + e_t)$$

where  $\Delta x_{t+1} = x_{t+1} - x_t$  for any variable  $x_t$ ,  $E[\cdot | I_t]$  is the expectation operator conditional on  $I_t$ , the information available to the economic agents at time  $t$ , and a positive constant  $b$  ( $0 \leq b \leq 1$ ) is the adjustment coefficient. The idea behind (3) is that the domestic price slowly adjusts toward its PPP level of  $p_t^* + e_t$ , while it adjusts instantaneously to the expected change in its PPP level. The adjustment speed is slow (fast) when  $b$  is close to zero (one). From (3),

$$(16.57) \quad \Delta p_{t+1} = d + b(p_t^* + e_t - p_t) + \Delta p_{t+1}^* + \Delta e_{t+1} + \varepsilon_{t+1}$$

where  $d = b\mu$ ,  $\varepsilon_{t+1} = E_t[p_{t+1}^* + e_{t+1}] - (p_{t+1}^* + e_{t+1})$ . Hence  $\varepsilon_{t+1}$  is a one-period ahead forecasting error, and  $E[\varepsilon_{t+1} | I_t] = 0$ . (4) can be referred to as the structural

gradual adjustment equation which implies a first order AR structure for the real exchange rate. To see this, let  $s_t = p_t^* + e_t - p_t$  be the log real exchange rate. Then (4) implies

$$(16.58) \quad s_{t+1} = -d + (1-b)s_t - \varepsilon_{t+1}$$

We define the half-life of the real exchange rate as the number of periods required for a unit shock to dissipate by one half in (5). Without measurement errors,  $b$  can be estimated by OLS directly from (4). In the presence of measurement errors, IV are necessary.

Let the money demand equation and the Uncovered Interest Parity (UIP) condition be

$$(16.59) \quad m_t = \theta_m + p_t - hi_t$$

$$(16.60) \quad i_t = i_t^* + E[e_{t+1}|I_t] - e_t$$

where  $m_t$  is the log nominal money supply minus the log real national income,  $i_t$  ( $i_t^*$ ) is the nominal interest rate in the domestic (foreign) country. In (6), we are assuming that the income elasticity of money is one. From (6) and (7),

$$(16.61) \quad E[e_{t+1}|I_t] - e_t = (1/h)\{\theta_m + p_t - \omega_t - hE[(p_{t+1}^* - p_t^*)|I_t]\}$$

where  $\omega_t = m_t + hr_t^*$  and  $r_t^*$  is the foreign real interest rate,  $r_t^* = i_t^* - E[p_{t+1}^*|I_t] + p_t^*$ .

Following Mussa (1982), solving (3) and (8) as a system of stochastic difference equations

$$(16.62) \quad p_t = E[F_t|I_{t-1}] - \sum_{j=1}^{\infty} (1-b)^j \{E[F_{t-j}|I_{t-j}] - E[F_{t-j}|I_{t-j-1}]\}$$

$$(16.63) \quad e_t = \frac{bh+1}{bh} E[F_t|I_t] - p_t^* - \frac{1}{bh} p_t$$

where  $F_t = (1-\delta) \sum_{j=0}^{\infty} \delta^j \omega_{t+j}$  and  $\delta = h/(1+h)$ . We assume that  $\omega_t$  is first difference stationary. Since  $\delta$  is a positive constant that is smaller than one, this implies that  $F_t$  is also first difference stationary. From (9) and (10),  $e_t + p_t^* - p_t = \frac{bh+1}{bh} \sum_{j=0}^{\infty} (1-b)^j \{E[F_{t-j}|I_{t-j}] - E[F_{t-j}|I_{t-j-1}]\}$ , which means  $e_t + p_t^* - p_t$  is stationary.<sup>7</sup>

For a structural ECM representation from the exchange rate model, we use Hansen and Sargent's (1980; 1982) formula for linear rational expectations models. From (16.63),

$$(16.64) \quad \Delta e_{t+1} = \frac{bh+1}{bh} (1-\delta) E\left[\sum_{j=0}^{\infty} \delta^j \Delta \omega_{t+j+1} | I_t\right] - \frac{1}{bh} \Delta p_{t+1} - \Delta p_{t+1}^* + \varepsilon_{e,t+1}$$

where  $\varepsilon_{e,t+1} = \frac{bh+1}{bh} [E(F_{t+1}|I_{t+1}) - E(F_{t+1}|I_t)]$ , so that the law of iterated expectation implies  $E[\varepsilon_{e,t+1}|I_t] = 0$ . The system method using Hansen and Sargent's (1982) method is applicable because this equation involves a discounted sum of expected future values of  $\Delta \omega_t$ .

Hansen and Sargent's (1982) method can be applied to this model by projecting the conditional expectation of the discounted sum,  $E[\delta^j \Delta \omega_{t+j+1} | I_t]$ , onto an econometrician's information set  $H_t$ . We take the econometrician's information set at  $t$ ,  $H_t$ , to be the one generated by linear functions of current and past values of  $\Delta p_t^*$ . For simplicity, we follow West (1987) in that we choose a single variable to generate the information set  $H_t$ . In terms of the orthogonality condition, any variable in  $I_t$  can be used for this purpose.<sup>8</sup> Replacing  $E[\sum_{j=0}^{\infty} \delta^j \Delta \omega_{t+j+1} | I_t]$  by the econometrician's linear forecast based on  $H_t$  in (11), we obtain

$$(16.65) \quad \Delta e_{t+1} = \frac{bh+1}{bh} (1-\delta) \widehat{E}\left[\sum_{j=0}^{\infty} \delta^j \Delta \omega_{t+j+1} | H_t\right] - \frac{1}{bh} \Delta p_{t+1} - \Delta p_{t+1}^* + u_{2,t+1}$$

where  $u_{2,t+1} = \varepsilon_{e,t+1} + \frac{bh+1}{bh}(1-\delta)E[(\sum_{j=0}^{\infty} \delta^j \Delta\omega_{t+j+1}|I_t) - \widehat{E}(\sum_{j=0}^{\infty} \delta^j \Delta\omega_{t+j+1}|H_t)]$  and  $\widehat{E}[u_{2,t+1}|H_t] = 0$ . Following Hansen and Sargent (1980, 1982) we obtain (See appendix A.)

$$(16.66) \quad \widehat{E}\left[\sum_{j=0}^{\infty} \Delta\omega_{t+j+1}|H_t\right] = \xi_1 \Delta p_t^* + \xi_2 \Delta p_{t-1}^* + \dots + \xi_p \Delta p_{t-p+1}^*$$

A system of four equations will be<sup>9</sup>:

$$(16.67) \quad \Delta p_{t+1} = d + \Delta p_{t+1}^* + \Delta e_{t+1} - b(p_t - p_t^* - e_t) + u_{1,t+1}$$

$$(16.68) \quad \Delta e_{t+1} = -\frac{1}{bh} \Delta p_{t+1} - \Delta p_{t+1}^* + \alpha \xi_1 \Delta p_t^* + \alpha \xi_2 \Delta p_{t-1}^* + \dots + \alpha \xi_p \Delta p_{t-p+1}^* + u_{2,t+1}$$

$$(16.69) \quad \Delta p_{t+1}^* = \beta_1 \Delta p_t^* + \beta_2 \Delta p_{t-1}^* + \dots + \beta_p \Delta p_{t-p+1}^* + u_{3,t+1}$$

$$(16.70) \quad \Delta\omega_{t+1} = \gamma_1 \Delta p_t^* + \gamma_2 \Delta p_{t-1}^* + \dots + \gamma_{p-1} \Delta p_{t-p+2}^* + u_{4,t+1}$$

where  $\alpha = \frac{bh+1}{bh}(1-\delta)$  and  $u_{1,t+1} = \varepsilon_{t+1}$  with a set of nonlinear restrictions imposed by (16.66),

$$(16.71) \quad \gamma(\delta)[1 - \delta\beta(\delta)]$$

$$\xi_j = \delta\gamma(\delta)[1 - \delta\beta(\delta)]^{-1}(\beta_{j+1} + \delta\beta_{j+1} + \dots + \delta^{p-j}\beta_p) + (\gamma_j + \delta\gamma_j + \dots + \delta^{p-j}\gamma_p)$$

for  $j = 1, \dots, p$ . We call (16.67) the gradual adjustment equation, and (16.68)-(16.70) the Hansen and Sargent equations. Given the data for  $[\Delta p_{t+1}, \Delta e_{t+1}, \Delta p_{t+1}^*, \Delta\omega_{t+1}]'$ , GMM can be applied to the system of four equations, (14)-(17).<sup>10</sup>

It is instructive to observe the relationship between the structural ECM and the reduced form ECM in the exchange rate model (See appendix B.). Comparing **G** and **B** shows that the speed of adjustment coefficient for the domestic price is

$b$  in the structural model, while it is  $b^2h/(bh + 1)$  in the reduced form model.  $b$  in the structural form is not a deep structural parameter, unlike parameters of a production function or a utility function. However, it is clearly a parameter of interest because it determines the half-life of the real exchange rate. The reduced form speed of adjustment coefficient is a nonlinear function of  $b$ , and thus cannot be directly compared with the half-life estimates in the literature.

## 16.6 The System Method

Since standard methods of estimating (16.54) may not recover the structural parameters of interest in  $\alpha^*$ , Kim, Ogaki, and Yang (2001) propose a system method based on GMM that does not require restrictions on  $\mathbf{B}_0$ .

To apply the system method to (14)-(17) of the exchange rate model, we need data for  $\Delta\omega_t$ , which requires knowledge of  $h$ . Even though  $h$  is unknown, a cointegrating regression can be applied to money demand if money demand is stable in the long-run, as in Stock and Watson (1993). For this purpose, we augment the model as follows:

$$(16.72) \quad m_t = \theta_m + p_t - hi_t + \zeta_{m,t}$$

where  $\zeta_{m,t}$  is assumed to be stationary so that money demand is stable. By redefining  $m_t$  as  $m_t - \zeta_{m,t}$ , the same equations as those in section 3.2 are obtained. For the measurement of  $\Delta\omega_t$ , the *ex ante* foreign real interest rate can be replaced by the *ex post* value because of the Law of Iterated Expectations. Using (16.72), we obtain

$$(16.73) \quad \Delta\omega_{t+1} = \Delta p_{t+1} - h\Delta i_{t+1} + h\Delta i_{t+1}^* - h(\Delta p_{t+2}^* - \Delta p_{t+1}^*)$$

With this expression,  $\Delta\omega_t$  can be measured from price and interest rate data once  $h$  is

obtained, even if data for the monetary aggregate and national income are unavailable.

We have now obtained a system of four equations, (16.67)-(16.70). Because  $E[u_{i,t}|I_{t-\tau}] = 0$  and  $\widehat{E}[u_{i,t}|H_t] = 0$ , we obtain a vector of IV  $\mathbf{z}_{1,t}$  in  $I_{t-\tau}$  for  $u_{1,t}$  and  $\mathbf{z}_{i,t}$  in  $H_t$  for  $u_{i,t}$  ( $i = 2, 3, 4$ ).<sup>11</sup> Using the moment conditions  $E[z_{i,t}u_{i,t}] = 0$  for  $i = 1, \dots, 4$  we form a GMM estimator, imposing the Hansen-Sargent restrictions and the other cross-equation restrictions implied by the model.<sup>12</sup> Given estimates of cointegrating vectors from the first step, this system method provides more efficient estimators than Kim's (2004) single equation method as long as the restrictions implied by the model are true.<sup>13</sup> The cross-equation restrictions can be tested by Wald, Likelihood Ratio (LR) type, and Lagrange Multiplier (LM) tests in the GMM framework (see Ogaki, 1993). When restrictions are nonlinear, LR and LM tests are known to be more reliable than Wald tests.

## 16.7 Tests for the Number of Cointegrating Vectors

Johansen's (1988; 1991) maximum likelihood (ML) estimation is based on an error correction representation:

$$(16.74) \quad \Delta \mathbf{y}_t = \boldsymbol{\delta}_\epsilon + \boldsymbol{\alpha}\boldsymbol{\beta}'\mathbf{y}_{t-1} + \mathbf{A}_1^*\Delta\mathbf{y}_{t-1} + \cdots + \mathbf{A}_{p-1}^*\Delta\mathbf{y}_{t-p+1} + \boldsymbol{\epsilon}_t,$$

where  $\mathbf{y}_t$  and  $\boldsymbol{\epsilon}_t$  are  $n \times 1$  vectors of random variables,  $\boldsymbol{\alpha}$  and  $\boldsymbol{\beta}$  are  $n \times r$  matrices of real numbers, and  $\mathbf{A}_i^*$ 's are  $n \times n$  matrices of real numbers. The first term  $\boldsymbol{\alpha}\boldsymbol{\beta}'\mathbf{y}_{t-1}$  is called an error correction term.<sup>12</sup> Engle and Granger (1987) show that first difference stationary  $\mathbf{y}_t$  has a possibly infinite order error correction representation with a

<sup>12</sup>Johansen uses an error correction term  $\boldsymbol{\alpha}\boldsymbol{\beta}'\mathbf{y}_{t-p}$  instead of more conventional  $\boldsymbol{\alpha}\boldsymbol{\beta}'\mathbf{y}_{t-1}$ . However, these two representations can be shown to be equivalent.

nonzero  $\alpha$  under general regularity conditions if  $\mathbf{y}_t$  is cointegrated with  $r$  linear independent cointegrating vectors. The columns of  $\beta$  are these cointegrating vectors. It should be noted that Johansen's assumption that the error correction representation of finite order can be very restrictive in some applications. For example, Gregory, Pagan, and Smith (1993) show that linear quadratic economic models with adjustment costs imply moving average terms in the error correction representation. Phillips's (1991) ML estimation method may be useful in these circumstances.

Johansen makes an additional assumption that  $\epsilon_t$  is normally distributed and derives a maximum likelihood estimator for  $\beta$ . In his procedure, all parameters are jointly estimated and his estimators are asymptotically efficient. Another way to estimate an error correction representation is to use Engle and Granger's (1987) two step estimation method. In the first step, cointegrating vectors are estimated. For example, if there is only one linear independent cointegrating vector, it can be estimated by OLS. Other efficient estimators may be used in this first step. Then the rest of the parameters in the error correction representation are estimated in the second step. Since cointegrating vector estimators converge faster than  $\sqrt{T}$ , the first step estimation does not affect the asymptotic distributions of the second step estimators. In the second step, only stationary variables are involved, so standard econometric theory can be used. See 16.C for Johansen's maximum likelihood estimation and the cointegration rank test for detail.

Johansen's (1988; 1991) likelihood ratio tests and Stock and Watson's (1988a) tests for common trends are often used to determine the number of cointegrating vectors in a system. These tests take the null hypothesis that a  $n \times 1$  vector process  $\mathbf{y}_t$  has  $r \geq 0$  linear independent cointegrating vectors (or it has  $n - r$  common stochastic

trends) against the alternative that it has  $k > r$  linear independent cointegrating vectors (or it has  $n - k$  common stochastic trends). Hence if  $r = 0$ , these statistics test the null hypothesis of no cointegration against the alternative of cointegration.

Podivinsky's (1998) Monte Carlo results suggest that there can be severe size distortion problem with Johansen's tests when the sample size is small. For example, when there is no cointegrating vector in the data generation process and when asymptotic critical values are used, he finds a tendency for the test with the null hypothesis of  $r = 0$  to overreject and the test with the null hypothesis of  $r \leq 1$  to underreject.

## 16.8 How Should an Estimation Method be Chosen?

There exist many estimation and testing methods for cointegration. It is advisable for an applied researcher to try at least two methods and check sensitivity of empirical results. When the researcher chooses a main method to be used, the following considerations naturally come to mind.

### 16.8.1 Are Short-Run Dynamics of Interest?

If, in addition to cointegrating vectors, the short-run dynamics are of interest, then it seems (at least conceptually) natural to estimate short-run dynamics and cointegrating vectors simultaneously. For example, this process can be done by applying Johansen's ML method to estimate an error correction model.

On the other hand, the researcher is often interested in the cointegrating vector but not in short-run dynamics (see, e.g., Atkeson and Ogaki, 1996; Clarida, 1994, 1996; Ogaki, 1992). In such cases, it is desirable to avoid making unnecessary assumptions about short-run dynamics. An estimation method that uses a nonparametric

method to estimate long-run covariance parameters such as CCR is natural in these circumstances.

### 16.8.2 The Number of the Cointegrating Vectors

In some empirical applications, the researcher may have many economic variables and may not have any guidance from economic models about which variables may be cointegrated. In such applications, tests for the number of cointegrating vectors are useful. It should be noted, however, that these tests may not have very good small sample properties because of the near observational equivalence problem discussed in Section 13.5. For this reason, it is desirable to use economic models to give some a priori information about which variables should be cointegrated.

In some applications, an economic model implies that there exist two or more linearly independent cointegrating vectors. In this case of multiple cointegrating vectors in a cointegrating regression, neither OLS nor CCR can be used to identify cointegrating vectors. Tests for the null of cointegration based on CCR discussed above also assume that there is only one cointegrating vector and hence cannot be used. However, it is sometimes possible to use a priori information from economic models to handle multiple cointegrating vectors with the CCR methodology.<sup>13</sup> Johansen's ML method has an advantage that it allows multiple cointegrating vectors. However, as pointed out by Park (1990) and Pagan (1995) among others, cointegrating vectors may not be identified even by the Johansen's ML method.

---

<sup>13</sup>See Kakkar and Ogaki (1993) for an example of an empirical application.

### 16.8.3 Small Sample Properties

It is known that Johansen's ML estimates and test results can be very sensitive to the choice of the order of autoregression in empirical applications (see, e.g., Stock and Watson, 1993). Therefore, it is important to check sensitivity of empirical results with respect to the order of autoregression when Johansen's method is used. This sensitivity may be related to the fact that Johansen's estimator for a normalized cointegrating vector has a very large mean square error when the sample size is small (see Park and Ogaki, 1991). Gonzalo (1993) also reports this property even though he emphasizes that Johansen's estimator has good small sample properties when the sample size is increased. Podivinsky's (1998) result that Johansen's likelihood ratio tests have severe size distortion problems in some circumstances discussed in Section 16.7 may be due to these observations.

Park and Ogaki (1991) find that the CCR estimator typically has smaller mean square errors than Johansen's ML estimator when the prewhitening method is used. Han and Ogaki (1991) find that Park's tests for the null of cointegration have reasonable small sample properties.

To improve small sample properties of CCR estimators, iterations on the estimation of the long-run covariance parameters are recommended. In empirical applications of CCR, OLS is typically used as an initial estimator. Since OLS coincides with CCR when there is no correlation between the disturbance term and the first difference of the regressors at all leads and lags, the initial OLS may be called the first stage CCR. The second stage CCR is obtained from the long-run covariance parameters calculated from the first stage CCR estimates. The third stage CCR is obtained from the long-run covariance parameters calculated from the second stage CCR es-

timates, and so on. Park and Ogaki (1991) report that the small sample properties of the third stage CCR estimator are typically better than those of the second stage CCR estimator. On the other hand, the fourth stage CCR estimator sometimes had a significantly larger mean square error. For Park's tests for the null of cointegration to be consistent, it is necessary to bound both the eigenvalues of the VAR prewhitening coefficient matrices and the bandwidth parameter estimate. For example, while using the first order VAR for prewhitening, Han and Ogaki (1991) bound the singular values of the VAR coefficient matrix by 0.99 and the bandwidth parameter by the square root of the sample size. When the variables are cointegrated, the CCR estimators have better small sample properties without these bounds. Consequently, they recommend reporting the third stage CCR estimates without the bounds imposed and the fourth stage CCR test results with the bounds imposed.

## Appendix

### 16.A Estimation of the Model with Long-Run Restrictions

The three variable model in KPSW highlights a real-business-model with permanent productivity shocks. Under the assumption of constant returns to scale, a production function with stochastic trends can be described as

$$(16.A.1) \quad y_t = \log \lambda_t + 1 - \theta k_t$$

$$(16.A.2) \quad \log \lambda_t = \mu_\lambda + \log \lambda_{t-1} + \xi_t$$

where  $y_t$  and  $k_t$  denote output per capita and capital stock per capita, respectively, in logarithms. Total productivity,  $\lambda_t$ , follows a logarithmic random walk, and  $\xi_t$

is *iid* with mean zero and variance  $\sigma^2$ . Let  $c_t$  and  $i_t$  be consumption per capita and investment per capita, respectively. In the steady state, output, consumption and investment have the same growth rate of  $\frac{\mu_\lambda + \xi_t}{\theta}$  which can be interpreted as a common stochastic trend. Thus, the ‘great ratios’,  $c_t - y_t$  and  $i_t - y_t$ , follow stationary stochastic processes, implying  $y_t, c_t$  and  $i_t$  are cointegrated with one common trend, or equivalently, with two cointegrating relations. Therefore, there exists only one permanent innovation,  $v_{1t}^k$  that can be interpreted as a productivity shock,  $\xi_t$ . Let  $\mathbf{x}_t = (y_t, c_t, i_t)'$ , then  $\Phi(1)$  in (16.25) becomes

$$(16.A.3) \quad \Phi(1) = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

Since  $\Phi(1)$  is normalized, the first column in 16.A.3 captures the long run effects of a unit shock of  $v_t^1$ .<sup>14</sup> It is straightforward to estimate structural parameters following a scheme described in Section 16.3.1 where  $k = 1$ ,  $\hat{\mathbf{A}} = (1 \ 1 \ 1)'$  and  $\Pi = 1$ .

To incorporate nominal shocks, a six-variable model is considered in KPSW. First, money demand has the following relation

$$(16.A.4) \quad m_t - p_t = \beta_y y_t - \beta_R R_t + u_t$$

where  $m_t - p_t$  is the logarithm of real balances,  $R_t$  is the nominal interest rate, and  $u_t$  is the money-demand disturbance. Second, the Fisher equation is considered to introduce nominal shocks

$$(16.A.5) \quad R_t = r_t + E_t \Delta p_{t+1}$$

where  $r_t$  is the *ex ante* real interest rate and  $p_t$  is the logarithm of the price level. Six variables  $(y_t, c_t, i_t, m_t - p_t, R_t, \Delta p_t)$  follow an  $I(1)$  process and exhibit cointegrating

<sup>14</sup> $v_{1t}^k$  is equal to  $\frac{\xi_t}{\theta}$  so that standard deviation of  $v_{1t}^k$  is equal to  $\frac{\sigma}{\theta}$ .

relationships. It has already been shown that there are two cointegrating relations among three variables  $(y_t, c_t, i_t)$ . An additional cointegrating relationship is captured by the money demand equation in (16.A.4) provided that money-demand disturbance is stationary. Consequently, there exist three cointegrating relationships, reflecting that the system can be described by three stochastic common trends. Letting  $\mathbf{x}_t = (y_t, c_t, i_t, m_t - p_t, R_t, \Delta p_t)'$ , three permanent shocks consist of a real balance shock, a neutral inflation shock, and a real interest shock so that  $\mathbf{A}$  is constructed as

$$(16.A.6) \quad \mathbf{A} = \hat{\mathbf{A}}\mathbf{\Pi} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & \phi_1 \\ 1 & 0 & \phi_2 \\ \beta_y & -\beta_R & -\beta_R \\ 0 & 1 & 1 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ \pi_{21} & 1 & 0 \\ \pi_{31} & \pi_{32} & 1 \end{bmatrix}$$

KPSW assumed  $\hat{\mathbf{A}}$  to be known, and constructed the parameters in  $\hat{\mathbf{A}}$  by the estimates from Dynamic OLS in each cointegrating equation. It is notable that these two cointegrating relationships are used as  $c - y = \phi_1(R - \Delta p)$  and  $i - y = \phi_2(R - \Delta p)$  provided that the real interest rate follows a nonstationary process. This assumption implies that the ‘great ratios’ exhibit permanent shifts from a permanent real interest shock.<sup>15</sup> The issue on nonstationarity of real interest is in order. The null hypothesis that the *ex post* real interest rate<sup>16</sup> has a unit root is investigated using the Dickey-Fuller test, and is not rejected at the 10% significance level. This model is a benchmark in KPSW.

This property, in turn, implies that  $\phi_1$  and  $\phi_2$  are zero since regression of the  $I(0)$  variable on the  $I(1)$  variable gives the estimate of zero from the theoretical

<sup>15</sup>A higher real interest rate raises the consumption-output ratio and lowers the investment-output ratio, which implies that  $\phi_1$  is positive and  $\phi_2$  is negative.

<sup>16</sup>Three nominal interest rates are used in King *et al.* (1989); *three month U.S. Treasury bills, an average rate on four to six month commercial paper, and the yield on a portfolio of high-grade longer term corporate bonds.*

viewpoint.<sup>17</sup> KPSW also investigate sensitive analysis other than the benchmark model. First, the coefficients,  $\phi_1$  and  $\phi_2$ , are set equal to zero. This modification, however, does not affect the main results in the benchmark model. Second, assuming that real interest rates are stationary, a model with four cointegrating relationships is considered, where two stochastic common trends are interpreted as a real balance shock and a neutral inflation shock. In this case,  $\hat{\mathbf{A}}$  is constructed as

$$(16.A.7) \quad \hat{\mathbf{A}} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ \beta_y & -\beta_R \\ 0 & 1 \\ 0 & 1 \end{bmatrix}$$

The main conclusions, however, in the benchmark model are still robust after this modification.

This section explains how we can construct  $\hat{\mathbf{A}}$  from the estimates of cointegrating vectors. Engle and Granger (1987) showed:

$$(16.A.8) \quad \boldsymbol{\beta}'\boldsymbol{\Psi}(1) = \mathbf{0},$$

which by the property of cointegration implies that  $\boldsymbol{\beta}'\mathbf{x}_t$  is stationary. It follows from  $\boldsymbol{\Phi}(1) = \boldsymbol{\Psi}(1)\boldsymbol{\Phi}_0$  and (16.25) that

$$(16.A.9) \quad \boldsymbol{\beta}'\mathbf{A} = \mathbf{0} \quad \text{or} \quad \boldsymbol{\beta}'\hat{\mathbf{A}} = \mathbf{0}.$$

This property enables one to choose  $\hat{\mathbf{A}} = \boldsymbol{\beta}_\perp$  after re-ordering  $\mathbf{x}_t$  conformably with  $\boldsymbol{\beta}_\perp$ , in which  $\boldsymbol{\beta}_\perp$  is an  $n \times k$  orthogonal matrix of cointegrating vectors,  $\boldsymbol{\beta}$ , satisfying  $\boldsymbol{\beta}'\boldsymbol{\beta}_\perp = \mathbf{0}$ . Johansen (1995) proposed a method to choose  $\boldsymbol{\beta}_\perp$  by:

$$(16.A.10) \quad \boldsymbol{\beta}_\perp = (\mathbf{I}_n - \mathbf{S}(\boldsymbol{\beta}'\mathbf{S})^{-1}\boldsymbol{\beta}')\mathbf{S}_\perp,$$

<sup>17</sup> $\phi_1$  and  $\phi_2$  are estimated as 0.0033(0.0022) -0.0028(0.0050), respectively, where values in parentheses are standard errors, implying coefficients are not significantly different from zero.

where  $\mathbf{S}$  is an  $n \times r$  selection matrix,  $(\mathbf{I}_r \ \mathbf{0})'$ , and  $\mathbf{S}_\perp$  is an  $n \times k$  selection matrix,  $(\mathbf{0} \ \mathbf{I}_k)'$ . Note that  $\boldsymbol{\beta}$  is identified up to the space spanned by  $\boldsymbol{\alpha}$  and  $\boldsymbol{\beta}$ . This condition does not necessarily mean that each cointegrating vector is identified, because  $\boldsymbol{\alpha}\boldsymbol{\beta}' = \boldsymbol{\alpha}\mathbf{F}\mathbf{F}^{-1}\boldsymbol{\beta}' = \tilde{\boldsymbol{\alpha}}\tilde{\boldsymbol{\beta}}'$ , i.e., any linear combination of each cointegrating vector is a cointegrating vector. The model does not require the identification of each cointegrating vector. Park (1990) argues that the identification condition is not required a priori but is necessary for proper interpretation of the estimated results.

Since  $\boldsymbol{\beta}_\perp$  is normalized so that the last  $k \times k$  submatrix is an identity matrix, one should *re-arrange* the variables  $\mathbf{x}_t$  conformably in order to maintain Blanchard and Quah (1989)-type long-run restrictions. Alternatively, one may *re-normalize*  $\boldsymbol{\beta}_\perp$  as shown below. Consider the six-variable model in KPSW, for instance. Let  $\mathbf{x}_t$  be  $(y_t, c_t, i_t, m_t - p_t, R_t, \Delta p_t)'$ , in which  $m_t - p_t$  is the logarithm of the real balance,  $R_t$  is the nominal interest rate, and  $p_t$  is the logarithm of the price level. KPSW noted that there are three permanent shocks: a real balanced growth shock, a neutral inflation shock, and a real interest shock. We impose long-run restrictions that a neutral inflation shock has no long-run effect on output, and that a real interest rate shock has no long-run effect on either output or the inflation rate. These restrictions imply a specific form of  $\hat{\boldsymbol{\beta}}_\perp$  as in:

$$(16.A.11) \quad \mathbf{A} = \hat{\boldsymbol{\beta}}_\perp \boldsymbol{\Pi} = \begin{bmatrix} 1 & 0 & 0 \\ \times & \times & \times \\ \times & \times & \times \\ \times & \times & \times \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ \pi_{21} & 1 & 0 \\ \pi_{31} & \pi_{32} & 1 \end{bmatrix},$$

where  $\times$  denotes that those parameters are not restricted other than  $\boldsymbol{\beta}'\hat{\boldsymbol{\beta}}_\perp = 0$ . From

$\mathbf{A} = \hat{\mathbf{A}}\boldsymbol{\Pi}$ , we can choose  $\hat{\mathbf{A}}$  using:<sup>18</sup>

$$(16.A.12) \quad \hat{\mathbf{A}} = \hat{\boldsymbol{\beta}}_{\perp}.$$

## 16.B Monte Carlo Integration

The literature on confidence intervals for impulse response estimates is well explained by Kilian (1998), which can be categorized by the following three traditional methods: the asymptotic interval method (see Lütkepohl, 1990), the parametric Monte Carlo integration method (see Doan, 1992; Sims and Zha, 1999), and the nonparametric bootstrap interval method (see Runkle, 1987). We provide the Monte Carlo integration method used in KPSW.<sup>19</sup>

It is convenient to rewrite the reduce-form VECM in (16.17) as:

$$(16.B.13) \quad \begin{aligned} \Delta \mathbf{x}'_t &= \boldsymbol{\delta}'_{\epsilon} + \mathbf{x}'_{t-1} \boldsymbol{\beta} \boldsymbol{\alpha}' + \sum_{i=1}^{p-1} \Delta \mathbf{x}'_{t-i} \mathbf{A}_i^{*'} + \boldsymbol{\epsilon}'_t \\ &= \mathbf{X}'_t \boldsymbol{\theta}' + \boldsymbol{\epsilon}'_t \end{aligned}$$

where  $\mathbf{X}'_t = (1, \mathbf{x}'_{t-1} \boldsymbol{\beta}, \Delta \mathbf{x}'_{t-1}, \dots, \Delta \mathbf{x}'_{t-p+1})$ , and  $\boldsymbol{\theta}' = (\boldsymbol{\delta}_{\epsilon}, \boldsymbol{\alpha}, \mathbf{A}_1^*, \dots, \mathbf{A}_{p-1}^*)$ . Stacking (16.B.13) for  $t = 1, \dots, T$ , the model is represented by the following matrix form:

$$(16.B.14) \quad \mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \mathbf{U}$$

Assuming that  $u_t$  is i.i.d. and normally distributed, Zellner (1971) finds that  $\boldsymbol{\Sigma}$  follows the Normal-inverse Wishart posterior distribution, with the prior,  $f(\text{vec}(\boldsymbol{\theta}), \boldsymbol{\Sigma}) \sim |\boldsymbol{\Sigma}|^{-\frac{n+1}{2}}$ :

$$(16.B.15) \quad \boldsymbol{\Sigma}^{-1} \sim \text{Wishart}((T\boldsymbol{\Sigma}_0))^{-1}, T) \quad \text{with given } \boldsymbol{\Sigma}_0,$$

<sup>18</sup>KPSW, instead, assume that  $\hat{\mathbf{A}}$  is known *a priori*, which is estimated by dynamic OLS in each cointegrating equation.

<sup>19</sup>Kilian (1998) examines the accuracy of these confidence intervals in the small samples, and proposes the bootstrap-after-bootstrap method. He finds from Monte Carlo simulations that his method is the best, the Monte Carlo integration method is the second best, the asymptotic interval is the third, and the standard bootstrap interval method is the worst.

and

$$(16.B.16) \quad \boldsymbol{\theta} \sim N(\boldsymbol{\theta}_0, \boldsymbol{\Sigma} \otimes (\mathbf{X}'\mathbf{X})^{-1}),$$

where  $\boldsymbol{\theta}_0$  and  $\boldsymbol{\Sigma}_0$  are the estimates of  $\boldsymbol{\theta}$  and  $\boldsymbol{\Sigma}$ , respectively, from OLS or MLE.

The algorithm for estimating confidence intervals of impulse responses is as follows:

1. Estimate (16.17) and let  $\boldsymbol{\beta}_0$ ,  $\boldsymbol{\theta}_0$  and  $\boldsymbol{\Sigma}_0$  be these estimates.
2. Let  $\mathbf{A}$  be a lower triangular matrix of Choleski decomposition of  $(\mathbf{X}'\mathbf{X})^{-1}$ .
3. Let  $\mathbf{S}^{-1}$  be a lower triangular matrix of Choleski decomposition of  $\boldsymbol{\Sigma}_0^{-1}$ .
4. Generate  $n \times T$  random numbers,  $\mathbf{w}_b$ , from the normal distribution,  $N(0, \frac{1}{T})$ .
5. Generate  $(n(p-1) + r + 1) \times n$  random numbers,  $\mathbf{u}_b$ , from the standard normal distribution,  $N(0, 1)$ .
6. Let  $\mathbf{r}_b = \mathbf{w}_b' \mathbf{S}^{-1}$ , and get  $\boldsymbol{\Sigma}_b^{-1} = \mathbf{r}_b' \mathbf{r}_b$ .
7. Let  $\mathbf{S}_b$  be a lower triangular matrix of Choleski decomposition of  $\boldsymbol{\Sigma}_b$ .
8. Let  $\boldsymbol{\theta} = \boldsymbol{\theta}_0 + \mathbf{e}_b$ , in which  $\mathbf{e}_b = \mathbf{A} \mathbf{u}_b \mathbf{S}_b'$ . Then,  $\boldsymbol{\theta} \sim N(\boldsymbol{\theta}_0, \boldsymbol{\Sigma}_b \otimes (\mathbf{X}'\mathbf{X})^{-1})$ .<sup>20</sup>
9. Draw impulse responses,  $\mathbf{ir}_b$ , as described in Section 16.3.3.

---

<sup>20</sup>Note that  $\text{var}(\mathbf{e}_b) = \text{var}(\text{vec}(\mathbf{e}_b)) = \text{var}((\mathbf{S}_b \otimes \mathbf{A})\text{vec}(\mathbf{u}_b)) = \mathbf{S}_b \mathbf{S}_b' \otimes \mathbf{A} \mathbf{A}' = \boldsymbol{\Sigma}_b \otimes (\mathbf{X}'\mathbf{X})^{-1}$ . RATS uses  $\text{vec}(\mathbf{e}_b) = (\mathbf{S}_b \otimes \mathbf{I}_{n(p-1)+r+1})\text{vec}(\mathbf{A} \mathbf{u}_b)$ , which is the same as what this text uses. Note that  $(\mathbf{S}_b \otimes \mathbf{A})\text{vec}(\mathbf{u}_b) = \text{vec}(\mathbf{A} \mathbf{u}_b \mathbf{S}_b') = (\mathbf{S}_b \otimes \mathbf{I}_n)\text{vec}(\mathbf{A} \mathbf{u}_b)$ , in which  $\text{vec}(\mathbf{A} \mathbf{B} \mathbf{C}) = (\mathbf{C}' \otimes \mathbf{A})\text{vec}(\mathbf{B})$  is used for transformation.

10. Repeat 4 ~ 9,  $B$  times, and calculate 95% upper and lower bands of impulse responses using<sup>21</sup>

$$(16.B.17) \quad Upper = \frac{1}{B} \sum_{b=1}^B \mathbf{ir}_b + 2 \left( \frac{1}{B} \sum_{b=1}^B \mathbf{ir}_b^2 - \left( \frac{1}{B} \sum_{b=1}^B \mathbf{ir}_b \right)^2 \right)^{\frac{1}{2}}$$

and

$$(16.B.18) \quad Lower = \frac{1}{B} \sum_{b=1}^B \mathbf{ir}_b - 2 \left( \frac{1}{B} \sum_{b=1}^B \mathbf{ir}_b^2 - \left( \frac{1}{B} \sum_{b=1}^B \mathbf{ir}_b \right)^2 \right)^{\frac{1}{2}}.$$

## 16.C Johansen's Maximum Likelihood Estimation and Cointegration Rank Tests

To see Johansen's method in detail, consider the VAR(p) model

$$(16.C.19) \quad \mathbf{y}_t = \boldsymbol{\delta}_\epsilon + \mathbf{A}_1 \mathbf{y}_{t-1} + \cdots + \mathbf{A}_p \mathbf{y}_{t-p} + \boldsymbol{\epsilon}_t,$$

where  $\mathbf{y}_t$  is an  $n \times 1$  vector of variables assumed to be  $I(1)$ . If  $\mathbf{y}_t$  is cointegrated, then there exists the following VECM representation proposed by Engle and Granger (1987):

$$(16.C.20) \quad \Delta \mathbf{y}_t = \boldsymbol{\delta}_\epsilon + \boldsymbol{\alpha} \boldsymbol{\beta}' \mathbf{y}_{t-1} + \mathbf{A}_1^* \Delta \mathbf{y}_{t-1} + \cdots + \mathbf{A}_{p-1}^* \Delta \mathbf{y}_{t-p+1} + \boldsymbol{\epsilon}_t,$$

where  $\boldsymbol{\alpha}$  and  $\boldsymbol{\beta}$  have full column rank of  $r$ , the number of cointegrating vectors.

We can concentrate on  $\boldsymbol{\alpha}$  and  $\boldsymbol{\beta}$  from a partial regression:

$$(16.C.21) \quad \text{Regress } \Delta \mathbf{y}_t \text{ on } \mathbf{1}, \Delta \mathbf{y}_{t-1}, \cdots, \Delta \mathbf{y}_{t-p+1} \rightarrow \text{Get residuals: } \mathbf{R}_{0t}$$

$$(16.C.22) \quad \text{Regress } \mathbf{y}_{t-1} \text{ on } \mathbf{1}, \Delta \mathbf{y}_{t-1}, \cdots, \Delta \mathbf{y}_{t-p+1} \rightarrow \text{Get residuals: } \mathbf{R}_{kt}$$

Then, we have a concentrated regression:

$$(16.C.23) \quad \mathbf{R}_{0t} = \boldsymbol{\alpha} \boldsymbol{\beta}' \mathbf{R}_{kt} + \boldsymbol{\epsilon}_t$$

<sup>21</sup>Note that we fix cointegrating vectors,  $\boldsymbol{\beta}$ , and generate parameters from a normal distribution,  $N(\boldsymbol{\theta}_0, \boldsymbol{\Sigma}_b \otimes (\mathbf{X}'\mathbf{X})^{-1})$ . Note also that we do not update  $\mathbf{S}$ .

For notational convenience, let

$$(16.C.24) \quad \mathbf{S}_{ij} = \frac{1}{T} \sum_{t=1}^T \mathbf{R}_{it} \mathbf{R}'_{jt}, \quad i, j = 0, k$$

Note that  $\boldsymbol{\alpha}$  can be easily estimated from (16.C.23) provided that  $\boldsymbol{\beta}$  is known:

$$(16.C.25) \quad \begin{aligned} \hat{\boldsymbol{\alpha}}' &= (\boldsymbol{\beta}' \mathbf{R}'_k \mathbf{R}_k \boldsymbol{\beta})^{-1} \boldsymbol{\beta}' \mathbf{R}'_k \mathbf{R}_0 \\ &= (\boldsymbol{\beta}' \mathbf{S}_{kk} \boldsymbol{\beta})^{-1} \boldsymbol{\beta}' \mathbf{S}_{k0}. \end{aligned}$$

Johansen (1988) estimates  $\boldsymbol{\beta}$  using MLE. Consider MLE for

$$(16.C.26) \quad \mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{U}, \quad u_t \sim N(\mathbf{0}, \boldsymbol{\Sigma}).$$

Then, the log likelihood of (16.C.26) is

$$(16.C.27) \quad \log L = -\frac{T}{2} \log 2\pi - \frac{T}{2} \log |\boldsymbol{\Sigma}| - \frac{1}{2} (\mathbf{Y} - \mathbf{X}\mathbf{B})' \boldsymbol{\Sigma}^{-1} (\mathbf{Y} - \mathbf{X}\mathbf{B})$$

The FOC of (16.C.27) for  $\boldsymbol{\Sigma}$  is:

$$(16.C.28) \quad \hat{\boldsymbol{\Sigma}} = \frac{1}{T} (\mathbf{Y} - \mathbf{X}\mathbf{B})' (\mathbf{Y} - \mathbf{X}\mathbf{B})$$

Plug (16.C.28) in (16.C.27), then we get a concentrated likelihood:

$$(16.C.29) \quad \log L = \text{constant} - \frac{T}{2} \log |\hat{\boldsymbol{\Sigma}}|,$$

which is proportional to

$$(16.C.30) \quad L_{max} = |\hat{\boldsymbol{\Sigma}}|^{-\frac{T}{2}}.$$

Let  $L(\boldsymbol{\beta}) = |\hat{\boldsymbol{\Sigma}}|^{-\frac{T}{2}}$ . Then,

$$(16.C.31) \quad \begin{aligned} |L(\boldsymbol{\beta})|^{-\frac{2}{T}} &= |\hat{\boldsymbol{\Sigma}}| \\ &= \left| \frac{1}{T} (\mathbf{R}_0 - \mathbf{R}_k \boldsymbol{\beta} \boldsymbol{\alpha}')' (\mathbf{R}_0 - \mathbf{R}_k \boldsymbol{\beta} \boldsymbol{\alpha}') \right| \\ &= \left| \frac{1}{T} (\mathbf{R}_0 \mathbf{R}_0 - \boldsymbol{\alpha} \boldsymbol{\beta}' \mathbf{R}'_k \mathbf{R}_k \boldsymbol{\beta} \boldsymbol{\alpha}') \right| \\ &= |\mathbf{S}_{00} - \mathbf{S}_{0k} \boldsymbol{\beta} (\boldsymbol{\beta}' \mathbf{S}_{kk} \boldsymbol{\beta})^{-1} \boldsymbol{\beta}' \mathbf{S}_{k0}| \end{aligned}$$

So,

$$\begin{aligned}
 (16.C.32) \quad \max_{\beta} L(\beta) &\Leftrightarrow \min_{\beta} |\mathbf{S}_{00} - \mathbf{S}_{0k}\beta(\beta'\mathbf{S}_{kk}\beta)^{-1}\beta'\mathbf{S}_{k0}| \\
 &\Leftrightarrow \min_{\beta} |\beta'\mathbf{S}_{kk}\beta - \beta'\mathbf{S}_{k0}\mathbf{S}_{00}^{-1}\mathbf{S}_{0k}\beta| \frac{|\mathbf{S}_{00}|}{|\beta'\mathbf{S}_{kk}\beta|} \\
 &\Leftrightarrow \max_{\beta} \frac{|\beta'\mathbf{S}_{kk}\beta|}{|\beta'(\mathbf{S}_{kk} - \mathbf{S}_{k0}\mathbf{S}_{00}^{-1}\mathbf{S}_{0k})\beta|} \frac{1}{|\mathbf{S}_{00}|}
 \end{aligned}$$

At the second line, we use the following formula:

$$(16.C.33) \quad \begin{vmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{vmatrix} = |\mathbf{A}| |\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B}| = |\mathbf{D}| |\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C}|$$

Thus,

$$(16.C.34) \quad |\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C}| = |\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B}| \frac{|\mathbf{A}|}{|\mathbf{D}|},$$

where  $\mathbf{A} = \mathbf{S}_{00}$ ,  $\mathbf{B} = \mathbf{S}_{0k}\beta$ ,  $\mathbf{C} = \beta'\mathbf{S}_{k0}$ , and  $\mathbf{D} = \beta'\mathbf{S}_{kk}\beta$ . Note also that FOC for

$$(16.C.35) \quad \max_{\mathbf{x}} \frac{\mathbf{x}'\mathbf{A}\mathbf{x}}{\mathbf{x}'\mathbf{B}\mathbf{x}} \quad (\equiv \lambda)$$

is

$$(16.C.36) \quad (\mathbf{A} - \lambda\mathbf{B})\mathbf{x} = \mathbf{0},$$

where  $\lambda$  is an eigenvalue, and  $\mathbf{x}$  is an eigenvector. Therefore, (16.C.32) becomes an eigenvalue problem. Let

$$(16.C.37) \quad \lambda_0 = \max_{\beta} \frac{|\beta'\mathbf{S}_{kk}\beta|}{|\beta'(\mathbf{S}_{kk} - \mathbf{S}_{k0}\mathbf{S}_{00}^{-1}\mathbf{S}_{0k})\beta|}.$$

Then, the FOC is

$$\begin{aligned}
 (16.C.38) \quad &(\mathbf{S}_{kk} - \lambda_0(\mathbf{S}_{kk} - \mathbf{S}_{k0}\mathbf{S}_{00}^{-1}\mathbf{S}_{0k}))\beta = \mathbf{0} \\
 &\Leftrightarrow ((1 - \lambda_0)\mathbf{S}_{kk} + \lambda_0(\mathbf{S}_{k0}\mathbf{S}_{00}^{-1}\mathbf{S}_{0k}))\beta = \mathbf{0} \\
 &\Leftrightarrow (\lambda_0(\mathbf{S}_{k0}\mathbf{S}_{00}^{-1}\mathbf{S}_{0k}) - (\lambda_0 - 1)\mathbf{S}_{kk})\beta = \mathbf{0} \\
 &\Leftrightarrow (\mathbf{S}_{k0}\mathbf{S}_{00}^{-1}\mathbf{S}_{0k} - (1 - \frac{1}{\lambda_0})\mathbf{S}_{kk})\beta = \mathbf{0} \\
 &\Leftrightarrow (\mathbf{S}_{k0}\mathbf{S}_{00}^{-1}\mathbf{S}_{0k} - \lambda\mathbf{S}_{kk})\beta = \mathbf{0},
 \end{aligned}$$

where  $\lambda = 1 - \frac{1}{\lambda_0}$ . Note that  $\lambda$  and  $\boldsymbol{\beta}$  are an eigenvalue and an eigenvector of  $\mathbf{S}_{kk}^{-1}\mathbf{S}_{k0}\mathbf{S}_{00}^{-1}\mathbf{S}_{0k}$ , respectively. Therefore, our maximization problem is reduced to find an eigenvalue and eigenvector of  $\mathbf{S}_{kk}^{-1}\mathbf{S}_{k0}\mathbf{S}_{00}^{-1}\mathbf{S}_{0k}$ .

Having estimated the model, we can construct the cointegration rank tests as follows. From (16.C.30), (16.C.32) and (16.C.37), we get

$$(16.C.39) \quad |L_{max}(\boldsymbol{\beta})|^{-\frac{2}{T}} = |\mathbf{S}_{00}| \prod_{i=1}^r \frac{1}{\lambda_{0i}}$$

$$(16.C.40) \quad L_{max}(\boldsymbol{\beta}) = -\frac{T}{2} |\mathbf{S}_{00}| \prod_{i=1}^r (1 - \lambda_i)$$

Therefore, we get the LR test (or Trace test) as:

$$(16.C.41) \quad \begin{aligned} LR &= -2 \log \frac{L_{max}(H_0 = r)}{L_{max}(H_1 = n)} \\ &= -T \sum_{i=r+1}^n \log(1 - \lambda_i) \end{aligned}$$

and the maximum eigenvalue test (or  $\lambda_{max}$  test) as:

$$(16.C.42) \quad \begin{aligned} \lambda_{max} &= -2 \log \frac{L_{max}(H_0 = r)}{L_{max}(H_1 = r + 1)} \\ &= -T \log(1 - \lambda_{r+1}). \end{aligned}$$

Note that the alternative hypothesis is different in each test. For large values of test statistics, we reject the null hypothesis that there exist  $r$  cointegrating vectors,  $H_0 = r$ . Johansen (1995) gives the critical values, and Osterwald-Lenum (1992) provides revised critical values.

Johansen (1995) considers five models with respect to data properties as well as cointegrating relations as follows: i) a model with a quadratic trend in  $\mathbf{y}_t$  (hflag=1):

$$(16.C.43) \quad \Delta \mathbf{y}_t = \boldsymbol{\delta}_\epsilon + \boldsymbol{\rho}_0 t + \boldsymbol{\alpha} \boldsymbol{\beta}' \mathbf{y}_{t-1} + \mathbf{A}_1^* \Delta \mathbf{y}_{t-1} + \cdots + \mathbf{A}_{p-1}^* \Delta \mathbf{y}_{t-p+1} + \boldsymbol{\epsilon}_t,$$

ii) a model with a linear trend in  $\mathbf{y}_t$  (hflag=2), in which deterministic cointegration is not satisfied:

$$(16.C.44) \quad \Delta \mathbf{y}_t = \boldsymbol{\delta}_\epsilon + \boldsymbol{\rho}_0 t + \boldsymbol{\alpha} \boldsymbol{\beta}' \mathbf{y}_{t-1} + \mathbf{A}_1^* \Delta \mathbf{y}_{t-1} + \cdots + \mathbf{A}_{p-1}^* \Delta \mathbf{y}_{t-p+1} + \boldsymbol{\epsilon}_t,$$

iii) a model with a linear trend in  $\mathbf{y}_t$  (hflag=3), in which deterministic cointegration is satisfied (cotrended):

$$(16.C.45) \quad \Delta \mathbf{y}_t = \boldsymbol{\delta}_\epsilon + \boldsymbol{\alpha} (\boldsymbol{\beta}' \mathbf{y}_{t-1} + \boldsymbol{\rho}_1 t) + \mathbf{A}_1^* \Delta \mathbf{y}_{t-1} + \cdots + \mathbf{A}_{p-1}^* \Delta \mathbf{y}_{t-p+1} + \boldsymbol{\epsilon}_t,$$

iv) a model with no trend in  $\mathbf{y}_t$  (hflag=4):

$$(16.C.46) \quad \Delta \mathbf{y}_t = \boldsymbol{\alpha} (\boldsymbol{\beta}' \mathbf{y}_{t-1} + \boldsymbol{\rho}_0) + \mathbf{A}_1^* \Delta \mathbf{y}_{t-1} + \cdots + \mathbf{A}_{p-1}^* \Delta \mathbf{y}_{t-p+1} + \boldsymbol{\epsilon}_t,$$

and v) a model with no trend in  $\mathbf{y}_t$  (hflag=5):

$$(16.C.47) \quad \Delta \mathbf{y}_t = \boldsymbol{\alpha} \boldsymbol{\beta}' \mathbf{y}_{t-1} + \mathbf{A}_1^* \Delta \mathbf{y}_{t-1} + \cdots + \mathbf{A}_{p-1}^* \Delta \mathbf{y}_{t-p+1} + \boldsymbol{\epsilon}_t.$$

Johansen (1995) illustrates how to estimate restricted cointegrating vectors. Consider a trivariate model with two cointegrating vectors. Let  $\mathbf{y}_t = (\mathbf{y}_{1t}, \mathbf{y}_{2t}, \mathbf{y}_{3t})'$  and  $\boldsymbol{\beta} = [\boldsymbol{\beta}_1 | \boldsymbol{\beta}_2]$ . One may impose a restriction of  $\boldsymbol{\beta}_{11} = \boldsymbol{\beta}_{13}$  using  $\mathbf{H}_1 \boldsymbol{\varphi}_1 = \boldsymbol{\beta}_1$  and  $\mathbf{H}_2 \boldsymbol{\varphi}_2 = \boldsymbol{\beta}_2$ , where  $\mathbf{H}_i$  is an  $n \times (n - q_i)$  matrix,  $\boldsymbol{\varphi}_i$  is an  $(n - q_i) \times 1$  matrix, and  $q_i$  is the number of restrictions on each cointegrating vector. In this particular example, letting

$$(16.C.48) \quad \mathbf{H}_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ -1 & 0 \end{bmatrix}, \quad \mathbf{H}_2 = \mathbf{I}_3$$

gives the following restrictions:

$$(16.C.49) \quad \mathbf{H}_1 \boldsymbol{\varphi}_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ -1 & 0 \end{bmatrix} \begin{bmatrix} \varphi_{11} \\ \varphi_{12} \end{bmatrix} = \begin{bmatrix} \varphi_{11} \\ \varphi_{12} \\ -\varphi_{11} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\beta}_{11} \\ \boldsymbol{\beta}_{12} \\ \boldsymbol{\beta}_{13} \end{bmatrix}.$$

**References**

- ATKESON, A., AND M. OGAKI (1996): "Wealth-Varying Intertemporal Elasticities of Substitution: Evidence from Panel and Aggregate Data," *Journal of Monetary Economics*, 38, 507–534.
- BLANCHARD, O. J., AND D. QUAH (1989): "The Dynamic Effects of Aggregate Demand and Supply Disturbances," *American Economic Review*, 79(4), 655–673.
- CAMPBELL, J. Y., AND R. J. SHILLER (1988): "Interpreting Cointegrated Models," *Journal of Economic Dynamics and Control*, 12, 505–522.
- CLARIDA, R. H. (1994): "Cointegration, Aggregate Consumption, and the Demand for Imports: A Structural Econometric Investigation," *American Economic Review*, 84(1), 298–308.
- (1996): "Consumption, Import Prices, and the Demand for Imported Consumer Durables: A Structural Econometric Investigation," *Review of Economics and Statistics*, 78, 369–374.
- DOAN, T. A. (1992): *RATS User's Manual, Version 4*. Estima, Evanston, IL.
- ENGLE, R. F., AND C. GRANGER (1987): "Co-Integration and Error Correction: Representation, Estimation, and Testing," *Econometrica*, 55, 251–276.
- ENGLE, R. F., AND B. S. YOO (1987): "Forecasting and Testing in Co-Integrated Systems," *Journal of Econometrics*, 35(1), 143–159.
- FISHER, L. A., P. L. FACKLER, AND D. ORDEN (1995): "Long-run Identifying Restrictions for an Error-Correction Model of New Zealand Money, Prices and Output," *Journal of International Money and Finance*, 14(1), 127–147.
- GALÍ, J. (1992): "How Well Does the IS-LM Model Fit Postwar U.S. Data?," *Quarterly Journal of Economics*, 107(2), 709–738.
- (1999): "Technology, Employment, and the Business Cycle: Do Technology Shocks Explain Aggregate Fluctuations?," *American Economic Review*, 89(1), 249–271.
- GONZALO, J. (1993): "Cointegration and Aggregation," *Ricerche Economiche*, 47(3), 281–291.
- GREGORY, A. W., A. R. PAGAN, AND G. SMITH (1993): "Estimating Linear Quadratic Models with Integrated Processes," in *Models, Methods and Applications of Econometrics*, ed. by P. C. B. Phillips, pp. 220–239. Basil Blackwell, Oxford.
- HAMILTON, J. D. (1994): *Time Series Analysis*. Princeton University Press, Princeton.
- HAN, H.-L., AND M. OGAKI (1991): "Consumption, Income, and Cointegration: Further Analysis," RCER Working Papers No. 305.
- HANSEN, L. P., AND T. J. SARGENT (1980): "Formulating and Estimating Dynamic Linear Rational Expectations Models," *Journal of Economic Dynamics and Control*, 2(1), 7–46.
- (1982): "Instrumental Variables Procedures for Estimating Linear Rational Expectations Models," *Journal of Monetary Economics*, 9(3), 263–296.
- JANG, K. (2001a): "Impulse Response Analysis with Long Run Restrictions on Error Correction Models," Working Paper No. 01-04, Department of Economics, Ohio State University.

- (2001b): “A Structural Vector Error Correction Model with Short-Run and Long-Run Restrictions,” Manuscript.
- JANG, K., AND M. OGAKI (2001): “The Effects of Monetary Policy Shocks on Exchange Rates: A Structural Vector Error Correction Model Approach,” Working Paper No. 01-02, Department of Economics, Ohio State University.
- JOHANSEN, S. (1988): “Statistical Analysis of Cointegration Vectors,” *Journal of Economic Dynamics and Control*, 12, 231–254.
- (1991): “Estimation and Hypothesis Testing of Cointegration Vectors in Gaussian Vector Autoregressive Models,” *Econometrica*, 59(6), 1551–1580.
- (1995): *Likelihood-Based Inference in Cointegrated Vector Autoregressive Models*. Oxford University Press, Oxford.
- KAKKAR, V., AND M. OGAKI (1993): “Real Exchange Rates and Nontradables,” Manuscript.
- KILIAN, L. (1998): “Small-Sample Confidence Intervals for Impulse Response Functions,” *Review of Economics and Statistics*, 80(2), 218–230.
- KIM, H., AND M. OGAKI (2009): “Purchasing Power Parity and the Taylor Rule,” Working Paper No. 09-03, Department of Economics, Ohio State University.
- KIM, J. (2004): “Half-Lives of Deviations from PPP: Contrasting Traded and Nontraded Components of Consumption Baskets,” *Review of International Economics*, 12(1), 162–168.
- KIM, J., M. OGAKI, AND M.-S. YANG (2001): “Structural Error Correction Models: Instrumental Variables Methods and Application to an Exchange Rate Model,” Working Paper No. 01-01, Department of Economics, Ohio State University.
- (2007): “Structural Error Correction Models: A System Method for Linear Rational Expectations Models and an Application to an Exchange Rate Model,” *Journal of Money, Credit, and Banking*, 39(8), 2057–2075.
- KING, R. G., C. I. PLOSSER, J. H. STOCK, AND M. W. WATSON (1989): “Stochastic Trends and Economic Fluctuations,” Manuscript.
- (1991): “Stochastic Trends and Economic Fluctuations,” *American Economic Review*, 81(4), 810–840.
- LASTRAPES, W. D., AND G. SELGIN (1995): “The Liquidity Effect: Identifying Short-Run Interest Rate Dynamics Using Long-Run Restrictions,” *Journal of Macroeconomics*, 17(3), 387–404.
- LÜTKEPOHL, H. (1990): “Asymptotic Distributions of Impulse Response Functions and Forecast Error Variance Decompositions of Vector Autoregressive Models,” *Review of Economics and Statistics*, 72(1), 116–125.
- LÜTKEPOHL, H., AND H.-E. REIMERS (1992): “Impulse response analysis of cointegrated systems,” *Journal of Economic Dynamics and Control*, 16, 53–78.
- MELLANDER, E., A. VREDIN, AND A. WARNE (1992): “Stochastic Trends and Economic Fluctuations in a Small Open Economy,” *Journal of Applied Econometrics*, 7(4), 369–394.

- MUSSA, M. (1982): "A Model of Exchange Rate Dynamics," *Journal of Political Economy*, 90(1), 74–104.
- OGAKI, M. (1992): "Engel's Law and Cointegration," *Journal of Political Economy*, 100(5), 1027–1046.
- (1993): "Generalized Method of Moments: Econometric Applications," in *Handbook of Statistics: Econometrics*, ed. by G. S. Maddala, C. R. Rao, and H. D. Vinod, vol. 11, chap. 17, pp. 455–488. North-Holland, Amsterdam.
- OGAKI, M., AND J. Y. PARK (1997): "A Cointegration Approach to Estimating Preference Parameters," *Journal of Econometrics*, 82(1), 107–134.
- OSTERWALD-LENUM, M. (1992): "A Note with Quantiles of the Asymptotic Distribution of the Maximum Likelihood Cointegration Rank Test Statistics," *Oxford Bulletin of Economics and Statistics*, 54(3), 461–471.
- PAGAN, A. R. (1995): "Three Econometric Methodologies: An Update," in *Surveys in Econometrics*, ed. by L. T. Oxley, C. J. Roberts, D. A. R. George, and S. T. Sayer, pp. 30–41. Basil Blackwell.
- PAGAN, A. R., AND J. C. ROBERTSON (1995): "Structural Models of the Liquidity Effect," Manuscript.
- (1998): "Structural Models of the Liquidity Effect," *Review of Economics and Statistics*, 80, 202–217.
- PARK, J. Y. (1990): "Maximum Likelihood Estimation of Simultaneous Cointegrated Models," Manuscript, Cornell University.
- PARK, J. Y., AND M. OGAKI (1991): "VAR Prewhitening to Estimate Short-Run Dynamics: On Improved Method of Inference in Cointegrated Models," RCER Working Paper No. 281.
- PHILLIPS, P. C. B. (1991): "Optimal Inference in Cointegrated Systems," *Econometrica*, 59, 283–306.
- PODIVINSKY, J. M. (1998): "Testing Misspecified Cointegrating Relationships," *Economics Letters*, 60(1), 1–9.
- QURESHI, H. (2008): "Explosive Roots in Level Vector Autoregressive Models," Working Papers 08-02, Ohio State University, Department of Economics.
- RUNKLE, D. E. (1987): "Vector Autoregressions and Reality," *Journal of Business and Economic Statistics*, 5, 437–442.
- SHAPIRO, M. D., AND M. W. WATSON (1988): "Sources of Business Cycle Fluctuations (with Comments)," in *NBER Macroeconomic Annual*, ed. by S. Fischer, vol. 3, pp. 111–156. Cambridge: MIT Press.
- SIMS, C. A. (1980): "Macroeconomics and Reality," *Econometrica*, 48, 1–48.
- SIMS, C. A., AND T. ZHA (1999): "Error Bands for Impulse Responses," *Econometrica*, 67(5), 1113–1156.

- STOCK, J. H., AND M. W. WATSON (1988a): "Testing for Common Trends," *Journal of the American Statistical Association*, 83(404), 1097–1107.
- (1988b): "Variable Trends in Economic Time Series," *Journal of Economic Perspectives*, 2(3), 147–74.
- (1993): "A Simple Estimator of Cointegrating Vectors in Higher Order Integrated Systems," *Econometrica*, 61(4), 783–820.
- WARNE, A. (1991): "A Common Trends Model: Identification, Estimation and Asymptotics," Manuscript.
- WEST, K. D. (1987): "A Specification Test for Speculative Bubbles," *Quarterly Journal of Economics*, 102, 553–580.
- ZELLNER, A. (1971): *An Introduction to Bayesian Inference in Econometrics*. Wiley, New York.

# Chapter 17

## PANEL AND CROSS-SECTIONAL DATA

Many recent macroeconomic applications use panel and cross-sectional data. For example, macroeconomic hypotheses are tested in micro data sets for households, industries, and business firms and in aggregate data sets for many countries. This chapter focuses on econometric issues that are particularly relevant for macroeconomic applications.

### 17.1 Generalized Method of Moments

This section discusses GMM from the cross-sectional average rather than from the time series average as in Chapter 9. The method here can be applied to both cross-sectional and panel data with many cross-sectional observations and to those with a relatively small number of observations over time. Given cross-sectional data for  $\mathbf{x}_i$ , let  $\mathbf{b}_0$  be a  $p$ -dimensional vector of the parameters to be estimated, and  $f(\mathbf{x}_i, \mathbf{b})$  a  $q$ -dimensional vector of functions. We refer to  $\mathbf{u}_i = f(\mathbf{x}_i, \mathbf{b}_0)$  as the disturbance of GMM. We assume that  $\mathbf{x}_i$  is i.i.d. Consider the (unconditional) moment restrictions

$$(17.1) \quad E(f(\mathbf{x}_i, \mathbf{b}_0)) = 0.$$

Note that  $E(\mathbf{u}_i \mathbf{u}_j') = 0$  for  $i \neq j$ . Suppose that a law of large numbers can be applied to  $f(\mathbf{x}_i, \mathbf{b})$  for all admissible  $\mathbf{b}$ , so that the sample mean of  $f(\mathbf{x}_i, \mathbf{b})$  converges to its population mean:

$$(17.2) \quad \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^N f(\mathbf{x}_i, \mathbf{b}) = E(f(\mathbf{x}_i, \mathbf{b}))$$

with probability one (or in other words, almost surely). The basic idea of GMM estimation is to mimic the moment restrictions (17.2) by minimizing a quadratic form of the sample means

$$(17.3) \quad J_N(\mathbf{b}) = \left\{ \frac{1}{N} \sum_{i=1}^N f(\mathbf{x}_i, \mathbf{b}) \right\}' W_N \left\{ \frac{1}{N} \sum_{i=1}^N f(\mathbf{x}_i, \mathbf{b}) \right\}$$

with respect to  $\mathbf{b}$ ; where  $W_N$  is a positive definite matrix, which satisfies

$$(17.4) \quad \lim_{N \rightarrow \infty} W_N = W_0$$

with probability one for a positive definite matrix  $W_0$ . The matrices  $W_N$  and  $W_0$  are both referred to as the distance or weighting matrix. The GMM estimator,  $\mathbf{b}_N$ , is the solution of the minimization problem (17.3). Under fairly general regularity conditions, the GMM estimator  $\mathbf{b}_N$  is a consistent estimator for arbitrary distance matrices. The optimal choice of the distance matrix is  $W_0 = E(\mathbf{u}_i \mathbf{u}_i')^{-1}$ .

The GMM for cross-sectional data can be applied to panel data with large  $N$  and short  $T$  in order to allow for a general serial correlation structure. Let  $\mathbf{x}_{it}$  be a random vector of economic variables for an individual  $i$  at period  $t$  and  $f_t(\mathbf{x}_{it}, \mathbf{b})$  be a  $q^*$ -dimensional vector of functions, and let  $\mathbf{u}_{it} = f_t(\mathbf{x}_{it}, \mathbf{b}_0)$ . Let  $q = Tq^*$ ,  $\mathbf{x}_i = (\mathbf{x}'_{i1}, \dots, \mathbf{x}'_{iT})'$  and  $f(\mathbf{x}_i, \mathbf{b}) = (f_1(\mathbf{x}_{i1}, \mathbf{b})', \dots, f_T(\mathbf{x}_{iT}, \mathbf{b})')'$ . In this framework,

$E(\mathbf{u}_i \mathbf{u}'_i)$  can be estimated by  $\frac{1}{N} \sum_{i=1}^N \mathbf{u}_i \mathbf{u}'_i$ . Since

$$(17.5) \quad E(\mathbf{u}_i \mathbf{u}'_i) = \begin{bmatrix} E(\mathbf{u}_{i1} \mathbf{u}'_{i1}) & \cdots & E(\mathbf{u}_{i1} \mathbf{u}'_{iT}) \\ \vdots & & \vdots \\ E(\mathbf{u}_{iT} \mathbf{u}'_{i1}) & \cdots & E(\mathbf{u}_{iT} \mathbf{u}'_{iT}) \end{bmatrix},$$

where some entries of  $E(\mathbf{u}_i \mathbf{u}'_i)$  represent autocovariances of  $\mathbf{u}_{it}$ . Thus a general form of serial correlation is allowed by stacking disturbance terms with different dates as different disturbance terms rather than treating them as different observations of one disturbance term. Unlike GMM for the time series average in Chapter 9, there is no need to use kernel estimators to allow for a general form of serial correlation.

## 17.2 Tests of Risk Sharing

As in Chapter 7, consider an economy with a single good, in which the current and past values of a random vector  $\mathbf{x}_t$  generate the information set  $I_t$ , which is available to the economic agents. The random vector  $\mathbf{H}_t = [\mathbf{x}'_0, \mathbf{x}'_1, \dots, \mathbf{x}'_t]'$  summarizes  $I_t$ . Let  $Prob(\mathbf{H}_t)$  denote the probability of  $\mathbf{H}_t$ . For simplicity, we assume that the economy ends at date  $T$ , and that there exist  $N$  possible values of  $\mathbf{H}_T$ .

We assume that consumer  $h$  maximizes the lifetime utility function

$$(17.6) \quad U^h = \sum_{t=0}^T \sum_{\mathbf{H}_t} Prob(\mathbf{H}_t) \beta^t u(C_t^h(\mathbf{H}_t)),$$

where  $\beta$  is a discount factor,  $u(\cdot)$  is the utility function, and  $C_t^h(\mathbf{H}_t)$  is the consumption at date  $t$  with history  $\mathbf{H}_t$ . As a bench mark case, we assume that there exists a complete set of contingent security markets at date 0. Assuming the existence of a complete set of markets, we obtain

$$(17.7) \quad \frac{\beta^1 Prob(\mathbf{H}_{t+1}) mu(C_{t+1}^h(\mathbf{H}_{t+1}))}{Prob(\mathbf{H}_t) mu(C_t^h(\mathbf{H}_t))} = \frac{P_{t+1}(\mathbf{H}_{t+1})}{P_t(\mathbf{H}_t)}$$

which we call the *state-by-state intertemporal first order condition*.

The first order condition (17.7) implies that the ratio of the marginal utilities,  $\frac{mu(C_{t+1}^h(\mathbf{H}_{t+1}))}{mu(C_t^h(\mathbf{H}_t))}$  is identical for all consumers for all possible histories. When this condition is satisfied, consumers are said to be completely risk sharing. The hypothesis of complete risk sharing has been tested by Altug and Miller (1990), Deaton (1990), Cochrane (1991), Mace (1991), Townsend (1994), and Hayashi, Altonji, and Kotlikoff (1996) among others.

The implications of complete risk sharing on consumption depend on the functional form of the utility function. With an isoelastic utility function,  $u(C_t) = \frac{C_t^{1-\alpha}-1}{1-\alpha}$ ,  $mu(C_t^h) = (C_t^h)^{-\alpha}$ . Hence complete risk sharing implies that consumption growth,  $\frac{C_{t+1}^h(\mathbf{H}_{t+1})}{C_t^h(\mathbf{H}_t)}$ , is identical for all consumers for all possible histories. With a constant absolute risk aversion utility function,  $u_t(C_t) = \exp(\alpha C_t)$ ,  $mu_t = \alpha \exp(\alpha C_t)$ . Hence (17.7) implies that  $\exp(\alpha(C_{t+1}^h(\mathbf{H}_{t+1}) - C_t^h(\mathbf{H}_t)))$  is identical for all consumers in all possible histories. Therefore, complete risk sharing implies that the change in consumption is identical for all consumers.

These implications hold exactly without any errors. For tests of complete risk sharing with household data, errors are introduced either as preference shocks or measurement errors. Since measurement errors are likely to be important for household data, we consider them.

With the isoelastic utility function, assume that consumption is measured with multiplicative errors:  $C_t^{mh} = C_t^h e_t^h$ , where  $C_t^{mh}$  is the measured level of consumption. Then let  $\phi_t$  be the logarithm of the growth rate of consumption that is common to all consumers:  $\ln(C_{t+1}^h) - \ln(C_t^h) = \phi_t$ . Substituting  $\ln(C_t^h) = \ln(C_t^{mh}) - \ln(e_t^h)$  into

this equation, we obtain

$$(17.8) \quad \ln(C_{t+1}^{mh}) - \ln(C_t^{mh}) = \phi_t + e_t^h,$$

where  $e_t^h = -\ln(e_{t+1}^h) + \ln(e_t^h)$ . Consider the regression

$$(17.9) \quad \ln(C_{t+1}^{mh}) - \ln(C_t^{mh}) = bd_t + \mathbf{x}_t^{h'} \mathbf{a} + e_t^h,$$

where  $d_t$  is a time dummy and  $\mathbf{x}_t^h$  contains variables that are uncorrelated with the logarithm of the measurement error in consumption. Typically, income growth of consumer  $h$  is used as  $\mathbf{x}_t^h$ . Wealth, unemployment, and sickness are other examples. The null hypothesis of complete risk sharing can be tested by testing  $\mathbf{a} = \mathbf{0}$ .

With the exponential utility function, assume that consumption is measured with additive errors:  $C_t^{mh} = C_t^h + e_t^h$ , where  $C_t^{mh}$  is the measured level of consumption. Then let  $\phi_t$  be the common first difference of consumption. Then

$$(17.10) \quad C_{t+1}^{mh} - C_t^{mh} = \phi_t + e_t^h,$$

where  $e_t^h = -e_{t+1}^h + e_t^h$ . Consider the regression

$$(17.11) \quad C_{t+1}^{mh} - C_t^{mh} = bd_t + \mathbf{x}_t^{h'} \mathbf{a} + e_t^h,$$

where  $d_t$  is a time dummy and  $\mathbf{x}_t^h$  contains variables that are uncorrelated with the measurement errors in consumption. Then the null hypothesis of complete risk sharing can be tested by testing  $\mathbf{a} = \mathbf{0}$ .

### 17.3 Decreasing Relative Risk Aversion and Risk Sharing

Ogaki and Zhang (2001) argue that decreasing relative risk aversion is more plausible than constant relative risk aversion and increasing relative risk aversion. A parsimo-

nious parameterization of the utility function which contains decreasing, constant, and increasing relative risk aversion as special cases is

$$(17.12) \quad u(C_t) = \frac{1}{1-\alpha}((C_t - \gamma)^{1-\alpha} - 1)$$

which is called the Hyperbolic Absolute Risk Aversion (HARA) utility function.

Then the relative risk aversion coefficient is

$$(17.13) \quad -\frac{u''C_t^h}{u'}\alpha\left(1 - \frac{\gamma}{C_t^h}\right)^{-1}.$$

Thus relative risk aversion is decreasing (increasing) in consumption if  $\gamma$  is positive (negative).

For the HARA utility function  $mu(C_t^h) = (C_t^h - \gamma)^{-\alpha}$ . Hence the complete risk sharing hypothesis implies that  $C_t^h - \gamma$  grows at the same rate for all consumers. Let  $\phi_t$  be the common growth rate:

$$(17.14) \quad \frac{C_{t+1}^h - \gamma}{C_t^h - \gamma} = \phi_t.$$

Assume that consumption is measured with additive errors:  $C_t^{mh} = C_t^h + e_t^h$  where  $C_t^{mh}$  is the measured level of consumption. Multiplying both sides of (17.14) by  $C_t^h - \gamma$ , substituting  $C_t^h = C_t^{mh} - e_t^h$ , and rearranging terms, we obtain

$$(17.15) \quad C_{t+1}^{mh} - \phi_t C_t^{mh} + (\phi_t - 1)\gamma = \nu_t^h,$$

where

$$(17.16) \quad \nu_t^h = e_{t+1}^h - e_t^h.$$

Let  $\mathbf{z}_t^h$  be a vector of instrumental variables that are uncorrelated with the consumption measurement errors. Then GMM can be applied to the moment conditions that  $E(\mathbf{z}_t^h \nu_t^h) = \mathbf{0}$ .

## 17.4 Euler Equation Approach

As in Chapter 7, the state-by-state intertemporal first order condition can be used to derive the Euler equation

$$(17.17) \quad \frac{E(\beta mu(C_{t+1}^h)R_{t+1}|I_t)}{mu(C_t^h)} = 1$$

for any asset return,  $R_{t+1}$ .

Imagine that a panel data set of  $C_t^h$  and  $R_t$  is available for  $t = 1, \dots, T$  and  $h = 1, \dots, N$ . In order to estimate and test the Euler equation with the panel data set, it is important to distinguish the time average and the cross-sectional average. In many panel data sets,  $N$  is large but  $T$  is small. Chamberlain (1984) criticized the use of such a panel data set for the Euler equation approach by pointing out a difficulty in such applications. This difficulty is often referred to as *Chamberlain's critique*.

For example, assume that the intraperiod utility function is  $u(C_t) = \frac{C_t^{1-\alpha}-1}{1-\alpha}$ , so that  $mu(C_t^h) = (C_t^h)^{-\alpha}$ , and the Euler equation is

$$(17.18) \quad E[\beta(\frac{C_{t+1}^h}{C_t^h})^{-\alpha}R_{t+1}|I_t] = 1.$$

Removing the conditional expectation yields

$$(17.19) \quad \beta(\frac{C_{t+1}^h}{C_t^h})^{-\alpha}R_{t+1} - 1 = e_t^h,$$

where  $E(e_t^h|I_t) = 0$ . It should be noted that  $E(e_t^h|I_t) = 0$  does not imply that the probability limit of the cross-sectional average of  $e_t^h$  is zero even though it implies that the probability limit of the time-series average of  $e_t^h$  is zero. In order to see this, recall that consumption growth is identical for all consumers under complete risk

sharing. Hence (17.19) implies that  $e_t^h$  is identical, and  $\frac{1}{N} \sum_{h=1}^{\infty} e_t^h = e_t^1$  for any  $N$ . In a panel data set with large  $N$  and small  $T$ , an appropriate asymptotic theory fixes  $T$  and drives  $N$  to infinity to derive asymptotic results. In this example, the estimators based on  $E(e_t^h | I_t) = 0$  are inconsistent because  $\frac{1}{N} \sum_{h=1}^{\infty} e_t^h$  does not converge to zero in probability when  $N$  is driven to infinity. This example illustrates Chamberlain's critique.

## 17.5 Panel Unit Root Tests

Panel data allows researchers to effectively increase the number of observations. Levin, Lin, and Chu (2002) developed unit root tests for panel data. Their null hypothesis is that all series in the panel data are difference stationary against all series are stationary. Their test is a panel version of the Augmented Dickey-Fuller test. For a panel data set of a variable  $x_{i,t}$  for  $i = 1, \dots, N$  and  $t = 1, \dots, T$ , they consider  $N$  time series regressions of the form:

$$(17.20) \quad \Delta \tilde{x}_{i,t} = \theta_i + \mu_i t + \rho \tilde{x}_{i,t-1} + \beta_{i,1} \Delta \tilde{x}_{i,t-1} + \dots + \beta_{i,p} \Delta \tilde{x}_{i,t-p} + \nu_t,$$

where  $\tilde{x}_{i,t} = x_{i,t} - (1/N) \sum_{i=1}^N x_{i,t}$ . Here the cross-sectional average is subtracted from  $x_{i,t}$  in each period in order to take into account the cross-sectional dependence or a common time effect. It should be noted that they assumed that  $\rho$  is common to all  $i$  under both null and alternative hypotheses. Their test statistic, which is basically the t-statistic for  $\rho = 0$ , is called the adjusted t-statistic. When  $N$  and  $T$  go to infinity, the test statistic has an asymptotic standard normal distribution. Im, Pesaran, and Shin (2003) relaxed Levin and Lin's assumption that  $\rho$  is common to all  $i$ . Their test

is based on regressions

$$(17.21) \quad \Delta\tilde{x}_{i,t} = \theta_i + \mu_i t + \rho_i \tilde{x}_{i,t-1} + \beta_{i,1} \Delta\tilde{x}_{i,t-1} + \cdots + \beta_{i,p} \Delta\tilde{x}_{i,t-p} + \nu_t.$$

For their test, the null hypothesis is that  $\rho_i = 0$  for all  $i$ , and the alternative hypothesis is that  $\rho_i < 0$  at least one  $i$ . Their test statistic is based on the average of the  $t$ -statistics for the hypothesis that  $\rho_i = 0$ . Its asymptotic distribution is the standard normal distribution.

Maddala and Wu (1999) also relaxed Levin and Lin's assumption that  $\rho$  is common to all  $i$ . Their test statistic is based on the  $p$ -values and can be used for an unbalanced panel in which  $T$  is different for different  $i$ . However, this test is computationally more involved than the other two tests mentioned above because the  $p$ -values need to be computed by simulations for each application.

The alternative hypothesis of both Im, Pesaran, and Shin's and Maddala and Wu's tests is that at least one series is stationary. Therefore, rejection of the null hypothesis should not be regarded as evidence that all series are stationary unless there is a reason to believe that all series are either difference stationary or stationary.

Most panel unit root tests assume that the error terms are cross-sectionally uncorrelated. If this assumption is violated, then the tests can show severe size distortions (see, e.g., O'Connell, 1998). A certain degree of cross-sectional dependence can be removed by subtracting the cross-sectional mean for each time period. However, if the true cross-sectional dependence exhibits substantial heterogeneity, then this method will not work very well. Moreover, if the series share a common stochastic trend, then the subtraction of the cross-sectional mean can transform a difference stationary series into a stationary series. A recent work by Chang (2000) has solved this problem.

The tests described so far take difference stationarity as the null hypothesis. There are tests for the null hypothesis of stationarity for panel data. Nyblom and Harvey (2000) extended Kwiatkowski, Phillips, Schmidt, and Shin (1992, KPSS for short) test for stationarity to panel data. The null hypothesis is that all series in the panel are stationary, and the alternative hypothesis is that at least one of them is difference stationary. Choi (2000) extended Park and Choi's (1988) G test to panel data. Choi (2000) reports Monte Carlo results that the panel G test is more powerful than the panel KPSS test for most data generation processes.

## 17.6 Cointegration and Panel Data

Pedroni (2001) developed residual based tests for the null hypothesis of no cointegration for panel data while allowing for estimated slope coefficients to vary across individual members of panel. Pedroni (2000) and Phillips and Moon (1999) extended Phillips and Hansen's (1990) fully modified OLS estimator to panel data. Mark and Sul (2002) extended the dynamic OLS technique to panel data. The dynamic OLS estimator is much computationally simpler to calculate in the panel data setting. These estimators assume that the regression errors are cross-sectionally uncorrelated after removing common time effects. Seemingly unrelated cointegration techniques explained in Chapter 15 (see, e.g., Mark, Ogaki, and Sul, 2003) can be used to allow for a general form of cross-sectional dependence in regression errors. However, these techniques cannot be used when  $N$  is large because too many free parameters for cross-sectional dependence need to be estimated.

## Exercises

**17.1** Suppose that each consumer maximizes the identical lifetime utility function

$$(17.E.1) \quad U = \sum_{t=0}^T \sum_{e_t} Prob(e_t) \beta^t U(c_t)$$

at time 0 in an Arrow-Debreu complete market, where  $e_t = (s_0, \dots, s_t)$  is the history of the economy,  $s_t \in \{1, \dots, S\}$  is the state of the economy at  $t$ , and  $Prob(e_t)$  denotes the probability of  $e_t$  conditioned on  $e_0$ . The intra-period utility function is assumed to be

$$(17.E.2) \quad U(c_t) = \frac{\{c_t - \gamma\}^{1-\alpha} - 1}{1-\alpha}$$

where  $c_t$  is consumption at time  $t$

- (a) Write down a complete market budget constraint.
- (b) Derive a parameterized formula for a state-by-state intertemporal first order condition for  $c_t$  and  $c_{t+1}$ . Discuss the complete risk sharing implication of the first order condition. Then use the first order condition to derive an asset pricing formula for an asset that pays off  $d_{t+1}$  at  $t+1$  ( $d_{t+1}$  varies depending on  $e_{t+1}$ ).
- (c) Imagine that you have panel data set for  $\{c_t^h : t = 1, \dots, T, h = 1, \dots, N\}$  and real bond returns  $\{R_t : t = 1, \dots, T\}$  (without measurement error) in this village. Suppose that these variables are stationary. Discuss how you set up the GMM estimation to estimate  $\beta, \alpha$ , and  $\gamma$  in this case, assuming  $T = 200$  and  $N = 300$ . If  $T = 2$  and  $N = 300$ , do you think that you can use the GMM to estimate these parameters for this model? Explain your answer.

- (d) Now assume that there exist multiplicative measurement errors with unknown serial correlation in the consumption data  $\{c_t^h : t = 1, \dots, T, h = 1, \dots, N\}$  in this panel data set of the following form:

$$(17.E.3) \quad c_t^h - \gamma = (c_t^{h*} - \gamma)\epsilon_t^h$$

where  $c_t^{h*}$  is the true consumption and  $\ln(\epsilon_t^h)$  has mean zero and is uncorrelated across the consumers and with any income variables. Also assume that there are no asset return data and that  $T = 6$  and  $N = 300$ . Discuss how you set up GMM estimation to estimate  $\gamma$  (parameterized disturbance and weighting matrix). In particular, discuss why the expected value of the parameterized GMM disturbance is zero.

- (e) Now assume that there exist additive measurement errors with unknown serial correlation in the consumption data  $\{c_t^h : t = 1, \dots, T, h = 1, \dots, N\}$  in this panel data set of the following form:

$$(17.E.4) \quad c_t^h = c_t^{h*} + \epsilon_t^h$$

where  $c_t^{h*}$  is the true consumption and  $\epsilon_t^h$  has mean zero and is uncorrelated across consumers and with any income variables. Also assume that there are no asset return data and that  $T = 6$  and  $N = 300$ . Discuss how you set up GMM estimation to estimate  $\gamma$  in terms of the parameterized disturbance. In particular, discuss why the expected value of the parameterized GMM disturbance is zero. You do not have to explain the weighting matrix for this question.

## References

- ALTUG, S., AND R. A. MILLER (1990): "Household Choices in Equilibrium," *Econometrica*, 58, 543–570.

- CHAMBERLAIN, G. (1984): "Panel Data," in *Handbook of Econometrics*, ed. by Z. Griliches, and M. D. Intriligator, vol. II, chap. 22, pp. 1248–1318. North Holland, Amsterdam.
- CHANG, Y. (2000): "Bootstrap Unit Root Tests in Panels with Cross-Sectional Dependency," Manuscript.
- CHOI, C.-Y. (2000): "Panel Unit Root Tests Under the Null Hypothesis of Stationarity and Confirmatory Analysis with Applications to PPP and the Convergence Hypothesis," Ph.D. thesis, The Ohio State University.
- COCHRANE, J. H. (1991): "A Simple Test of Consumption Insurance," *Journal of Political Economy*, 99, 957–976.
- DEATON, A. (1990): "On Risk, Insurance, and Intra-Village Smoothing," Manuscript, Princeton University.
- HAYASHI, F., J. ALTONJI, AND L. KOTLIKOFF (1996): "Risk-Sharing Between and Within Families," *Econometrica*, 64, 261–294.
- IM, K. S., M. H. PESARAN, AND Y. SHIN (2003): "Testing for Unit Roots in Heterogeneous Panels," *Journal of Econometrics*, 115, 53–74.
- KWIATKOWSKI, D., P. C. B. PHILLIPS, P. SCHMIDT, AND Y. C. SHIN (1992): "Testing the Null Hypothesis of Stationarity against the Alternative of a Unit-Root - How Sure are We that Economic Time-Series Have a Unit-Root," *Journal of Econometrics*, 54(1–3), 159–178.
- LEVIN, A., C.-F. LIN, AND C.-S. J. CHU (2002): "Unit Root Tests in Panel Data: Asymptotic and Finite-Sample Properties," *Journal of Econometrics*, 108(1), 1–24.
- MACE, B. J. (1991): "Full Insurance in the Presence of Aggregate Uncertainty," *Journal of Political Economy*, 99, 928–956.
- MADDALA, G. S., AND S. WU (1999): "A Comparative Study of Unit Root Tests with Panel Data and a New Simple," *Oxford Bulletin of Economics and Statistics*, 61, 631–652.
- MARK, N. C., M. OGAKI, AND D. SUL (2003): "Dynamic Seemingly Unrelated Cointegrating Regression," NBER Technical Working Paper No. 292.
- MARK, N. C., AND D. SUL (2002): "Cointegration Vector Estimation by Panel DOLS and Long-Run Money Demand," NBER Technical Working Paper No. 287.
- NYBLOM, J., AND A. HARVEY (2000): "Testing for Common Stochastic Trends," *Econometric Theory*, 16, 176–199.
- O'CONNELL, P. G. J. (1998): "The Overvaluation of Purchasing Power Parity," *Journal of International Economics*, 44(1), 20.
- OGAKI, M., AND Q. ZHANG (2001): "Decreasing Relative Risk Aversion and Tests of Risk Sharing," *Econometrica*, 69(2), 515–526.
- PARK, J. Y., AND B. CHOI (1988): "A New Approach to Testing for a Unit Root," CAE Working Paper No. 88-23, Cornell University.
- PEDRONI, P. (2000): "Fully Modified OLS for Heterogeneous Cointegrated Panels," in *Advances in Econometrics*, ed. by B. H. Baltagi, vol. 15, pp. 93–130. Emerald Group Publishing Limited.

——— (2001): “Purchasing Power Parity Tests In Cointegrated Panels,” *Review of Economic Studies*, 83(4), 727–731.

PHILLIPS, P. C. B., AND B. E. HANSEN (1990): “Statistical Inference in Instrumental Variables Regression with I(1) Processes,” *Review of Economic Studies*, 57, 99–125.

PHILLIPS, P. C. B., AND H. R. MOON (1999): “Linear Regression Limit Theory for Nonstationary Panel Data,” *Econometrica*, 67(5), 1057–1112.

TOWNSEND, R. M. (1994): “Risk and Insurance in Village India,” *Econometrica*, 62, 539–591.

# Appendix A

## INTRODUCTION TO GAUSS

The purpose of this appendix is to give a quick introduction to GAUSS. For more complete information on GAUSS, see the GAUSS manual.

### A.1 Starting and Exiting GAUSS

#### A.1.1 The Windows Version

For the Windows version of GAUSS, click the icon  for GAUSS and you will be in the COMMAND mode of GAUSS. In the COMMAND mode, you can execute screen-resident programs.

To exit GAUSS, click  at the right upper corner.

#### A.1.2 The DOS Version

For the DOS version of GAUSS, from the DOS prompt, type GAUSS  to start GAUSS (for some versions of GAUSS, you may have to type GAUSSI or GAUSS386 instead of GAUSS). You will be in the COMMAND mode of GAUSS. You will see the GAUSS prompt, >>. In the COMMAND mode, you can execute screen-resident programs.

To exit the DOS version of GAUSS, press , then type  at the prompt.

## A.2 Running a Program Stored in a File from the COMMAND Mode

From the command mode type `RUN FILENAME.EXP` and hit  to run a file named `FILENAME.EXP`, for example.

## A.3 Editing a File

From the COMMAND mode, type `EDIT FILENAME.EXP` and hit  to edit a file named `FILENAME.EXP`, for example. You will be in the EDIT mode of GAUSS. You can edit the file with a full screen editor.

You can move around the file using arrow keys and   keys that are usually at the right of the keyboard. You can edit the file by deleting letters using  key and typing in letters.

## A.4 Rules of Syntax

This section lists some of the general rules of syntax for GAUSS programs.

### A.4.1 Statements

A GAUSS program consists of a series of statements. A statement is a complete expression or command. Statements in GAUSS end with a semicolon.

### A.4.2 Case

GAUSS does not distinguish between upper and lower case except inside double quotes.

### A.4.3 Comments

Comments can be placed inside `/*` and `*/`, which can nest other comments or inside `@` and `@`, which cannot nest other comments.

### A.4.4 Symbol Names

The names of matrices, strings, procedures, and functions can be up to eight characters long. The characters must be alphanumerical or the underscore. The first character must be alphabetic or an underscore. Note that you cannot use some names that are already used by GAUSS. It is often a good idea to use unusual names in your programs to avoid potential problems.

## A.5 Reading and Storing Data

```
LOAD X[n,m]=FILENAME.DAT;
```

reads in data stored in a ASCII file named FILENAME.DAT, for example. This data file should contain data separated by spaces in the form of an  $n \times m$  matrix.

If  $X$  is a matrix of numbers in GAUSS,

```
SAVE XFILE=X;
```

stores  $X$  into a file named XFILE.FMT. Then you can read in the data again by

```
LOAD X=XFILE;
```

## A.6 Operators

### A.6.1 Operators for Matrix Manipulations

*Assignment operator:*

```
Y=3;
```

assigns the value 3 to the  $1 \times 1$  matrix  $Y$ .

*Indexing operator:*

Brackets [ ] are used to index matrices. It is very important to note that parentheses ( ) are used for different purposes in GAUSS such as indicating the dimensions of a matrix or to take arguments for commands or functions.

```
Y=X[3,3];
```

assigns the 3-3 element of  $X$  to  $Y$ . Commas are used to separate row indices from column indices. A vector can take one argument.

*Period:*

Dots are used in brackets to signify “all rows” or “all columns”.

```
Y=X[.,3];
```

assigns the third column of  $X$  to  $Y$ .

*Colon:*

A colon is used within brackets to create a continuous range of indices.

```
Y=X[1:5, .];
```

*Transpose operator:*

' transposes matrices.

*Vertical Concatenation:*

| is used to concatenate two matrices vertically.

```
Z=X|Y;
```

*Horizontal Concatenation:*

$\sim$  is used to concatenate two matrices horizontally.

$$Z=X\sim Y;$$

## A.6.2 Numeric Operators

*Usual Operators:*

Usual operators in GAUSS work according to standard rules of matrix algebra. For example,  $*$  is the operator for matrix multiplication, and

$$Y=X*Z;$$

performs matrix multiplication when  $X$  and  $Z$  are conformable in the sense of matrix algebra.

Usual operators include  $*$ ,  $+$ ,  $-$ , and Kronecker product ( $.*$ ).  $y=x.*.z$  results in a matrix in which every element in  $x$  has been multiplied by the matrix  $z$ . For example,

$$x = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}, z = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \text{ then } x.*.z = \begin{bmatrix} 1 & 0 & 2 & 0 \\ 0 & 1 & 0 & 2 \\ 3 & 0 & 4 & 0 \\ 0 & 3 & 0 & 4 \end{bmatrix}.$$

*Element by Element Operators:*

In some applications,  $X$  is an  $m \times n$  matrix, and  $Y$  is an  $m \times 1$  vector, and it is convenient to multiply each row of  $Y$  with each of the  $n$  elements in each row of  $X$ . The Element by Element Operators allow you to perform such operations. For example,  $.*$  is the element by element multiplication operator, and

$$Z=X.*Y$$

performs the operation described above.

Other element by element operators are the following:

- ./ Element by element division:  
 $Y=X./Z;$
- ^ Element by element exponentiation:  
 $Y=X^Z;$
- + Element by element addition.  
 $Y=X+Z;$
- - Element by element subtraction.  
 $Y=X-Y;$

## A.7 Commands

### A.7.1 Functions

The following is a short list of useful functions. See the GAUSS manual for other useful functions.

- `cols(x)`: with a matrix  $x$  returns the number of columns of  $x$ .
- `diag(x)`: with an  $M \times M$  matrix  $x$  returns a column vector of the diagonal of  $x$ .
- `eye(N)`: returns an  $N \times N$  identity matrix.
- `ln(x)`: with an  $M \times N$  matrix  $x$  returns an  $M \times N$  matrix of the natural logarithm of all elements in  $x$ .
- `meanc(x)`: with an  $M \times N$  matrix  $x$  returns an  $N \times 1$  vector of the means of the columns of  $x$ .
- `int(x)`: with an  $M \times N$  matrix  $x$  returns the  $M \times N$  matrix of the largest integer which is smaller than or equal to each element of  $x$ .

- `invpd(x)`: with a symmetric, positive definite  $N \times N$  matrix  $x$  returns the inverse of  $x$ .
- `inv(x)`: with an  $N \times N$  matrix  $x$  returns the inverse of  $x$ .
- `ones(M,N)`: returns an  $M \times N$  matrix of ones. For example, `x=ones(3,2)`; will create  $x = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{bmatrix}$ .
- `sqrt(x)`: with an  $M \times N$  matrix  $x$  returns an  $M \times N$  matrix of the square roots of all elements of  $x$ .
- `reshape(x,r,c)`: with an  $N \times K$  matrix  $x$ , and two scalars  $r$  and  $c$ , returns an  $r \times c$  matrix created from the elements of  $x$ . The elements in  $x$  are first stored in row major order, and then the first  $c$  elements are put into the first row of the created matrix, the second in the second row, and so on. For example, `y=reshape(x,4,3)` for  $x = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 7 & 8 & 9 & 10 & 11 & 12 \end{bmatrix}$  creates  $y = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \\ 10 & 11 & 12 \end{bmatrix}$ .
- `rows(x)`: with a matrix  $x$  returns the number of rows of  $x$ .
- `zeros(M,N)`: returns an  $M \times N$  matrix of zeros.

### A.7.2 Printing

```
PRINT X Y;
```

will print matrices  $X$  and  $Y$  to the screen. Instead of matrices, you can print words inside double quotes:

```
PRINT "This will be printed";
```

You can use ? instead of PRINT:

```
? X Y;
```

### A.7.3 Preparing an Output File

```
OUTPUT FILE = FILENAME.OUT RESET;
```

allows you to write the output of PRINT statements to a file named FILENAME.OUT, for example. To print out or edit the output file, you have to close the output file by the

```
OUTPUT OFF;
```

command. Most of the programs in the GMM package contain this statement near their end. However, if your program does not reach its end because of errors, you have to issue this command from the COMMAND mode to close the file to check the output file.

## A.8 Procedure

## A.9 Examples

# Appendix B

## COMPLEX VARIABLES, THE SPECTRUM, AND LAG OPERATOR

In this Appendix, we review some basic results of complex variables, and their applications to the lag operator and spectral analysis. Section B.1 collects standard results of complex variables without proofs. Since most results in textbooks of complex variables are not relevant for our purpose, it is useful to collect the results used in macroeconometrics. Section B.2 gives examples of Hilbert spaces on  $\mathbb{C}$ . We will use a Hilbert space in this section in order to define the spectrum and give a foundation for using lag operator methods. Section B.3 uses these results to prove important results involving the lag operator such as convergence conditions for infinite order AR and MA representations and invertibility conditions of AR and MA representations. The results relating to a removable singular point in Section B.1 are used to derive the Beveridge-Nelson decomposition (Section 13.2) and the nonlinear restrictions by the linear rational expectations models presented in Chapter 16. Section B.3 presents some results for the spectrum, using the tools developed in Sections B.1 and B.2.<sup>1</sup>

---

<sup>1</sup>Some of the results in Sections B.2 and B.3 can be found in Sargent (1987). The main difference between Sargent's presentation and the presentation here lies in the difference in the convergence

## B.1 Complex Variables

### B.1.1 Complex Numbers

A complex number  $z = x + iy$  can be defined as ordered pairs  $(x, y)$  of real numbers, where  $i$  is a pure imaginary number that satisfies  $i^2 = -1$ . The real numbers  $x$  and  $y$  are known as the *real* and *imaginary parts* of  $z$ , respectively. It is natural to associate the complex number with a point in the plane whose Cartesian coordinates are  $x$  and  $y$ . In other words, each complex number corresponds to just one point. When used for the purpose of displaying the numbers  $z = x + iy$  geometrically, the  $xy$  plane is called the *complex plane*  $C$ . We denote the complex number which corresponds to the origin of the complex plane by  $0$ .

The *absolute value*, or *modulus*, of a complex number  $z = x + iy$  is defined as  $\sqrt{x^2 + y^2}$  and is denoted by  $|z|$ . The *complex conjugate* of a complex number  $z = x + iy$  is defined as the complex number  $x - iy$  and is denoted by  $\bar{z}$ . An important identity relating the conjugate of a complex number  $z$  to its absolute value is  $z\bar{z} = |z|^2$ .

A *circle with center at*  $z_0$  and radius  $\epsilon$  is  $\{z : z \text{ is complex number such that } |z - z_0| = \epsilon\}$ . The interior points of the circle are called the  $\epsilon$  *neighborhood* of  $z_0$ . For any real number  $\theta$ , it is convenient to define  $e^{i\theta}$ , or  $\exp(i\theta)$ , by

$$(B.1) \quad e^{i\theta} = \cos \theta + i \sin \theta.$$

Then  $\overline{e^{i\theta}} = \cos \theta - i \sin \theta = e^{-i\theta}$ , and  $|e^{i\theta}| = \sqrt{e^{i\theta} e^{-i\theta}} = 1$ . Hence  $e^{i\theta}$  represents the circle with the center at the origin and radius of one. This circle is called the *unit*

---

concept for the  $z$  transform. Our definition allows us to use the results in Section B.1 for the  $z$  transform, which can be used to prove various results such as the condition for invertibility of a lag polynomial in terms of the zeros of the  $z$  transform.

*circle.* We can express any complex number in *exponential form*:

$$(B.2) \quad z = re^{i\theta}.$$

### B.1.2 Analytic Functions

For a sequence of complex numbers  $\{z_i\}_{i=1}^{\infty}$  and an infinite series of complex numbers  $\sum_{i=1}^{\infty} z_i$ , convergence and divergence are defined in the same way as those of real numbers except that the distance for complex numbers is used for the definitions. The series  $\sum_{i=1}^{\infty} z_i$  is *absolutely convergent* if the series  $\sum_{i=1}^{\infty} |z_i|$  of real numbers converges. Absolute convergence of a series of complex numbers implies the convergence of that series.

A complex-valued *function*  $f$ , defined on a set of complex numbers  $D$ , assigns a complex number  $f(z)$  to each  $z$  in  $D$ . The set  $D$  is the *domain of definition* of  $f$ . A specific value of  $z$  for which  $f(z) = 0$  is called a *zero* of a function  $f$ .

If  $n$  is a nonnegative integer, and if  $a_0, a_1, a_2, \dots, a_n$  are complex constants, where  $a_n \neq 0$ , the function  $P(z) = a_0 + a_1z + a_2z^2 + \dots + a_nz^n$  is a *polynomial* of degree  $n$ . Any polynomial of degree  $n$  has precisely  $n$  zeros as in the following proposition:

**Proposition B.1.1** (*The Fundamental Theorem of Algebra*) For any polynomial of degree  $n$ ,  $P(z) = a_0 + a_1z + a_2z^2 + \dots + a_nz^n$  where  $n \geq 1$ , there exist  $n$  complex numbers  $z_1, z_2, \dots, z_n$ , such that

$$P(z) = a_n(z - z_1)(z - z_2) \cdots (z - z_n).$$



Here  $z_i$  is a zero of  $P(z)$ , and a root of  $P(z) = 0$ . Note that  $z_i$  may be equal to  $z_j$  for some  $j$ .

The limits, continuity, derivatives, and differentiability of functions are defined in the same way as those of real-valued functions of a real variable except that the distance for complex numbers is used for the definitions. For example, for a function  $f$  with domain  $S$

$$(B.3) \quad \lim_{z \rightarrow z_0} f(z) = w_0$$

means that for each positive number  $\epsilon$  there is a positive number  $\delta$  such that  $|f(z) - w_0| < \epsilon$  whenever  $0 < |z - z_0| < \delta$  and  $z \in S$ . Similarly, the *derivative* of  $f$  at  $z_0$  is defined by

$$(B.4) \quad f'(z_0) = \lim_{z \rightarrow z_0} \frac{f(z) - f(z_0)}{z - z_0},$$

provided this limit exists. The function  $f$  is said to be *differentiable* at  $z_0$  if its derivative at  $z_0$  exists. Since it is possible to approach  $z_0$  from many directions on the complex plane, the differentiability of functions of complex numbers is in a sense stricter than the differentiability of functions of real numbers as in the next example:

**Example B.1.1** Let  $f(z) = |z|^2$ . Churchill and Brown (1984, p.40) show that  $f(z)$  is differentiable only at the origin. For  $z = x + iy$ , let the real and imaginary parts of  $f(z)$  be  $u(x, y)$  and  $v(x, y)$ :  $f(z) = u(x, y) + iv(x, y)$ . Then  $u(x, y) = x^2 + y^2$ , and  $v(x, y) = 0$ . Hence even when the real and imaginary components of a function of a complex variable have continuous derivatives at  $z_0$ , the function may not be differentiable there. ■

Since the definition of a derivative in (B.4) is identical to that of the derivative of a real-valued function of a real variable, most of the basic differentiation formulas

remain valid for functions of complex variables. For example, if  $n$  is a positive integer,  $\frac{dz^n}{dz} = nz^{n-1}$ . This formula remains valid when  $n$  is a negative integer as long as  $z \neq 0$ .

A function  $f$  of the complex number  $z$  is *analytic* at a point  $z_0$  if its derivative exists not only at  $z_0$  but also at each point  $z$  in some neighborhood of  $z_0$ . An *entire function* is a function that is analytic at each point in the entire complex plane. Every polynomial is an *entire function*.

If two functions  $f(z)$  and  $g(z)$  are analytic in a domain  $D$ , then their sum and their product are both analytic in  $D$ . The quotient  $\frac{f(z)}{g(z)}$  is also analytic in  $D$  provided that  $g(z) \neq 0$  for any  $z$  in  $D$ . Hence the quotient  $\frac{P(z)}{Q(z)}$  of two polynomials is analytic in any domain throughout which  $Q(z) \neq 0$ .

The following three propositions are important for our purposes. See Churchill and Brown (1984, p.113, p.126, and p.153) for proofs.

**Proposition B.1.2** Let a function  $f$  be analytic at a point  $z_0$  of  $f$ . There is a neighborhood of  $z_0$  throughout which  $f$  has no other zeros, unless  $f$  is identically zero. That is, the zeros of an analytic function which is not identically zero are isolated. ■

**Proposition B.1.3** If a function  $f$  is analytic at a point, then its derivatives of all orders exist and are themselves analytic there. ■

**Proposition B.1.4** (*Taylor's Theorem*) Let  $f$  be analytic everywhere inside a circle  $C$  with center at  $z$  and radius  $R$ . Then at each point  $z$  inside  $C$ ,

$$(B.5) \quad \begin{aligned} f(z) &= f(z_0) + \frac{f'(z_0)}{1!}(z - z_0) + \frac{f''(z_0)}{2!}(z - z_0)^2 + \cdots \\ &+ \frac{f^{(n)}(z_0)}{n!}(z - z_0)^n + \cdots \end{aligned}$$

■

The special case of series (B.5) when  $z_0 = 0$  is called the *Maclaurin series*.

**Example B.1.2** This example provides a Maclaurin series representation. Let  $f(z) = \frac{1}{1-az}$  for a nonzero real number  $a$ . Then  $f(z)$  is analytic on the complex plane except for  $z = a^{-1}$ .

$$(B.6) \quad f^{(n)}(z) = \frac{n!a^n}{(1-az)^{n+1}}$$

At each point  $z$  such that  $|z| < |a^{-1}|$ ,

$$\frac{1}{1-az} = 1 + az + a^2z^2 + \cdots + a^nz^n + \cdots .$$

■

Let  $S(z)$  be a power series:

$$S(z) = \sum_{n=0}^{\infty} a_n z^n .$$

See Churchill and Brown (1984, p.137 and p.143) for proofs of the following two propositions.

**Proposition B.1.5** If the power series converges when  $z = z_1$  ( $z_1 \neq 0$ ), it is absolutely convergent for every value of  $z$  such that  $|z| < |z_1|$ . ■

The greatest circle about the origin such that the series converges at each point inside is called the *circle of convergence* of the power series.

**Proposition B.1.6** The power series  $S(z)$  represents a function that is analytic at every point in the interior of its circle of convergence. ■

If  $S(z)$  converges for  $z$  such that  $|z| < R$ , then  $S(z - z_0)$  is analytic for  $z$  such that  $|z - z_0| < R$  because it is a composite function of two analytic functions.

When  $f(z)$  is analytic for all  $z$  such that  $|z - z_0| < R$  but fails to be analytic at  $z_0$ , then we cannot apply Taylor's theorem at that point. However, we can find a series representation for  $f(z)$  involving both positive and negative powers of  $z - z_0$ . If  $f(z)$  is analytic in the domain of all points  $z$  such that  $R_1 \leq |z - z_0| \leq R_2$ , then

$$(B.7) \quad f(z) = \sum_{n=0}^{\infty} a_n(z - z_0)^n + \sum_{n=0}^{\infty} b_n(z - z_0)^{-n}$$

in the domain. The series here is called a *Laurent series*. A series representation of this type is unique (see Churchill and Brown, 1984, pp. 132-134 and p.148).

When all the coefficients  $b_n$  in (B.7) are zero, the point  $z_0$  is called a *removable singular point* of  $f$ . In this case, the Laurent series (B.7) contains only nonnegative powers of  $z - z_0$ . If we define  $f(z)$  as  $a_0$  at  $z_0$ , the function becomes analytic at  $z_0$ .

Suppose that a function can be written in the form

$$(B.8) \quad f(z) = \frac{g(z)}{z - z_0},$$

where  $g(z)$  is analytic everywhere inside a circle  $C$  with center at  $z_0$  and radius  $R$ . Then at each point  $z$  inside  $C$ ,  $f(z)$  is analytic for all  $z$  except for  $z = z_0$ . From the Taylor series

$$(B.9) \quad \begin{aligned} g(z) &= g(z_0) + \frac{g'(z_0)}{1!}(z - z_0) + \frac{g''(z_0)}{2!}(z - z_0)^2 + \dots \\ &+ \frac{g^{(n)}(z_0)}{n!}(z - z_0)^n + \dots \end{aligned}$$

It follows that

$$(B.10) \quad \begin{aligned} f(z) &= \frac{g(z_0)}{z - z_0} + \frac{g'(z_0)}{1!} + \frac{g''(z_0)}{2!}(z - z_0) + \dots \\ &+ \frac{g^{(n)}(z_0)}{n!}(z - z_0)^{n-1} + \dots \end{aligned}$$

Then  $a$  is a nonzero real number, and  $S(z)$  is a polynomial or a power series which converges for all  $z$  such that  $|z| < R$  for some  $R$ . Hence if  $g(z_0) = 0$ , then  $z_0$  is a removable singular point of  $f(z)$ .

## B.2 Hilbert Spaces on $C$

In Appendix 3.A, it was noted that the complex plane,  $C$ , can be used as the set of scalars  $K$  for a vector space, and therefore for a Hilbert space. This section gives examples of Hilbert spaces for which  $K = C$ . The space of complex-valued random variables explained in Example B.2.4 and  $L^2(Prob)$  of real-valued random variables explained in Appendix 3.A are the two Hilbert spaces we use in this book.

**Example B.2.1** The complex plane,  $C$ , is a vector space on  $K = C$  with addition and scalar multiplication defined in the usual way. When the norm of a complex number is defined as its absolute value,  $C$  is a Banach space. When the inner product is defined as  $(x|y) = x\bar{y}$ ,  $C$  is a Hilbert space. ■

**Example B.2.2** Vectors in the space consist of sequences of  $n$  complex numbers,  $C^n$ , is a vector space on  $C$  when  $\mathbf{x} + \mathbf{y}$  for  $\mathbf{x} = (x_1, x_2, \dots, x_n)'$  and  $\mathbf{y} = (y_1, y_2, \dots, y_n)'$  is defined by  $(x_1 + y_1, x_2 + y_2, \dots, x_n + y_n)'$  and  $\alpha\mathbf{x}$  for  $\alpha$  in  $C$  is defined by  $(\alpha x_1, \alpha x_2, \dots, \alpha x_n)'$ . When we define a norm of  $\mathbf{x}$  as  $\|\mathbf{x}\| = \sqrt{\sum_{i=1}^n |x_i|^2}$ ,  $C^n$  is a Banach space. When we define  $(\mathbf{x}|\mathbf{y}) = \sum_{i=1}^n x_i \bar{y}_i$ ,  $C^n$  is a Hilbert space on  $C$ . ■

**Example B.2.3** The space  $l_2$  consists of all sequences of complex numbers  $\{x_1, x_2, \dots\}$  for which  $\sum_{i=1}^{\infty} |x_i|^2 < \infty$ . The inner product of elements  $\mathbf{x} = \{x_1, x_2, \dots\}$  and  $\mathbf{y} = \{y_1, y_2, \dots\}$  in  $l_2$  is defined as  $(\mathbf{x}|\mathbf{y}) = \sum_{i=1}^{\infty} x_i \bar{y}_i$ . With this inner product,  $l_2$  is a Hilbert space on  $C$ . ■

**Example B.2.4** On the interval  $[-\pi, \pi]$ , use the uniform distribution to define the probability of the  $\sigma$ -field of the Borel sets in the interval. On this probability space, consider a complex-valued random variable  $z = x + iy$  where  $x$  and  $y$  are real-valued random variables on  $[-\pi, \pi]$ . Define

$$E(z) = E(x) + iE(y) = \frac{1}{2\pi} \int_{-\pi}^{\pi} x(\lambda) d\lambda + i \frac{1}{2\pi} \int_{-\pi}^{\pi} y(\lambda) d\lambda$$

Let  $L^2[-\pi, \pi] = \{h: h \text{ is a complex valued random variable on } [-\pi, \pi] \text{ and } E(|h|^2) < \infty\}$ . Then with an inner product defined by  $(h_1|h_2) = E(h_1\bar{h}_2)$ ,  $L^2[-\pi, \pi]$  is a Hilbert space. As in  $L^2(Prob)$ , if two different random variables  $h_1$  and  $h_2$  satisfy  $E[|h_1 - h_2|^2] = 0$ , then we view  $h_1$  and  $h_2$  as the same element in this space.<sup>2</sup> ■

## B.3 Spectrum

This section defines the spectral density for a linearly regular covariance stationary process. We will first consider stochastic processes of random variables. Then we will consider stochastic processes of random vectors.

Imagine that we are given a white noise process  $\{v\}_{t=-\infty}^{\infty}$  on a probability space  $(S, \mathcal{F}, Prob)$  that satisfies  $E(v_t^2) = \sigma_v^2$  and  $E(v_t v_s) = 0$  for  $t \neq s$ . It is convenient to normalize this white noise process by defining  $e_t = \frac{v_t}{\sigma_v}$ . Then  $\{e_t\}_{t=-\infty}^{\infty}$  is an orthonormal sequence in  $L^2(Prob)$  because it satisfies  $\|e_t\| = \sqrt{E(e_t^2)} = 1$  and  $(e_t|e_s) = E(e_t e_s) = 0$  for  $t \neq s$ . Let  $b(L) = b_0 + b_1 L + b_2 L^2 + \dots$  be a series in the lag operator. Then from Proposition 3.A.5,  $b(L)e_t$  converges to an element in  $L^2(Prob)$  if and only if  $\{b_j\}_{j=1}^{\infty}$  is square summable, that is,  $\sum_{j=1}^{\infty} |b_j|^2 < \infty$ .

<sup>2</sup>For our purpose, it is convenient to view an element of  $L^2[-\pi, \pi]$  as a complex-valued random variable when the uniform distribution is given on  $[-\pi, \pi]$ . In many books, this interpretation is not given, and elements in  $L^2[-\pi, \pi]$  are considered complex-valued functions,  $f$ , which are measurable on  $[-\pi, \pi]$ .

Given an orthonormal sequence  $\{e_t\}_{t=-\infty}^{\infty}$  in  $L^2(Prob)$ , imagine that we are interested in certain properties of  $b(L)e_t$  for various series in the lag operator  $b(L) = b_0 + b_1L + b_2L^2 + \dots$  such as convergence of  $b(L)e_t$  and the autocovariance of  $b(L)e_t$ . As long as these properties do not depend on the probability space, we can choose a probability space that makes studying these properties convenient. As we will see, it is convenient to consider a random variable and an orthonormal sequence in  $L^2[-\pi, \pi]$  in which the probability is given by the uniform distribution on  $[-\pi, \pi]$ .

For this purpose, we consider a sequence  $\{u_t\}_{t=-\infty}^{\infty}$  in  $L^2[-\pi, \pi]$  where  $u_t(\lambda) = \exp(-i\lambda t) = \cos(\lambda t) - i \sin(\lambda t)$ . Then  $|u_t(\lambda)| = 1$  for all  $\lambda$  in  $[-\pi, \pi]$ , so that  $\|u\| = \sqrt{E(|u_t|^2)} = 1$ . If  $t \neq s$ ,

$$\begin{aligned}
 \text{(B.11)} \quad (u_t|u_s) &= \frac{1}{2\pi} \int_{-\pi}^{\pi} \exp(-i\lambda t) \overline{\exp(-i\lambda s)} d\lambda \\
 &= \frac{1}{2\pi} \int_{-\pi}^{\pi} \exp(-i\lambda t) \exp(i\lambda s) d\lambda \\
 &= \frac{1}{2\pi} \int_{-\pi}^{\pi} \exp(i\lambda(s-t)) d\lambda \\
 &= \frac{1}{2\pi} \int_{-\pi}^{\pi} [\cos(\lambda(s-t)) + i \sin(\lambda(s-t))] d\lambda \\
 &= 0.
 \end{aligned}$$

Thus  $\{u_t\}_{t=-\infty}^{\infty}$  is an orthonormal sequence.

Given  $b(L)e_t = \sum_{j=0}^{\infty} b_j e_{t-j}$ , consider a process  $b(L)u_t = \sum_{j=0}^{\infty} b_j u_{t-j}$  in  $L^2[-\pi, \pi]$ . From Proposition 3.A.5,  $b(L)e_t$  and  $b(L)u_t$  converge if and only if  $\{b_j\}$  is square summable. Hence  $b(L)e_t$  converges if and only if  $b(L)u_t$  converges.

Let  $M$  be the closed subspace in  $L^2[-\pi, \pi]$  generated by  $\{u_t\}_{t=-\infty}^{\infty}$ . From Propo-

sition 3.A.6, for any element  $y$  in  $M$ ,

$$(B.12) \quad y = \sum_{j=0}^{\infty} c_j \exp(i\lambda j)$$

$$(B.13) \quad c_j = \frac{1}{2\pi} \int_{-\pi}^{\pi} y(\lambda) \exp(-i\lambda j) d\lambda$$

where  $c_j$  is the Fourier coefficient for  $u(-j) = \exp(i\lambda j)$  and  $\{c_j\}$  is square summable.

When  $\{b_j\}_{j=0}^{\infty}$  is square summable, let  $x_t = b(L)e_t$ . Then the autocovariance  $\Phi(k) = E(x_t x_{t-k})$  is given by  $\Phi(k) = \lim_{n \rightarrow \infty} E \left[ (\sum_{j=0}^n b_j e_{t-j}) (\sum_{j=0}^n b_j e_{t-k-j}) \right] = \sum_{j=k}^{\infty} b_j b_{j-k}$ , where the last equality can be proved by the continuity of the inner product (Proposition 3.A.2).

Define the autocovariance of order  $k$ ,  $\Phi(k)$  for  $h_t = \sum_{j=0}^{\infty} b_j \exp(-i\lambda(t-j)) = \sum_{j=0}^{\infty} b_j \exp(i\lambda j) \exp(-i\lambda t) = h_0 \exp(-i\lambda t)$  as

$$(B.14) \quad \Phi(k) = E(h_t \bar{h}_{t-k}).$$

Then  $\Phi(k) = \sum_{j=k}^{\infty} b_j b_{j-k}$ . Thus the autocovariance of  $h_t$  coincides with that of  $x_t$ .

A simple expression for  $\Phi(k)$  can be obtained in  $L^2[-\pi, \pi]$ :

$$\begin{aligned} \Phi(k) &= E(h_0 \bar{h}_{0-k}) \\ &= \frac{1}{2\pi} \int_{-\pi}^{\pi} h_0(\lambda) \overline{h_0(\lambda)} \exp(-i\lambda k) d\lambda \\ &= \frac{1}{2\pi} \int_{-\pi}^{\pi} h_0(\lambda) \overline{h_0(\lambda)} \exp(i\lambda k) d\lambda \end{aligned}$$

Hence

$$(B.15) \quad \Phi(k) = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(\lambda) \exp(i\lambda k) d\lambda$$

where

$$(B.16) \quad \begin{aligned} f(\lambda) &= h_0(\lambda) \overline{h_0(\lambda)} \\ &= \left[ \sum_{j=0}^{\infty} b_j \exp(i\lambda j) \right] \left[ \sum_{j=0}^{\infty} b_j \exp(-i\lambda j) \right] \end{aligned}$$

is the spectral density.

For a vector process  $\mathbf{x}_t = B(L)\mathbf{v}_t = \sum_{j=0}^{\infty} B_j \mathbf{v}_{t-j}$  where  $\mathbf{x}_t$  and  $\mathbf{v}_t$  are  $p \times 1$  vectors and  $B_j$  is a  $p \times p$  matrix, we consider a matrix process  $\mathbf{h}_t = \sum_{j=0}^{\infty} B_j \exp(i\lambda(t-j))$ . Then define  $\Phi(k) = E(\mathbf{x}_t \mathbf{x}'_{t-k})$  for  $\mathbf{x}_t$  and  $\Phi(k) = E(\mathbf{h}_t \overline{\mathbf{h}'_{t-k}})$  for  $\mathbf{h}_t$ . Then for both  $\mathbf{x}_t$  and  $\mathbf{h}_t$ ,  $\Phi(k) = \sum_{j=k}^{\infty} B_j B'_{j-k}$ , and

$$(B.17) \quad \Phi(k) = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(\lambda) \exp(i\lambda k) d\lambda$$

where

$$(B.18) \quad \begin{aligned} f(\lambda) &= \frac{1}{2\pi} \mathbf{h}_0(\lambda) \overline{\mathbf{h}_0(\lambda)}' \\ &= \sum_{j=0}^{\infty} B_j \exp(i\lambda(t-j)). \end{aligned}$$

## B.4 Lag Operators

In this section, we will apply the results of the previous sections to polynomials and series of the lag operator. We will first consider stochastic processes of random variables. Then we will consider stochastic processes of random vectors.

Given  $x_t = b(L)e_t = \sum_{j=0}^{\infty} b_j e_{t-j}$  with an orthonormal  $e_t$  in  $L^2(Prob)$ , we consider  $h_t(\lambda) = \sum_{j=0}^{\infty} b_j \exp(-i\lambda(t-j)) = \sum_{j=0}^{\infty} b_j \exp(i\lambda j) \exp(-i\lambda t) = h_0(\lambda) \exp(-i\lambda t)$  in  $L^2[-\pi, \pi]$  as in the previous section. As we will see, there is a one-to-one mapping that preserves the distance between the closed linear space generated by  $\{u_t\}_{t=-\infty}^{\infty}$  and that by  $\{e_t\}_{t=-\infty}^{\infty}$ . Moreover,  $L^n h_t(\lambda) = h_{t-n}(\lambda) = h_t \exp(i\lambda n)$ . Hence applying the lag operator  $n$  times to the stochastic process  $h_t$  corresponds with multiplying a complex number  $h_t$  by  $\exp(i\lambda)$   $n$  times. For these reasons, we can study various properties of  $b(L)e_t$  by studying the power series  $b(z) = \sum_{j=0}^{\infty} b_j z^j$  of a complex variable  $z$ . For example, if  $b(z)$  converges on the unit circle, then  $b(\exp(i\lambda))$  converges

for each  $\lambda$  in  $[-\pi, \pi]$ . This point-wise convergence in turn implies the convergence of  $h_0(\lambda) = b_j \exp(i\lambda j)$  in  $L^2[-\pi, \pi]$  by the Bounded Convergence Theorem.

Let  $\{e_t\}_{t=-\infty}^\infty$  be an orthonormal sequence in  $L^2(Prob)$ , and let a sequence  $\{u_t\}_{t=-\infty}^\infty$  in  $L^2[-\pi, \pi]$  be defined by  $u_t(\lambda) = \exp(-i\lambda t) = \cos(\lambda t) - i \sin(\lambda t)$  as in the previous section. Given  $b(L)e_t = \sum_{j=0}^\infty b_j e_{t-j}$ , consider a process  $b(L)u_t = \sum_{j=0}^\infty b_j u_{t-j}$  in  $L^2[-\pi, \pi]$ . From Proposition 3.A.5,  $b(L)e_t$  and  $b(L)u_t$  converge if and only if  $\{b_j\}$  is square summable. Hence  $b(L)e_t$  converges if and only if  $b(L)u_t$  converges.

From these results, we obtain the following proposition which gives a convenient sufficient condition for  $b(L)e_t$  to be defined.

**Proposition B.4.1** Let  $\{e_t\}_{t=-\infty}^\infty$  be a white noise stochastic process with  $E(e_t^2) = 1$ . Suppose that  $b(z) = \sum_{j=0}^\infty b_j z^j$  converges for  $z = z_1$  such that  $|z_1| > 1$ . Then  $\sum_{j=0}^N b_j e_{t-j}$  converges in mean square to a random variable with a finite second moment  $y_t$  as  $N \rightarrow \infty$  and  $y_t = b(L)e_t$  is a covariance stationary process.

**Proof** From Proposition B.1.5,  $b(z) = \sum_{j=0}^\infty b_j z^j$  is converges for  $|z| = 1$ , and hence  $b(\exp(i\lambda))$  converges for each  $\lambda$  in  $[-\pi, \pi]$  in  $C$ . Let  $s_N(\lambda) = \sum_{j=0}^N b_j \exp(i\lambda j)$ , and let  $h_0(\lambda) = b(\exp(i\lambda))$  be the limit of  $s_N(\lambda)$  in  $C$ . For each  $\lambda$ ,  $\lim_{N \rightarrow \infty} |s_N(\lambda) - h_0(\lambda)| = 0$ . Hence by Lebesgue's dominated convergence theorem,

$$\lim_{N \rightarrow \infty} \int_{-\pi}^{\pi} |s_N(\lambda) - h_0(\lambda)| = \int_{-\pi}^{\pi} \lim_{N \rightarrow \infty} |s_N(\lambda) - h_0(\lambda)| = 0,$$

which implies that  $\sum_{j=0}^N b_j \exp(-i\lambda j)$  converges in  $L^2[-\pi, \pi]$  to  $h_0(\lambda)$ . Hence  $\{b_j\}_{j=0}^\infty$  is square summable. To see that  $y_t$  is covariance stationary, note that  $E(y_t) = b_0 E(e_t)$  does not depend on  $t$ . Since the inner product in  $L^2$  is continuous,  $E(y_t y_{t-\tau}) = \lim_{N \rightarrow \infty} E((\sum_{j=0}^N b_j e_{t-j})(\sum_{j=0}^N b_j e_{t-j-\tau}))$ . Since  $e_t$  is covariance stationary,  $E((\sum_{j=0}^N b_j e_{t-j})(\sum_{j=0}^N b_j e_{t-j-\tau}))E((\sum_{j=0}^N b_j e_{t-j})^2)$  does not depend on  $t$ . ■

In this book,  $b(L)e_t$  is taken to mean the limit of  $\sum_{j=0}^N b_j e_{t-j}$  in  $L^2(Prob)$  as  $N \rightarrow \infty$ .

**Example B.4.1** Let  $b(L) = 1 + aL + a^2L^2 + \cdots + a^nL^n + \cdots$ . If  $|a| < 1$ , then  $b(z)$  converges for  $z = z_1$  where  $z_1$  is a real number such that  $1 < z_1 < a^{-1}$ . Hence  $b(L)e_t$  can be defined in  $L^2(Prob)$ . ■

Proposition B.4.1 gives a sufficient condition for  $b(L)e_t$  to be covariance stationary.

The next proposition gives a sufficient condition for  $b(L)e_t$  to be strictly stationary.

**Proposition B.4.2** Let  $\{e_t\}_{t=-\infty}^{\infty}$  be a strictly stationary white noise process with finite second moments. Suppose that  $b(z) = \sum_{j=0}^{\infty} b_j z^j$  converges for  $z = z_1$  such that  $|z_1| > 1$ . Then  $y_t = b(L)e_t$  is a strictly stationary process.

**Proof** Let  $s_{Nt} = \sum_{j=0}^N b_j e_{t-j}$ . Then  $s_{Nt}$  converges in mean square to  $y_t$  as  $N \rightarrow \infty$ . Therefore,  $s_{Nt}$  converges in probability to  $y_t$ , and hence it converges in distribution to  $y_t$ . Let  $F_{Nt}(\zeta)$  be the distribution function of  $s_{Nt}$ , and  $F_t(\zeta)$  be the distribution function of  $y_t$ . Then  $F_{t+\tau}(\zeta) = \lim_{N \rightarrow \infty} F_{Nt}(\zeta) = F_t(\zeta)$  except for the discontinuity points of  $F_{t+\tau}(\zeta)$  and  $F_t(\zeta)$  because  $e_t$  is strictly stationary. There are only countably many discontinuity points, and the distribution function is right continuous. Therefore,  $F_{t-\tau}(\zeta) = F_t(\zeta)$  for all  $\zeta$ . Similar arguments can be made to show that the joint distribution function of  $(y_t, y_{t+1}, \cdots, y_{t+k})$  does not depend on  $t$ . ■

## References

- CHURCHILL, R. V., AND J. W. BROWN (1984): *Complex Variables and Applications*. McGraw Hill, New York, 4th edn.
- SARGENT, T. J. (1987): *Macroeconomic Theory*. Academic Press, New York, 2nd edn.

# Appendix C

## ANSWERS TO SELECTED QUESTIONS

### Answers to Chapter 2

**2.1** The six states of the world are,  $s_1 = [300, 10, 300]$ ,  $s_2 = [300, 10, 150]$ ,  $s_3 = [300, 5, 300]$ ,  $s_4 = [300, 5, 150]$ ,  $s_5 = [150, 5, 300]$ , and  $s_6 = [150, 5, 150]$ . Let I be the information set generated by  $X_1 = (Y_1, i_1)$ , and let  $\mathcal{F}$  be the partition that represents the same information as I. Then,  $\mathcal{F} = \{\Lambda_1, \Lambda_2, \Lambda_3\}$ , where  $\Lambda_1 = \{s_1, s_2\}$ ,  $\Lambda_2 = \{s_3, s_4\}$ , and  $\Lambda_3 = \{s_5, s_6\}$ . Similarly, Let J be the information set generated by  $Y_1$ , and let  $\mathcal{G}$  be the partition that represents the same information as J. Then,  $\mathcal{G} = \{\Lambda_4, \Lambda_5\}$ , where  $\Lambda_4 = \{s_1, s_2, s_3, s_4\}$  and  $\Lambda_5 = \{s_5, s_6\}$ . Using (2.3),  $Pr(s_1|s \in \Lambda_1) = \frac{0.15}{0.15+0.05} = \frac{3}{4}$  and  $Pr(s_2|s \in \Lambda_1) = \frac{0.05}{0.15+0.05} = \frac{1}{4}$ . Hence  $E(Y_2|s \in \Lambda_1) = 300 \times \frac{3}{4} + 150 \times \frac{1}{4} = 262.5$ . Similarly,  $Pr(s_3|s \in \Lambda_2) = \frac{2}{5}$ ,  $Pr(s_4|s \in \Lambda_2) = \frac{3}{5}$ ,  $Pr(s_5|s \in \Lambda_3) = \frac{1}{3}$ ,  $Pr(s_6|s \in \Lambda_3) = \frac{2}{3}$ ,  $E(Y_2|s \in \Lambda_2) = 210$ , and  $E(Y_2|s \in \Lambda_3) = 200$ . Hence the random variable  $E(Y_2|I)$  is given by

$$E(Y_2|I)(s) = \begin{cases} 262.5 & \text{if } s \in \Lambda_1 \\ 210 & \text{if } s \in \Lambda_2 \\ 200 & \text{if } s \in \Lambda_3 \end{cases}$$

Similar computations yield

$$E(Y_2|J)(s) = \begin{cases} 225 & \text{if } s \in \Lambda_4 \\ 200 & \text{if } s \in \Lambda_5 \end{cases}$$

Now, we need to verify that  $E(Y_2|J)(s) = E(E(Y_2|I)|J)(s)$  for all  $s \in S$ .  $E(Y_2|I)(s)$  is given above, while  $E(E(Y_2|I)|J)(s)$  can be computed as following:  $Pr(s \in \Lambda_1|s \in \Lambda_4) = \frac{0.15+0.05}{0.15+0.05+0.20+0.30} = \frac{2}{7}$  and  $Pr(s \in \Lambda_2|s \in \Lambda_4) = \frac{0.20+0.30}{0.15+0.05+0.20+0.30} = \frac{5}{7}$ . Therefore,  $E(E(Y_2|I)|s \in \Lambda_4) = 262.5 \times \frac{2}{7} + 210 \times \frac{5}{7} = 225$ , while  $Pr(s \in \Lambda_3|s \in \Lambda_5) = 1$  so that  $E(E(Y_2|I)|s \in \Lambda_5) = 200$  as above. In summary,

$$E(E(Y_2|I)|J)(s) = \begin{cases} 225 & \text{if } s \in \Lambda_4 \\ 200 & \text{if } s \in \Lambda_5 \end{cases}$$

which is equivalent to  $E(Y_2|J)(s)$ .

**2.2** Let  $Y_t = AY_{t-1} + e_t$ ,  $E(e_t) = 0$ , and  $Y_0 = 0$ . Then,  $E(Y_1) = E(AY_0 + e_1) = 0$ ,  $E(Y_2) = E(AY_1 + e_2) = 0$ ,  $Var(Y_1) = Var(AY_0 + e_1) = \sigma^2$ , and  $Var(Y_2) = Var(AY_1 + e_2) = (A^2 + 1)\sigma^2$ . Therefore,  $Y_t$  is not strictly stationary if  $A \neq 0$ .

**2.3** Let  $Y_t = AY_{t-1} + e_t$ ,  $E(e_t) = 0$ , and  $Y_0 \sim N(0, \frac{\sigma^2}{1-A^2})$ . Then,  $E(Y_1) = E(AY_0 + e_1) = 0$ ,  $E(Y_2) = E(AY_1 + e_2) = 0$ ,  $Var(Y_1) = Var(AY_0 + e_1) = A^2 \frac{\sigma^2}{1-A^2} + \sigma^2 = \frac{\sigma^2}{1-A^2}$ , and  $Var(Y_2) = Var(AY_1 + e_2) = Var(A^2Y_0 + Ae_1 + e_2) = A^4 \frac{\sigma^2}{1-A^2} + A^2\sigma^2 + \sigma^2 = \frac{\sigma^2}{1-A^2}$ . Similarly, we can show that  $Var(Y_t) = \frac{\sigma^2}{1-A^2}$  for any  $t$ . Finally,  $Cov(Y_t, Y_{t+k}) = Cov(Y_t, A^kY_t + A^{k-1}e_{t+1} + \dots + e_{t+k}) = A^k \frac{\sigma^2}{1-A^2}$  for any  $t$ . Therefore  $Y_t$  is covariance stationary. Furthermore,  $Y_t$  is strictly stationary, because  $Y_t$  follows the normal distribution, in which the first and second moments completely determine the distribution.

**2.4**  $E(e_{t+1}|I_t) = E(Y_{t+1} - Y_t|I_t) = E(Y_{t+1}|I_t) - Y_t$ . Since  $Y_t$  is a martingale adapted to  $I_t$ ,  $I_t$  is a sequence of increasing information sets,  $E(Y_{t+1}|I_t) = Y_t$ , and  $e_t$  is in  $I_t$ . Therefore,  $E(e_{t+1}|I_t) = 0$  so that  $e_t$  is a martingale difference sequence.

**2.5** Let  $e_t$  be a covariance stationary martingale difference sequence. By definition of covariance stationarity, the mean and variance of  $e_t$  are finite and constant over time. Since  $e_t$  is a martingale difference sequence,  $E(e_{t+1}|I_t) = 0$ . Using the law of iterated expectations, we also have  $E(e_{t+k}|I_t) = E(E(e_{t+k}|I_{t+k-1}|I_t)) = 0$  for any  $k > 0$ . Therefore,  $E(e_t) = E(e_{t+1}) = E(E(e_{t+1}|I_t)) = 0$ . Second,  $E(e_t e_{t+k}) = E(E(e_t e_{t+k}|I_t)) = E(e_t E(e_{t+k}|I_t)) = 0$  for any  $k > 0$ . Therefore,  $e_t$  is white noise.

**2.6** Let  $e_t$  be an i.i.d. white noise process and  $I_t$  be the information set generated from  $\{e_t, e_{t-1}, \dots\}$ . Then  $e_t$  is in  $I_t$ , and  $I_t$  is an increasing sequence of information sets. Since  $e_{t+1}$  is independent of  $\{e_t, e_{t-1}, \dots\}$  by the definition of i.i.d.,  $E(e_{t+1}|I_t) = E(e_{t+1}) = E(e_t) = 0$ . Therefore,  $e_t$  is a martingale difference sequence.

### Answers to Chapter 3

**3.1** Let  $S_{t+1} = F_t + \epsilon_{t+1}$ , where  $F_t$  is in  $I_t$  and  $\epsilon_{t+1}$  is in  $I_{t+1}$ . Since  $F_t$  is orthogonal to  $\epsilon_{t+1}$ ,  $E(F_t \epsilon_{t+1}) = 0$ . Thus,  $E(S_{t+1}^2) = E(F_t^2) + E(\epsilon_{t+1}^2)$ , which implies  $E(S_{t+1}^2) \geq E(F_t^2)$  as  $E(\epsilon_{t+1}^2) \geq 0$ . It follows from  $E(S_{t+1}) = E(F_t)$  that  $Var(S_{t+1}) \geq Var(F_t)$ .

**3.2** Yes. Let  $A_t = i_{n,t} + \epsilon_t$ , where  $i_{n,t}$  is in  $I_t$  and  $\epsilon_t$  is in  $I_{t+1}$ . Since  $i_{n,t}$  is orthogonal to  $\epsilon_t$ ,  $E(i_{n,t} \epsilon_t) = 0$ . Thus,  $E(A_t^2) = E(i_{n,t}^2) + E(\epsilon_t^2)$ , which implies  $E(A_t^2) \geq E(i_{n,t}^2)$  as  $E(\epsilon_t^2) \geq 0$ . It follows from  $E(A_t) = E(i_{n,t})$  that  $Var(A_t) \geq Var(i_{n,t})$ .

**3.3** Let  $\nu_t = N_t - \hat{E}(N_t|H_t)$ . From the identity  $N_t = \hat{E}(N_t|H_t) + \nu_t$ , we get  $E(N_t^2) = E((\hat{E}(N_t|H_t))^2) + E(\nu_t^2)$  as  $\hat{E}(N_t|H_t)$  is orthogonal to  $\nu_t$ . Thus,  $E(N_t^2) \geq E((\hat{E}(N_t|H_t))^2)$ . Note that  $Var(N_t) = E(N_t^2) - (E(N_t))^2$ , and that  $E(N_t) = E(\hat{E}(N_t|H_t))$  by the law of iterated projections because unconditional expectation is the projection onto the set of constants. Since  $E(N_t^2) \geq E((\hat{E}(N_t|H_t))^2)$ ,  $Var(N_t) \geq E((\hat{E}(N_t|H_t))^2) - (E(\hat{E}(N_t|H_t)))^2 = Var(\hat{E}(N_t|H_t))$ . Thus,  $Var(N_t) \geq \eta$ .

### Answers to Chapter 4

**4.1** (i) Let  $E(u_t^2) = \sigma^2$ .  $E(x_t) = E(u_t) + 0.8E(u_{t-1}) = 0$  for any  $t$  because  $u_t$  is a white noise process. Similarly,  $E(x_t x_{t+k})$  is  $1.64\sigma^2$  if  $k = 0$ , is  $0.8\sigma^2$  if  $|k| = 1$ , and is 0 if  $|k| > 1$ . They are invariant over time so that  $x_t$  is covariance stationary. (ii) Let  $H_t$  be the linear information set generated by the current and past values of  $x_t$ , and  $H_t^u$  be the linear information set generated by the current and past values of  $u_t$ . Since  $|B| \leq 1$ ,  $u_t$  is fundamental so that  $H_t = H_t^u$ . (iii)  $\hat{E}(x_t|u_{t-1}, u_{t-2}, \dots) = \hat{E}(u_t + 0.8u_{t-1}|u_{t-1}, u_{t-2}, \dots) = 0.8u_{t-1}$ . (iv) Yes.  $\hat{E}(x_t|x_{t-1}, x_{t-2}, \dots) = \hat{E}(x_t|u_{t-1}, u_{t-2}, \dots) = 0.8u_{t-1}$  because  $H_t = H_t^u$ .

**4.2** (i) Let  $E(u_t^2) = \sigma^2$ .  $E(x_t) = E(u_t) + 1.2E(u_{t-1}) = 0$  for any  $t$  because  $u_t$  is a white noise process. Similarly,  $E(x_t x_{t+k})$  is  $2.44\sigma^2$  if  $k = 0$ , is  $1.2\sigma^2$  if  $|k| = 1$ , and is 0 if  $|k| > 1$ . They are invariant over time so that  $x_t$  is covariance stationary. (ii) Let  $H_t$  be the linear information set generated by the current and past values of  $x_t$ , and  $H_t^u$  be the linear information set generated by the current and past values of  $u_t$ . Since  $|B| > 1$ ,  $u_t$  is not fundamental so that  $H_t \neq H_t^u$ . (iii)  $\hat{E}(x_t|u_{t-1}, u_{t-2}, \dots) = \hat{E}(u_t + 1.2u_{t-1}|u_{t-1}, u_{t-2}, \dots) = 1.2u_{t-1}$ . (iv) No.  $\hat{E}(x_t|x_{t-1}, x_{t-2}, \dots) \neq \hat{E}(x_t|u_{t-1}, u_{t-2}, \dots) = 1.2u_{t-1}$  because  $H_t \neq H_t^u$ .

**4.3** (i) Let  $E(u_t^2) = \sigma^2$ .  $E(x_t) = E(u_t) + E(u_{t-1}) = 0$  for any  $t$  because  $u_t$  is a white noise process. Similarly,  $E(x_t x_{t+k})$  is  $2\sigma^2$  if  $k = 0$ , is  $\sigma^2$  if  $|k| = 1$ , and is 0 if  $|k| > 1$ . They are invariant over time so that  $x_t$  is covariance stationary. (ii) Let  $H_t$  be the linear information set generated by the current and past values of  $x_t$ , and  $H_t^u$  be the linear information set generated by the current and past values of  $u_t$ . Since  $|B| \leq 1$ ,  $u_t$  is fundamental so that  $H_t = H_t^u$ . (iii)  $\hat{E}(x_t|u_{t-1}, u_{t-2}, \dots) = \hat{E}(u_t + u_{t-1}|u_{t-1}, u_{t-2}, \dots) = u_{t-1}$ . (iv) Yes.  $\hat{E}(x_t|x_{t-1}, x_{t-2}, \dots) = \hat{E}(x_t|u_{t-1}, u_{t-2}, \dots) = u_{t-1}$  because  $H_t = H_t^u$ .

## Answers to Chapter 5

**5.1** We consider an  $s$ -period ahead forecast of  $X_t$ ,  $E(X_{t+s}|I_t)$ , and the forecast error,  $e_t = X_{t+s} - E(X_{t+s}|I_t)$ . (i) Since  $X_t$  is strictly stationary and ergodic,  $e_t$  is strictly stationary and ergodic. Now, we need to show that  $e_t$  has mean zero and  $E(|e_t|^2) < \infty$ . Note that  $E(e_t|I_t) = 0$ , and  $e_t$  is in the information set  $I_{t+s}$ . First,  $E(e_t) = E(E(e_t|I_t)) = 0$ . Second,  $E(|e_t|^2) < \infty$  since  $E(|X_t|^2) < \infty$ . (ii)  $E(e_t|e_{t-j}, e_{t-j-1}, \dots) = E(E(e_t|I_t)|e_{t-j}, e_{t-j-1}, \dots) = 0$  for any  $j \geq s$ . Thus,  $E(e_t|e_{t-j}, e_{t-j-1}, \dots)$  converges in mean square to 0 as  $j \rightarrow \infty$ . (iii) Let  $r_{tj} = E(e_t|e_{t-j}, e_{t-j-1}, \dots) - E(e_t|e_{t-j-1}, e_{t-j-2}, \dots)$  where  $e_t$  is in  $I_{t+s}$ . Note that  $r_{tj} = 0$  for any  $j \geq s$  based on (ii). Thus,  $\sum_{j=0}^{\infty} [E(r_{tj}^2)]^{\frac{1}{2}} = \sum_{j=0}^{s-1} [E(r_{tj}^2)]^{\frac{1}{2}} < \infty$  because  $s$  is finite and  $e_t$  has a finite second moment.

**5.2** We consider an  $s$ -period ahead forecast of  $X_t$ ,  $E(X_{t+s}|I_t)$ , and the forecast error,  $e_t = X_{t+s} - E(X_{t+s}|I_t)$ . Let  $\mathbf{Z}_t$  be a random vector with finite second moments in the information set  $I_t$ . Define  $\mathbf{f}_t = \mathbf{Z}_t e_t = g(e_t)$ . (i) Because  $X_t$  and  $\mathbf{Z}_t$  are stationary and ergodic,  $\mathbf{f}_t$  is strictly stationary and ergodic. Now, we need to show that  $\mathbf{f}_t$  is with mean zero and finite second moments. Note that  $E(e_t|I_t) = 0$  and  $e_t$  is in the information set  $I_{t+s}$  so that  $\mathbf{f}_t$  is in the information set  $I_{t+s}$ . Thus,  $E(\mathbf{f}_t|I_t) = E(\mathbf{Z}_t e_t|I_t) = E(\mathbf{Z}_t E(e_t|I_t)) = \mathbf{0}$ . First,  $E(\mathbf{f}_t) = E(E(\mathbf{f}_t|I_t)) = \mathbf{0}$ . Second,  $E(|\mathbf{f}_t|^2) < \infty$  since  $E(|X_t|^2) < \infty$ ,  $E(|\mathbf{Z}_t|^2) < \infty$ . (ii)  $E(\mathbf{f}_t|\mathbf{f}_{t-j}, \mathbf{f}_{t-j-1}, \dots) = E(E(\mathbf{f}_t|I_t)|\mathbf{f}_{t-j}, \mathbf{f}_{t-j-1}, \dots) = \mathbf{0}$  for any  $j \geq s$ . Thus,  $E(\mathbf{f}_t|\mathbf{f}_{t-j}, \mathbf{f}_{t-j-1}, \dots)$  converges in mean square to  $\mathbf{0}$  as  $j \rightarrow \infty$ . (iii) Let  $\mathbf{r}_{tj} = E(\mathbf{f}_t|\mathbf{f}_{t-j}, \mathbf{f}_{t-j-1}, \dots) - E(\mathbf{f}_t|\mathbf{f}_{t-j-1}, \mathbf{f}_{t-j-2}, \dots)$  where  $\mathbf{f}_t$  is in  $I_{t+s}$ . Note that  $\mathbf{r}_{tj} = \mathbf{0}$  for any  $j \geq s$  based on (ii). Thus,  $\sum_{j=0}^{\infty} [E(\mathbf{r}_{tj}' \mathbf{r}_{tj})]^{\frac{1}{2}} = \sum_{j=0}^{s-1} [E(\mathbf{r}_{tj}' \mathbf{r}_{tj})]^{\frac{1}{2}} < \infty$  because  $s$  is finite and  $\mathbf{f}_t$  has a finite second moment.

**5.3** Let  $f_t = (1 - L)e_t = e_t - e_{t-1}$  where  $e_t = \Psi(L)u_t$ . Then  $f_t$  has an MA representation that  $f_t = \Psi^*(L)u_t$  where  $\Psi^*(L) = (1 - L)\Psi(L)$ . By similar calculation as (5.12) in the text, we have  $\Psi = [\Psi^*(1)]^2 E(u_t^2) = [0 \times \Psi(1)]^2 E(u_t^2) = 0$ .

**5.4** A test is said to under-reject in small samples when the probability of rejecting the null hypothesis when it is true is smaller than the nominal size when the nominal critical value is used in small samples. In this case, the true critical value is smaller than the nominal critical value (provided that the test procedure is to reject when the test statistic or the absolute value of it is greater than the nominal critical value).

5.5 (a) The IV estimator is

$$\mathbf{b}_{IV} = \left( \sum_{t=1}^T \mathbf{z}_t \mathbf{x}_t' \right)^{-1} \sum_{t=1}^T \mathbf{z}_t y_t$$

(b) Let  $I_t$  be the information set generated by  $\{\mathbf{g}_{t-1}, \mathbf{g}_{t-2}, \dots\}$ . Since  $\mathbf{z}_t$  is in  $I_t$ ,  $E(\mathbf{g}_t | I_t) = E(\mathbf{z}_t e_t | I_t) = \mathbf{z}_t E(e_t | I_t) = \mathbf{0}$ . By the law of iterated expectations, we have  $E(\mathbf{g}_t | I_{t-1}) = E(E(\mathbf{g}_t | I_t) | I_{t-1}) = \mathbf{0}$ . Hence  $\mathbf{g}_t$  is a martingale difference sequence.

(c) Applying the strong law of Ergodic theorem to  $\mathbf{z}_t \mathbf{x}_t$  and  $\mathbf{z}_t e_t$ , we have

$$\begin{aligned} \frac{1}{T} \sum_{t=1}^T \mathbf{z}_t \mathbf{x}_t' &\xrightarrow{a.s.} E(\mathbf{z}_t \mathbf{x}_t') \\ \frac{1}{T} \sum_{t=1}^T \mathbf{z}_t e_t &\xrightarrow{a.s.} E(\mathbf{z}_t e_t) \end{aligned}$$

By Assumption (A5),  $\frac{1}{T} \sum_{t=1}^T \mathbf{z}_t \mathbf{x}_t'$  is nonsingular for large enough  $T$  with probability one, and therefore,

$$\begin{aligned} \mathbf{b}_{IV} - \boldsymbol{\beta} &= \left( \frac{1}{T} \sum_{t=1}^T \mathbf{z}_t \mathbf{x}_t' \right)^{-1} \frac{1}{T} \sum_{t=1}^T \mathbf{z}_t y_t - \boldsymbol{\beta} \\ &= \left( \frac{1}{T} \sum_{t=1}^T \mathbf{z}_t \mathbf{x}_t' \right)^{-1} \frac{1}{T} \sum_{t=1}^T \mathbf{z}_t (\mathbf{x}_t \boldsymbol{\beta} + e_t) - \boldsymbol{\beta} \\ &= \left( \frac{1}{T} \sum_{t=1}^T \mathbf{z}_t \mathbf{x}_t' \right)^{-1} \frac{1}{T} \sum_{t=1}^T \mathbf{z}_t e_t \xrightarrow{a.s.} E(\mathbf{z}_t \mathbf{x}_t')^{-1} E(\mathbf{z}_t e_t) = \mathbf{0}. \end{aligned}$$

(d) Applying the Ergodic stationary martingale differences central limit theorem to  $\mathbf{z}_t e_t$ , we obtain

$$\frac{1}{\sqrt{T}} \sum_{t=1}^T \mathbf{z}_t e_t \xrightarrow{d} N(\mathbf{0}, E(e_t^2 \mathbf{z}_t \mathbf{z}_t')).$$

By Assumption (A3),  $E(e_t^2 \mathbf{z}_t \mathbf{z}_t') = E(E(e_t^2 \mathbf{z}_t \mathbf{z}_t' | \mathbf{z}_t)) = \sigma^2 E(\mathbf{z}_t \mathbf{z}_t')$ . Hence

$$\sqrt{T}(\mathbf{b}_{IV} - \boldsymbol{\beta}) = \left( \frac{1}{T} \sum_{t=1}^T \mathbf{z}_t \mathbf{x}_t' \right)^{-1} \frac{1}{\sqrt{T}} \sum_{t=1}^T \mathbf{z}_t e_t \xrightarrow{d} E(\mathbf{z}_t \mathbf{x}_t')^{-1} N(\mathbf{0}, \sigma^2 E(\mathbf{z}_t \mathbf{z}_t')).$$

Hence  $\sqrt{T}(\mathbf{b}_{IV} - \boldsymbol{\beta}) \xrightarrow{d} N(\mathbf{0}, \sigma^2 E(\mathbf{z}_t \mathbf{x}_t')^{-1} E(\mathbf{z}_t \mathbf{z}_t') E(\mathbf{x}_t \mathbf{x}_t')^{-1})$ .

(e) (i) Prove that  $\mathbf{g}_t$  satisfies Gordin's conditions. First Gordin's condition is to show  $E(\mathbf{g}_t^2) < \infty$ . By Assumption (A2), it is satisfied.

Second Gordin's condition is to show that  $E(\mathbf{g}_t | \mathbf{g}_{t-j}, \mathbf{g}_{t-j-1}, \dots) \xrightarrow{m.s.} \mathbf{0}$  as  $j \rightarrow \infty$ . Because  $y_t$  is in  $I_{t+2}$ , so are both  $e_t$  and  $\mathbf{g}_t$  in  $I_{t+2}$ . Therefore, for  $j \geq 2$ ,

$$\begin{aligned} &E(\mathbf{g}_t | \mathbf{g}_{t-j}, \mathbf{g}_{t-j-1}, \dots) \\ &= E(E(\mathbf{g}_t | I_t) | \mathbf{g}_{t-j}, \mathbf{g}_{t-j-1}, \dots) \text{ (by law of iterated expectation)} \\ &= \mathbf{0} \text{ (since } E(\mathbf{g}_t | I_t) = E(\mathbf{z}_t e_t | I_t) = \mathbf{z}_t E(e_t | I_t) = \mathbf{0}\text{)}. \end{aligned}$$

Third Gordin's condition is to show that telescoping sum is finite. Since  $r_{tj} = E(\mathbf{g}_t | \mathbf{g}_{t-j}, \mathbf{g}_{t-j-1}, \dots) - E(\mathbf{g}_t | \mathbf{g}_{t-j-1}, \mathbf{g}_{t-j-2}, \dots) = \mathbf{0}$  for  $j \geq 2$ , the telescoping sum is finite. Hence it is satisfied.

(ii) We can apply the same logic to show that the IV estimator is consistent under (A1)–(A5).

(iii) To prove that IV estimator is asymptotically normally distributed, apply the Gordin and Hansen's central limit theorem to  $\mathbf{z}_t e_t$ . Then we obtain

$$\frac{1}{\sqrt{T}} \sum_{t=1}^T \mathbf{z}_t e_t \xrightarrow{d} N(\mathbf{0}, \mathbf{S})$$

where  $\mathbf{S} = \mathbf{\Gamma}_{-1} + \mathbf{\Gamma}_0 + \mathbf{\Gamma}_1$  and  $\mathbf{\Gamma}_j = E(e_t e'_{t-j} \mathbf{z}_t \mathbf{z}'_{t-j})$ . Hence

$$\sqrt{T}(\mathbf{b}_{IV} - \boldsymbol{\beta}) = \left(\frac{1}{T} \sum_{t=1}^T \mathbf{z}_t \mathbf{z}'_t\right)^{-1} \frac{1}{\sqrt{T}} \sum_{t=1}^T \mathbf{z}_t e_t \xrightarrow{d} E(\mathbf{z}_t \mathbf{z}'_t)^{-1} N(\mathbf{0}, \mathbf{S}).$$

Hence  $\sqrt{T}(\mathbf{b}_{IV} - \boldsymbol{\beta}) \xrightarrow{d} N(\mathbf{0}, E(\mathbf{z}_t \mathbf{z}'_t)^{-1} \mathbf{S} E(\mathbf{z}_t \mathbf{z}'_t)^{-1})$ .

### 5.6 (a)

$$t_k = \frac{\sqrt{n}(b_k - \bar{\beta}_k)}{\sqrt{s^2 \left[\frac{1}{n}(\mathbf{X}'\mathbf{X})^{-1}\right]_{kk}}}$$

The denominator converges in probability to the square root of  $\sigma^2[E(\mathbf{x}_t \mathbf{x}'_t)]_{kk}^{-1}$ , and the numerator converges almost surely to a normal random variable with mean zero and variance  $\sigma^2[E(\mathbf{x}_t \mathbf{x}'_t)]_{kk}^{-1}$ . Hence  $t_k$  converges in distribution to a standard normal random variable.

- (b) The  $t_k$  statistic has the exact  $t$  distribution with  $n - K$  degrees of freedom. Since the critical value based on the  $t$  distribution is always greater than 1.96 in finite samples, the test overrejects. The actual size can be obtained from a  $t$  distribution table. The actual size is larger than 10 percent if the cutoff point for Student's  $t$  distribution for the 0.05 right-hand tail probability is greater than 1.96. This property is true for  $df = 1, 2, 3, 4, 5$ , and not true if  $df$  is greater than 5. When  $n = 4, df = 1$ , and therefore the actual size is larger than 10 percent. If  $n = 8, df = 5$ , so this is still true. For  $n = 9, 10, 11$ , the  $df$  is greater than 5. Therefore, the actual size is smaller than 10 percent for  $n = 9, 10, 11$ .

### 5.7 (i) The Program

```
new;
cls;
outfile = "741c-hw2.out";
output file = ^outfile reset;
$outfile;
"By Kyungho Jang and Masao Ogaki";
datestr(0);
timestr(0);
seedno = 36481111;
rndseed seedno;
format /rdn 8,0;
"Seed Number Used: rndseed = ";; seedno;
format /rdn 12,4;
n = 26; @ sample size @
n_mc = 500; @ # of replication for Monte Carlo study @
df_e = 4; @ degree of freedom of e_t @

@ Critical value with 5% significance level with df = 25 @
nc = 1.96; @ Normal Distribution @
tc = 2.060; @ Student's t distribution @

@ Monte Carlo Study @
```

```

tnm1 = zeros(n_mc,1); @ vector of t values under H0 using t1 @
tam1 = zeros(n_mc,1); @ vector of t values under H1 using t1 @
tnm2 = zeros(n_mc,1); @ vector of t values under H0 using t2 @
tam2 = zeros(n_mc,1); @ vector of t values under H1 using t2 @

i = 1;
do while i <= n_mc;
    @ Generate independent series @
    dat = rndn(n,2+df_e);
    x = dat[:,1];
    z = dat[:,2];
    q = sumc((dat[:,3:2+df_e]^2)');
    e = z ./ sqrt(q/df_e);

    @ Under H0: beta = 0, Construct y @
    yn = e;
    @ Under H1: beta = 0.15, Construct y @
    ya = x*0.15 + e;

    @ Estimation @
    k = cols(x);
    df1 = n - k; @ for t1 @
    df2 = n; @ for t2 @

    bn = invpd(x'x)*x'yn; @ b under H0 @
    ba = invpd(x'x)*x'ya; @ b under H1 @

    un = yn - x*bn; @ Residuals under H0 @
    ua = ya - x*ba; @ Residuals under H1 @

    sign2_1 = un'un/df1; @ Variance of e_t under H0 for t1 @
    siga2_1 = ua'ua/df1; @ Variance of e_t under H1 for t1 @

    sign2_2 = un'un/df2; @ Variance of e_t under H0 for t2 @
    siga2_2 = ua'ua/df2; @ Variance of e_t under H1 for t2 @

    sbn1 = sqrt(sign2_1*invpd(x'x)); @ Standard error of b under H0 for t1 @
    sba1 = sqrt(siga2_1*invpd(x'x)); @ Standard error of b under H1 for t1 @

    sbn2 = sqrt(sign2_2*invpd(x'x)); @ Standard error of b under H0 for t2 @
    sba2 = sqrt(siga2_2*invpd(x'x)); @ Standard error of b under H1 for t2 @

    tnm1[i] = bn ./ sbn1; @ t values under H0 using t1 @
    tam1[i] = ba ./ sba1; @ t values under H1 using t1 @

    tnm2[i] = bn ./ sbn2; @ t values under H0 using t2 @
    tam2[i] = ba ./ sba2; @ t values under H1 using t2 @

    i = i + 1;
endo;

```

```

@ Sort the results with absolute value @
tnm1 = sortc(abs(tnm1),1);
tam1 = sortc(abs(tam1),1);

tnm2 = sortc(abs(tnm2),1);
tam2 = sortc(abs(tam2),1);

"";
"***** Under H0: b = 0 *****";
"";
"Ex. 1-(a): Based on Normal Distribution";
"Estimated true size with 5% critical value ";
meanc(tnm1 .> nc);

"";
"Ex. 1-(b): Based on t Distribution with df = 25";
"Estimated true size with 5% critical value ";
meanc(tnm1 .> tc);

"";
"Ex. 1-(c): True Critical Value of the t test for the 5% significance level";
"Estimated true 5% critical value ";
etcv1 = tnm1[int(n_mc*0.95)];
etcv1;

"";
"***** Under H1: b = 0.15 *****";
"";
"Ex. 1-(d): Based on t Distribution with df = 25";
" Estimated power with the nominal critical value ";
meanc(tam1 .> tc);

"";
"Ex. 1-(e): Based on t Distribution with df = 25";
" Estimated size corrected power";
meanc(tam1 .> etcv1);

"";
"***** Under H0: b = 0 *****";
"";
"Ex. 2-(a): Based on Normal Distribution";
"Estimated true size with 5% critical value ";
meanc(tnm2 .> nc);

"";
"Ex. 2-(b): Based on t Distribution with df = 25";
"Estimated true size with 5% critical value ";
meanc(tnm2 .> tc);

"";

```

```

"Ex. 2-(c): True Critical Value of the t test for the 5% significance level";
"Estimated true 5% critical value ";
etcv2 = tnm2[int(n_mc*0.95)];
etcv2;

"";
"***** Under H1: b = 0.15 *****";
"";
"Ex. 2-(d): Based on t Distribution with df = 25";
" Estimated power with the nominal critical value ";
meanc(tam2 .> tc);

"";
"Ex. 2-(e): Based on t Distribution with df = 25";
" Estimated size corrected power";
meanc(tam2 .> etcv2);

end;

```

### (ii) The Output

```

741c-hw2.out
By Kyungho Jang and Masao Ogaki
5/28/02
1:23:09
Seed Number Used: rndseed = 36481111

***** Under H0: b = 0 *****

Ex. 1-(a): Based on Normal Distribution
Estimated true size with 5% critical value
0.0600

Ex. 1-(b): Based on t Distribution with df = 25
Estimated true size with 5% critical value
0.0580

Ex. 1-(c): True Critical Value of the t test for the 5% significance level
Estimated true 5% critical value
2.1213

***** Under H1: b = 0.15 *****

Ex. 1-(d): Based on t Distribution with df = 25
Estimated power with the nominal critical value
0.0960

Ex. 1-(e): Based on t Distribution with df = 25
Estimated size corrected power
0.0880

***** Under H0: b = 0 *****

```

Ex. 2-(a): Based on Normal Distribution  
 Estimated true size with 5% critical value  
 0.0680

Ex. 2-(b): Based on t Distribution with df = 25  
 Estimated true size with 5% critical value  
 0.0600

Ex. 2-(c): True Critical Value of the t test for the 5% significance level  
 Estimated true 5% critical value  
 2.1634

\*\*\*\*\* Under H1: b = 0.15 \*\*\*\*\*

Ex. 2-(d): Based on t Distribution with df = 25  
 Estimated power with the nominal critical value  
 0.1040

Ex. 2-(e): Based on t Distribution with df = 25  
 Estimated size corrected power  
 0.0880

### (iii) Explanation

- (a) It is better to use the t distribution than normal distribution, because the sample size ( $n = 26$ ) is not large enough.
- (b) There is no difference between t1 and t2, because the size corrected power of t1 (0.0880) is the same as that of t2 (0.0880).

## Answers to Chapter 6

- 6.1** (a) Let  $e_t = X_{t+3} - E(X_{t+3} | I_t)$  and  $I_t$  be an information set generated by the current and past  $\mathbf{Y}_t$  which contains  $X_t$ . Then  $e_t$  is in  $I_{t+3}$  and  $E(e_t | I_t) = 0$ . Note that  $E(e_t e_{t-j}) = E[E(e_t e_{t-j} | I_t)] = E[e_{t-j} E(e_t | I_t)] = 0$  for any  $j \geq 3$ . Thus, this case is of known order of serial correlation, which is 2. (i) The long run variance of  $e_t$  is  $\Omega = \sum_{\tau=-2}^2 \Phi(\tau)$ , where  $\Phi(\tau) = E(e_t e_{t-\tau})$ . (ii) Since the order of serial correlation is known, we should use the truncated kernel estimator with  $S_T = 3$ . If the estimate of long run variance is not positive definite, then we can use Modified Durbin's method or kernel HAC for the estimation of the long run variance of  $e_t$ . Note that zero restrictions should not be used in the latter method.
- (b)  $\mathbf{Z}_t$  is in  $I_t$  and  $\mathbf{f}_t = \mathbf{Z}_t e_t$ . Then  $\mathbf{f}_t$  is in  $I_{t+3}$  and  $E(\mathbf{f}_t | I_t) = E(\mathbf{Z}_t e_t | I_t) = \mathbf{Z}_t E(e_t | I_t) = \mathbf{0}$ . Note that  $E(\mathbf{f}_t \mathbf{f}'_{t-j}) = E[E(\mathbf{f}_t \mathbf{f}'_{t-j} | I_t)] = E[E(\mathbf{f}_t | I_t) \mathbf{f}'_{t-j}] = \mathbf{0}$  for any  $j \geq 3$ . Thus, this is the case of known order of serial correlation, which is 2. (i) The long run variance of  $\mathbf{f}_t$  is  $\Omega = \sum_{\tau=-2}^2 \Phi(\tau)$ , where  $\Phi(\tau) = E(\mathbf{f}_t \mathbf{f}'_{t-\tau})$ . (ii) Since the order of serial correlation is known, we should use the truncated kernel estimator with  $S_T = 3$ . If the estimate of long run variance is not positive definite, then we can use Modified Durbin's method or kernel HAC for the estimation of long run variance of  $\mathbf{f}_t$ . Note that the former method is not reliable when

the number of elements in  $\mathbf{f}_t$  is large compared with the sample size. Note also that zero restrictions should not be used in the latter method.

### Answers to Chapter 8

- 8.1 (a)** Let  $\mathbf{z}_t = (m_t, y_t)'$  be a two dimensional covariance stationary process. We say that  $y$  fails to *Granger-cause*  $m$  if for all  $s > 0$ ,

$$(8.E.2) \quad \hat{E}(m_{t+s}|m_t, m_{t-1}, \dots, y_t, y_{t-1}, \dots) = \hat{E}(m_{t+s}|m_t, m_{t-1}, \dots).$$

We also say that  $y_t$  is not *linearly informative* about future  $m$ , or  $m$  is exogenous in the time series sense with respect to  $y$ .

- (b)** One can test the null hypothesis that  $y$  fails to Granger-cause  $m$  by applying OLS to

$$(8.E.3) m_{t+1} = \delta_{\epsilon_2} + a_1 m_t + a_2 m_{t-1} + \dots + a_p m_{t-p+1} + b_1 y_t + b_2 y_{t-1} + \dots + b_p y_{t-p+1} + \epsilon_{2t}$$

with the null hypothesis

$$(8.E.4) \quad H_0 : b_i = 0 \text{ for } i = 1, \dots, p.$$

- (c)** Consider a model in which consumers and firms increase their demand for money when they expect future real GDP to increase. In such a model money can Granger-cause real GDP because money supply responds to the future expected values of real GDP.
- (d)** First, define the orthogonalized impulse response function. Consider the Wold representation for  $\mathbf{z}_t$ :

$$\begin{aligned} \mathbf{z}_t &= \boldsymbol{\mu} + \boldsymbol{\epsilon}_t + \boldsymbol{\Psi}_1 \boldsymbol{\epsilon}_{t-1} + \boldsymbol{\Psi}_2 \boldsymbol{\epsilon}_{t-2} + \dots \\ &= \boldsymbol{\mu} + \boldsymbol{\Psi}(L) \boldsymbol{\epsilon}_t. \end{aligned}$$

Define  $\boldsymbol{\Sigma}_\epsilon = E(\boldsymbol{\epsilon}_t \boldsymbol{\epsilon}_t')$ . Assume that  $\boldsymbol{\Sigma}_\epsilon$  is positive definite. Then there exists a unique lower triangular matrix  $\boldsymbol{\Phi}_0$  with 1's along the principal diagonal and a unique diagonal matrix  $\boldsymbol{\Lambda}$  with positive entries along the principal diagonal such that

$$(8.E.5) \quad \boldsymbol{\Sigma}_\epsilon = \boldsymbol{\Phi}_0 \boldsymbol{\Lambda} \boldsymbol{\Phi}_0'.$$

Let

$$(8.E.6) \quad \mathbf{e}_t = \boldsymbol{\Phi}_0^{-1} \boldsymbol{\epsilon}_t.$$

Then  $E(\mathbf{e}_t \mathbf{e}_t') = \boldsymbol{\Phi}_0^{-1} \boldsymbol{\Sigma}_\epsilon (\boldsymbol{\Phi}_0^{-1})' = \boldsymbol{\Lambda}$ , which is diagonal. Because

$$(8.E.7) \quad \boldsymbol{\epsilon}_t = \boldsymbol{\Phi}_0 \mathbf{e}_t,$$

$\mathbf{z}_t$  has a MA representation in terms of  $\mathbf{e}_t$ :

$$\begin{aligned} \mathbf{z}_t &= \boldsymbol{\mu} + \boldsymbol{\Phi}_0 \mathbf{e}_t + \boldsymbol{\Phi}_1 \mathbf{e}_{t-1} + \boldsymbol{\Phi}_2 \mathbf{e}_{t-2} + \dots \\ &= \boldsymbol{\mu} + \boldsymbol{\Phi}(L) \mathbf{e}_t. \end{aligned}$$

where  $\boldsymbol{\Phi}_s = \boldsymbol{\Psi}_s \boldsymbol{\Phi}_0$ . Let  $e_{j,t}$  be the the  $j$ -th element of  $\mathbf{e}_t$ , and  $\phi_{s,ij}$  be the  $(i, j)$ -th element of  $\boldsymbol{\Phi}_s$ , then

$$(8.E.8) \quad \frac{\partial y_{i,t+s}}{\partial e_{j,t}} = \phi_{s,ij}.$$

A plot of (8.E.8) as a function of  $s > 0$  is the orthogonalized impulse response function. Second, discuss conditions for  $\mathbf{B}_0$  under which the orthogonalized impulse response function represents the effects of each element of  $\mathbf{e}_t$  on  $\mathbf{z}_{t+s}$ . Let  $\mathbf{e}_t = \mathbf{B}_0\boldsymbol{\epsilon}_t$ . We need to identify  $\mathbf{B}_0$  so that  $n^2$  restrictions are necessary (not sufficient) for identification. The first assumption is that structural shocks are assumed to be orthogonal to each other so that the covariance matrix of  $\mathbf{e}_t$  is a diagonal matrix. This assumption gives  $\frac{n(n-1)}{2}$  conditions. Additional  $n$  conditions are given by the second normalization conditions that diagonal components are set to be one. Therefore, we need  $\frac{n(n-1)}{2}$  conditions for identification as follows:

- (i) Following Sims (1980), assume that  $\mathbf{B}_0$  is lower triangular. This assumption gives  $\frac{n(n-1)}{2}$  necessary conditions. From  $\mathbf{e}_t = \mathbf{B}_0\boldsymbol{\epsilon}_t$ ,  $\boldsymbol{\Lambda} = \mathbf{B}_0\boldsymbol{\Sigma}_\epsilon\mathbf{B}'_0$ . Therefore,

$$(8.E.9) \quad \boldsymbol{\Phi}_0\boldsymbol{\Lambda}\boldsymbol{\Phi}'_0 = \boldsymbol{\Sigma}_\epsilon.$$

Let  $\mathbf{P}$  be a lower triangular matrix of Cholesky decomposition of  $\boldsymbol{\Sigma}_\epsilon$  so that  $\mathbf{P}\mathbf{P}' = \boldsymbol{\Sigma}_\epsilon$ . From  $\boldsymbol{\Phi}_0\boldsymbol{\Lambda}^{\frac{1}{2}} = \mathbf{P}$ , it follows that

$$(8.E.10) \quad \boldsymbol{\Phi}_0 = \mathbf{P}\boldsymbol{\Lambda}^{-\frac{1}{2}}.$$

In this case,  $\mathbf{B}_0 = \boldsymbol{\Phi}_0^{-1}$ .

- (ii) Assume that  $\mathbf{B}_0$  is not lower triangular but has only  $\frac{n(n-1)}{2}$  unknown parameters. From  $\mathbf{e}_t = \mathbf{B}_0\boldsymbol{\epsilon}_t$

$$(8.E.11) \quad \boldsymbol{\Lambda} = \mathbf{B}_0\boldsymbol{\Sigma}_\epsilon\mathbf{B}'_0.$$

Blanchard and Watson (1986) directly solve (8.E.11) and get  $\mathbf{C}_0$ .

- (iii) When  $\mathbf{z}_t$  is  $I(1)$ , Blanchard and Quah (1989) considers a model with long run restrictions. Consider the structural form of

$$(8.E.12) \quad \Delta\mathbf{z}_t = \boldsymbol{\Phi}(L)\mathbf{e}_t$$

and the reduced form of

$$(8.E.13) \quad \Delta\mathbf{z}_t = \boldsymbol{\Psi}(L)\boldsymbol{\epsilon}_t.$$

where  $\mathbf{z}_t$  is  $I(1)$ . Assumption that  $\boldsymbol{\Phi}(1)$  is lower triangular gives  $\frac{n(n-1)}{2}$  necessary conditions. From  $\boldsymbol{\Phi}(1)\mathbf{e}_t = \boldsymbol{\Psi}(1)\boldsymbol{\epsilon}_t$

$$(8.E.14) \quad \boldsymbol{\Phi}(1)\boldsymbol{\Lambda}\boldsymbol{\Phi}(1)' = \boldsymbol{\Psi}(1)\boldsymbol{\Sigma}_\epsilon\boldsymbol{\Psi}(1)'.$$

Let  $\mathbf{P}$  be a lower triangular matrix of Cholesky decomposition of  $\boldsymbol{\Psi}(1)\boldsymbol{\Sigma}_\epsilon\boldsymbol{\Psi}(1)'$  so that  $\mathbf{P}\mathbf{P}' = \boldsymbol{\Psi}(1)\boldsymbol{\Sigma}_\epsilon\boldsymbol{\Psi}(1)'$ . From  $\boldsymbol{\Phi}(1)\boldsymbol{\Lambda}^{\frac{1}{2}} = \mathbf{P}$ , it follows that

$$(8.E.15) \quad \boldsymbol{\Phi}(1) = \mathbf{P}\boldsymbol{\Lambda}^{-\frac{1}{2}}$$

and from  $\boldsymbol{\Phi}(1) = \boldsymbol{\Psi}(1)\boldsymbol{\Phi}_0$

$$(8.E.16) \quad \boldsymbol{\Phi}_0 = \boldsymbol{\Psi}(1)^{-1}\boldsymbol{\Phi}(1).$$

**8.2** (a) True. (b) False. (c) True. (d) False. (e) False. (f) True.

Table C.1: GMM Results

	$\theta$	$\ln A$	$\rho_A$	$\sigma_\epsilon$	$A_y$	$\ln \gamma$	$\delta$	$\alpha$	$\sigma_i$
parameters	5.1199	0.1936	0.9663	0.0119	8.5722	0.0041	0.0209	0.6553	0.0427
standard errors	0.0590	0.1441	0.0250	0.0009	0.0201	0.0003	0.0003	0.0059	0.0044

Table C.2: Data moments and model moments

	Model Moments	Std. Error	Data Moments	Std. Error	Wald Test	P-value
$\sigma_i$	0.0880	0.0197	0.0427	0.0044	5.3699	0.0205

### Answers to Chapter 9

- 9.3** (a) The GMM estimates are given by the following Table C.1.
- (b) The data moments and model moments are given by Table C.2. The Wald test shows that we can reject the null hypothesis that these two moments are the same.
- (c) (The program) The program we need to modify is as follows:

```

nf=1 ; @<<<<<<@

fc = ((-cyratio/iyratio)^(alpha/iyratio)); @<<<<<<@ fx = ((1-alpha)/iyratio);
@<<<<<<@ fe = (1/iyratio); @<<<<<<@

mm = sd[6,1];@<<<<@

dm=b[9,1]; @<<<<<@

bgm=1|1|0.9|0.01|1|.01|.02|0.6|.01 ; @<<<<@

nw=9; @<<<<@           @ # of disturbance terms in w(t); scalar @

w9=hi[2:114,1]^2-b[9,1]^2 ; @<<<<@ retp(w1~w2~w3~w4~w5~w6~w7~w8~w9) ; @<<<<<<@

```

### Answers to Chapter 10

- 10.1** (i)  $\text{mas} = 0$  for HS, BG, and EZ since there is no serial correlation, while  $\text{mas} = 1$  for FC since there is one-lag serial correlation.
- (ii) (HS) Strength: Theoretically,  $\beta$  and  $\gamma$  have economic meanings of the discount rate and CRRA, respectively. Empirically,  $\frac{C_{t+1}}{C_t}$  tends to be stationary, which is necessary for GMM. Weakness: Theoretically, the utility function is time separable. It is easy to handle but might not be true. Empirically, the estimate of  $\gamma$  is so low that we can not explain such a high degree of risk aversion.

Since the model uses consumption data, it has a measurement error problem and a time aggregation problem.

(BG) Strength: Theoretically,  $\gamma$  has an economic meaning of CRRA. Empirically, it is free of a measurement error problem and a time aggregation problem. Weakness: Theoretically, the model has to assume that consumption is martingale to obtain constant  $E(\frac{C_{t+1}}{C_t} | I_t)$ . It is also hard to interpret  $\beta^*$ . Empirically, there occurs an identification problem when  $R_{t+1} = R_{t+1}^m$  in that we have a trivial solution of  $\beta^* = 1$  and  $\alpha = 1$ . It is also subject to Roll's critique since  $R_{t+1}^m$  is a value weighted return.

(EZ) Strength: Theoretically, it is a general expression. Empirically, we do not have to concern ourselves with stationarity. Weakness: Empirically, it has a measurement error problem and a time aggregation problem since it uses consumption data. It is also subject to Roll's critique since  $R_{t+1}^m$  is a value weighted return. There occurs an identification problem when  $R_{t+1} = R_{t+1}^m$  in that we have a trivial solution of  $\beta^+ = 1, \eta = -1$  and  $\theta = 0$ .

(FC) Strength: Theoretically, it can account for habit formation, which is plausible. Empirically, habit formation may help solve the equity premium puzzle. Weakness: Theoretically,  $S_{t+1}$  can be negative if  $a_1 > 1$ . Empirically, we need to modify the formula to be stationary since it may have an identification problem.

**10.2 (a) (i)** Let  $w_t = E(\sum_{i=1}^{\infty} \beta^i d_{t+i} | I_t) - \hat{E}(\sum_{i=0}^{\infty} \beta^i d_{t+i} | H_t)$ , then the present value formula is rewritten by

$$(10.E.1) \quad p_t = \hat{E}\left(\sum_{i=0}^{\infty} \beta^i d_{t+i} | H_t\right) + w_t.$$

where  $\hat{E}(w_t | H_t) = 0$ . From  $\hat{E}(d_t | H_{t-1}) = \phi d_{t-1}$ , we get  $\hat{E}(d_{t+i}) = \phi^i d_t$ , and (10.E.1) becomes

$$(10.E.2) \quad \begin{aligned} p_t &= \sum_{i=1}^{\infty} \beta^i \hat{E}(d_{t+i} | H_t) + w_t \\ &= \sum_{i=1}^{\infty} (\beta\phi)^i d_t + w_t \\ &= \frac{\beta\phi}{1 - \beta\phi} d_t + w_t. \end{aligned}$$

Thus,  $\delta = \frac{\beta\phi}{1 - \beta\phi}$ . Note that  $\hat{E}(w_t | H_t) = \hat{E}(\sum_{i=1}^{\infty} \beta^i d_{t+i} | H_t) - \hat{E}(\sum_{i=0}^{\infty} \beta^i d_{t+i} | H_t) = 0$  by the law of iterated expectations, while  $E(w_t | I_t) = E(\sum_{i=1}^{\infty} \beta^i d_{t+i} | I_t) - \hat{E}(\sum_{i=0}^{\infty} \beta^i d_{t+i} | H_t) \neq 0$ . Because  $I_t$  is bigger than  $H_t$  and agents generally use a non-linear forecasting rule whereas econometricians use a linear forecasting rule, it is impossible to prove that  $E(w_t | I_t) = 0$ .

- (ii) Since  $w_t$  is not necessarily in  $H_{t+1}$ , generally  $E(w_t w_{t+1}) \neq 0$ . To see this, suppose that  $p_t$  is in  $H_{t+1}$ . Then, from (10.E.4),  $w_t$  is in  $H_{t+1}$ . Since  $\hat{E}(w_{t+i} | H_{t+1}) = 0$  for  $i \geq 1$  due to the law of iterated projection, it follows from the orthogonality condition that  $E(w_t w_{t+i} | H_{t+1}) = 0$  for  $i \geq 1$ . Therefore,  $w_t$  is serially uncorrelated with the additional assumption that  $p_t$  is in  $H_{t+1}$ . However, this additional assumption is not realistic, because  $p_t$  is the expectation of future  $d_t$  conditional on the information set that is generated by a *nonlinear* function of  $\{d_t, d_{t-1}, d_{t-2}, \dots\}$ .
- (iii) We can exploit the three equations to estimate  $\beta$  and  $\phi$  in the framework of Generalized Method of Moments, imposing the restriction on  $\delta$  we derived.

$$(10.E.3) \quad p_t = \beta E(p_{t+1} + d_{t+1} | I_t)$$

$$(10.E.4) \quad \hat{E}(d_t | \mathbf{H}_{t-1}) = \phi d_{t-1}$$

$$(10.E.5) \quad p_t = \delta d_t + w_t$$

From the equation (10.E.3),

$$(10.E.6) \quad p_t = \beta(p_{t+1} + d_{t+1}) + u_{t+1},$$

where  $E(u_{t+1} | \mathbf{I}_t) = 0$ . Let  $\mathbf{z}_{2t}$  be a random variable in  $\mathbf{I}_t$ , then we obtain a orthogonality condition

$$(10.E.7) \quad E(\mathbf{z}_{2t} u_{t+1}) = 0$$

From the equation (10.E.4),

$$(10.E.8) \quad d_{t+1} = \phi d_t + v_{t+1},$$

where  $\hat{E}(v_{t+1} | \mathbf{H}_{t+1}) = 0$ . Note that  $\hat{E}(v_{t+1} | \mathbf{H}_t) = 0$  due to the law of iterated projection. Let  $\mathbf{z}_{1t}$  be a random variable in  $\mathbf{H}_t$ , then we can obtain the second orthogonality condition

$$(10.E.9) \quad E(\mathbf{z}_{1t} v_{t+1}) = 0$$

From the equation (10.E.5),  $\hat{E}(w_t | \mathbf{H}_t) = 0$  and we can get third orthogonality condition

$$(10.E.10) \quad E(\mathbf{z}_{1t} w_t) = 0$$

Based on equations (10.E.7), (10.E.9), and (10.E.10), we have the following moment conditions with a restriction  $\delta = \frac{\beta\phi}{1-\beta\phi}$ :

$$(10.E.11) \quad E(f(\mathbf{x}_t, \beta, \phi)) = \mathbf{0},$$

where, parameterized disturbances are

$$f(\mathbf{x}_t, \beta, \phi) = \begin{bmatrix} \mathbf{z}_{2t}(p_t - \beta(p_{t+1} + d_{t+1})) \\ \mathbf{z}_{1t}(d_{t+1} - \phi d_t) \\ \mathbf{z}_{1t}(p_t - \frac{\beta\phi}{1-\beta\phi} d_t) \end{bmatrix}$$

where  $\mathbf{x}_t = (d_t, p_t)'$  and valid instrument variables are  $\mathbf{z}_{1t}$  which is in  $\mathbf{H}_t$  and  $\mathbf{z}_{2t}$  which is in  $\mathbf{I}_t$ . In order to compute the long-run covariance matrix, one should use either pre-whitened QS kernel or VARHAC estimator since  $w_t$  has an unknown order of serial correlation.

- (iv) One can use the Wald test, LM test, or LR test to test the restriction. Under a set of regularity conditions, these tests have the same asymptotic  $\chi^2(q)$  distribution, in which  $q$  is the number of restrictions, in particular  $q = 1$  in this example. The LM test and LR test are better than the Wald test because the latter not only has poor small sample properties but also depends on parameterization of nonlinear restrictions.
- (b) Adding and subtracting  $\sum_{i=1}^{\infty} \beta^i d_t$  in the right-hand side of the present value formula,  $p_t = \sum_{i=1}^{\infty} \beta^i E(d_{t+i} | \mathbf{I}_t)$  yields,

$$p_t - \frac{\beta}{1-\beta} d_t = \sum_{i=1}^{\infty} \beta^i E(d_{t+i} - d_t | \mathbf{I}_t)$$

As  $d_t$  is a difference stationary process,  $d_{t+i} - d_t$  is stationary for  $i > 0$ . Thus, the right hand side is stationary. Hence we obtain a stationarity restriction that  $p_t - \frac{\beta}{1-\beta} d_t$  is stationary. This restriction implies that  $p_t$  and  $d_t$  are cointegrated with a cointegrating vector  $(1, -\frac{\beta}{1-\beta})$ .

- 10.3** (a) (i) Let  $w_t = \frac{1}{1-\alpha} E(\sum_{i=0}^{\infty} (\frac{\alpha}{\alpha-1})^i m_{t+i} | I_t) - \frac{1}{1-\alpha} \hat{E}(\sum_{i=0}^{\infty} (\frac{\alpha}{\alpha-1})^i m_{t+i} | H_t)$ , then (10.E.8) is rewritten by

$$(10.E.12) \quad p_t = \frac{1}{1-\alpha} \hat{E}\left(\sum_{i=0}^{\infty} \left(\frac{\alpha}{\alpha-1}\right)^i m_{t+i} | H_t\right) + w_t.$$

where  $\hat{E}(w_t | H_t) = 0$ . From (10.E.9), we get  $\hat{E}(m_{t+i}) = \phi^i m_t$ , and (10.E.12) becomes

$$(10.E.13) \quad \begin{aligned} p_t &= \frac{1}{1-\alpha} \hat{E}\left(\sum_{i=0}^{\infty} \left(\frac{\alpha\phi}{\alpha-1}\right)^i m_t | H_t\right) + w_t \\ &= \frac{1}{1-\alpha} \sum_{i=0}^{\infty} \left(\frac{\alpha\phi}{\alpha-1}\right)^i m_t + w_t \\ &= \frac{1}{1-\alpha + \alpha\phi} m_t + w_t. \end{aligned}$$

Thus,  $\delta = \frac{1}{1-\alpha + \alpha\phi}$ .

- (ii)  $w_t$  is serially correlated in general. To see why, suppose that  $p_t$  is in  $H_{t+1}$ . Then, from (10.E.10),  $w_t$  is in  $H_{t+1}$ . Since  $\hat{E}(w_{t+i} | H_{t+1}) = 0$  for  $i \geq 1$ , it follows from the orthogonality condition that  $E(w_t w_{t+i} | H_{t+1}) = 0$  for  $i \geq 1$ . Therefore,  $w_t$  is serially uncorrelated with the additional assumption that  $p_t$  is in  $H_{t+1}$ . However, this additional assumption is not realistic, because  $p_t$  is the expectation of future  $m_t$  conditional on the information set that is generated by a *nonlinear* function of  $\{m_t, m_{t-1}, m_{t-2}, \dots\}$ .
- (iii) Let  $u_t = p_t - E(p_t | I_{t-1})$ . From (10.E.7) we get

$$(10.E.14) \quad p_t - p_{t-1} = \frac{1}{\alpha} (m_{t-1} - p_{t-1}) + u_t.$$

Since  $E(u_t | I_{t-1}) = 0$  and  $m_{t-1} - p_{t-1}$  is in  $I_{t-1}$ ,  $E(u_t (m_{t-1} - p_{t-1})) = 0$  by the orthogonality condition so that we get an unbiased estimator of  $\alpha$  by taking the reciprocal of the OLS estimate of (10.E.14).

- (iv) From  $\hat{E}(w_t | H_t) = 0$ ,  $\hat{E}(v_{t+1} | H_t) = 0$ , and  $E(u_{t+1} | I_t) = 0$ , we get the following moment conditions with a restriction,  $\delta = \frac{1}{1-\alpha + \alpha\phi}$ :

$$(10.E.15) \quad \begin{aligned} f(\mathbf{x}_t, \alpha, \phi) &= \begin{bmatrix} \mathbf{z}_{1t} \left( p_t - \frac{1}{1-\alpha + \alpha\phi} m_t \right) \\ \mathbf{z}_{1t} (m_{t+1} - \phi m_t) \\ \mathbf{z}_{2t} \left( p_t - p_{t-1} - \frac{1}{\alpha} (m_{t-1} - p_{t-1}) \right) \end{bmatrix} \\ E(f(\mathbf{x}_t, \alpha, \phi)) &= \mathbf{0}, \end{aligned}$$

where  $\mathbf{x}_t = (m_t, p_t)'$ ,  $\mathbf{z}_{1t}$  is in  $H_t$ , and  $\mathbf{z}_{2t}$  is in  $I_t$ . One can use GMM to estimate parameters using the above moment conditions. In order to compute the long-run covariance matrix, one should use either a prewhitened QS kernel or VARHAC estimator since  $w_t$  has an unknown order of serial correlation.

- (v) One can use the Wald test, LM test, or LR test to test the restriction. Under a set of regularity conditions, these tests have the same asymptotic  $\chi^2(q)$  distribution, in which  $q$  is the number of restrictions, in particular  $q = 1$  in this example. The LM test and LR test are better than the Wald test because the latter not only has poor small sample properties but also depends on parameterization of nonlinear restrictions.
- (b) Since  $m_t$  is a DSP,  $p_t$  is also a DSP from (10.E.8). Note that  $p_{t+i} - p_t$  is stationary for  $i > 0$ . Thus,  $E(p_{t+1} - p_t | I_t)$  is stationary. Therefore,  $m_t - p_t$  is stationary from (10.E.7) so that  $m_t$  and  $p_t$  are cointegrated with a known vector  $(1, -1)'$ .

## Answers to Chapter 15

- 15.1 (a)** The first order condition implies that the relative price is equal to the marginal rate of substitution. So,  $P_{2t} = \frac{MU_2}{MU_1}$ , where  $MU_2 = \beta^t \sigma_2 e^{\theta t} (S_{2t}^{-\alpha_2} + \beta e^\theta \delta S_{2,t+1}^{-\alpha_2})$  and  $MU_1 = \beta^t C_{1t}^{-\alpha_1}$ . Thus,

$$\begin{aligned} P_{2t} &= \sigma_2 e^{\theta t} E_t \left[ \frac{(e^{\theta t} (C_{2t} + \delta C_{2,t-1}))^{-\alpha_2} + \beta e^\theta \delta (e^{\theta(t+1)} (C_{2,t+1} + \delta C_{2t}))^{-\alpha_2}}{C_{1t}^{-\alpha_1}} \right] \\ &= \sigma_2 e^{(1-\alpha_2)\theta t} \frac{(C_{2t} + \delta C_{2,t-1})^{-\alpha_2} + \beta e^{(1-\alpha_2)\theta} \delta E_t (C_{2,t+1} + \delta C_{2t})^{-\alpha_2}}{C_{1t}^{-\alpha_1}}. \end{aligned}$$

Therefore,  $\frac{P_{2t} C_{1t}^{-\alpha_1}}{C_{2t}^{-\alpha_2} e^{(1-\alpha_2)\theta t}} = \sigma_2 (1 + \delta \frac{C_{2,t-1}}{C_{2t}})^{-\alpha_2} + \beta e^{(1-\alpha_2)\theta} \delta E_t (\frac{C_{2,t+1}}{C_{2t}} + \delta)^{-\alpha_2}$ .

- (b)** Note that  $\frac{S_{2t}}{C_{2t}} = \frac{e^{\theta t} (C_{2t} + \delta C_{2,t-1})}{C_{2t}} = e^{\theta t} (1 + \delta \frac{C_{2,t-1}}{C_{2t}})$ . So,  $\ln \frac{S_{2t}}{C_{2t}} = \theta t + \ln(1 + \delta \frac{C_{2,t-1}}{C_{2t}})$ , where the second component is stationary since  $\frac{C_{2,t-1}}{C_{2t}}$  is assumed to be stationary. Therefore,  $\ln \frac{S_{2t}}{C_{2t}}$  is trend stationary.
- (c)** Definitions: A set of variables is stochastically cointegrated if stochastic trends are eliminated by a linear combination of difference stationary variables. If the linear combination eliminates both stochastic and deterministic trends, the deterministic cointegration restriction is satisfied. To consider the following cases, take a log of the last equation in (a) and denote  $p_{2t} = \ln P_{2t}$ ,  $c_{1t} = \ln C_{1t}$ , and  $c_{2t} = \ln C_{2t}$ . Then,  $p_{2t} - \alpha_1 c_{1t} + \alpha_2 c_{2t} - (1 - \alpha_2)\theta t = \ln \sigma_2 + \ln(1 + \delta \frac{C_{2,t-1}}{C_{2t}})^{-\alpha_2} + \beta e^{(1-\alpha_2)\theta} \delta E_t (\frac{C_{2,t+1}}{C_{2t}} + \delta)^{-\alpha_2}$ , where the right hand side is stationary since  $\frac{C_{2,t-1}}{C_{2t}}$  and  $\frac{C_{2,t+1}}{C_{2t}}$  are stationary. Therefore, the left hand side is also stationary. We also assume that  $c_{1t}$  and  $c_{2t}$  are not cointegrated, which implies  $p_{2t}$  is nonstationary.
- (i)** Case 1: If  $\theta = 0$  and  $c_{it}$  is difference stationary for  $i = 1, 2$ , then  $p_{2t} - \alpha_1 c_{1t} + \alpha_2 c_{2t}$  is stationary. Therefore,  $p_{2t}$ ,  $c_{1t}$ , and  $c_{2t}$  are cointegrated with a cointegrating vector  $(1, -\alpha_1, \alpha_2)'$ , and the deterministic cointegration restriction is satisfied. By the property of cointegration,  $\alpha_1$  and  $\alpha_2$  are identified.
- (ii)** Case 2: If  $\theta \neq 0$  and  $c_{it}$  is difference stationary for  $i = 1, 2$ , then  $p_{2t} - \alpha_1 c_{1t} + \alpha_2 c_{2t}$  is trend stationary. Therefore,  $p_{2t}$ ,  $c_{1t}$ , and  $c_{2t}$  are cointegrated with a cointegrating vector  $(1, -\alpha_1, \alpha_2)'$ , and the deterministic cointegration restriction is not satisfied. By the property of cointegration,  $\alpha_1$ ,  $\alpha_2$  and  $\theta$  are identified.
- (iii)** Case 3: If  $\theta = 0$  and  $c_{1t}$  is difference stationary and  $c_{2t}$  is stationary, then  $p_{2t} - \alpha_1 c_{1t}$  is stationary. Therefore,  $p_{2t}$ , and  $c_{1t}$  are cointegrated with a cointegrating vector  $(1, -\alpha_1)'$ , and the deterministic cointegration restriction is satisfied. By the property of cointegration,  $\alpha_1$  is identified.
- (iv)** Case 4: If  $\theta \neq 0$  and  $c_{1t}$  is difference stationary and  $c_{2t}$  is stationary, then  $p_{2t} - \alpha_1 c_{1t}$  is trend stationary. Therefore,  $p_{2t}$ , and  $c_{1t}$  are cointegrated with a cointegrating vector  $(1, -\alpha_1)'$  and the deterministic cointegration restriction is not satisfied. By the property of cointegration,  $\alpha_1$  is identified.
- (v)** Case 5: If  $\theta = 0$  and  $c_{1t}$  is difference stationary and  $c_{2t}$  is trend stationary with a nonzero linear trend, then  $p_{2t} - \alpha_1 c_{1t}$  is trend stationary. Therefore,  $p_{2t}$ , and  $c_{1t}$  are cointegrated with a cointegrating vector  $(1, -\alpha_1)'$ , and the deterministic cointegration restriction is not satisfied. On the other hand,  $p_{2t}$ ,  $c_{1t}$ , and  $c_{2t}$  are cotrended with a cotrending vector  $(1, -\alpha_1, \alpha_2)'$ , and the deterministic cointegration restriction is not satisfied. By the property of cointegration,  $\alpha_1$  and  $\alpha_2$  are identified.

- (vi) Case 6: If  $\theta \neq 0$  and  $c_{1t}$  is difference stationary and  $c_{2t}$  is trend stationary with a nonzero linear trend, then  $p_{2t} - \alpha_1 c_{1t}$  is trend stationary. Therefore,  $p_{2t}$ , and  $c_{1t}$  are cointegrated with a cointegrating vector  $(1, -\alpha_1)'$ , and the deterministic cointegration restriction is not satisfied. By the property of cointegration,  $\alpha_1$  is identified.

**15.2 (a)** GPQ tests do not reject the null of trend stationary at 5% significance level. This implies that nondurables and durables are trend stationary.

Table C.3: GPQ tests

	$\ln C_{1t}$	$\ln C_{2t}$
G(1,2)	3.7606 (0.0525)	0.6232 (0.4298)
G(1,3)	3.7647 (0.1522)	0.6256 (0.7314)

Note: The numbers in the parenthesis denote p-values.

- (b) The ADF test for nondurables does not reject the null of difference stationary at 5% significance level, while the test for durables rejects the null of difference stationary at 5% significance level. This implies that nondurables are difference stationary and durables are trend stationary.

Table C.4: ADF tests

	$\ln C_{1t}$	$\ln C_{2t}$
Coefficient	0.9391 (-2.3967)	0.8670 (-3.4090)
lag-length	11	11

Note: The ADF test statistics are computed by a regression equation with a constant and a time trend. The lag-length is chosen following Campbell and Perron (1991) with maximum lag length 20. The numbers in the parenthesis denote t-statistics.

- (c) CCR estimation shows that  $\alpha_1 = 1.9474$  and  $\alpha_2 = 0.9629$ , while the Wald test rejects the null hypothesis that the coefficients are equal to one.  $H(p, q)$  tests results are mixed. The deterministic cointegration restriction is rejected at 5% significance level.  $H(1, 2)$  does not reject the null of cointegration, while  $H(1, 3)$  rejects the null at 5% significance level.

.11

## Answers to Chapter 10.11

## Answers to Chapter 17

**17.1 (a)** Denote  $H_t = e_t | e_0$ . Then, the budget constraint is given by

$$(17.E.5) \quad \sum_{t=0}^T \sum_{H_t} p(H_t) c(H_t) \leq \sum_{t=0}^T \sum_{H_t} p(H_t) c^*(H_t),$$

where  $c_t$  is consumption,  $c_t^*$  is endowments, and  $p_t$  is the price of consumption goods.

Table C.5: CCR estimation and H(p,q) tests

Estimation		
Coefficients	$\ln P_{2t}$	$\ln C_{2t}$
Inferred Coefficients	$\frac{1}{\alpha_1} = 0.5135(0.1219)$	$\frac{\alpha_2}{\alpha_1} = 0.4945(0.0398)$
$H_0$ : coefficients are the same as 1	$\alpha_1 = 1.9474$	$\alpha_2 = 0.9629$
	Wald test	$p$ -value
	532.8906	0.0000
Cointegration Tests	Test statistics	$p$ -value
H(0,1)	5.2972	0.0214
H(1,2)	0.2711	0.6026
H(1,3)	15.4890	0.0004

Note: The numbers in the parenthesis denote standard errors.

- (b) Let  $\mathcal{L} = \sum_{t=0}^T \sum_{H_t} \text{Prob}(H_t) \beta^t u_t - \lambda (\sum_{t=0}^T \sum_{H_t} p(H_t) c(H_t) - \sum_{t=0}^T \sum_{H_t} p(H_t) c^*(H_t))$ , then the FOCs are given by

$$(17.E.6) \quad \begin{aligned} c_t &: \text{Prob}(H_t) \beta^t (c_t - \gamma)^{-\alpha} = \lambda P(H_t) \\ c_{t+1} &: \text{Prob}(H_{t+1}) \beta^{t+1} (c_{t+1} - \gamma)^{-\alpha} = \lambda P(H_{t+1}). \end{aligned}$$

It follows from (17.E.5) that

$$(17.E.7) \quad \frac{p(H_{t+1})}{p(H_t)} = \beta \frac{\text{Prob}(H_{t+1})}{\text{Prob}(H_t)} \left( \frac{c_{t+1} - \gamma}{c_t - \gamma} \right)^{-\alpha}.$$

Therefore,  $\frac{c_{t+1} - \gamma}{c_t - \gamma}$  is identical for all consumers and for all history, which implies complete risk sharing. This in turn implies that consumption grows at the same rate for all consumers. From the arbitrage condition

$$(17.E.8) \quad \begin{aligned} v(H_t) &= \frac{\sum_{H_{t+1}|H_t} p(H_{t+1}) d(H_{t+1})}{p(H_t)} \\ &= \sum_{H_{t+1}|H_t} \beta \frac{\text{Prob}(H_{t+1})}{\text{Prob}(H_t)} \left( \frac{c_{t+1} - \gamma}{c_t - \gamma} \right)^{-\alpha} d(H_{t+1}) \end{aligned}$$

the asset pricing formula is given by

$$(17.E.9) \quad v_t = E\left(\beta \left(\frac{c_{t+1} - \gamma}{c_t - \gamma}\right)^{-\alpha} d_{t+1} | \mathbb{I}_t\right)$$

or

$$(17.E.10) \quad 1 = E_t\left(\beta \left(\frac{c_{t+1} - \gamma}{c_t - \gamma}\right)^{-\alpha} R_{t+1} | \mathbb{I}_t\right),$$

where  $R_{t+1} = \frac{d_{t+1}}{v_t}$ .

- (c) Let  $\epsilon_t^h = \beta \left(\frac{c_{t+1} - \gamma}{c_t - \gamma}\right)^{-\alpha} R_{t+1} - 1$ , where  $h$  denotes each household. Then,  $E(\epsilon_t^h | \mathbb{I}_t) = 0$ . From complete risk sharing, each household is identical so that  $\epsilon_t^h = \epsilon_t^1$  for all  $h$ . Therefore,  $\lim_{T \rightarrow \infty} \frac{1}{T} \sum_t \epsilon_t^h = \lim_{T \rightarrow \infty} \frac{1}{T} \sum_t \epsilon_t^1 = 0$ , while  $\lim_{N \rightarrow \infty} \frac{1}{N} \sum_h \epsilon_t^h = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_h \epsilon_t^1 = \epsilon_t^1 \neq 0$ . Therefore the GMM estimator is not consistent unless  $T$  is large enough. Let  $\mathbf{f}_t = \mathbf{z}_t \epsilon_t$ , where  $\mathbf{z}_t$  is in  $\mathbb{I}_t$ . For example, one can choose  $\mathbf{z}_t = (1, \frac{c_t}{c_{t-1}}, \frac{c_{t-1}}{c_{t-2}}, R_t, R_{t-1})'$ . Note

also that  $\epsilon_t$  is serially uncorrelated because  $\epsilon_t$  is in  $I_{t+1}$ . Thus,  $\mathbf{f}_t$  is also serially uncorrelated. One can use the GMM to estimate the parameters using

$$(17.E.11) \quad \text{Min}\left\{\frac{1}{T} \sum_{t=1}^T f(\mathbf{x}_t, \mathbf{b})\right\}' \mathbf{W}_T \left\{\frac{1}{T} \sum_{t=1}^T f(\mathbf{x}_t, \mathbf{b})\right\}.$$

The optimal choice of the weighting matrix is  $\mathbf{W}_T = \mathbf{\Omega}^{-1}$ , where the long-run covariance matrix is given by  $\frac{1}{T} \sum_{t=1}^T \mathbf{f}_t \mathbf{f}_t'$  since  $\mathbf{f}_t$  is serially uncorrelated.

- (d) Note that one can not use the disturbances in (c) since  $R_t$  is not available. Instead, one can use the property of complete risk sharing. Let  $\phi_{t+1}$  denote

$$(17.E.12) \quad \begin{aligned} \phi_{t+1} &= \frac{c_{t+1}^{h*} - \gamma}{c_t^{h*} - \gamma} \\ &= \frac{c_{t+1}^h - \gamma}{c_t^h - \gamma} \frac{\epsilon_t^h}{\epsilon_{t+1}^h}. \end{aligned}$$

By taking log of (17.E.12), we get

$$(17.E.13) \quad \log \epsilon_{t+1}^h - \log \epsilon_t^h = \log(c_{t+1}^h - \gamma) - \log(c_t^h - \gamma) - \log \phi_{t+1}.$$

Let  $\tilde{\epsilon}_{t+1}^h = \log(c_{t+1}^h - \gamma) - \log(c_t^h - \gamma) - \log \phi_{t+1}$ . Then,  $\lim_{N \rightarrow \infty} \frac{1}{N} \sum_h \tilde{\epsilon}_{t+1}^h = 0$  since  $\log \epsilon_t^h$  and  $\log \epsilon_{t+1}^h$  have mean zero. Thus, we have  $E(\mathbf{z}_{t+1} \tilde{\epsilon}_{t+1}^h) = \mathbf{0}$ , where  $\mathbf{z}_{t+1} = (1, y^p, \Delta y_{t+1})'$ ,  $y^p$  is a proxy variable of permanent income, and  $y_t$  is an income variable. In particular,  $y_{t+1}$  is taken difference to use stationary instruments. Since  $T = 6$ , we have the following 5 parameterized disturbances:

$$(17.E.14) \quad \begin{aligned} f(\mathbf{x}^h, \mathbf{b}) &= \begin{bmatrix} \mathbf{z}_2^h \tilde{\epsilon}_2^h \\ \mathbf{z}_3^h \tilde{\epsilon}_3^h \\ \mathbf{z}_4^h \tilde{\epsilon}_4^h \\ \mathbf{z}_5^h \tilde{\epsilon}_5^h \\ \mathbf{z}_6^h \tilde{\epsilon}_6^h \end{bmatrix} \\ E(f(\mathbf{x}^h, \mathbf{b})) &= \mathbf{0}, \end{aligned}$$

where  $\mathbf{b} = (\phi_2, \phi_3, \phi_4, \phi_5, \phi_6, \gamma)'$ . We have 15 moment restrictions and 6 unknown parameters, One can use the GMM to estimate parameters using

$$(17.E.15) \quad \text{Min}\left\{\frac{1}{N} \sum_{h=1}^N f(\mathbf{x}^h, \mathbf{b})\right\}' \mathbf{W}_N \left\{\frac{1}{N} \sum_{h=1}^N f(\mathbf{x}^h, \mathbf{b})\right\}.$$

The optimal choice of the weighting matrix is  $\mathbf{W}_N = \mathbf{\Omega}^{-1}$ , where the long-run covariance matrix is given by  $\frac{1}{N} \sum_{h=1}^N \mathbf{f}^h \mathbf{f}^{h'}$  since  $\mathbf{f}^h$  is uncorrelated across consumers.

- (e) Let  $\phi_{t+1}$  denote

$$(17.E.16) \quad \phi_{t+1} = \frac{c_{t+1}^{h*} - \gamma}{c_t^{h*} - \gamma}.$$

From (17.E.16) we have

$$(17.E.17) \quad (c_{t+1}^h - \epsilon_{t+1}^h - \gamma) = \phi_{t+1} (c_t^h - \epsilon_t^h - \gamma)$$

or

$$(17.E.18) \quad \epsilon_{t+1}^h - \phi_{t+1} \epsilon_t^h = c_{t+1}^h - \phi_{t+1} c_t^h - \gamma + \phi_{t+1} \gamma.$$

Let  $\tilde{\epsilon}_{t+1}^h = c_{t+1}^h - \phi_{t+1}c_t^h - \gamma + \phi_{t+1}\gamma$ . Then,  $\lim_{N \rightarrow \infty} \frac{1}{N} \sum_h \tilde{\epsilon}_{t+1}^h = 0$  since  $\epsilon_t^h$  and  $\epsilon_{t+1}^h$  have mean zero. Thus, we have  $E(\mathbf{z}_{t+1}^h \tilde{\epsilon}_{t+1}^h) = \mathbf{0}$ , where  $\mathbf{z}_{t+1}^h = (1, y^p, \Delta y_{t+1})'$ ,  $y^p$  is a proxy variable of permanent income, and  $y_t$  is an income variable. In particular, we take the first difference of  $y_{t+1}$  to use stationary instruments. Since  $T = 6$ , we have the following 5 parameterized disturbances:

$$(17.E.19) \quad f(\mathbf{x}^h, \mathbf{b}) = \begin{bmatrix} \mathbf{z}_2^h \epsilon_2^h \\ \mathbf{z}_3^h \epsilon_3^h \\ \mathbf{z}_4^h \epsilon_4^h \\ \mathbf{z}_5^h \epsilon_5^h \\ \mathbf{z}_6^h \epsilon_6^h \end{bmatrix}$$

$$E(f(\mathbf{x}^h, \mathbf{b})) = \mathbf{0},$$

where  $\mathbf{b} = (\phi_2, \phi_3, \phi_4, \phi_5, \phi_6, \gamma)'$ . We have 15 moment restrictions and 6 unknown parameters. One can use the GMM to estimate parameters using

$$(17.E.20) \quad \text{Min} \left\{ \frac{1}{N} \sum_{h=1}^N f(\mathbf{x}^h, \mathbf{b}) \right\}' \mathbf{W}_N \left\{ \frac{1}{N} \sum_{h=1}^N f(\mathbf{x}^h, \mathbf{b}) \right\}.$$

The optimal choice of the weighting matrix is  $\mathbf{W}_N = \mathbf{\Omega}^{-1}$ , where the long-run covariance matrix is given by  $\frac{1}{N} \sum_{h=1}^N \mathbf{f}^h \mathbf{f}^{h'}$  since  $\mathbf{f}^h$  is uncorrelated across consumers.