SOME ASPECTS OF ESTIMATION FOR VECTOR TIME SERIES MODELS

By

MOSES OLUFEMI SALAU

A Thesis submitted to the
AUSTRALIAN NATIONAL UNIVERSITY
for the Degree of
DOCTOR OF PHILOSOPHY

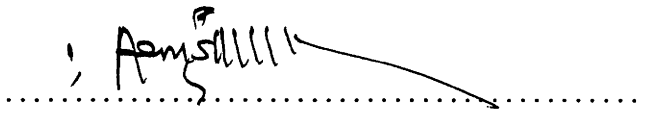
MARCH, 1993

DECLARATION

The results of Chapter 6 have appeared in Communications in Statistics volume 22 issue numbered 2 of 1993, and those of Chapters 3 and 4 are currently being reviewed for publication.

Part of the results of Chapter 4 were presented by the author at the eleventh Australian Statistical Association Conference in July, 1992, and the results of Chapter 5 are under review for publication in Biometrika.

Except where otherwise stated, the work presented in this thesis is my own.



Acknowledgements

It is a pleasure to thank my Supervisor, Dr. D. S. Poskitt, for introducing me to vector time series; for his enthusiastic supervision and continuing interest in the problems addressed in this thesis. I am greatly indebted to Professor E. J. Hannan and Dr. D. F. Nicholls, who are members of my supervisory panel, for stimulating discussions and helpful suggestions on several issues that arose during the preparation of this thesis.

I would like to thank a number of people who have also read the draft chapters of the thesis. In particular, my deep sense of appreciation goes to Dr. A. Welsh, Dr. J. Robertson, Dr. R.A. Adewusi, L. Kaggwa and my colleague, M. B. Obwona, for their help and suggested improvements.

Special thanks goes to my wife (Margaret) and children (Titilayo, Ayodeji and Adetokunbo) for their unfailing support and especially for bringing me back to reality when my mind has been on this thesis.

I gratefully acknowledge the financial support of the Australian Commonwealth Scholarship and Fellowship Award.

Finally, I express profound gratitude to God for giving me the fortitude to bear the death of my son who died when this research was in progress.

Contents

Summary

This thesis is primarily concerned with some aspects of estimation for vector autoregressive moving average models which are in their appropriate echelon canonical forms. We restrict attention to the most straightforward part of the modelling procedure, namely, the estimation for fixed values of the Kronecker indices of the structural parameters using ordinary least squares, Gaussian (maximum) likelihood and generalized least squares methods, respectively. Our primary objective here is to give a systematic account of these procedures for handling data and also to provide a thorough exposition of the mathematical details that underlie the techniques. In addition to these abstract mathematical derivations, emphasis will be placed on the practical aspects of the procedures. The discussion of these various issues is organized into six chapters as follows:

In Chapter 1 we introduce the class of models and assumptions upon which the results obtained in the thesis are based, and the justification for adopting an echelon structure for such models is also provided. This introductory chapter concludes with a description of the identification procedures for echelon canonical forms. Chapter 2 considers the estimation of the structural parameters using maximum (Gaussian) likelihood procedure and the asymptotic properties of the corresponding estimators are presented. In the evaluation of the parameter estimates, however, explicit expressions are derived for the gradient vector and (approximate) Hessian matrix of the log likelihood function in relatively simple terms.

Chapter 3 commences with a procedure for evaluating the least squares estimators. Consistency and asymptotic normality results are established. Chapter 4 assesses the asymptotic relative efficiency of the Gaussian and least squares estimators via the variance-covariance matrices of the limiting normal distributions obtained in Chapters 2 and 3, respectively. Situations under which substantial loss or gain in efficiency could be expected are discussed and illustrated with some numerical examples.

Chapter 5 is devoted to a detailed discussion of the generalized least squares (GLS) procedure for parameter estimation. In particular, the theoretical aspect of the relationship between the GLS and Gaussian estimation methods is investigated and the asymptotic convergence of the GLS estimator to the Gaussian estimator is established. Also, an alternative numerical method for implementing the GLS procedure is proposed and some simulation results are presented to illustrate the theory.

Finally, in Chapter 6, a method for generating a stable spectral factor from an unstable $v \times v$ full rank polynomial operator using closed form algebraic manipulations is proposed. An application of the technique is illustrated and the implementation of the method in the statistical context of system estimation is discussed.

Notation

We now list some of the symbols and abbreviations used in the text. Those symbols that are not standard are either explained now or will be explained when they first appear in the text.

Symbol	Interpretation
T	sample size
$\mathbf{I}_v$	$v \times v$ identity matrix
δ_{st}	Kronecker delta function,
	$\delta_{st} = 1$ if $s = t$ and 0 otherwise
Q_T	$(\log\log T/T)^{\frac{1}{2}}$
H_T	$(\log T)^{\tau}$, $\tau < \infty$
θ	vector of unknown parameters
$\Re$	a field of real numbers
$\mathbf{A} = [a_{ij}]_{i,j=1,\ldots,v}$	a $v \times v$ matrix with a_{ij} as the element in the i^{th} row
	and j^{th} column of $\mathbf{A}$
$\|\mathbf{A}\|$	Euclidean matrix norm of $\mathbf{A}$, $\|\mathbf{A}\| = (\sum_i \sum_j a_{ij}^2)^{\frac{1}{2}}$
$\|\mathbf{A}\|_\infty$	Banach (row sum) norm of $\mathbf{A}$, $\max_i(\sum_j a_{ij})$
$\mathcal{E}$	expectation operator
$\mathbf{A}^\dagger$	Moore-Penrose generalized inverse of the matrix $\mathbf{A}$
$\mathbf{A}^{-1}$	inverse of matrix $\mathbf{A}$
$\hat{\mathbf{A}}$	estimate of $\mathbf{A}$
x'	transpose of vector x

$\mathbf{A}'$, $\mathbf{A}^*$	matrix transpose, conjugate transpose
$tr(\mathbf{A})$, $tr\,\mathbf{A}$	trace of $\mathbf{A}$
$\det(\mathbf{A})$, $\det\,\mathbf{A}$	determinant of $\mathbf{A}$
$adj(\mathbf{A})$, $adj\,\mathbf{A}$	adjoint (or adjugate) of $\mathbf{A}$
$\lambda_{min}(\mathbf{A})$	minimum eigen value of the matrix $\mathbf{A}$
$\lambda_{max}(\mathbf{A})$	maximum eigen value of the matrix $\mathbf{A}$
$\square$	end of proof; end of statement
$\in$	belongs to
$\rightarrow$	tends to (approaches)
$v(T) = O(C_T)$	$\limsup_{T\rightarrow\infty} \lvert v(T)\rvert / C_T < \infty$ almost surely
	for $v(T)$ a sequence of random variable,
	C_T a sequence of constants
$v(T) = o(C_T)$	$\limsup_{T\rightarrow\infty} \lvert v(T)\rvert / C_T = 0$ almost surely
$v(T) = O_p(C_T)$	$\limsup_{T\rightarrow\infty} \lvert v(T)\rvert / C_T < \infty$ in probability
$v(T) = o_p(C_T)$	$\limsup_{T\rightarrow\infty} \lvert v(T)\rvert / C_T = 0$ in probability
$\mathcal{F}_t$	$\sigma-$algebra generated by $\varepsilon(s)$, $s \leq t$
$diag(a_1, \ldots, a_v)$	the diagonal matrix with a_j in the j^{th} position
$\mathbf{A} \otimes \mathbf{B}$	Kronecker product of matrices $\mathbf{A}$ and $\mathbf{B}$.
$vec\,(\mathbf{A})$, $vec\,\mathbf{A}$	the vectorized form of a $v \times v$ matrix $\mathbf{A}$,
	the resultant $v^2 \times 1$ vector is
	$(a_{11}, \ldots, a_{v1}; a_{12}, \ldots, a_{v2}, \ldots, a_{1v}, \ldots, a_{vv})'$
$vec\,(\mathbf{ABC})$	$(\mathbf{C}' \otimes \mathbf{A})\,vec\,\mathbf{B}$

Abbreviation	**Meaning**
AR	autoregressive
ARMA	autoregressive moving average
ARMAX	autoregressive moving average with exogenous variable

a.s.	almost surely (with probability 1)
CLT	central limit theorem
GLS	generalized least squares
LS	least squares
LSE	least squares estimate
MLE	maximum likelihood estimate
op. cit.	*opere citato* – in the earlier work of the author(s) cited
et. al.	*et alii* — and others

Chapter 1

Introduction

1.1 ARMA Representation

The use of finitely parameterized models to characterize the behaviours of time series data assumed to be generated by a stochastic process has received a wide coverage in both statistical and engineering literature. An important class of such stochastic models is the autoregressive moving average (ARMA) models. One reason for the great practical importance of ARMA models is that every regular stationary process can be approximated with arbitrary accuracy by an ARMA process and that only a finite number of (real-valued) parameters is needed to describe (up to second moments) an ARMA process. In the remainder of this section we define the ARMA system and briefly discuss its salient characteristics which in turn form the technical background for the problems addressed in this thesis. In particular, some terminology and assumptions will be introduced, and the conditions relating to the uniqueness and stability of such systems are discussed.

Let $\mathbf{y}(t)$ be an observed time-invariant process of v components, $y_i(t)$, $i = 1, \ldots, v$, satisfying a set of difference equations of the form

$$\sum_{j=0}^{p} \mathbf{A}(j)\mathbf{y}(t-j) = \sum_{j=0}^{q} \mathbf{M}(j)\varepsilon(t-j), \ t \in \mathcal{Z}. \tag{1.1.1}$$

Here $\mathcal{Z}$ denotes the set of integers and t is the time index. The sequence $\mathbf{y}(t)$ is often called the output or vector of endogenous variables and the unobserved stochastic inputs, $\varepsilon(t)$, are also v-dimensional. It is always assumed that the $\varepsilon(t)$ in (1.1.1) are egordic processes satisfying

$$\mathcal{E}\{\varepsilon(t)\} = 0, \quad \mathcal{E}\{\varepsilon(t)\varepsilon(s)'\} = \delta_{st}\mathbf{\Sigma} \tag{1.1.2}$$

where $\mathbf{\Sigma}$ is a positive definite symmetric matrix with δ_{st} denoting the Kronecker's delta (i. e. $\delta_{st} = 1$ for $s = t$ and zero otherwise). The assumption $\mathcal{E}\{\varepsilon(t)\} = 0$ means that in practice the data will need to be mean corrected but since that adjustment will not make any difference to the asymptotic results we obtain later, we ignore it for simplicity. The integers p, q (or other integers introduced to replace them)

will be called the order of the system. The coefficient matrices $\mathbf{A}(j)$, $\mathbf{M}(j) \in \Re^{v \times v}$ specifying (1.1.1) will be called the system parameters and the remaining $\frac{v}{2}(v+1)$ variance parameters specify $\mathbf{\Sigma}$.

Interpreting z as the unit lag operator, i.e. $z\mathbf{y}(t) = \mathbf{y}(t-1)$, (1.1.1) can be written more compactly as

$$\mathbf{A}(z)\mathbf{y}(t) = \mathbf{M}(z)\varepsilon(t) \tag{1.1.3}$$

where the generating functions (or z-transforms) $\mathbf{A}(z) = \sum_{j=0}^{p} \mathbf{A}(j)z^j$ and $\mathbf{M}(z) = \sum_{j=0}^{q} \mathbf{M}(j)z^j$ are assumed to satisfy

$$\det \mathbf{A}(z) \neq 0, \quad |z| \leq 1 \tag{1.1.4}$$

and

$$\det \mathbf{M}(z) \neq 0, \quad |z| < 1 \tag{1.1.5}$$

with det denoting the determinant of the indicated argument. A difference equation system (1.1.1) fulfilling $\det \mathbf{A}(z) \not\equiv 0$ where the inputs satisfy (1.1.2) with $\mathbf{A}(0) = \mathbf{M}(0)$ but not necessarily equal to the identity matrix is called an ARMA system. Adding a term $\sum_{j=0}^{r} \mathbf{D}(j)\mathbf{x}(t-j)$ on the right-hand side of (1.1.1) where the $\mathbf{x}(t)$ are observed (or controlled) inputs (exogenous variables) gives an ARMAX system, an acronym for autoregressive moving average with exogenous components. The addition of these observed inputs to (1.1.1) would cause no essential changes to our considerations; thus for the sake of simplicity we will primarily consider the ARMA case. For further discussion on ARMAX models see, for example, Hannan, Dunsmuir and Deistler (1980) and more recently, Hannan and Deistler (1988) provides a comprehensive review of the subject-matter.

Functions of the form $\mathbf{A}(z)$ and $\mathbf{M}(z)$ satisfying (1.1.4) and (1.1.5) are called stable and miniphase, respectively. When (1.1.4) holds, it means that we can rewrite (1.1.1) as

$$\mathbf{y}(t) = \sum_{j=0}^{\infty} \mathbf{K}(j)\varepsilon(t-j), \quad \mathbf{K}(0) = \mathbf{I}_v \tag{1.1.6}$$

where the $\mathbf{K}(j)$ are generated by the matrix transfer functions, whose elements are rational functions,

$$\mathbf{K}(z) = \sum_{j=0}^{\infty} \mathbf{K}(j) z^j = \mathbf{A}(z)^{-1} \mathbf{M}(z). \tag{1.1.7}$$

Because of (1.1.4) and (1.1.5) $\mathbf{K}(z)$ is composed of functions analytic for $|z| < 1$ and $\det \mathbf{K}(z) \neq 0, \ |z| < 1$. Thus, (1.1.7) form a convergent power series expansion in a suitable neighbourhood of zero and (1.1.6) constitutes a causal solution. The expression (1.1.7) is said to provide a matrix fraction description of $\mathbf{K}(z)$ see, for example, Hannan and Deistler (1988, p.37).

The condition (1.1.5) ensures that the $\varepsilon(t)$ are the linear innovations (Hannan (1970), chap. III), i.e. the errors from the best linear predictor, $\mathbf{y}(t|t-1)$, of $\mathbf{y}(t)$ from $\mathbf{y}(s)$, $s < t$ and thus (1.1.6) corresponds to Wold's representation (Rosanov (1967, Section II.2.3)). For the asymptotic results we obtain later it may be necessary to strengthen the assumptions (1.1.2). Denote by $\mathcal{F}_t$, the σ- algebra of events determined by $\varepsilon(s)$, $s \leq t$. Then the following conditions which henceforth will be referred to as Condition A are assumed to hold.

Condition A:

For all $1 \leq i, j, k, \ell \leq v$ and $-\infty < t < \infty$,

$\mathbf{A}_1$ $\quad \mathcal{E}\{\varepsilon_i(t)|\mathcal{F}_{t-1}\} = 0 \ \ a.s$

$\mathbf{A}_2$ $\quad \mathcal{E}\{\varepsilon(t)\varepsilon(t)'|\mathcal{F}_{t-1}\} = \mathbf{\Sigma} \ \ a.s$

$\mathbf{A}_3$ $\quad \mathcal{E}\{|\varepsilon_i(t)\varepsilon_j(t)\varepsilon_k(t)\varepsilon_\ell(t)|\} < \infty.$

The essential feature of Condition A, of course, is that the classical theory still goes through even though the innovation sequence, $\varepsilon(t)$, are no longer assumed to be independent. Hence, the asymptotic results that we obtain later will be of a more general nature than if independence assumptions were made. We now briefly examine the implications of Condition A in relation to the analysis of linear time series models. For these class of models, Condition A provides a natural and important weakening of the usual assumption that the innovation sequence, $\varepsilon(t)$, are serially

independent. The first, $\mathbf{A}_1$, is just the condition that $\varepsilon(t)$ be a sequence of martingale differences (which is clearly satisfied if the $\varepsilon(t)$ are independent and identically distributed). As explained in Hannan and Heyde (1972, p. 2059) $\mathbf{A}_1$ is equivalent to the assertion that the best predictor of $\mathbf{y}(t)$ conditioning on $\mathbf{y}(s)$, $s < t$, is equal to the best linear predictor, best being defined in terms of minimizing mean square error (see also Hannan and Deistler (1988, Theorem (1.4.2)).

Theorem (1.1.1) [Hannan and Deistler (1988)]:

The necessary and sufficient condition for the best linear predictor of the stationary process $\mathbf{y}(t)$*, with finite mean square, to be the best predictor is that the linear prediction errors* $\varepsilon(t)$ *should satisfy* $\mathbf{A}_1$.

The condition $\mathbf{A}_2$ says that the $\varepsilon(t)$ behave as if they were independent up to second moments. This is needed in order that simple formulae for the covariances of the estimators in their limiting distributions should result. We should however note that condition $\mathbf{A}_2$ is not most desirable except on the basis of Gaussanity since it is a requirement that would seem difficult to justify from the observed properties of the output. A feasible alternative is to assume only that $\mathcal{E}\{\varepsilon(t)\varepsilon(t)'|\mathcal{F}_{t-1}\} < \infty$ *a.s.* In some cases condition $\mathbf{A}_2$ can be replaced by the weaker condition

$$\lim_{k\to\infty} \mathcal{E}\{\varepsilon(t)\varepsilon(t)'|\mathcal{F}_{t-k}\} = \mathcal{E}\{\varepsilon(t)\varepsilon(t)'|\mathcal{F}_{-\infty}\} = \mathbf{\Sigma}\ a.s \tag{1.1.8}$$

where $\mathcal{F}_{-\infty} = \cap\mathcal{F}_t$. The only difficulty with assuming (1.1.8) and not condition $\mathbf{A}_2$ is that the asymptotic covariance matrix in the limiting distribution will not then be the classical one and can often become very complicated (see Hannan and Deistler (1988), p. 109). However, condition $\mathbf{A}_2$ or its close alternative will be sufficient for our investigation.

For the current study it may be helpful to relate the rational transfer function $\mathbf{K}(z)$ to the spectral theory of stationary processes in the following way. Let $\mathbf{\Gamma}_{yy}(r) = \mathcal{E}\{\mathbf{y}(t)\mathbf{y}(t+r)'\}$ denote the autocovariance of the process $\mathbf{y}(t)$. From the representation (1.1.6) and assumption (1.1.2) or Condition A we can express $\mathbf{\Gamma}_{yy}(r)$

as

$$\mathbf{\Gamma}_{yy}(r) = \sum_{j=0}^{\infty} \mathbf{K}(j)\mathbf{\Sigma}\mathbf{K}(j+r)', \quad \mathbf{\Gamma}_{yy}(r) = \mathbf{\Gamma}_{yy}(-r)', \quad r \geq 0.$$

Thus if $\mathbf{f}_y(\omega)$ is the function with Fourier coefficients $\mathbf{\Gamma}_{yy}(r)$ then $\mathbf{y}(t)$ has a spectral density of the form

$$\mathbf{f}_y(\omega) = (2\pi)^{-1} \sum_{r=-\infty}^{\infty} \mathbf{\Gamma}_{yy}(r)e^{-ir\omega} = (2\pi)^{-1}\mathbf{K}(e^{i\omega})\mathbf{\Sigma}\mathbf{K}(e^{i\omega})^*,$$

$$\mathbf{K}(e^{i\omega}) = \mathbf{A}(e^{i\omega})^{-1}\mathbf{M}(e^{i\omega}), \; z = e^{i\omega}, \; \omega \in [-\pi, \pi] \qquad (1.1.9)$$

where $\mathbf{f}_y(\omega)$ is a rational matrix in the sense that each of its entries is the quotient of trigonometric polynomials and an asterisk has been used here and elsewhere to indicate transposition combined with conjugation. From the relationship (1.1.9) the covariance matrix $\mathbf{\Gamma}_{yy}(r)$ can be recovered from $\mathbf{f}_y(\omega)$ through the inverse Fourier transform as

$$\mathbf{\Gamma}_{yy}(r) = (2\pi)^{-1} \int_{-\pi}^{\pi} e^{ir\omega}\mathbf{K}(e^{i\omega})\mathbf{\Sigma}\mathbf{K}(e^{i\omega})^*. \qquad (1.1.10)$$

However, the problem of finding $\mathbf{K}(e^{i\omega})$ and $\mathbf{\Sigma}$ such that (1.1.9) holds is often referred to as the spectral factorization problem. Thus, we present the following result.

Theorem (1.1.2) [Rosanov (1967)]:

Any rational and almost everywhere non-singular spectral density matrix $\mathbf{f}_y(\omega)$ *may be uniquely factored as in (1.1.9), where* $\mathbf{K}(z)$ *is rational in* z*, analytic within a circle containing the closed unit disc,* $\det \mathbf{K}(z) \neq 0$, $|z| < 1$ *and* $\mathbf{K}(0) = \mathbf{I}_v$ *and where* $\mathbf{\Sigma} > 0$.

Under the conditions of Theorem (1.1.2) it is clear that $\mathbf{K}(z)$ and $\mathbf{\Sigma}$ can be uniquely determined from a complete realization given that $\mathbf{y}(t)$ is egordic. However, we do not have a complete realization, hence the statistical problem of estimating these parameters. In particular, $\mathbf{A}(z)$ and $\mathbf{M}(z)$ are not uniquely determined from a given $\mathbf{K}(z)$ and this causes the problems of description for ARMA systems. Since (1.1.1) can be regarded as an approximation to (1.1.6), a key issue therefore is to understand the structure and parameterization of (1.1.1) for a given linear process

$\mathbf{y}(t)$. This will now be studied in the next section.

1.2 Canonical structure for ARMA models

As indicated in the last section, there is an inherent non-uniqueness in the parameterization $[\mathbf{A}(z) : \mathbf{M}(z)]$ for the process (1.1.1). An infinite set of matrix pairs $[\mathbf{A}(z) : \mathbf{M}(z)]$ will yield via (1.1.1) a process $\mathbf{y}(t)$ with the same transfer function (1.1.7) and thus for fixed $\mathbf{\Sigma}$ the same spectral function (1.1.9). To get a better sense of this problem, let us pre-multiply both sides of (1.1.1) by an arbitrarily non-singular matrix or more precisely by a matrix polynomial in z. This will yield a class of models that have identical covariance matrix structures. This is true even when the order say n of $\mathbf{K}(z)$ has been fixed see, e.g. Kalman, Falb and Arbib (1969, p. 286). Thus, without further restrictions, the class of models associated with (1.1.1) is not identifiable in the sense that we cannot uniquely determine the values of p and q, and the coefficient matrices $\mathbf{A}(i),\ i = 1,\ldots,p$ and $\mathbf{M}(j), j = 1,\ldots,q$ from the second order properties of the process $\mathbf{y}(t)$. The problem of identifiability for vector ARMA models has been extensively studied see, for example, Hannan (1969, 1979), Akaike (1976) and the references therein. This problem, of course, is directly linked to the problem of the description of a set of canonical forms for a set of transfer function matrices of the form (1.1.7), i.e. a bijective relation between a parameter space and a set of transfer functions (or impulse responses). Such canonical forms are sought in either matrix fraction description, that is, as a pair of polynomial matrices $[\mathbf{A}(z) : \mathbf{M}(z)]$ such that (1.1.7) holds or in state space form. For conciseness we restrict discussion to the construction via matrix fraction description but refer to Solo (1986), Aoki (1987), Hannan and Deistler (1988) and Tsay (1989b) for some illuminating discussions on the state space forms.

As has previously been noted, the decomposition in the right-hand side of (1.1.7) is known as the matrix fraction description of the transfer function $\mathbf{K}(z)$,

often abbreviated as *mfd.* Two *mfd*'s having the same transfer function $\mathbf{K}(z)$ are called observationally equivalent. It is well known (Deistler and Hannan (1981)) that observational equivalence gives a partitioning of the set of all (feasible) *mfd's* into equivalence classes. Again the *mfd* is not unique. In order to parameterize an ARMA system (1.1.1) we have to prescribe a suitable representative, a so-called canonical form, for each equivalence class and this leads to a study of canonical *mfd's.* Before continuing with this development we pause here to bring up some facts from the algebra of polynomial and rational matrices that are of direct relevance to the subsequent discussion. Some account of the matrix theory involved is given in MacDuffee (1956), Wolovich (1974) and Kailath (1980) with the latter two emphasizing applications to linear systems.

The first form of redundancy that can arise is if $\mathbf{A}(z)$ and $\mathbf{M}(z)$ have a common left divisor that disappears when $\mathbf{A}(z)^{-1}\mathbf{M}(z)$ is formed. That is, if $\mathbf{C}(z)$ is a matrix polynomial such that $\mathbf{A}(z) = \mathbf{C}(z)\mathbf{A}_1(z)$, $\mathbf{M}(z) = \mathbf{C}(z)\mathbf{M}_1(z)$ and $\mathbf{A}_1(z)$, $\mathbf{M}_1(z)$ are matrix polynomials, then $\mathbf{C}(z)$ is termed a common left divisor of $\mathbf{A}(z)$ and $\mathbf{M}(z)$. The matrix $\mathbf{C}(z)$ is often called the greatest common left divisor for $\mathbf{A}(z)$ and $\mathbf{M}(z)$ if it is a common left divisor and any other common left divisor (c.l.d) $\mathbf{D}(z)$ has $\mathbf{C}(z)$ as right multiple, i.e. $\mathbf{C}(z) = \mathbf{D}(z)\mathbf{D}_1(z)$ (see MacDuffee (1956), p. 35). Any other (c.l.d) of $\mathbf{A}(z)$ and $\mathbf{M}(z)$, say $\mathbf{C}_1(z)$, is of the form $\mathbf{C}_1(z) = \mathbf{C}(z)\mathbf{U}(z)$ where $\mathbf{U}(z)$ is unimodular. A polynomial matrix $\mathbf{U}(z)$ is called unimodular if $\det \mathbf{U}(z)$ is a non-zero constant. Thus, this unimodular matrix is precisely a matrix of polynomials for which $\mathbf{U}(z)^{-1}$ is also a matrix of polynomials.

Some degree of uniqueness in the *mfd* in (1.1.7) can be achieved by requiring the polynomial matrix $[\mathbf{A}(z) : \mathbf{M}(z)]$ to be left coprime.

Lemma (1.2.1) [Rosenbrock (1970)]:

The matrix pair $[\mathbf{A}(z) : \mathbf{M}(z)]$ *is said to be left coprime if and only if any one of the following conditions holds:*

(i) The matrix pair $[\mathbf{A}(z) : \mathbf{M}(z)]$ *has full row rank* v *for every* z,

(ii) $\det \mathbf{A}(z)$ *has minimal degree among all observationally equivalent* mfd's.

Thus we have the following important characterization of the equivalence classes.

Theorem (1.2.1) [Hannan (1969), Popov (1969), Rosenbrock (1970)]: *Two left coprime matrix fraction descriptions* $[\mathbf{A}(z) : \mathbf{M}(z)]$ *and* $[\bar{\mathbf{A}}(z) : \bar{\mathbf{M}}(z)]$ *are observationally equivalent if and only if there exists a unimodular matrix* $\mathbf{U}(z)$ *such that* $[\mathbf{A}(z) : \mathbf{M}(z)] = \mathbf{U}(z)\,[\bar{\mathbf{A}}(z) : \bar{\mathbf{M}}(z)]$.

It follows from Theorem (1.2.1) that the degree of $\det \mathbf{A}(z)$, say n, is invariant for all observationally equivalent left prime pairs $[\mathbf{A}(z) : \mathbf{M}(z)]$ and n is often called the order (Forney (1975)) or the McMillan degree of $\mathbf{K}(z)$ (or of the system). From the foregoing, however, it should be understood that in general, factoring unimodular operator from $\mathbf{A}(z)$ and $\mathbf{M}(z)$ is unavoidable if no further constraints are imposed. Thus, to obtain uniqueness of left coprime operators we have to impose restrictions ensuring that the only feasible unimodular operator is $\mathbf{U}(z) = \mathbf{I}_v$. The condition $\mathbf{U}(z) = \mathbf{I}_v$ combined with the assumptions (1.1.4) and (1.1.5) will be referred to as the *elementary conditions* for identifiability (Hannan (1975)). These conditions, however, appear to fit any natural interpretation of the class of models (1.1.1) in that, setting $\mathbf{U}(z) = \mathbf{I}_v$ in Theorem (1.2.1) eliminates redundancies while stationarity and miniphase assumptions (1.1.4) and (1.1.5) ensure that $\mathbf{y}(t)$ depends only on $\varepsilon(t)$, $s \leq t$. However, more is needed for a canonical choice for an ARMA process as we shall see in the subsequent developments.

In the situation considered in this thesis, there are no *a priori* restrictions on the system parameters and identifiability is achieved by selecting a particular canonical form from each observationally equivalent class. For instance, we may desire a member that often leads to a difference equation with the smallest order among all the difference equations of the form (1.1.1) corresponding to some equivalent class. This particular member can be called a canonical form since it possesses some special properties and all other members in the equivalence class can be generated

from this member. In this context, the term canonical form refers to a well specified ARMA model in which

1. $\mathbf{A}(z)$ and $\mathbf{M}(z)$ are left coprime,

2. the model contains no redundant parameters and

3. the orders of the polynomial involved are as small as possible so that the total number of parameters that require estimation is minimized.

Several methods of prescribing canonical forms for (1.1.1) such that properties 1 to 3 are satisfied are readily available see, for example, the Hermite's normal form (Rosenbrock (1970), Dickson, Kailath and Morf (1974), Kailath (1980)); the scalar component model representation (Tiao and Tsay (1989), Tsay (1991)); and the echelon canonical form which was first introduced in Popov (1969) (see also Forney (1975)). Herein, however, we shall restrict attention to an echelon canonical form for brevity. One useful way of prescribing an echelon canonical structure for matrix pair $[\mathbf{A}(z) : \mathbf{M}(z)]$ is to start from the Hankel matrices

$$\mathbf{H}_\infty = \begin{pmatrix} \mathbf{K}(1) & \mathbf{K}(2) & \mathbf{K}(3) & \dots \\ \mathbf{K}(2) & \mathbf{K}(3) & \mathbf{K}(4) & \dots \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \end{pmatrix}$$

of the transfer function $\mathbf{K}(z) = \sum_{j=0}^{\infty} \mathbf{K}(j)z^j$. Here $\mathbf{K}(j)$ will be called a *sub-block* of v rows and v columns. The importance of $\mathbf{H}_\infty$ can be seen from the, almost obvious, fact that the best r-step ahead predictor is $\mathbf{y}(t+r|t) = \sum_{j=0}^{\infty} \mathbf{K}(j)\varepsilon(t+r-j), r \geq 0$ so that $\mathbf{H}_\infty$ has, as the r^{th} row of blocks, the coefficient blocks in that prediction. The problem then is how to obtain unique integer parameters that govern the canonical *mfd* for $\mathbf{K}(z)$ given the Hankel matrices $\mathbf{H}_\infty$. To provide some discussion in this direction let $h(j,i)$ be the i^{th} row of the j^{th} block of rows of $\mathbf{H}_\infty$ so that $i = 1,\dots,v;\ j = 1,2,\dots$.

Now consider selecting a basis for the space spanned by the rows of $\mathbf{H}_\infty$ such that

$$If\ h(k,j),\ k>1\ is\ a\ basis\ vector\ then\ h(k-1,j)\ is\ also\ a\ basis\ vector. \quad (1.2.1)$$

A selection of basis set satisfying (1.2.1) can be described by a partition $\nu = \{n_1,\ldots,n_v\}$ of n (i.e. $\sum_{i=1}^{v} n_i = n,\ n_i > 0$) when the basis given by ν is of the form

$$h(j,i),\ j=1,\ldots,n_i;\ i=1,\ldots,v. \quad (1.2.2)$$

Usually there will be many such partitions and, as can be easily checked, at least $\begin{pmatrix} n+v-1 \\ v-1 \end{pmatrix}$ partitions are possible. A special selection is obtained if we seek for the first (in natural order) basis of rows of $\mathbf{H}_\infty$ for which (1.2.2) is the first linearly independent set of rows of $\mathbf{H}_\infty$ when these are ordered as

$$h(1,1),\ h(1,2),\ldots,h(1,v),\ h(2,1),\ h(2,2),\ldots. \quad (1.2.3)$$

Thus, (1.2.2) gives a nice selection in the sense that $h(n_i+1,i)$ is a linear combination of the basis of rows preceeding it in the ordering (1.2.3). The corresponding integer parameters $\nu = (n_1,\ldots,n_v)$ that uniquely determine these first linearly independent rows of $\mathbf{H}_\infty$ are called the Kronecker (structural) indices. For an alternative description of the Hankel matrix and some illustrations of how the Kronecker indices are chosen see, for example, Solo (1986) and Tsay (1989b). The other approaches that have been suggested for specifying a parsimonious vector ARMA model using canonical variate analysis can be found in Akaike (1974a, 1975, 1976) with Cooper and Wood (1982) providing a modification of this, and that of the scalar component model technique suggested by Tiao and Tsay (1989). For a discussion of the relationship between scalar component and Kronecker index approaches see, e.g. Tsay (1989a).

In most cases it is more convenient to give the parameterization directly in terms of ARMA system parameters and to make the structure theory simpler it is

much easier to consider the transfer function

$$\tilde{\mathbf{K}}(z) = \mathbf{K}(z^{-1}) - \mathbf{I}_v = \sum_{j=1}^{\infty} \mathbf{K}(j) z^{-j}$$

rather than $\mathbf{K}(z)$ where, by our convention, z^{-1} should be interpreted as a forward-shift operator. Thus, $\tilde{\mathbf{K}}(z)$ has the advantage of being strictly proper, that is, all denominator degrees are higher than the numerator degrees. This follows from the fact that $\lim_{z\to\infty} \tilde{\mathbf{K}}(z) = 0$ and the properness of $\tilde{\mathbf{K}}(z)$ is equivalent to causality of $\mathbf{K}(z)$. Let $\mathcal{M}(n)$ be the set of all strictly proper, rational transfer function $\tilde{\mathbf{K}}(z)$ of order n. If we let Θ_ν be the set of such $\tilde{\mathbf{K}}(z)$ whose dynamical (Kronecker) indices are $\nu = \{n_1, \ldots, n_v\}$ with $n = \sum_{i=1}^{v} n_i$ then it is known that $\{\mathbf{\Theta}_\nu | n_1 + n_2 + \ldots + n_v = n\}$ form a disjoint partition of $\mathcal{M}(n)$ and that each $\mathbf{\Theta}_\nu$ can be mapped homeomorphically into an open subset of Euclidean space $\Re^{d(\nu)}$ via the parameterization in terms of $[\tilde{\mathbf{A}}(z) : \tilde{\mathbf{M}}(z)]$ where

$$\tilde{\mathbf{A}}(z) = \sum_{j=0}^{p} \tilde{\mathbf{A}}(j) z^{-j}$$

and

$$\tilde{\mathbf{M}}(z) = \sum_{j=0}^{q} \tilde{\mathbf{M}}(j) z^{-j} \tag{1.2.4}$$

with $d(\nu)$ given by (1.2.9) below. If we let $\tilde{A}_{ij}(z)$, $\tilde{M}_{ij}(z)$ denote the $(i,j)^{th}$ element of $\tilde{\mathbf{A}}(z)$ and $\tilde{\mathbf{M}}(z)$, $i,j = 1,\ldots,v$, respectively then for each $\mathbf{K}(z^{-1}) - \mathbf{I}_v \in \mathbf{\Theta}_\nu$ there is a unique *mfd* $[\tilde{\mathbf{A}}(z) : \tilde{\mathbf{M}}(z)]$ of $\tilde{\mathbf{K}}(z)$ where $\tilde{\mathbf{A}}(z)$ and $\tilde{\mathbf{M}}(z)$ are left coprime matrices of polynomials with

$$\tilde{A}_{ij}(z) = \sum_{k=0}^{n_{ij}-1} \tilde{a}_{ij}(k) z^{-k}, \; i \neq j;$$

$$\tilde{A}_{ii}(z) = z^{-k} + \sum_{k=0}^{n_i-1} \tilde{a}_{ii}(k) z^{-k}, \; \tilde{a}_{ii}(n_i) = 1$$

and in $\tilde{\mathbf{M}}(z)$,

$$\tilde{M}_{ij}(z) = \sum_{k=0}^{n_i-1} \tilde{m}_{ij}(k) z^{-k} \tag{1.2.5}$$

such that for $\mathbf{K}(0) = \mathbf{I}_v$, $\mathbf{K}(z)$ is uniquely determined from $\tilde{\mathbf{K}}(z)$ and a matrix fraction description (*mfd*) $[\mathbf{A}(z) : \mathbf{M}(z)]$ of $\mathbf{K}(z)$ is obtained from an *mfd* $[\tilde{\mathbf{A}}(z) : \tilde{\mathbf{M}}(z)]$ of $\tilde{\mathbf{K}}(z)$ in a unique way via the transformation

$$[\mathbf{A}(z) : \mathbf{M}(z)] = \{diag(z^{n_1}, \ldots, z^{n_v})\}(\tilde{\mathbf{A}}(z),\ \tilde{\mathbf{A}}(z) + \tilde{\mathbf{M}}(z)) \tag{1.2.6}$$

where n_i is the maximum degree of the i^{th} row of $[\tilde{\mathbf{A}}(z) : \tilde{\mathbf{M}}(z)]$ and

$$n_{ij} = \begin{cases} \min(n_i + 1, n_j), & i > j \\ \min(n_i, n_j), & i \leq j. \end{cases}$$

It should be mentioned that $[\tilde{\mathbf{A}}(z) : \tilde{\mathbf{M}}(z)]$ has the same input-output behaviour as the ARMA system $[\mathbf{A}(z) : \mathbf{M}(z)]$. Clearly, (1.2.6) defines a one-to-one relation between $[\mathbf{A}(z) : \mathbf{M}(z)]$ and $[\tilde{\mathbf{A}}(z) : \tilde{\mathbf{M}}(z)]$, and as a result we associate the same $\mathbf{H}_\infty$ with $\tilde{\mathbf{K}}(z)$. The Kronecker indices $\nu = \{n_1, \ldots, n_v\}$ and $[\tilde{\mathbf{A}}(z) : \tilde{\mathbf{M}}(z)]$ as just described constitute a complete set of invariants for all $\tilde{\mathbf{K}}(z)$ equivalence class.

Lemma (1.2.2) [Hannan and Deistler (1988)]: *Assume the relationship (1.2.6) holds. If* $[\tilde{\mathbf{A}}(z) : \tilde{\mathbf{M}}(z)]$ *is left prime, then the corresponding* $[\mathbf{A}(z) : \mathbf{M}(z)]$ *is left prime if and only if* $\det \mathbf{A}(0) \neq 0$.

Given that $\mathbf{\Theta}_\nu$ corresponds to the set of elements in $\mathcal{M}(n)$ for which the rows of (1.2.2) are the first linearly independent rows met as one moves down the matrix $\mathbf{H}_\infty$ imposes additional restrictions on the elements $\tilde{a}_{ij}(k)$, viz:

$$\begin{aligned} \tilde{a}_{ij}(k) &= 0,\ k > n_i - 1,\ j > i; \\ \tilde{a}_{ij}(k) &= 0,\ k > n_i,\ j \leq i \end{aligned} \tag{1.2.7}$$

and thus the highest degree in the i^{th} row of $\tilde{\mathbf{A}}(z)$ is n_i. As a result, the maximum degree of the i^{th} row of $[\tilde{\mathbf{A}}(z) - \tilde{\mathbf{M}}(z)]$ is $n_i - 1$ because $\tilde{\mathbf{K}}(z)$ is strictly proper. Next we summarize the degree relationships in the *mfd* $[\tilde{\mathbf{A}}(z) : \tilde{\mathbf{M}}(z)]$ in the following theorem (see e.g. Deistler (1985)).

Theorem (1.2.2):

If $[\tilde{\mathbf{A}}(z) : \tilde{\mathbf{M}}(z)]$ *is the matrix fraction description of* $\tilde{\mathbf{K}}(z) \in \Theta_\nu$ *and* $[\tilde{\mathbf{A}}(z) : \tilde{\mathbf{M}}(z)]$ *are as defined in (1.2.4) satisfying (1.2.5) with additional restrictions (1.2.7), where* $\nu = \{n_1, \ldots, n_v\}$, *then*

1. $[\tilde{\mathbf{A}}(z) : \tilde{\mathbf{M}}(z)]$ *are left coprime, and*

2. $\tilde{\mathbf{A}}_{ii}(z)$ *are monic polynomials (i.e. the leading coefficient is equal to unity) with*

   (i) $\delta(\tilde{A}_{ii}(z)) = n_i$ *where* δ *is to be read as "degree of"*

   (ii) $\delta(\tilde{A}_{ij}(z)) \leq \delta(\tilde{A}_{ii}(z)) = n_i,\ j \leq i$

   (iii) $\delta(\tilde{A}_{ij}(z)) < n_i,\ j > i$

   (iv) $\delta(\tilde{A}_{ji}(z)) < n_i,\ j \neq i$

   (v) $\delta(\tilde{M}_{ij}(z) - \tilde{A}_{ij}(z)) < n_i,\ i, j = 1, \ldots, v.$

The occurrence of $\tilde{M}_{ij}(z) - \tilde{A}_{ij}(z)$ in (v) is due to the fact that it is $\tilde{\mathbf{K}}(z) = \tilde{\mathbf{A}}(z)^{-1}(\tilde{\mathbf{M}}(z) - \tilde{\mathbf{A}}(z))$ that is strictly proper as earlier indicated. Thus, every $[\tilde{\mathbf{A}}(z) : \tilde{\mathbf{M}}(z)]$ satisfying properties (1) and (2) of Theorem (1.2.2) is said to be in echelon (ARMA canonical) form. If $[\tilde{\mathbf{A}}(z) : \tilde{\mathbf{M}}(z)]$ is the echelon *mfd* of $\tilde{\mathbf{K}}(z)$ then $[\mathbf{A}(z) : \mathbf{M}(z)]$ as defined in (1.2.6) is said to be in (reversed) echelon form for $\tilde{\mathbf{K}}(z)$. Thus, the (reversed) echelon form possesses the following properties that uniquely define the matrix pair $[\mathbf{A}(z) : \mathbf{M}(z)]$:

$$
\begin{aligned}
A_{ii}(z) &= 1 + \sum_{k=1}^{n_i} a_{ii}(k) z^k,\ i = 1, \ldots, v, \\
A_{ij}(z) &= \sum_{k=n_i - n_{ij}+1}^{n_i} a_{ij}(k) z^k,\ i \neq j, \\
M_{ii}(z) &= 1 + \sum_{k=1}^{n_i} m_{ii}(k) z^k,\ i = 1, \ldots, v
\end{aligned}
$$

and

$$
M_{ij}(z) = \sum_{k=0}^{n_i} m_{ij}(k) z^k,\ i \neq j. \tag{1.2.8}
$$

Using property (1.2.8) it is obvious that $[\mathbf{A}(z) : \mathbf{M}(z)]$ has row degrees n_i, $i = 1,\ldots,v$, and that $\mathbf{A}(0) = \mathbf{M}(0)$ is lower triangular with units down the leading diagonal. It suffices now to note that the number of coefficients not restricted to be zero or one in $\mathbf{A}(z)$ is

$$d_\alpha = \sum_{i=1}^{v}\sum_{j=1}^{v} n_{ij} = n + \sum\sum_{i<j}^{v}\{min(n_i,n_j) + min(n_i,n_j+1)\}, \quad n = \sum_{i=1}^{v} n_i$$

and in $\mathbf{M}(z)$ there are an additional $\sum_{i=1}^{v}\sum_{j=1}^{v} n_i = nv$ giving a total of

$$d(\nu) = n(v+1) + \sum\sum_{i<j=1}^{v}\{min(n_i,n_j) + min(n_i,n_j+1)\} \tag{1.2.9}$$

freely varying parameters. Also, $\delta(\det \mathbf{A}(z)) + \delta(\det \mathbf{M}(z)) \leq 2n$. These features are fairly evident from (1.2.8).

The ARMA representation specified by (1.2.8) is complete in the sense that all the unknown parameters in $\mathbf{A}(z)$ and $\mathbf{M}(z)$ are identified and that each individual polynomial is specifically given. Moreover, it is apparent from the above discussion that the echelon structure guarantees uniqueness of the vector ARMA representation. In other words, if a vector ARMA representation is in echelon form the representation is unique within the class of all observationally equivalent *mfd's*. Also, for any stable, invertible ARMA representation (1.1.1) there exists an equivalent echelon form. Thus, if we let the matrix pair $[\mathbf{A}(z) : \mathbf{M}(z)]$ in (1.1.1) be in its appropriate (reversed) echelon form then it follows that $p = q = \max(n_1,\ldots,n_v)$ with the coefficient matrices $[\mathbf{A}(j), \mathbf{M}(j)]$, $j = 1,\ldots,p$ satisfying the restrictions imposed by (1.2.8). Setting $p = \max(n_1,\ldots,n_v)$ follows from the fact that n_i is the maximum order of $A_{ij}(z)$ and $M_{ij}(z)$.

In the rest of this thesis we shall be concerned with ARMA representation specified by (1.2.8) and some virtues for adopting this structure in the current study will now be discussed. Of special interest are the consequences of this representation for estimation. The modelling strategy offered by (1.2.8) permits considerable parsimony in parameterization, thereby aiding computational tractability and associated

precision in parameter estimates. In particular, the structure specified by (1.2.8) often involves fewer freely varying parameters in comparison with other forms of model (1.1.1). As will be shown in Chapter 2, having as few free parameters as possible is important to ease the numerical problems in maximizing the likelihood function and perhaps enhances a gain in the efficiency of the parameter estimates. Finally, to avoid excessive notation $\mathbf{y}(t)$ will henceforth be employed to denote both a given process and a realized value of that process. The particular meaning of the symbol will be clear from the usage.

1.3 Structure determination for echelon forms

Section (1.2) provided some discussion of how to determine the order, n, of the system or the set of Kronecker indices, ν, from the abstract mathematical framework. In most applications these integer-valued parameters are not known *a priori* and will have to be determined from the observed values of $\mathbf{y}(t)$. A choice is likely to be made from a range of values of ν that are regarded, *a priori*, as being appropriate, the choice in itself depends on the estimated characteristics of the model for a given ν.

In this section methods are discussed for estimating the Kronecker indices consistently from a given set of observations, $\mathbf{y}(t) = (y_1(t), \ldots, y_v(t))'$, $t = 1, \ldots, T$. In particular, attention will be restricted to the specification procedures suggested by Hannan and Kavalieris (1984) and Poskitt (1992), respectively. In our discussion of these procedures, we shall highlight only the fundamental ideas (see Hannan and Deistler (1988, chapters 5, 6, and 7) and Poskitt (1992) for extensive expositions). Before proceeding it may be worth emphasizing that the data is assumed not to exhibit any apparent deviation from stationarity and that T observations are available on the stationary process $\mathbf{y}(t)$. For notational convenience, it will also be assumed that this process has zero mean.

In theory, the Kronecker indices, $\nu = \{n_1, \ldots, n_v\}$, can be determined by a

maximum likelihood procedure but, as we shall see below, less costly techniques will be preferred in practice. The choice of ν that best describe the data at hand is made by minimizing the criterion function

$$Cr(\nu) = \log\det\hat{\mathbf{\Sigma}}(\nu) + d(\nu)\frac{C(T)}{T} \tag{1.3.1}$$

where $d(\nu)$, given by (1.2.9), is the number of freely varying parameters that require estimation. Here $\hat{\mathbf{\Sigma}}(\nu)$ is the maximum likelihood estimate of $\mathbf{\Sigma}$ from a model defined by the Kronecker indices, $\hat{\nu} = \{\hat{n}_1, \ldots, \hat{n}_v\}$, and $\hat{\nu}$ is chosen to minimize (1.3.1) over a suitable range of values of ν. The constant $C(T)$ has to be prescribed and various choices of this quantity have been considered. When $C(T) = 2$, (1.3.1) is known as Akaike Information Criterion (AIC) and for $C(T) = \log T$ (1.3.1) is called Bayesian Information Criterion (BIC). AIC was introduced in the univariate case by Akaike (1969, 1974b). BIC was introduced by Akaike (1976) and Schwarz (1978) from an argument based on Bayesian framework. It has also been justified on another basis in Rissanen (1976). Another choice of (1.3.1) which uses a borderline penalty term, $C(T) = 2\log\log T$, was introduced in Hannan and Quinn (1979). It is implicit in the result of Hannan (1981) that the consistency or inconsistency of the criterion function (1.3.1) depends on the choice of the sequence $C(T)$. Hannan and Deistler (1988, Chapt. 5, Section 5) show that a criterion such as (1.3.1) provides a consistent estimator of the true set of Kronecker indices if $C(T)$ is a non-decreasing function of T satisfying $C(T) \to \infty$ and $\frac{C(T)}{T} \to 0$ as $T \to \infty$ and the true data generating process satisfies some weak conditions (cf. Condition A). If, in addition, $\frac{C(T)}{2\log\log T} > 1$ as $T \to \infty$, the criterion provides a strongly consistent estimator of the true Kronecker indices.

A major difficulty for using the criterion function (1.3.1) is the requirement to employ the log-likelihood a number of times before arriving at the model(s) that best describe the data. This is costly and moreover is computationally burdensome because the likelihood function, as is shown in Chapter 2, will be non-linear in

the parameters and iterative optimization algorithms have to be employed. Since we just need an estimator of $\mathbf{\Sigma}(\nu)$ for the evaluation of model selection criterion such as (1.3.1) an obvious modification of the procedure is to use an estimator that avoids the non-linear optimization problem. Such an estimator, as we shall fully demonstrate in Chapter 3, may be obtained from a procedure based on linear least squares. To motivate discussion in this direction let us assume that $\hat{\varepsilon}_h(t)$, the estimates of the innovation sequence $\{\varepsilon(t)\}$, have been obtained by fitting a long autoregression,

$$\sum_{j=0}^{h} \hat{\mathbf{\Phi}}_h(j)\mathbf{y}(t-j) = \hat{\varepsilon}_h(t), \quad \hat{\mathbf{\Phi}}_h(0) = \mathbf{I}_v,$$

to data for $h \leq (\log T)^\tau$, $\tau < \infty$. Using $\hat{\varepsilon}_h(t)$ a full search procedure for the optimal Kronecker indices can be carried out by regressing each component of $\mathbf{y}(t)$ on the appropriate components of $\mathbf{y}(t) - \hat{\varepsilon}_h(t)$, $\mathbf{y}(t-j), \hat{\varepsilon}_h(t-j)$. Of course, the selection of the components involved in these regressions is determined by the Kronecker indices of the model being fitted.

Let $\tilde{\mathbf{\Sigma}}(\nu)$ be the residual mean square from the above regression. The Kronecker indices $\nu = \{n_1, \ldots, n_v\}$ are estimated as the argument minimizing the criterion function

$$Cr(\nu) = \log\det\tilde{\mathbf{\Sigma}}(\nu) + d(\nu)\frac{C(T)}{T}. \tag{1.3.2}$$

We could consider minimizing (1.3.2) over all sets of Kronecker indices $\nu = \{n_1, \ldots, n_v\}$ with $n_i \leq N$, where N is a prespecified upper bound for the Kronecker indices but, because of the large number of models to be considered even when v is quite small, some short-cut techniques have been proposed. The first one that is immediately presented below is inspired by Hannan and Kavalieris (1984). We have chosen to present this procedure in the most straightforward form although these authors discuss a number of computational simplifications that could be made. See Hannan and Deistler (1988, chapt. 6) for an elaborate discussion. For ease of reference, we shall refer to this method as Procedure A.

Procedure A:

Step 1: Evaluate $Cr(s,\ldots,s)$ for $s = 0,1,2,\ldots,N$ using criterion function (1.3.2) and call the value s that minimizes this criterion function as $n^{(1)}$ where, as earlier mentioned, the integer N is *a priori* specified.

Step 2: Fix all other indices at $n^{(1)}$ and vary the last index between 0 and $n^{(1)}$ to evaluate $Cr(n^{(1)},\ldots,n^{(1)},n_v)$ where $n_v = 0,1,2,\ldots,n^{(1)}$. Denote the value of n_v that minimizes (1.3.2) by $\hat{n}_v$.

Step 3: Fix n_v at $\hat{n}_v$ and consider

$$Cr(n^{(1)},\ldots,n^{(1)},n_{v-1},\hat{n}_v)$$

for $n_{v-1} = 0,1,\ldots,n^{(1)}$. Choose the value of n_{v-1} that minimizes the prescribed criterion function. Continue in this fashion until the overall minimum is found.

More generally, $\hat{n}_s,\ s \leq v$, is chosen such that

$$Cr(n^{(1)},\ldots,n^{(1)},\hat{n}_s,\ldots,\hat{n}_v) = min\{Cr[(n^{(1)},\ldots,n^{(1)},n_s,\hat{n}_{s+1},\ldots,\hat{n}_v)]\}$$

for $n_s = 0,\ldots,n^{(1)}$. The main feature of this procedure is that the number of echelon manifolds to be considered has been reduced from $(N+1)^v$ to at most $(N+1)+v(N+1)$ and hence brings about large computational savings. In addition, if $n^{(1)}$ is small the number of models to be considered may be substantially reduced. It should be mentioned, however, that this procedure may not always lead to the overall minimum but under the stated conditions for the model selection criterion the procedure does produce consistent estimates of ν.

An alternative procedure, which we refer to as Procedure B, is suggested in Poskitt (1992).

Procedure B:

The essential feature of this procedure is that evaluations are based on separate

least squares estimation of each of the v equations contained in (1.1.1) with $\hat{\varepsilon}_h(t)$ replacing $\varepsilon(t)$ and $p = \max(n_i)$, $i = 1, \ldots, v$. A model selection criterion of the form

$$Cr_i(\nu) = \log \hat{\sigma}_i^2(\nu) + \frac{C(T)}{T} d_i(\nu), \; i = 1, \ldots, v \tag{1.3.3}$$

is then evaluated for each of the v equations separately. Here $T\hat{\sigma}_i^2(\nu)$ is the residual sum of squares of the i^{th} equation and $d_i(\nu)$ denotes the number of freely varying (or functional) parameters in the i^{th} equation, and $C(T)$, as before, is a real-valued non-negative possibly stochastic function of T. It needs to be mentioned that (1.3.3) is the single equation analogue of the criterion function (1.3.2). The various steps in the procedure are now summarized as follows:

Note that for $\nu = \{0, \ldots, 0\}$, $\hat{\sigma}_i^2(0) = T^{-1} \sum_{t=1}^{T} y_i(t)^2$. Then, evaluate $Cr_i(\nu)$ and choose the estimates of the Kronecker indices, $\hat{n}_i$, $i = 1, \ldots, v$ according to the following rule:

If $Cr_i([0, \ldots, 0]) \geq Cr_i([1, \ldots, 1])$ for all $i = 1, \ldots, v$, compute $Cr_i([2, \ldots, 2])$, $i = 1, \ldots, v$, and compare with $Cr_i([1, \ldots, 1])$. If the $Cr_i([1, \ldots, 1])$ are all greater than $Cr_i([2, \ldots, 2])$ proceed to $Cr_i([3, \ldots, 3])$ and continue in this way until such a point when the inequality fails.

Suppose at some stage the inequality

$$Cr_i([s-1, \ldots, s-1]) \geq Cr_i([s, \ldots, s])$$

fails to hold for all i, choose $\hat{n}_i = s - 1$. The $\hat{n}_i$ so obtained is fixed in all the subsequent steps. The procedure is continued by increasing the remaining indices and comparing the criteria for those equations for which the Kronecker indices are not yet fixed.

In order to make the procedure a bit more transparent it may be helpful to consider an example. Assume we have a four-dimensional system, $\mathbf{y}(t) = (y_1(t), \ldots, y_4(t))'$. In this case $v = 4$ and consequently, $\nu = (n_1, \ldots, n_4)$. As a first step of the procedure $Cr_i([0,0,0,0])$ and $Cr_i([1,1,1,1])$ are evaluated for $i = 1, 2, 3, 4$. Suppose

$Cr_i([0,0,0,0]) \geq$

$Cr_i([1,1,1,1])$ for $i = 1,2,3,4$. Then evaluate $Cr_i([2,2,2,2])$, $i = 1,2,3,4$.

Now assume

$$Cr_1([1,1,1,1]) < Cr_1([2,2,2,2])$$

and that

$$Cr_i([1,1,1,1]) \geq Cr_i([2,2,2,2])$$

for $i = 2,3,4$. Then $\hat{n}_1 = 1$ is fixed and $Cr_i([1,2,2,2])$ is compared with $Cr_i([1,3,3,3])$ for $i = 2,3,4$. Again, suppose $Cr_i([1,2,2,2]) \geq Cr_i([1,3,3,3])$ for $i = 2,3$ but $Cr_4([1,2,2,2]) < Cr_4([1,3,3,3])$. Then fix $\hat{n}_4 = 2$.

We are now left to fix $\hat{n}_2$ and $\hat{n}_3$, respectively. Continuing in the same fashion we compare $Cr_i([1,2,2,2])$ with $Cr_i([1,3,3,2])$, $i = 2,3$ but note that the values of n_1 and n_4 are now fixed in $Cr_i([\cdot])$ at 1 and 2, respectively. If $Cr_i([1,2,2,2]) \geq Cr_i([1,3,3,2])$ for all $i = 2,3$ then evaluate $Cr_i([1,4,4,2])$. Now suppose $Cr_2([1,3,3,2]) \geq Cr_2([1,4,4,2])$ and $Cr_3([1,3,3,2]) < Cr_3([1,4,4,2])$. Then we fix $\hat{n}_3 = 3$ and compare $Cr_2([1,4,3,2])$ with $Cr_2([1,5,3,2])$. Continue in this way until $\hat{n}_2$ is fixed.

It is important to note, however, that for each Kronecker index only the first local minimum of the corresponding criterion is searched for. For moderate or large systems the present procedure has the advantage of involving only very reasonable amount of computations and more importantly, the Kronecker indices are consistently estimated. The conditions under which the consistency of the Kronecker indices is achieved are provided in Poskitt (op. cit).

From the discussion of this section, it is apparent that consistent and feasible strategies for estimating the Kronecker indices exist in practice. Unfortunately, the relative performance of Procedures A and B in small finite samples is unknown. Given this state of affairs it is difficult to give well-founded recommendations as to which strategy to adopt in any particular situation. However, it may be advisable to

try these alternative strategies in conjunction with different model selection criteria, and compare the resulting models and the implications for the subsequent analysis.

Chapter 2

Maximum Likelihood Estimation for ARMA Models

2.1 Introduction

This chapter is concerned with the maximum likelihood estimation procedures for analysing data, $\mathbf{y}(t)$, $t = 1,\ldots,T$, assumed to be generated by a member of the class of ARMA processes (1.1.1). The specific situation considered is one in which the structural (Kronecker) indices, $\nu = (n_1,\ldots,n_v)$, generating the system are known *a priori* or have been consistently estimated as indicated in Section (1.3) of the previous chapter and that the system parameters are estimated by maximizing a particular likelihood function.

The class of ARMA models generated by (1.1.1) was introduced in Chapter 1 as

$$\sum_{j=0}^{p} \mathbf{A}(j)\mathbf{y}(t-j) = \sum_{j=0}^{p} \mathbf{M}(j)\varepsilon(t-j) \tag{2.1.1}$$

where again the process $\mathbf{y}(t)$ is stationary, and the unobservable input $\{\varepsilon(t)\}$ is a strict stationary martingale difference sequence satisfying Condition A. In (2.1.1), $p = \max(n_1,\ldots,n_v)$ is an observability index, and that $\mathbf{A}(j)$ and $\mathbf{M}(j)$ are $v \times v$ matrices while $\mathbf{y}(t)$ and $\varepsilon(t)$ are v-dimensional vectors. To avoid redundancy and non-uniqueness in (2.1.1) we require the matrix pair $[\mathbf{A}(z) : \mathbf{M}(z)]$ to be left coprime and in (reversed) echelon form where $[\mathbf{A}(z) : \mathbf{M}(z)]$ are as defined in (1.1.3). We shall assume that all calculations are made with mean corrected quantities so that $\mathbf{y}(t)$ is the residual from such adjustment. The asymptotic properties of the parameter estimates so obtained (as $T \to \infty$) will not be affected by this adjustment.

The problem of parameter estimation for models such as (1.1.1) has a long history in the time series literature, having been considered by Hannan (1970), Wilson (1973), Nicholls (1976), Kohn (1978), Jenkins and Alavi (1981), and in a more general context, Hannan and Kavalieris (1984), and Hannan and Deistler (1988), amongst others. The estimation procedures proposed are generally based on optimization of a likelihood function, constructed as if the data were from a Gaussian process. With the model structure (Kronecker indices) assumed known

our interest is in describing a procedure for evaluating the efficient estimates of $\mathbf{A}(1),\ldots,\mathbf{A}(p):\mathbf{M}(1),\ldots,\mathbf{M}(p)$ and $\boldsymbol{\Sigma}$ in relation to (2.1.1).

The rest of the chapter is structured as follows. The next section sets out notation and provides discussion on various simplifying approximations to the likelihood function. Section (2.3) outlines the efficient parameter estimation via a modified Newton-Raphson procedure and discusses some practical aspects of the employed technique. This is followed in Section (2.4) by the presentation of the asymptotic properties of the estimators.

2.2 The Likelihood function

For maximum likelihood (ML) estimation the likelihood function is needed. We will now give some simplifying approximations to the likelihood function and for ease of exposition it will be necessary to first introduce notation. Employing the convention adopted in Poskitt (1992), we set $\zeta_p(z)' = (z,\ldots,z^p)$, $\bar{\alpha} = vec\{\mathbf{A}(1),\ldots,\mathbf{A}(p)\}$, $\bar{\beta} = vec\{\mathbf{A}(0)-\mathbf{I}_v\}$ and $\bar{\lambda} = vec\{\mathbf{M}(1),\ldots,\mathbf{M}(p)\}$. Thus, the elements of $\bar{\alpha}$, for example, are obtained by stacking the columns of $\mathbf{A}(1)$ followed by those of $\mathbf{A}(2)$, and so on. The vectors $\bar{\beta}$ and $\bar{\lambda}$ are constructed in a similar way. The vector $\bar{\beta}$ occurs as a result of the matrix pair $[\mathbf{A}(z):\mathbf{M}(z)]$ being in (reversed) echelon form in which case $\mathbf{A}(0)-\mathbf{I}_v = \mathbf{M}(0)-\mathbf{I}_v$ could have non-zero elements below the main diagonal. Let $\mathbf{S}_{\alpha(\nu)}$ be a $d_\alpha \times pv^2$ selection matrix where d_α denotes the number of freely varying parameters in $\bar{\alpha}$. Typically, each row of $\mathbf{S}_{\alpha(\nu)}$ has a single element equal to unity to indicate the element of the vector $\bar{\alpha}$ that is being selected and the remaining elements are null. Denoting α as a vector of parameters that are not restricted to be zero in $\bar{\alpha}$, it follows that $\alpha = \mathbf{S}_{\alpha(\nu)}\bar{\alpha}$. Similarly, $\beta = \mathbf{S}_{\beta(\nu)}\bar{\beta}$ where $\mathbf{S}_{\beta(\nu)}$ selects the non-zero elements of $\mathbf{A}(0)$ below the leading diagonal in $\bar{\beta}$ and $\lambda = \mathbf{S}_{\lambda(\nu)}\bar{\lambda}$, $\mathbf{S}_{\lambda(\nu)}$ being defined in a manner analogous to $\mathbf{S}_{\alpha(\nu)}$ and $\mathbf{S}_{\beta(\nu)}$. A straightforward matrix manipulation shows that $\mathbf{S}_{\alpha(\nu)}\mathbf{S}'_{\alpha(\nu)} = \mathbf{I}_{d_\alpha}$ and that, $\mathbf{S}_{\alpha(\nu)}$ and $\mathbf{S}'_{\alpha(\nu)}$ are reflexive

generalized inverses and hence $\bar{\alpha} = \mathbf{S}'_{\alpha(\nu)}\alpha$. Also, $\mathbf{S}_{\beta(\nu)}$ and $\mathbf{S}_{\lambda(\nu)}$ in joint operation with their respective transpositions carry similar interpretations. To fix ideas, the elements of the parameter matrices in (2.1.1) are organized in one vector of length $d(\nu) = d_\alpha + d_\beta + d_\lambda$ given by $\theta = (\alpha' : \beta' : \lambda')'$ where, by construction, α, β and λ contain the elements of $[\mathbf{A}(z) : \mathbf{M}(z)]$ that are not restricted to be either zero or one with d_β and d_λ denoting the length of vectors β and λ, respectively. Appendix (A.1, p.41) provides an illustration of the basic concepts and ideas presented in this paragraph.

Our main objective in this section, however, is to consider efficient methods of estimating the unknown parameter values θ and $\mathbf{\Sigma}$. The estimation methods considered are based on the Gaussian likelihood though Gaussianity assumption is not necessary for the validity of the methods or for the asymptotic theory as long as all second moments are assumed to exist. Let $\mathbf{y}_T = (\mathbf{y}(1)', \ldots, \mathbf{y}(T)')'$ be a realization of T observations from a system generating (2.1.1) and define $\mathbf{\Theta} = (\theta : vech\,\mathbf{\Sigma})$ as the parameter space associated with the class of models (2.1.1) under (1.1.4), (1.1.5) and over all positive-definite matrices $\mathbf{\Sigma}$. The estimation method is based on maximizing the likelihood of $\mathbf{y}_T$, which under Gaussian assumptions omitting constant terms and for $\theta \in \mathbf{\Theta}$, is given by

$$\begin{aligned} \mathcal{L}_T(\theta) &= -2T^{-1}(\log \textit{ Gaussian likelihood}) \\ &= T^{-1}\log\det \mathbf{G}_T(\theta) + T^{-1}\mathbf{y}'_T\mathbf{G}_T(\theta)^{-1}\mathbf{y}_T \end{aligned} \tag{2.2.1}$$

where the covariance matrix $\mathbf{G}_T(\theta) = \mathcal{E}(\mathbf{y}_T\mathbf{y}'_T)$ has as its $(m,n)^{th}$ block of $v \times v$ elements the matrix

$$\mathbf{\Gamma}_{yy}(m-n) = \int_{-\pi}^{\pi} \mathbf{e}^{i(m-n)}\mathbf{f}_y(\omega)d\omega$$

with the spectral density matrix, $\mathbf{f}_y(\omega)$, given by (1.1.9). Thus the maximum likelihood estimate (MLE) of θ is obtained by minimizing (2.2.1). If $\mathbf{y}(t)$ is not Gaussian, as in the present case, it makes sense to regard (2.2.1) or some approximation to it as a measure of the goodness of fit of the model to the data. Under this circumstance

(2.2.1), as in econometric literature, is often referred to as *quasi-maximum likelihood function* see, for example, Phillips (1976, p. 450).

The main difficulty in using (2.2.1) to obtain MLE of θ arises, of course, from the need to compute $\det \mathbf{G}_T(\theta)$ and $\mathbf{G}_T(\theta)^{-1}$ directly. To obviate this problem a simpler form of the likelihood function can be obtained by expressing $\mathbf{G}_T(\theta)$ in terms of one-step predictors $\mathbf{y}_\theta(t|t-1)$, which are easily calculated recursively. A natural estimation procedure for (2.1.1) therefore is to maximize the objective function

$$\mathcal{L}_T(\theta, \mathbf{\Sigma}) = \frac{1}{2} \log \det \mathbf{\Sigma} + \frac{1}{2} tr \mathbf{\Sigma}^{-1} \mathbf{S}_T(\theta) \tag{2.2.2}$$

which, as indicated, leads to a simpler optimization problem where

$$\mathbf{S}_T(\theta) = T^{-1} \sum_{t=1}^{T} \mathbf{e}_\theta(t) \mathbf{e}_\theta(t)'$$

and

$$\begin{aligned} \mathbf{e}_\theta(t) &= \sum_{j=0}^{p} \mathbf{A}(j)\mathbf{y}(t-j) - \sum_{j=0}^{p} (\mathbf{M}(j) - \delta_{0j}\mathbf{I}_v)\mathbf{e}_\theta(t-j) \\ &= \mathbf{M}(z)^{-1}\mathbf{A}(z)\mathbf{y}(t), \ t = 1, \ldots, T, \end{aligned}$$

represents an estimable function of the unobserved inputs, $\{\varepsilon(t)\}$. The appendage θ is used to denote the dependence of the residuals $\mathbf{e}_\theta(t)$, $t = 1, \ldots, T$, on the value of the parameter vector θ. To avoid technical difficulties we shall, henceforth, strengthen the miniphase assumption (1.1.5) to

$$\det \mathbf{M}(z) \neq 0, \ |z| \leq 1. \tag{2.2.3}$$

In practice, only $\mathbf{y}_T$ is available to estimate θ so that the pre-period values, $\{\mathbf{y}(t), \mathbf{e}_\theta(t)\}, t \leq 0$, are not available and thus represent unknown quantities. Some suggestions have been made for specifying these initial values. One possibility is to use some $\mathbf{y}(t)$ values at the beginning of the sample as the initial observations $(\mathbf{y}(-p+1), \ldots, \mathbf{y}(0))$ while $(\mathbf{e}(-p+1), \ldots, \mathbf{e}(0))$ are set to their unconditional expectation which is zero as in Wilson (1973) or Reinsel (1976). The method of back

forecasting (Hillmer and Tiao (1979)) could also be used to estimate these pre-period values. However, a much simpler approach is to modify the likelihood function (2.2.2) by treating presample values as zero, that is, $\{\mathbf{y}(t), \mathbf{e}_\theta(t)\} = 0,\ t \leq 0$. For efficiency the use of back-forecasting for parameter estimation is important for models that are close to being non-stationary and especially for series that are relatively short. The assumption of zero inputs for observation before $t = 1$ is referred to as prewindowing in the engineering literature see, for example, Ljung and Soderstrom (1983). Asymptotically this approximation makes no difference since its effect is usually negligible in most samples when T is sufficiently large. However, this approximate maximum likelihood method has been shown (at least in the scalar case) to be less adequate for small sample sizes, particularly when the moving average operator of the model is close to being non-miniphase. See Pagan and Nicholls (1976), and Box and Jenkins (1976, p. 219) for some discussion. In such situations it is preferrable to consider the maximization of the exact likelihood function. Procedures for deriving the exact likelihood for vector ARMA models have been studied by Osborn (1977) for the pure moving average case, and by Phadke and Kedem (1978), Nicholls and Hall (1979), Hillmer and Tiao (1979), and in a more general case, Hannan, Dunsmuir and Deistler (1980). More recently, interest has centered on the use of the Kalman filter as a computationally efficient means of deriving the likelihood function for this and other time series models. See, for example, Solo (1984), Shea (1984), Gardner, Harvey and Phillips (1980), and Ansley and Kohn (1983), who show how to incorporate the possibility of missing and aggregated data. However, the exact closed form of the likelihood function is complicated and in general very slow to compute see, Tiao and Box (1981) for some discussion.

In many practical situations, the choice $\mathcal{L}_T(\theta, \boldsymbol{\Sigma})$ is only one choice of an objective function to be maximised and in this view, the exact maximum likelihood procedure may not always be required. Thus, since the minimizing value of $\boldsymbol{\Sigma}$ for given θ is $\boldsymbol{\Sigma}_T(\theta) = T^{-1}\sum_{t=1}^{T} \mathbf{e}_\theta(t)\mathbf{e}_\theta(t)'$ then following Kohn (1978) the optimization

of (2.2.2) can be effected by minimizing the concentrated likelihood function

$$\mathcal{L}_T(\theta) = \frac{1}{2} \log \det \mathbf{\Sigma}_T(\theta) \tag{2.2.4}$$

where, as noted earlier, $\mathbf{e}_\theta(t), t = 1, \ldots, T$ are the observed residuals associated with the model and is recursively generated via the relation

$$\mathbf{e}_\theta(t) = \sum_{j=0}^{t-1} \mathbf{\Phi}(j)\mathbf{y}(t-j) \tag{2.2.5}$$

where $\sum_{j=0}^{\infty} \mathbf{\Phi}(j)z^j = \mathbf{\Phi}(z) = \mathbf{M}(z)^{-1}\mathbf{A}(z)$ and that $(\mathbf{y}(t), \mathbf{e}_\theta(t)) = 0, \quad t \leq 0$. The effect of starting up the process with $\mathbf{y}(t) = \mathbf{e}_\theta(t) = 0$ for $t \leq 0$ is quite easily seen in (2.2.5), namely, the infinite order pure autoregressive representation is truncated at lag $t-1$ for $\mathbf{y}(t)$. Such a truncation has little effect if the sample size is large and the zeros of $\det \mathbf{M}(z)$ are not close to the unit circle. This suggests that the likelihood approximation in (2.2.4) will improve as the sample size gets large and will become exact as $T \to \infty$. If $\hat{\theta}_G$ minimizes $\mathcal{L}_T(\theta)$ then $\mathbf{\Sigma}_T(\hat{\theta}_G)$ is taken as the estimate of $\mathbf{\Sigma}$. Thus, $\hat{\theta}_G$ and $\mathbf{\Sigma}_T(\hat{\theta}_G)$ are referred to as the approximate Gaussian maximum likelihood estimators of θ_0 and $\mathbf{\Sigma}_0$, respectively, where the zero subscript is used here, as it will be below, to indicate the true parameter values. These estimators differ from those that would be obtained from the exact Gaussian (maximum) likelihood function in their treatment of initial conditions. Since we will be concerned predominantly with large sample properties in the following we will indulge in the luxury of working with the criterion function (2.2.4).

It is perhaps worth noting that if the model is uniquely parameterized, the likelihood function has a locally unique maximum. This property is of obvious importance to guarantee the existence and uniqueness of the Gaussian estimators. In the scalar case ($v = 1$), for example, Astrom and Soderstrom (1974) show that $\lim_{T\to\infty} \mathcal{L}_T(\theta)$ has a unique local minimum at θ_0. In general, however, the likelihood function may have more than one local minimum or may not even have a minimum in the region of minimization. A more detailed discussion of the properties of the

likelihood function can be found in Deistler and Potscher (1984). However, $\mathcal{L}_T(\hat{\theta}_G)$ always has a well defined meaning as $\inf \mathcal{L}_T(\theta)$, $\theta \in \Theta$ where inf denotes the infimum of the indicated argument. For the minimization of (2.2.4), it can be seen from (2.2.5) that $\Sigma_T(\theta)$ will in general be non-linear in θ so that $\hat{\theta}_G$ needs to be solved by iterative techniques. In the next section we discuss the evaluation of $\hat{\theta}_G$ via (2.2.4).

2.3 Maximization of the likelihood function

In this section we describe the explicit calculation of asymptotically efficient estimates, $\hat{\theta}_G$, of the system parameters θ, given that the true model structure (Kronecker indices) are known. There are a number of different approaches to numerical optimization but the main method we consider is one based on the modified Newton-Raphson (Gauss-Newton) procedure. Our emphasis on this optimization scheme is due to the fact that it gives some theoretical insight into estimation in a non-linear framework and moreover, is the heart of many general purpose computer packages for fitting ARMA processes.

2.3.1 The Gauss-Newton Scheme

Here we show how to obtain efficient estimates of the parameter θ via Gauss-Newton procedure. In so doing, we give an explicit evaluation of the gradient vector and (approximate) Hessian matrix (i.e. matrix of second derivatives) of the log likelihood in relatively simple terms. To introduce the procedure consider the criterion function (2.2.4),

$$\mathcal{L}_T(\theta) = \frac{1}{2} \log \det \Sigma_T(\theta)$$

over some permissible parameter space Θ, to obtain $\hat{\theta}_G$, an estimate of an unknown parameter value θ, and consequently, $\Sigma(\hat{\theta}_G)$.

To obtain the Gaussian estimator, $\hat{\theta}_G$, the standard technique is to solve the

first-order conditions

$$\frac{\partial \mathcal{L}_T(\theta)}{\partial \theta} = T^{-1} \sum_{t=1}^{T} \frac{\partial \mathbf{e}_\theta(t)'}{\partial \theta} \mathbf{\Sigma}_T(\theta)^{-1} \mathbf{e}_\theta(t) = 0$$

but in general these equations are highly non-linear in θ and are extremely difficult, if not impossible, to evaluate analytically so recourse must be made to either numerical optimization methods or iterative techniques. A common iterative algorithm to determine $\hat{\theta}_G$ is the Newton-Raphson procedure (e.g. Kowalik and Osborne (1968, p. 65), Judge et al (1985, Section B), Kashyap and Rao (1976)) which is defined by

$$\hat{\theta}_G^{(j+1)} = \hat{\theta}_G^{(j)} - \left[\left(\frac{\partial^2 \mathcal{L}_T(\theta)}{\partial \theta \partial \theta'} \right)^{-1} \frac{\partial \mathcal{L}_T(\theta)}{\partial \theta} \right]_{\theta = \hat{\theta}_G^{(j)}}, \; j \geq 0$$

with $\hat{\theta}_G^{(0)}$ being some consistent estimator given prior to the commencement of the iterations and $\hat{\theta}_G^{(j)}$ denotes the j^{th} step estimate of θ. Since $\frac{\partial^2 \mathcal{L}_T(\theta)}{\partial \theta \partial \theta'}$ might be too complicated to be computed, a modified version of the Newton's procedure consists of replacing this quantity by the matrix

$$\mathbf{\Delta}_T(\theta_G) = T^{-1} \sum_{t=1}^{T} \frac{\partial \mathbf{e}_\theta(t)'}{\partial \theta} \mathbf{\Sigma}_T(\theta)^{-1} \frac{\partial \mathbf{e}_\theta(t)}{\partial \theta'},$$

where the expected value of $\mathbf{\Delta}_T(\theta_G)$ is the Fisher's information matrix, which can be obtained as the limit in probability as $T \to \infty$ of $-T^{-1} \frac{\partial^2 \mathcal{L}_T}{\partial \theta \partial \theta'}$, the limit being computed under the assumption that the observations $\mathbf{y}_T$ obey the model with parameters θ. This modification gives rise to the Gauss-Newton iterative scheme

$$\hat{\theta}_G^{(j+1)} = \hat{\theta}_G^{(j)} - \mathbf{\Delta}_T(\theta)^{-1} \sum_{t=1}^{T} \frac{\partial \mathbf{e}_\theta(t)'}{\partial \theta} \mathbf{\Sigma}_T(\theta)^{-1} \mathbf{e}_\theta(t) \mid_{\theta = \hat{\theta}_G^{(j)}}, \; j \geq 0. \tag{2.3.1}$$

This has the advantage that it only contains first derivatives of $\mathbf{e}_\theta(t)$, and which can be readily evaluated. If $T^{\frac{1}{2}}(\hat{\theta}_G^{(0)} - \theta_0) = O_p(1)$ then $\hat{\theta}_G^{(j)}$ for all $j \geq 1$ and in particular $\hat{\theta}_G^{(1)}$ will be consistent and asymptotically equivalent, at least in the scalar case, to the maximum likelihood estimator. (See Fuller (1976, Chapter 5) for detailed exposition and proof). Thus, it follows that the iterates $\hat{\theta}_G^{(j)}$ will converge to θ_0 with increasing j as $T \to \infty$. To simplify the computation, we may therefore

prefer to choose as the estimate of θ_0, $\hat{\theta}_G^{(j)}$, for some small positive j, rather than iterating (2.3.1) to convergence. Whether this is done or not will depend on the sample size and the class of models that is being considered. This appears to have an intuitively appealing interpretation when we observe that the parameter adjustment $\hat{\theta}_G^{(j+1)} - \hat{\theta}_G^{(j)}$ from (2.3.1) can be interpreted as the coefficients in a regression of $\mathbf{e}_\theta(t)$ on $-\frac{\partial \mathbf{e}_\theta(t)}{\partial \theta'} = -(\frac{\partial \mathbf{e}_\theta(t)}{\partial \alpha'} : \frac{\partial \mathbf{e}_\theta(t)}{\partial \beta'} : \frac{\partial \mathbf{e}_\theta(t)}{\partial \lambda'})$ with weighting matrix $\mathbf{\Sigma}_T(\theta)$. This last observation suggests that we only need to have an explicit expression for evaluating the regressor variables $\frac{\partial \mathbf{e}_\theta(t)}{\partial \theta'}$ in order to complete calculations yielding the fully efficient estimates. Now proceeding in a manner analogous to Poskitt and Tremayne (1982, p. 116), but noting that $\mathbf{A}(0) = \mathbf{M}(0)$ does not necessarily equal an identity matrix, we find, using (2.2.5), that

$$\begin{aligned} \frac{\partial \mathbf{e}_\theta(t)}{\partial \alpha'} &= \frac{\partial}{\partial \alpha'}\{\mathbf{M}(z)^{-1}\mathbf{A}(z)\mathbf{y}(t)\} \\ &= \frac{\partial}{\partial \alpha'}\{vec(\mathbf{M}(z)^{-1}\mathbf{A}(z)\mathbf{y}(t))\} \end{aligned} \tag{2.3.2}$$

since $\mathbf{e}_\theta(t) = vec\{\mathbf{e}_\theta(t)\}$. Now employing the rule $vec(\mathbf{ABC}) = (\mathbf{C}' \otimes \mathbf{A})vec\,\mathbf{B}$, we observe that (2.3.2) can be rewritten as

$$\begin{aligned} \frac{\partial \mathbf{e}_\theta(t)}{\partial \alpha'} &= \frac{\partial}{\partial \alpha'} vec(\mathbf{M}(z)^{-1}\mathbf{A}(z)\mathbf{y}(t)) \\ &= (\mathbf{y}(t)' \otimes \mathbf{M}(z)^{-1})\frac{\partial}{\partial \alpha'} vec\mathbf{A}(z) \\ &= (\mathbf{y}(t)' \otimes \mathbf{M}(z)^{-1})(\zeta_p(z)' \otimes \mathbf{I}_{v^2})\mathbf{S}'_{\alpha(\nu)} \\ &= \{\zeta_p(z)' \otimes \mathbf{y}(t)' \otimes \mathbf{M}(z)^{-1}\}\mathbf{S}'_{\alpha(\nu)} \end{aligned} \tag{2.3.3}$$

where

$$\begin{aligned} \frac{\partial}{\partial \alpha'}(vec\mathbf{A}(z)) &= \frac{\partial}{\partial \alpha'}\{\sum_{j=1}^{p} vec\mathbf{A}(j)z^j + vec\ \mathbf{A}(0)\} \\ &= \frac{\partial}{\partial \alpha'}[\{(z, \ldots, z^p) \otimes \mathbf{I}_{v^2}\}\{vec(\mathbf{A}(1) : \ldots : \mathbf{A}(p))\} + vec\mathbf{A}(0)] \\ &= \frac{\partial}{\partial \alpha'}[\{(z, \ldots, z^p) \otimes \mathbf{I}_{v^2}\}\mathbf{S}'_{\alpha(\nu)}\alpha + vec\mathbf{A}(0)] \\ &= (\zeta_p(z)' \otimes \mathbf{I}_{v^2})\mathbf{S}'_{\alpha(\nu)}. \end{aligned}$$

Completely analogous manipulations applied to $\frac{\partial \mathbf{e}_\theta(t)}{\partial \lambda'}$ lead to the expression

$$\frac{\partial \mathbf{e}_\theta(t)}{\partial \lambda'} = \{-\zeta_p(z)' \otimes \mathbf{e}_\theta(t)' \otimes \mathbf{M}^{-1}(z)\}\mathbf{S}'_{\lambda(\nu)}. \tag{2.3.4}$$

It now remains to obtain a similar expression for $\frac{\partial \mathbf{e}_\theta(t)}{\partial \beta'}$. By applying the result

$$\frac{\partial \mathbf{D}^{-1}}{\partial \mathbf{a}} = -\mathbf{D}^{-1}\frac{\partial \mathbf{D}}{\partial \mathbf{a}}\mathbf{D}^{-1} \tag{2.3.5}$$

for some non-singular matrix $\mathbf{D}$ and a column vector $\mathbf{a}$ in conjunction with the chain rule for matrix differentiation (see, for example, Lutkephol (1991, p. 469)) we get

$$\begin{aligned}
\frac{\partial \mathbf{e}_\theta(t)}{\partial \beta'} &= [vec(\mathbf{M}^{-1}(z)\frac{\partial \mathbf{M}(z)}{\partial \beta'}\mathbf{M}^{-1}(z)\mathbf{A}(z)\mathbf{y}(t)) + vec\{\mathbf{M}^{-1}(z)\frac{\partial \mathbf{A}(z)}{\partial \beta'}\mathbf{y}(t)\}] \\
&= -\{\mathbf{e}_\theta(t)' \otimes \mathbf{M}^{-1}(z)\}\frac{\partial}{\partial \beta'}vec\,\mathbf{M}(z) + \{\mathbf{y}(t)' \otimes \mathbf{M}^{-1}(z)\}\frac{\partial}{\partial \beta'}vec\,\mathbf{M}(z) \\
&= \{(\mathbf{y}(t) - \mathbf{e}_\theta(t))' \otimes \mathbf{M}^{-1}(z)\}(\frac{\partial}{\partial \beta'}vec\,\mathbf{M}(z)) \\
&= \{(\mathbf{y}(t) - \mathbf{e}_\theta(t))' \otimes \mathbf{M}^{-1}(z)\}\{(\mathbf{I}_v \otimes \mathbf{I}_v)\mathbf{S}'_{\beta(\nu)}\} \\
&= \{(\mathbf{y}(t) - \mathbf{e}_\theta(t))' \otimes \mathbf{M}^{-1}(z)\}\mathbf{S}'_{\beta(\nu)}.
\end{aligned} \tag{2.3.6}$$

Thus, putting

$$\eta(t) = (\mathbf{y}(t)' \otimes \mathbf{M}(z)^{-1}), \quad \xi(t) = (\mathbf{e}_\theta(t)' \otimes \mathbf{M}(z)^{-1}), \tag{2.3.7}$$

we may rewrite $\frac{\partial \mathbf{e}_\theta(t)}{\partial \theta'}$ using (2.3.4) – (2.3.6) via (2.3.7) as

$$\frac{\partial \mathbf{e}_\theta(t)}{\partial \theta'} = \{\zeta_p(z)' \otimes \eta(t) : \eta(t) - \xi(t) : -\zeta_p(z)' \otimes \xi(t)\}\mathbf{S}'_\nu \tag{2.3.8}$$

where $\mathbf{S}'_\nu = diag(\mathbf{S}'_{\alpha(\nu)} : \mathbf{S}'_{\beta(\nu)} : \mathbf{S}'_{\lambda(\nu)})$ and the $v \times v^2$ matrix (derivative) sequences $\eta(t)$ and $\xi(t)$ can be recursively generated via (2.3.7) as

$$\sum_{j=0}^{p}\mathbf{M}(j)(\eta(t-j) : \xi(t-j)) = (\mathbf{y}(t)' : \mathbf{e}_\theta(t)') \otimes \mathbf{I}_v \tag{2.3.9}$$

with $\eta(t) = \xi(t) = 0,\ t \leq 0$. Some other approaches for computing the derivatives have been considered in the literature see, for example, Kashyap (1970), Astrom

(1980), Wilson and Kumar (1982) and the references therein. Now let us rewrite (2.3.1) as

$$\begin{aligned}\boldsymbol{\Delta}_T(\hat{\theta}_G^{(j)})\hat{\theta}_G^{(j+1)} &= \{\boldsymbol{\Delta}_T(\theta)\theta - \sum_{t=1}^{T}\frac{\partial \mathbf{e}_\theta(t)'}{\partial\theta}\boldsymbol{\Sigma}_T(\theta)^{-1}\mathbf{e}_\theta(t)\}|_{\theta=\hat{\theta}_G^{(j)}},\ j \geq 0 \\ &= -\sum_{t=1}^{T}\frac{\partial \mathbf{e}_\theta(t)'}{\partial\theta}\boldsymbol{\Sigma}_T(\theta)^{-1}\{(-\frac{\partial \mathbf{e}_\theta(t)}{\partial\theta'}\theta + \mathbf{e}_\theta(t))\}|_{\theta=\hat{\theta}_G^{(j)}} \\ &= -\sum_{t=1}^{T}\frac{\partial \mathbf{e}_\theta(t)'}{\partial\theta}\boldsymbol{\Sigma}_T^{-1}(\theta)\{\mathbf{e}_\theta(t) + (\eta(t) - \xi(t))vec\,\mathbf{I}_v\}|_{\theta=\hat{\theta}_G^{(j)}}\end{aligned} \tag{2.3.10}$$

where we have employed the relatively easily verified fact that $-\frac{\partial \mathbf{e}_\theta(t)}{\partial\theta'}\theta = \{\eta(t)\,vec\mathbf{I}_v - \xi(t)\,vec\,\mathbf{I}_v\}$. (See Appendix (A.2, p. 43) for a complete derivation). Thus, $\hat{\theta}_G^{(j+1)}$ is seen to be the vector of regression coefficients obtained from the multivariate regression of $(\eta(t) - \xi(t))\,vec\,\mathbf{I}_v + \mathbf{e}_\theta(t)$ on $-\frac{\partial \mathbf{e}_\theta(t)}{\partial\theta'}$ where these quantities are evaluated at the point $\hat{\theta}_G^{(j)}$ given by the previous iterate. There is an argument for iterating (2.3.1) or, its equivalent, (2.3.10) to obtain the fully efficient estimates, $\hat{\theta}_G$. For finite T, the residual $\hat{\mathbf{e}}_\theta(t)$ may not be exactly linear in the parameter θ and a single adjustment will not immediately produce the least squares values. Instead the adjusted values are substituted as new iterates and the process repeated until the specified convergence criterion occurs. Although there can be no improvement asymptotically, it may be useful to iterate (2.3.1) or equivalently (2.3.10) for finite T and usually one or two iterations may be worthwhile. Typically, convergence is usually faster if consistent estimators, such as may be obtained at the identification stage, are used at the commencement of iterations.

To complete the calculations it is therefore necessary to specify how such an initial value $\hat{\theta}_G^{(0)}$ is to be obtained. Following the technique suggested in Hannan and Kavalieris (1984) we find that $\hat{\theta}_G^{(0)}$ is achieved in two stages. The precise details of this procedure, together with the asymptotic properties of the estimator $\hat{\theta}_G^{(0)}$ so obtained, are provided in Chapter 3. For ease of reference in the subsequent chapters, the procedure for evaluating the initial value $\hat{\theta}_G^{(0)}$ together with the Gauss-Newton

stage will be referred to as the Gaussian estimation scheme. A detailed discussion of the Gauss-Newton procedure for estimating model (2.1.1) and its natural extension can be found in Hannan and Kavalieris (1984). See also Hannan and Deistler (1988, p. 294)).

The Gaussian estimator also has a natural generalized least squares representation as shown in Reinsel, Basu and Yap (1992). First, let the symbol $[\cdot]_{t=1}^{T}$ denotes a matrix (or vector) formed by T consecutive observations of the indicated argument from $t = 1$ to T. Employing this nomenclature leads to an expression for (2.3.9) in matrix form as

$$\mathbf{C}_G[\eta(t) : \xi(t)]_{t=1}^{T} = [\mathbf{y}(t)' \otimes \mathbf{I}_v : \mathbf{e}_\theta(t)' \otimes \mathbf{I}_v]_{t=1}^{T} \tag{2.3.11}$$

where $\mathbf{C}_G$ is a $Tv \times Tv$ lower triangular block Toeplitz matrix given by

$$\mathbf{C}_G = \begin{bmatrix} \mathbf{M}(0) & \mathbf{0} & \dots & \mathbf{0} & \dots & \dots & \mathbf{0} \\ \mathbf{M}(1) & \mathbf{M}(0) & \dots & \mathbf{0} & \dots & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \mathbf{M}(p) & \mathbf{M}(p-1) & \dots & \mathbf{M}(0) & \ddots & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{M}(p) & \dots & \mathbf{M}(1) & \mathbf{M}(0) & \dots & \mathbf{0} \\ \vdots & \mathbf{0} & \ddots & \ddots & \ddots & \ddots & \vdots \\ \mathbf{0} & \dots & \mathbf{0} & \mathbf{M}(p) & \dots & \dots & \mathbf{M}(0) \end{bmatrix}. \tag{2.3.12}$$

Setting $\mathbf{Z}_G = [\frac{\partial \mathbf{e}_\theta(t)}{\partial \theta'}]_{t=1}^{T}$, and using (2.3.11) it follows that $-\frac{\partial \mathbf{e}_\theta(t)}{\partial \theta'}$ can be written in matrix form as $\mathbf{C}_G\mathbf{Z}_G = \mathbf{X}_G$ where

$$\mathbf{X}_G = [\{-\zeta_p(z)' \otimes \mathbf{y}(t)' \otimes \mathbf{I}_v : -(\mathbf{y}(t)' - \hat{\mathbf{e}}_\theta(t)') \otimes \mathbf{I}_v : \zeta_p(z)' \otimes \hat{\mathbf{e}}_\theta(t)' \otimes \mathbf{I}_v\}\mathbf{S}_\nu']_{t=1}^{T}.$$

Similarly, put $\mathbf{W}_G = [\hat{\mathbf{e}}_\theta(t) + (\hat{\eta}(t) - \hat{\xi}(t))vec\mathbf{I}_v]_{t=1}^{T}$ and define $\mathbf{Y}_G$ via the relationship $\hat{\mathbf{C}}_G\mathbf{W}_G = \mathbf{Y}_G$. Hence $\mathbf{Y}_G = \hat{\mathbf{C}}_G[\hat{\mathbf{e}}_\theta(t)]_{t=1}^{T} + [\mathbf{y}(t) - \hat{\mathbf{e}}_\theta(t)]_{t=1}^{T}$ and

$$\begin{aligned} \hat{\theta}_G^{(j+1)} &= \{\mathbf{Z}_G'(\mathbf{I}_T \otimes \mathbf{\Sigma}_T(\theta)^{-1})\mathbf{Z}_G\}^{-1}\{\mathbf{Z}_G'(\mathbf{I}_T \otimes \mathbf{\Sigma}_T(\theta)^{-1})\mathbf{W}_G\}|_{\theta=\hat{\theta}_G^{(j)}},\ j \geq 0 \\ &= \{\mathbf{X}_G'\mathbf{C}_G'^{-1}(\mathbf{I}_T \otimes \mathbf{\Sigma}_T(\theta)^{-1})\mathbf{C}_G^{-1}\mathbf{X}_G\}^{-1}\{\mathbf{X}_G'\mathbf{C}_G'^{-1}(\mathbf{I}_T \otimes \mathbf{\Sigma}_T(\theta)^{-1})\mathbf{C}_G^{-1}\mathbf{Y}_G\}|_{\theta=\hat{\theta}_G^{(j)}} \\ &= \{\mathbf{X}_G'\hat{\mathbf{\Omega}}_G^{-1}\mathbf{X}_G\}^{-1}\{\mathbf{X}_G'\hat{\mathbf{\Omega}}_G^{-1}\mathbf{Y}_G\} \end{aligned} \tag{2.3.13}$$

where $\hat{\mathbf{\Omega}}_G = \hat{\mathbf{C}}_G(\mathbf{I}_T \otimes \mathbf{\Sigma}_T(\hat{\theta}_G))\hat{\mathbf{C}}'_G$, an estimate of $\mathbf{\Omega}_G$.

2.3.2 Modifications

In practice, it is often necessary to modify some aspects of the iterative scheme (2.3.1) in order for it to operate successfully. This becomes necessary because in finite sample, random flunctuations may cause the estimates of the parameters to occur outside the stability region and slow convergence of the iterates may also occur when the zero of the moving average operator is close to the boundary of the unit circle. In this section we now wish to discuss a number of modifications to the Gauss-Newton algorithm which are designed to handle these problems.

To implement the scheme (2.3.1) a check has to be made in (2.3.9) that the estimate of $\mathbf{M}(z)$ satisfies $\det \hat{\mathbf{M}}(z) \neq 0, |z| \leq 1$ so that the estimates of $\mathbf{e}_\theta(t)$, $\eta(t)$ and $\xi(t)$ can be formed. If this is not the case, these quantities will grow exponentially with t and consequently the calculations will break down due to numerical overflow. Procedures for checking the presence of zeros of an operator inside the unit circle and for reflecting those zeros that lead to instability on the unit circle are clearly required. Chapter 6 provides detailed discussion on this issue.

As we indicated in the last subsection, because the likelihood surface over which the search procedure is conducted may not be truly quadratic the scheme (2.3.1) has to be iterated. The iterations continue until some specified convergence criteria are reached. Some convergence criteria that are commonly used are the relative reduction in the objective function, the maximum change in the parameter values less than a specified level, or the number of iterations greater than a certain number. Slow convergence could occur if the zeros of $\det \hat{\mathbf{M}}(z)$ are close to $|z| = 1$ and consequently the effects due to starting the recursion (2.2.5) with $\mathbf{y}(t) = \mathbf{e}_\theta(t) = 0,\ t \leq 0$ in computing $\mathbf{e}_\theta(t)$ and initiating $\eta(t), \xi(t)$ to zero for $t \leq 0$ in (2.3.9) will die out very slowly. This effect can lead to poor estimates of $\mathbf{e}_\theta(t)$, $\eta(t)$ and $\xi(t)$ which in turn, can lead to a poor Gauss-Newton estimate, particularly in small samples. To

achieve faster convergence, many modifications of the Gauss-Newton scheme have been suggested. One of these modifications, which is designed to avoid the problem of overshooting and to help ensure that the objective function, $\det \boldsymbol{\Sigma}_T(\hat{\theta}_G)$, actually decreases at each iteration, is defined by

$$\hat{\theta}_G^{(j+1)} = \hat{\theta}_G^{(j)} - s^{(j)} \boldsymbol{\Delta}_T(\theta)^{-1} \sum_{t=1}^{T} \frac{\partial \mathbf{e}_\theta(t)'}{\partial \theta} \boldsymbol{\Sigma}_T(\theta)^{-1} \mathbf{e}_\theta(t)|_{\theta=\hat{\theta}_G^{(j)}}, \; j \geq 0$$

where $s^{(j)}$ is the scale factor (or step-length) at the j^{th} iterative step. It should be noted that the direction along which the the objective function decreases for a unit scale factor is only optimal if $\mathcal{L}_T(\theta)$ is truly quadratic. Since this condition cannot always be guaranteed, we stand to minimize the total computation time and ensure convergence of the Gauss-Newton procedure by selecting the scale factor in the given direction so that the objective function is decreased as much as possible. The success of this modification is mainly due to the fact that the Gauss-Newton direction is downhill so that each step of the modified iteration decreases $\mathcal{L}_T(\theta)$. For the reasons given earlier it may be worthwhile selecting s so that the zeros of $\det \mathbf{M}_s(z)$ are bounded some small distance away from the unit circle where $\mathbf{M}_s(z) = \mathbf{M}(0) + s\{\mathbf{M}(z) - \mathbf{M}(0)\}$. One possible way of determining s is to evaluate $\mathcal{L}_T(\theta, s)$ for $0 < s \leq 1$ and choose the value of s for which $\mathcal{L}_T(\theta, s)$ is minimized. For a detailed discussion of the various possible ways of choosing the scale factor s see, for example, Kowalik and Osborne (1968, p. 71), Bard (1974, p. 110 - 116), Kavalieris (1984), and Judge et al (1985, Section B2).

A different modification of the Gauss-Newton algorithm which is also designed to increase computational efficiency and enhance convergence, is due to Marquardt, Levenberg and Morrison (see Bard (1974, p. 94)) and is given by

$$\hat{\theta}_G^{(j+1)} = \hat{\theta}_G^{(j)} - \{\boldsymbol{\Delta}_T(\theta) + r_j \mathbf{I}\}^{-1} \{\sum_{t=1}^{T} \frac{\partial \mathbf{e}_\theta(t)'}{\partial \theta} \boldsymbol{\Sigma}_T(\theta)^{-1} \mathbf{e}_\theta(t)\}|_{\theta=\hat{\theta}_G^{(j)}}, \; j \geq 0$$

where $r_j > 0$ and $\mathbf{I}$ being an identity matrix of dimension $d(\nu)$. As some experience have shown, these modifications can result in improved convergence properties.

However, slow convergence or no convergence at all may be the consequence of using too high Kronecker indices.

2.4 Asymptotic Properties of Estimators

In this section the asymptotic properties of the estimators obtained from the minimization of (2.2.1), or equivalently (2.2.2), (2.2.4) are given. In using (2.2.4) all presample values are set equal to zero, but the error thereby introduced is transient and Dunsmuir and Hannan (1976) and Kohn (1978, 1979) have shown that the resulting estimators are asymptotically equivalent to the exact maximum likelihood ones. See also Nicholls (1976, 1977) and Anderson (1980). Hence, the asymptotic properties of the estimators obtained are the same.

We state the consistency results without proof but refer the reader, for example, to Dunsmuir and Hannan (1976), Kohn (1978), Hannan (1979), Rissanen and Caines (1979), and Hannan and Deistler (1988). Now let $\hat{\theta}_T$ (for fixed T) denote an argument minimizing $\mathcal{L}_T(\theta)$ and define $\hat{\mathbf{\Sigma}}_T$ as the corresponding estimate of the prediction error variance, $\mathbf{\Sigma}$.

Theorem (2.4.1): Consistency and Asymptotic Normality of the MLE *Assume* $\mathbf{y}(t)$ *is generated by an ARMA process satisfying (1.2.8) and that* $\hat{\theta}_T$ *and* $\hat{\mathbf{\Sigma}}_T$ *are the Gaussian (maximum likelihood) estimates of* θ_0 *and* $\mathbf{\Sigma}_0$*, respectively. If* $\varepsilon(t)$ *is egordic then*

$$\lim_{T\to\infty}(\hat{\theta}_T : \hat{\Sigma}_T) \quad \to \quad (\theta_0 : \mathbf{\Sigma}_0) \quad a.s.$$

In addition, if $\varepsilon(t)$ *satisfy Condition A then* $\sqrt{T}(\hat{\theta}_T - \theta_0)$ *converges in distribution to a zero mean Gaussian variate with variance-covariance matrix* $\mathbf{\Lambda}_G = (\mathbf{S}_\nu \mathbf{\Omega} \mathbf{S}_\nu')^{-1}$ *where the explicit form of* $\mathbf{\Omega}$ *is given by (2.4.2)* below.

In deriving the asymptotic normality result we need to evaluate $\mathbf{\Lambda}_G$. However, the exhibition of the elements of $\mathbf{\Lambda}_G$ is made difficult for two reasons. The first arises from the fact that the maximum lag depends on the row number and the second

from the fact that $\mathbf{A}(0) = \mathbf{M}(0)$ but is not necessarily equal to an identity. These difficulties are at least made conceptually simple by the introduction of the selection matrix, $\mathbf{S}_\nu$, into the estimation process. Thus, the variance-covariance matrix, $\mathbf{\Lambda}_G$, is obtained in the following way. Following Kohn (1978) it is readily shown that

$$\mathbf{\Lambda}_G = \left[\lim_{T\to\infty} T^{-1} \sum_{t=1}^{T} \frac{\partial \mathbf{e}_\theta(t)'}{\partial\theta} \mathbf{\Sigma}_T(\theta)^{-1} \frac{\partial \mathbf{e}_\theta(t)}{\partial\theta'} \right]^{-1} \tag{2.4.1}$$

and the derivatives along with $\mathbf{\Sigma}_T(\theta)$ are evaluated at the true parameter vector θ_0. The derivation of the variance-covariance matrix, $\mathbf{\Lambda}_G$, only involves the construction of the score vector $\frac{\partial \mathbf{e}_\theta(t)'}{\partial\theta}$ evaluated at $\theta = \theta_0$. Now employing expressions (2.3.4) - (2.3.6) evaluated at the true parameter values, θ_0, in (2.4.1) and observing that $\mathbf{y}(t)$ is expressible in terms of $\varepsilon(t)$ via Wold's decomposition as $\mathbf{y}(t) = \sum_{j\geq 0} \mathbf{K}(j)\varepsilon(t-j)$, $\mathbf{K}(z) = \mathbf{A}(z)^{-1}\mathbf{M}(z) = \sum_{j\geq 0} \mathbf{K}(j)z^j$, we find that

$$\begin{aligned} \mathbf{\Lambda}_G &= (\mathbf{S}_\nu \mathbf{\Omega} \mathbf{S}_\nu')^{-1}, \\ \mathbf{\Omega} &= (2\pi)^{-1} \int_{-\pi}^{\pi} (\mathbf{P} \otimes \mathbf{M}_0'^{-1} \mathbf{\Sigma}_0^{-\frac{1}{2}})(\mathbf{P} \otimes \mathbf{M}_0'^{-1} \mathbf{\Sigma}_0^{-\frac{1}{2}})^* d\omega \\ &= (2\pi)^{-1} \int_{-\pi}^{\pi} \{\mathbf{P}\mathbf{P}^* \otimes (\bar{\mathbf{M}}_0 \mathbf{\Sigma}_0 \mathbf{M}_0')^{-1}\} d\omega \end{aligned} \tag{2.4.2}$$

where

$$\mathbf{P}(z) = \begin{bmatrix} -\zeta_p(z) \otimes \mathbf{K}_0(z)\mathbf{\Sigma}_0^{\frac{1}{2}} \\ -(\mathbf{K}_0(z) - \delta_{0j}\mathbf{I}_v)\mathbf{\Sigma}_0^{\frac{1}{2}} \\ \zeta_p(z) \otimes \mathbf{\Sigma}_0^{\frac{1}{2}} \end{bmatrix}.$$

Herein, the argument of the integrand, $z = e^{i\omega}$, has been omitted for convenience. The symbols bar and asterisk have been used to denote the complex conjugate and complex conjugate transpose, respectively. See Dunsmuir and Hannan (1976, Section 4), and Hannan and Deistler (1988, p. 132) for a comparison.

Some remarks concerning Theorem (2.4.1) may be worthwhile. In the derivation of $\mathbf{\Lambda}_G$ no formal justification has been given for the assertion that the variance-covariance matrix in the limiting distribution is equal to the inverse of the Fisher's

information matrix. For a precise justification, reference should be made to Hannan (1970), Fuller (1976) or Crowther (1976).

It is perhaps worth emphasizing that in small samples the estimator obtained by minimizing (2.2.4) is not identical to the exact maximum likelihood estimator and this difference has been theoretically established see, for example, Dahlhaus (1988). There is also some Monte-Carlo evidence in the scalar case ($v = 1$) where the exact maximum likelihood estimators are preferrable in small sample situations see, for example, Ansley and Newbold (1980).

Appendix A

A.1 An Illustration

We now illustrate how the selection matrix, $\mathbf{S}_\nu = diag(\mathbf{S}_{\alpha(\nu)} : \mathbf{S}_{\beta(\nu)} : \mathbf{S}_{\lambda(\nu)})$, and the parameter vector θ introduced in Section (2.2) might be constructed from a given ARMA structure. To demonstrate this let us consider for expository purposes a bivariate ARMA process with Kronecker indices $n_1 = 2$ and $n_2 = 1$,

$$\mathbf{A}(z) = \begin{pmatrix} 1 & 0 \\ a_{21}(0) & 1 \end{pmatrix} + \begin{pmatrix} a_{11}(1) & 0 \\ a_{21}(1) & a_{22}(1) \end{pmatrix} z + \begin{pmatrix} a_{11}(2) & a_{12}(2) \\ 0 & 0 \end{pmatrix} z^2,$$

$$\mathbf{M}(z) = \begin{pmatrix} 1 & 0 \\ a_{21}(0) & 1 \end{pmatrix} + \begin{pmatrix} m_{11}(1) & m_{12}(1) \\ m_{21}(1) & m_{22}(1) \end{pmatrix} z + \begin{pmatrix} m_{11}(2) & m_{12}(2) \\ 0 & 0 \end{pmatrix} z^2.$$

Thus, since $p = \max(n_1, n_2) = 2$ it follows that

$$\begin{aligned} \bar{\alpha} &= vec(\mathbf{A}(1) : \mathbf{A}(2)) \\ &= (a_{11}(1),\ a_{21}(1),\ 0,\ a_{22}(1),\ a_{11}(2),\ 0,\ a_{12}(2),\ 0)' \\ \bar{\beta} &= vec(\mathbf{A}(0) - \mathbf{I}_v) \\ &= (0,\ a_{21}(0),\ 0,\ 0)' \end{aligned}$$

and

$$\begin{aligned}\bar{\lambda} &= vec(\mathbf{M}(1) : \mathbf{M}(2)) \\ &= (m_{11}(1),\ m_{21}(1),\ m_{12}(1),\ m_{22}(1),\ m_{11}(2),\ 0,\ m_{12}(2),\ 0)'.\end{aligned}$$

Then the corresponding selection matrices constructed in a manner described in Section (2.2) are

$$\mathbf{S}_{\alpha(\nu)} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix},$$

$$\mathbf{S}_{\beta(\nu)} = (0,1,0,0)$$

and

$$\mathbf{S}_{\lambda}(\nu) = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix},$$

respectively. By direct matrix multiplication it is readily verified that

$$\begin{aligned}\alpha &= \mathbf{S}_{\alpha(\nu)}\bar{\alpha} = (a_{11}(1),\ a_{21}(1),\ a_{22}(1),\ a_{11}(2),\ a_{12}(2))' \\ \beta &= \mathbf{S}_{\beta(\nu)}\bar{\beta} = a_{21}(0) \\ \lambda &= \mathbf{S}_{\lambda(\nu)}\bar{\lambda} = (m_{11}(1),\ m_{21}(1),\ m_{12}(1),\ m_{22}(1),\ m_{11}(2),\ m_{12}(2))'\end{aligned}$$

where, as easily checked, $d_\alpha = 5$, $d_\beta = 1$ $d_\lambda = 6$ and hence, the number of freely varying parameters, d_ν, is 12. The construction of the parameter vector θ as $\theta = (\alpha' : \beta' : \lambda')'$ is now straightforward.

A.2 Formula Derivation

We now wish to verify the validity of the substitution made for $-\frac{\partial \mathbf{e}_\theta(t)}{\partial \theta'}\theta$ in (2.3.10). First recall that

$$\frac{\partial \mathbf{e}_\theta(t)}{\partial \theta'} = (\frac{\partial \mathbf{e}_\theta(t)}{\partial \alpha'} : \frac{\partial \mathbf{e}_\theta(t)}{\partial \beta'} : \frac{\partial \mathbf{e}_\theta(t)}{\partial \lambda'})$$

and

$$\theta = (\alpha' : \beta' : \lambda')'.$$

Now consider

$$\begin{aligned} -\frac{\partial \mathbf{e}_\theta(t)}{\partial \theta'}\theta &= -(\frac{\partial \mathbf{e}_\theta(t)}{\partial \alpha'} : \frac{\partial \mathbf{e}_\theta(t)}{\partial \beta'} : \frac{\partial \mathbf{e}_\theta(t)}{\partial \lambda'}) \cdot (\alpha' : \beta' : \lambda')' \\ &= -(\frac{\partial \mathbf{e}_\theta(t)}{\partial \alpha'}\alpha + \frac{\partial \mathbf{e}_\theta(t)}{\partial \beta'}\beta + \frac{\partial \mathbf{e}_\theta(t)}{\partial \lambda'}\lambda). \end{aligned}$$

Substituting for $\eta(t)$ and $\xi(t)$ in $\frac{\partial \mathbf{e}_\theta(t)}{\partial \alpha'}$, $\frac{\partial \mathbf{e}_\theta(t)}{\partial \beta'}$ and $\frac{\partial \mathbf{e}_\theta(t)}{\partial \lambda'}$ using (2.3.7) gives an alternative expression for $-\frac{\partial \mathbf{e}_\theta(t)}{\partial \theta'}\theta$ as

$$\begin{aligned} -\frac{\partial \mathbf{e}_\theta(t)}{\partial \theta'}\theta &= -(\{\zeta_p(z)' \otimes \eta(t)\}\mathbf{S}'_{\alpha(\nu)}\alpha + \{\eta(t) - \xi(t)\}\mathbf{S}'_{\beta(\nu)}\beta \\ &\quad + \{\zeta_p(z)' \otimes \xi(t)\}\mathbf{S}'_{\lambda(\nu)}\lambda) \\ &= -(\{1 \otimes \eta(t)\}\{(\zeta_p(z)' \otimes \mathbf{I}_{v^2}) \cdot \begin{pmatrix} vec(\mathbf{A}(1)) \\ \vdots \\ vec(\mathbf{A}(p)) \end{pmatrix}\} \\ &\quad + \{\eta(t) - \xi(t)\}vec(\mathbf{A}(0) - \mathbf{I}_v) \\ &\quad - \{1 \otimes \xi(t)\}\{(\zeta_p(z)' \otimes \mathbf{I}_{v^2}) \begin{pmatrix} vec(\mathbf{M}(1)) \\ \vdots \\ vec(\mathbf{M}(p)) \end{pmatrix}\}) \\ &= -(\eta(t)\{vec\mathbf{A}(z) - vec\mathbf{A}(0)\} + \{\eta(t) - \xi(t)\}vec(\mathbf{A}(0) - \mathbf{I}_v) \\ &\quad - \xi(t)\{vec\,\mathbf{M}(z) - vec\,\mathbf{A}(0)\}) \\ &= -(\eta(t)\ vec\,\mathbf{A}(z) - \eta(t)\ vec\,\mathbf{I}_v + \xi(t)\ vec\,\mathbf{I}_v - \xi(t)\ vec\,\mathbf{M}(z)). \end{aligned}$$

Now using (2.3.7) we can rewrite the last expression by substituting for $\eta(t)$ in the first term and $\xi(t)$ in the last term to obtain

$$\begin{aligned} -\frac{\partial \mathbf{e}_\theta(t)}{\partial \theta'}\theta &= -(\{\mathbf{y}(t)' \otimes \mathbf{M}(z)^{-1}\} vec\, \mathbf{A}(z) - \eta(t)\ vec\, \mathbf{I}_v \\ &\quad + \xi(t) vec\, \mathbf{I}_v - \{\mathbf{e}_\theta(t)' \otimes \mathbf{M}(z)^{-1}\} vec\, \mathbf{M}(z)). \end{aligned}$$

Employing the rule $(\mathbf{C}' \otimes \mathbf{A}) vec\, \mathbf{B} = vec(\mathbf{ABC})$ this expression can be further simplified as

$$\begin{aligned} -\frac{\partial \mathbf{e}_\theta(t)}{\partial \theta'}\theta &= -(vec\{\mathbf{M}^{-1}(z)\mathbf{A}(z)\mathbf{y}(t)\} - \eta(t) vec \mathbf{I}_v + \xi(t) vec \mathbf{I}_v \\ &\quad - vec\{\mathbf{M}^{-1}(z)\mathbf{M}(z)\mathbf{e}_\theta(t)\}) \\ &= -(\mathbf{e}_\theta(t) - \eta(t) vec\, \mathbf{I}_v + \xi(t) vec\, \mathbf{I}_v - \mathbf{e}_\theta(t)) \\ &= \eta(t) vec\, \mathbf{I}_v - \xi(t) vec\, \mathbf{I}_v \end{aligned}$$

where the penultimate line follows from the basic definitions $\mathbf{e}_\theta(t) = \mathbf{M}^{-1}(z)\mathbf{A}(z)\mathbf{y}(t)$ and $\mathbf{M}^{-1}(z)\mathbf{M}(z) = \mathbf{I}_v$. Thus, we conclude that the substitution made for $-\frac{\partial \mathbf{e}_\theta(t)}{\partial \theta'}\theta$ in (2.3.10) is quite legitimate.

Chapter 3

Asymptotic Properties Of Least Squares Estimators

3.1 Introduction

The problem addressed in this chapter is motivated by a consideration of the maximum likelihood estimation for vector ARMA models. As was stressed in the previous chapter, it is important to commence the iterative non-linear optimization of the likelihood function with good initial (consistent) estimates of the system parameters. In theory, these preliminary estimates can be determined using the maximum likelihood procedure but the integer parameters (the Kronecker indices) that govern the structure of the underlying process are rarely, if ever, known, and will have to be determined from the data. Such a determination, as was demonstrated in Chapter 1 (Section (1.3)), involves estimating the parameters of a range of different models and because of the non-linear nature of the likelihood equations this requirement could impose a large computational burden. It is for this reason that practitioners usually apply maximum (Gaussian) likelihood technique only when they know which model specifications are worth estimating efficiently. It is therefore desirable to have good and simple preliminary methods of estimation which could be used to experiment with different model specifications and to produce accurate initial values for the iterative maximum likelihood procedure. This consideration is inspired by Hannan and Rissanen (1982) in the context of scalar time series models, although it has its origin in the earlier contribution by Durbin (1960) who introduced a three stage procedure to estimate the parameters of the ARMA model when the model specification is assumed known. This procedure has been extended to vector ARMA(X) case by Hannan and Kavalieris (1984). In the sequel, we will be concerned with the first two stages of this estimation scheme and in particular, restrict attention to the most straightforward part of modelling procedure, namely, the estimation for fixed values of the Kronecker indices, $\nu = \{n_1, \ldots, n_v)\}$, of the structural parameters using ordinary least squares method.

The rest of this chapter is organized into three sections. Section (3.2) describes

the main ideas of the estimation procedure. Some technical results associated with the first stage of the procedure are stated. Consistency and the asymptotic normality results for the least squares estimator are derived in Sections (3.3) and (3.4), respectively. The discussion of the algorithms that can be employed in the implementation of the first stage of the estimation scheme is contained in the appendix to this chapter (see p. 78).

3.2 The Least Squares Estimator

In this section we give a brief exposition of the stages involved in the determination of the least squares estimator of the parameter vector θ. The form of the model we shall be concerned with was given in Chapter 1 as

$$\sum_{j=0}^{p} \mathbf{A}(j)\mathbf{y}(t-j) = \sum_{j=0}^{p} \mathbf{M}(j)\varepsilon(t-j) \tag{3.2.1}$$

where the standard assumptions (1.1.4), (1.1.5), (1.2.8) and Condition A are assumed to hold. When (1.1.5) holds, $\mathbf{y}(t)$ has an infinite autoregressive representation

$$\sum_{j=0}^{\infty} \mathbf{\Phi}(j)\mathbf{y}(t-j) = \varepsilon(t), \quad \mathbf{\Phi}(0) = \mathbf{I}_v \tag{3.2.2}$$

where the generating function $\mathbf{\Phi}(z) = \sum_{j\geq 0} \mathbf{\Phi}(j)z^j$ is as in (2.2.5). For notational simplicity, we have neglected the mean corrections on the process $\mathbf{y}(t)$ but these would be needed in practice and, as earlier indicated, the results presented below will not be affected by that adjustment.

Now given that $\mathbf{y}(t)$ is sampled at T consecutive time points the following two stage estimation scheme may be used to obtain the least squares estimates of θ. Such an estimation scheme can be found in Poskitt (1992) and is a variant of the method proposed in Hannan and Kavalieris (1984). The first step of this procedure commences with an autoregression.

3.2.1 Stage 1: Autoregressive approximation

Consider an autoregressive approximation to the representation (3.2.2) of the form

$$\sum_{j=0}^{h} \mathbf{\Phi}_h(j)\mathbf{y}(t-j) = \varepsilon_h(t), \quad \mathbf{\Phi}_h(0) = \mathbf{I}_v \tag{3.2.3}$$

where $\mathcal{E}\{\varepsilon_h(t)\varepsilon_h(t)'\} = \mathbf{\Sigma}_h$. Let $\mathbf{y}_h(t) = \sum_{j=1}^{h} \mathbf{\Phi}_h(j)\mathbf{y}(t-j)$ denote the best linear predictor of $\mathbf{y}(t)$ based on $\mathbf{y}(t-j), j = 1, \ldots, h$ and best being defined in terms of the minimizing mean square error (see Theorem (1.1.1)). The precise nature of the procedures that are usually employed to solve (3.2.3) is provided in Appendix (B, p.77). Thus, the model fitted will be of the form

$$\sum_{j=0}^{h} \hat{\mathbf{\Phi}}_h(j)\mathbf{y}(t-j) = \hat{\varepsilon}_h(t), \quad \hat{\mathbf{\Phi}}_h(0) = \mathbf{I}_v \tag{3.2.4}$$

where h is chosen to minimize the criterion function

$$\log\{\det \hat{\mathbf{\Sigma}}_h\} + \frac{hv^2}{T}C(T), \quad h \leq H_T \tag{3.2.5}$$

where $\hat{\mathbf{\Sigma}}_h$ is an estimate of $\mathbf{\Sigma}_h$. H_T and $C(T)$ are prescribed sequences, the forms of which are provided later. As part of the statistical procedure, the order h will have to be determined from the available data. One possibility for determining h, the order of the approximating autoregression (3.2.4), is to use a criterion such as Akaike Information Criterion (AIC). This requires that we select that value of h for which (3.2.5) is minimum with $C(T) = 2$. Another possibility is via Bayesian Information Criterion (BIC) where h is chosen with reference to (3.2.5) and $C(T) = \log T$. The second term in (3.2.5) can be thought of as a penalty for heavily parameterized models. As was shown in Hannan and Kavalieris (1986, Table 1), if T is not very large then the values of h obtained in practice, using BIC, tend to be a great deal smaller. This result tends to suggest that it is likely that AIC will often be preferred but has the tendency to overestimate h (Jones (1975), Shibata (1976)). The reliability of AIC in a modelling situation has been justified on the Gaussian assumption by Shibata (1980) nevertheless Steinberg, Gasser and Franke (1985) in

an empirical study conducted on EEG data provided an example where BIC seems to perform better. However, at this point our objective is not to estimate the order h of a true model (if that ever exists) but rather to approximate a mixed model by an autoregression of sufficiently high order. Therefore, a rather liberal criterion might be appropriate for choosing this order. For a detailed discussion on the appropriateness of order selection, see Hannan and Deistler (1988, Chapters 5 and 6) and Lutkepohl (1985) with the latter author providing some simulation studies in connection with the vector autoregressive processes.

The $\hat{\mathbf{\Phi}}_h(j)$, $\hat{\varepsilon}_h(t)$ in (3.2.4) are estimates of the corresponding quantities $\mathbf{\Phi}_h(j)$, $\varepsilon_h(t)$ which are defined via a consideration of the best linear predictor, $\mathbf{y}_h(t)$. The statistical properties of the estimates associated with (3.2.4) have been treated elsewhere in the literature see, for example, Hannan and Kavalieris (1986). For ease of reference, it will suffice to merely collate those that are of direct relevance to our subsequent discussions.

We first consider the rate of convergence of the coefficients, $\hat{\mathbf{\Phi}}_h = (\hat{\mathbf{\Phi}}_h(1)', \ldots, \hat{\mathbf{\Phi}}_h(h)')'$. In particular, we look at the differences $\|\hat{\mathbf{\Phi}}_h - \mathbf{\Phi}_h\|$ where $\|\mathbf{B}\|$ is defined for $m \times n$ matrix $\mathbf{B}$ to be the Euclidean length of the mn dimensional vector consisting of all the components of the indicated matrix and $\mathbf{\Phi}_h$ is as $\hat{\mathbf{\Phi}}_h$ with estimated quantities being replaced by their theoretical counterparts.

Lemma (3.2.1) [Hannan and Kavalieris (1986)]

Assume $\mathbf{y}(t)$ *is generated by an ARMA process (3.2.1) and* $\varepsilon(t)$ *satisfying Condition A. Then for* $h \leq (\log T)^\tau$, $\tau < \infty$,

$$\sup_{1 \leq j \leq h} \|\hat{\mathbf{\Phi}}_h(j) - \mathbf{\Phi}_h(j)\| = O(Q_T) \quad a.s$$

where $Q_T = (\log \log T/T)^{\frac{1}{2}}$.

(Here and in the remainder of the thesis, the quantity Q_T will be assumed to be as defined in Lemma (3.2.1)).

The result of Lemma (3.2.1) leads us to examine the differences $\|\mathbf{\Phi}_h(j) - \mathbf{\Phi}_0(j)\|$,

$j \leq h$ where $\mathbf{\Phi}_0(j), j = 1, 2, \ldots, \infty$ are defined by the relation (3.2.2). The subscript zero is used here again to denote an evaluation at the true parameter values and the same nomenclature will be maintained in the remaining sections of this chapter. Since the process generating the data may not be truly an autoregression it may be useful to assume that $\mathbf{\Phi}_0(z) = \sum_{j=0}^{\infty} \mathbf{\Phi}_0(j) z^j$ has an equivalent representation $\mathbf{M}_0(z)^{-1}\mathbf{A}_0(z) = \{\det \mathbf{M}_0(z)\}^{-1}\mathbf{N}_0(z)$ where $\mathbf{N}_0(z)$ is a matrix of polynomials.

To motivate further discussion, suppose there exists an absolutely summable sequence $c(k)$ such that $\phi_h(e^{i\omega}) = \sum_{j=1}^{h} \phi_h(j) e^{ij\omega}$ satisfy

$$\int_{-\pi}^{\pi} e^{ik\omega} \phi_h(e^{i\omega}) f_y(\omega) d\omega = c(k), \; k = 1, \ldots, h \tag{3.2.6}$$

where $f_y(\omega)$, a spectral density matrix, is as defined in (1.1.9) with $\phi_h(e^{i\omega})$, $f_y(\omega)$ and c denoting the appropriate equivalent scalar quantities corresponding to $\mathbf{\Phi}_h(e^{i\omega})$, $\mathbf{f}_y(\omega)$ and C, respectively. Under some mild conditions on $f_y(\omega)$, Baxter (1963) shows that

$$\sum_{j=1}^{h} |\phi_h(j)| \leq \kappa \sum_{k=1}^{h} c(k) \tag{3.2.7}$$

and κ is independent of the sequence $c(k)$ or h. When the multivariate generalization of (3.2.7), (Findley(1992), Hannan and Deistler (1988)), is applied to the transfer function $\mathbf{\Phi}_0(e^{i\omega}) - \mathbf{\Phi}_h(e^{i\omega})$ one gets the following important result relating the differences $\mathbf{\Phi}_h(j) - \mathbf{\Phi}_0(j)$, $j = 1, \ldots, h$, to the tail of the series $\sum_{j=0}^{\infty} \mathbf{\Phi}_0(j)\mathbf{y}(t-j)$ where $\mathbf{\Phi}_0(e^{i\omega}) = \sum_{j=1}^{\infty} \mathbf{\Phi}_0(j) e^{ij\omega}$ and $\mathbf{\Phi}_h(e^{i\omega})$ is analogously defined.

Theorem (3.2.1) [Baxter's Inequality]

Let $\mathbf{f}_y(\omega) = \frac{1}{2\pi}\mathbf{K}_0(e^{i\omega})\mathbf{\Sigma}_0\mathbf{K}_0(e^{i\omega})^*$ *as in (1.1.9),* $\mathbf{K}_0(z) = \mathbf{A}_0(z)^{-1}\mathbf{M}_0(z)$ *under (1.1.4) and (1.1.5) with* $\mathbf{K}_0(z)^{-1}\mathbf{y}(t) = \mathbf{M}_0(z)^{-1}\mathbf{A}_0(z)\mathbf{y}(t) = \sum_{j\geq 0} \mathbf{\Phi}_0(j)\mathbf{y}(t-j)$. *Let* $\mathbf{\Phi}_h$ *be as defined above. Then there exist an integer* $\mathcal{N}$ *and a constant* κ *depending on* $\mathbf{f}_y(\omega)$ *such that for* $h \geq \mathcal{N}$,

$$\sum_{j=1}^{h} \|\mathbf{\Phi}_h(j) - \mathbf{\Phi}_0(j)\| \leq \kappa \sum_{j\geq h+1} \|\mathbf{\Phi}_0(j)\|. \tag{3.2.8}$$

We may now proceed to obtain a bound for the term on the right hand side of (3.2.8). Let ρ_0 be the modulus of the zero of $\det \mathbf{M}_0(z^{-1})$ nearest $|z| = 1$, then under the assumption (2.2.3), ρ_0 satisfies the inequality $0 < \rho_0 < 1$. By taking partial fraction expansion of $\{\det \mathbf{M}_0(z)\}^{-1}$ it can be readily verified that the $\|\mathbf{\Phi}_0(j)\|$ decreases geometrically at a rate determined by ρ_0. Indeed, $\|\mathbf{\Phi}_0(j)\| \leq \kappa\rho_0^j$. Thus, if we let $\ell(T) \leq h \leq H_T$ where $\ell(T) = \log T/(-2\log\rho_0)$ and in particular, set $h = \ell(T)$ we have

$$\begin{aligned}
\kappa \sum_{j=h+1}^{\infty} \|\mathbf{\Phi}_0(j)\| &\leq \kappa \sum_{j=h+1}^{\infty} \rho_0^j = \kappa\, \rho_0^h \sum_{j=1}^{\infty} \rho_0^j \\
&= constant \cdot exp\{h \log \rho_0\} \\
&= constant \cdot exp\{\frac{\log T}{-2\log\rho_0} \cdot \log\rho_0\} \\
&= constant \cdot T^{-\frac{1}{2}}. \qquad (3.2.9)
\end{aligned}$$

So far we have stated some results concerning the accuracy of the predictor coefficients, $\hat{\mathbf{\Phi}}_h$, in relation to an autoregressive approximation (3.2.3). It may also be of interest to know at what rate the convergence of quantities such as $T^{-1}\sum_t \varepsilon(t)\varepsilon(t-j)'$, $T^{-1}\sum_t \varepsilon(t)\mathbf{y}(t-s)', etc$, takes place. Thus the following results are of importance.

Theorem (3.2.2) [An, Chen and Hannan(1982)]

Let $\varepsilon(t)$ satisfy condition A and

$$\mathbf{y}(t) = \sum_{j=0}^{\infty} \mathbf{K}(j)\varepsilon(t-j), \ \sum_{j=0}^{\infty} \|\mathbf{K}(j)\|^2 < \infty, \ \mathbf{K}(0) = \mathbf{I}_v. \qquad (3.2.10)$$

Then for $H_T \leq (\log T)^\tau$, $\tau < \infty$,

$$\sup_{0\leq s\leq H_T} \|T^{-1} \sum_{t=s+1}^{T} \varepsilon(t)\mathbf{y}(t-s)' - \delta_{0s}\mathbf{\Sigma}\| = O(Q_T).$$

It follows that the autocovariances may also be bounded. Now define $\hat{\mathbf{\Gamma}}_{yy}(r) = T^{-1}\sum_{t=1}^{T-r} \mathbf{y}(t)\mathbf{y}(t+r)' = \hat{\mathbf{\Gamma}}_{yy}(-r)'$, $r = 0, 1, \ldots, T-1$ and $\mathbf{\Gamma}_{yy}(r) = \mathcal{E}\{\mathbf{y}(t)\mathbf{y}(t+r)'\}$.

Similarly, we shall let $\mathbf{\Gamma}_h$ denote a square matrix of $h \times h$ blocks whose $(j,k)^{th}$ $v \times v$ subblock of elements is given by $\mathbf{\Gamma}_{yy}(j-k)$, $j,k = 1,\ldots,h$ for integer $h > 0$ and that $\mathbf{G}_h$ is its estimate.

Theorem (3.2.3) [An, Chen and Hannan (1982)]

Assume conditions of Theorem (3.2.2) hold and that $\mathbf{K}(z) = \sum_{j=0}^{\infty} \mathbf{K}(j)z^j$ *is rational. Then for* $H_T \leq (\log T)^\tau$, $\tau < \infty$

$$\sup_{0 \leq k \leq H_T} \|\hat{\mathbf{\Gamma}}_{yy}(k) - \mathbf{\Gamma}_{yy}(k)\| = O(Q_T) \quad a.s$$

.

Results of this kind first appeared in An, Chen and Hannan (1982) for the scalar ARMA case although results based on independence assumptions about $\varepsilon(t)$ were given in Hannan (1980, 1981). These results were further improved and extended to vector ARMA (ARMAX) case by Hannan and Kavalieris (1983, 1984) under some weaker conditions. However, Condition A will suffice for our purpose.

From Theorem (3.2.3) it follows that, uniformly in h, $h \leq H_T$,

$$\|\mathbf{G}_h - \mathbf{\Gamma}_h\| = o(1) \tag{3.2.11}$$

since the typical element of (3.2.11) is of the form $\hat{\mathbf{\Gamma}}_{yy}(j,k) - \mathbf{\Gamma}_{yy}(j-k)$.

Lemma (3.2.2) [Hannan and Deistler (1988)]

Assume condition A holds. If $h \leq H_T$ *then*

$$0 < \kappa_1 \leq \lambda_{min}\{\mathbf{G}_h\} \leq \lambda_{max}\{\mathbf{G}_h\} \leq \kappa_2 < \infty \; a.s \;\; as \;\; T \to \infty$$

where $\lambda_{min}\{\mathbf{G}_h\}$, $\lambda_{max}\{\mathbf{G}_h\}$ *denote the smallest and largest eigen values of* $\mathbf{G}_h$. Consequently, it can be shown (Hannan and Kavalieris (1984)) that $\mathbf{G}_h^{-1}$, $\mathbf{\Gamma}_h^{-1}$ are bounded $a.s$ and uniformly in h.

Lemma (3.2.3) [Hannan and Kavalieris (1984)]

Under condition A and uniformly in $h \leq H_T = (\log T)^\tau$, $\tau < \infty$,

$$\|\mathbf{G}_h^{-1}\|_\infty \leq \kappa_1 < \infty, \quad \|\mathbf{\Gamma}_h^{-1}\|_\infty \leq \kappa_2 < \infty. \tag{3.2.12}$$

where κ_1 and κ_2 are some constants.

The last result of this section relates to the rate of convergence or more precisely the accuracy of prediction errors, $\hat{\varepsilon}_h(t)$. If h is sufficiently large, the residual $\{\hat{\varepsilon}_h(t)\}$ will approximate the innovation sequence, $\{\varepsilon(t)\}$, in a manner made explicit in Kavalieris (1984, Lemma (5.1.3)). See also Hannan and Deistler (1988, Theorem (6.6.5)). Thus, the convergence of quantities of the form $T^{-1}\sum_t(\hat{\varepsilon}_h(t)-\varepsilon(t))\,(\hat{\varepsilon}_h(t-j)-\varepsilon(t-j))'$ can therefore be analyzed.

Lemma (3.2.4) [Kavalieris (1984)]

Let $\mathbf{y}(t)$ be generated by an ARMA process (3.2.1) satisfying (1.1.4), (1.1.5) and (1.2.8). Assume $\varepsilon(t)$ satisfy condition A. If $\ell(T) \leq h \leq H_T = (\log T)^\tau$, $\tau < \infty$, then uniformly in $h \leq H_T$, for $j \geq 0$

$$T^{-1}\sum_{t=1}^{T}\{\hat{\varepsilon}_h(t)-\varepsilon(t)\}\,\{\hat{\varepsilon}_h(t-j)-\varepsilon(t-j)\}' = \delta_{0j}\frac{hv}{T}\mathbf{\Sigma} + o_p(\frac{hv}{T})$$

$$T^{-1}\sum_{t=1}^{T}\varepsilon(t)\{\hat{\varepsilon}_h(t-j)-\varepsilon(t-j)\}' = -\delta_{0j}\frac{hv}{T}\mathbf{\Sigma} + o_p(\frac{hv}{T}).$$

The essence of this stage, however, is to provide estimates of the unobserved innovation sequence $\{\varepsilon(t)\}$. In what follows, we shall be presenting a procedure that leads to the determination of the least squares estimators and which constitutes the second stage of the estimation scheme being employed in this chapter.

3.2.2 Stage 2: Least squares method

Suppose that observations $\mathbf{y}(t)$, $t = 1,\ldots,T$, are available on the output process of the system (3.2.1) and that the estimates of the unobserved innovation sequence $\{\varepsilon(t)\}$ have been estimated via the regression equations (3.2.4) using criterion function (3.2.5) possibly with $C(T)$ set equal to 2. As earlier indicated, the

Kronecker indices, $\nu = \{n_1, \ldots, n_v\}$, are fixed *a priori*. Now consider the multivariate least squares estimation of the true system parameters in the representation

$$\sum_{j=0}^{p} \mathbf{A}(j)\mathbf{y}(t-j) = \sum_{j=0}^{p} \mathbf{M}(j)\hat{\varepsilon}_h(t-j) \tag{3.2.13}$$

where, as before, $p = max(n_1, \ldots, n_v)$. The least squares estimates of the freely varying parameters in (3.2.13) are obtained by regressing each component of $\mathbf{y}(t)$ on certain components of $\mathbf{y}(t-j)$, $\mathbf{y}(t) - \hat{\varepsilon}_h(t)$, $\hat{\varepsilon}_h(t-j)$, $j = 1, \ldots, p$. One obvious way of carrying out this task is by adding and deleting variables from a regression as in Seber (1977, p. 338 - 342). Since the selection of the regressor variables for the i^{th} regression equation is governed by the maximum Kronecker index associated with the corresponding i^{th} row in (3.2.13) it may then be convenient to systematize these choices. Lutkepohl (1991, p. 265-270) and Poskitt (1992) provide some thinking in this direction. We will make use of such a device in our exposition which in turn provides an appropriate starting point for our present discussion. The attraction of such an approach is that the freely varying parameters in (3.2.13) can now be relatively and cheaply estimated by ordinary least squares method.

Before proceeding, it may be helpful to introduce some notation and results in a matrix algbera. Let α, β, λ, θ, $\mathbf{S}_{\alpha(\nu)}$, $\mathbf{S}_{\beta(\nu)}$ and $\mathbf{S}_{\lambda(\nu)}$ be as defined in Section (2.2) of Chapter 2. We now collate some additional notation and results that will be called upon in the subsequent developments. Employing the algebraic conventions for vec operators and tensor (or Kronecker) matrix product presented in Magnus and Neudecker (1988), and Henderson and Searle (1979), we have

$$\begin{aligned} vec\{\mathbf{A}(z)\} &= vec\{\sum_{j=1}^{p} \mathbf{A}(j)z^j\} + vec\{\mathbf{A}(0)\} \\ &= \{\zeta_p(z)' \otimes \mathbf{I}_{v^2}\} vec\{\mathbf{A}(1) : \ldots : \mathbf{A}(p)\} + vec\{\mathbf{A}(0)\} \\ &= (\zeta_p(z)' \otimes \mathbf{I}_{v^2})\mathbf{S}'_{\alpha(\nu)}\alpha + \mathbf{S}'_{\beta(\nu)}\beta + vec\,\mathbf{I}_v \end{aligned} \tag{3.2.14}$$

and similarly,

$$vec\{\mathbf{M}(z)\} = vec\{\sum_{j=1}^{p} \mathbf{M}(j)z^j\} + vec\{\mathbf{A}(0)\}$$

$$= (\zeta_p(z)' \otimes \mathbf{I}_{v^2})\mathbf{S}'_{\lambda(\nu)}\lambda + \mathbf{S}'_{\beta(\nu)}\beta + vec\,\mathbf{I}_v \qquad (3.2.15)$$

since $\mathbf{A}(0) = \mathbf{M}(0)$. To complete the preliminaries, let (3.2.13) be rewritten in the form $\mathbf{e}_\nu(t) = \mathbf{A}(z)\mathbf{y}(t) - (\mathbf{M}(z) - \mathbf{I}_v)\hat{\varepsilon}_h(t)$. Now applying the rule $vec\mathbf{ABC} = (\mathbf{C}' \otimes \mathbf{A})vec\mathbf{B}$ in conjuction with (3.2.14) and (3.2.15), and noting that a column vector is its own vectorization, we have

$$\begin{aligned} vec\{\mathbf{A}(z)\mathbf{y}(t)\} &= (\mathbf{y}(t)' \otimes \mathbf{I}_v)vec\mathbf{A}(z) \\ &= (\mathbf{y}(t)' \otimes \mathbf{I}_v)\{(\zeta_p(z)' \otimes \mathbf{I}_{v^2})\mathbf{S}'_{\alpha(\nu)}\alpha + \mathbf{S}'_{\beta(\nu)}\beta + vec\,\mathbf{I}_v\} \\ &= (1 \otimes Y_t)(\zeta_p(z)' \otimes \mathbf{I}_{v^2})\mathbf{S}'_{\alpha(\nu)}\alpha + (\mathbf{y}(t)' \otimes \mathbf{I}_v)\mathbf{S}'_{\beta(\nu)}\beta + \mathbf{y}(t) \\ &= \mathbf{y}(t) + (\zeta_p(z)' \otimes Y_t)\mathbf{S}'_{\alpha(\nu)}\alpha + (y(t)' \otimes \mathbf{I}_v)\mathbf{S}'_{\beta(\nu)}\beta \\ &= \mathbf{y}(t) + \mathbf{Y}_\nu(t)\alpha + (\mathbf{y}(t)' \otimes \mathbf{I}_v)\mathbf{S}'_{\beta(\nu)}\beta \qquad (3.2.16) \end{aligned}$$

where $\mathbf{Y}_\nu(t) = (\zeta_p(z)' \otimes \mathbf{y}(t)' \otimes \mathbf{I}_v)\mathbf{S}'_{\alpha(\nu)}$ and $Y_t = (\mathbf{y}(t)' \otimes \mathbf{I}_v)$. A completely analogous manipulation applied to $vec\{\mathbf{M}(z)\hat{\varepsilon}_h(t)\}$ yields

$$vec\{\mathbf{M}(z)\hat{\varepsilon}_h(t)\} = \hat{\varepsilon}_h(t) + \hat{\mathbf{E}}_\nu(t)\lambda + (\hat{\varepsilon}_h(t)' \otimes \mathbf{I}_v)\mathbf{S}'_{\beta(\nu)}\beta \qquad (3.2.17)$$

where $\hat{\mathbf{E}}_\nu(t) = (\zeta_p(z)' \otimes \hat{\varepsilon}_h(t)' \otimes \mathbf{I}_v)\mathbf{S}'_{\lambda(\nu)}$. Thus, $\mathbf{e}_\nu(t) = \mathbf{y}(t) + \mathbf{Y}_\nu(t)\alpha + \hat{\mathbf{F}}_\nu(t)\beta - \hat{\mathbf{E}}_\nu(t)\lambda$ where $\hat{\mathbf{F}}_\nu(t) = \{(\mathbf{y}(t) - \hat{\varepsilon}_h(t))' \otimes \mathbf{I}_v\}\mathbf{S}_{\beta(\nu)}$. The above constructions serve to provide tools for presenting the estimation of the least squares estimator in a straightforward fashion.

Now let $\hat{\mathbf{R}}_\nu(t)' = (-\mathbf{Y}_\nu(t) : -\hat{\mathbf{F}}_\nu(t) : \hat{\mathbf{E}}_\nu(t))$, $t = 1, \ldots, T$ denote the regressor variables. The regressor variables $\mathbf{y}(t) - \hat{\varepsilon}_h(t)$ have been included because it is assumed that the matrix pair $[\mathbf{A}(z) : \mathbf{M}(z)]$ is in (reversed) echelon form, $\mathbf{A}(0) = \mathbf{M}(0)$ and that $\mathbf{A}(0)$ needs not be an identity matrix. Thus, the least squares estimator is obtained by regressing $\mathbf{y}(t)$ on the regressor variables $\hat{\mathbf{R}}_\nu(t)'$ to obtain $\hat{\theta}_{LS}$, the estimate of $\theta = (\alpha' : \beta' : \lambda')'$, which minimizes the criterion function $\sum_{t=1}^{T} \|\mathbf{e}_\nu(t)\|^2 = \sum_{t=1}^{T} \|\mathbf{y}(t) - \hat{\mathbf{R}}_\nu(t)'\theta\|^2$ with respect to the $d(\nu) = d_\alpha + d_\beta + d_\lambda$ freely varying coefficients in $[\mathbf{A}(z): \mathbf{M}(z)]$. d_β and d_λ, as previously defined, are used to denote the dimension of vectors β and λ, respectively.

Following standard procedures, the least squares estimator, $\hat{\theta}_{LS}$, is obtained by solving the normal equations

$$T^{-1}\sum_{t}\hat{\mathbf{R}}_{\nu}(t)\hat{\mathbf{R}}_{\nu}(t)'\theta_{LS} = T^{-1}\sum_{t}\hat{\mathbf{R}}_{\nu}(t)\mathbf{y}(t) \tag{3.2.18}$$

and therefore $\hat{\theta}_{LS}$ is given by

$$\hat{\theta}_{LS} = \hat{X}_{\nu,T}^{-1}\hat{U}_{\nu,T} \tag{3.2.19}$$

where $\hat{X}_{\nu,T} = T^{-1}\sum_{t=1}^{T}\hat{\mathbf{R}}_{\nu}(t)\hat{\mathbf{R}}_{\nu}(t)'$ and $\hat{U}_{\nu,T} = T^{-1}\sum_{t=1}^{T}\hat{\mathbf{R}}_{\nu}(t)\mathbf{y}(t)$. Since by construction, $\hat{X}_{\nu,T}$ is a square matrix of order $d(\nu)$ with elements that are sums of squares and cross-products involving $\mathbf{y}(t)$ and $\hat{\varepsilon}_h(t)$ with finite expectation then $X_{\nu,T}^{-1}$ can be guaranteed to exist. Implicitly, the existence of matrix $X_{\nu,T}^{-1}$ follows from the identifiability conditions (see Chapter 1 (Section (1.2))) imposed on model (3.2.1). Thus (3.2.19) provides unique solution for $\hat{\theta}_{LS}$.

One final remark is necessary in connection with the above construction. The procedure seems to provide a convenient method for adding and deleting regression components, a common feature of the identification stage of the analysis of multivariate time series where it is often the case to estimate a number of alternative models before arriving at the one that best describes the data at hand. This is made possible by observing that the non-zero elements in the i^{th} row of $\hat{\mathbf{R}}_{\nu}(t)'$ constitute the regressor variables for the i^{th} equation. The simplicity of this method stems from the fact that the integer-valued parameters (Kronecker indices) are determined in terms of the individual equations as demonstrated in Procedure B of Chapter 1 and this in turn allows the parameter estimates corresponding to each equation to be evaluated independently of the other. Approximate standard errors can also be obtained from these separate estimations.

3.3 Strong Consistency

In this section the strong consistency property of the least squares estimator, $\hat{\theta}_{LS}$, is investigated. Before considering this however some preliminary notation and often used results will be introduced. For model (3.2.1), $\mathbf{y}(t)$, $\varepsilon(t)$, $t = 0, \pm 1, \ldots$ are v-dimensional, regular, stationary sequence having zero means and finite second moments. Since, under assumption (1.1.4), $\mathbf{y}(t)$ is expressible in terms of the innovation seqence, $\{\varepsilon(t)\}$, as in (1.1.6) with the spectral density matrix, $\mathbf{f}_y(\omega)$, given by (1.1.9). We now recall that the cross-covariances can be recovered from the corresponding spectral densities through the inverse Fourier transform. In particular, for any integer j,

$$\begin{aligned} \mathcal{E}\{\mathbf{y}(t)\mathbf{y}(t+j)'\} &= \mathbf{\Gamma}_{yy}(j) = \int_{-\pi}^{\pi} e^{ij\omega}\mathbf{f}_y(\omega)\, d\omega, \\ \mathcal{E}\{\mathbf{y}(t)\varepsilon(t+j)'\} &= \mathbf{\Gamma}_{y\varepsilon}(j) = \int_{-\pi}^{\pi} e^{ij\omega}\mathbf{K}(e^{i\omega})\mathbf{\Sigma}\, d\omega \end{aligned}$$

and similarly,

$$\mathcal{E}\{\varepsilon(t)\varepsilon(t+j)'\} = \mathbf{\Gamma}_{\varepsilon\varepsilon}(j) = \int_{-\pi}^{\pi} e^{ij\omega}\mathbf{\Sigma}\, d\omega. \tag{3.3.1}$$

For later convenience, it may be helpful to introduce the following cross-covariances:

$$\begin{aligned} \mathbf{G}_{yy}(r,s) &= T^{-1}\sum_{t'}^{T}\mathbf{y}(t-r)\mathbf{y}(t-s)' = \mathbf{G}_{yy}(s,r)', \\ \mathbf{G}_{y\varepsilon}(r,s) &= T^{-1}\sum_{t'}^{T}\mathbf{y}(t-r)\hat{\varepsilon}_h(t-s)' = \mathbf{G}_{y\varepsilon}(s,r)', \quad \textit{and} \\ \mathbf{G}_{\varepsilon\varepsilon}(r,s) &= T^{-1}\sum_{t'}^{T}\hat{\varepsilon}_h(t-r)\hat{\varepsilon}_h(t-s)' = \mathbf{G}_{\varepsilon\varepsilon}(s,r)', \end{aligned} \tag{3.3.2}$$

where $t' = max(r,s) + 1$. One point to note, of course, is that the sample cross-covariances (3.3.2) are not based on Toeplitz assumptions and hence depend on (r,s) but not on the difference $(r-s)$. The difference between the estimates based on Toeplitz calculations and those of (3.3.2) differ in the treatment of certain *end effects* which, as shown below, becomes asymptotically negligible. Thus, we state

the following lemma.

Lemma (3.3.1):

Let $\mathbf{y}(t)$ *and* $\varepsilon(t)$ *be related by model (3.2.1) under assumptions (1.1.4), (1.1.5) and* $\ell(T) \leq h \leq (\log T)^{\tau}$, $\tau < \infty$. *Then as* $T \to \infty$,

(i) $\mathbf{G}_{yy}(r,s) = \mathbf{\Gamma}_{yy}(r-s) + O(Q_T)$ *a.s*

(ii) $\mathbf{G}_{\varepsilon\varepsilon}(r,s) = \mathbf{\Gamma}_{\varepsilon\varepsilon}(r-s) + O(Q_T)$ *a.s*

(iii) $\mathbf{G}_{y\varepsilon}(r,s) = \mathbf{\Gamma}_{y\varepsilon}(r-s) + O(Q_T)$ *a.s.*

Proof: Now consider

$$\mathbf{G}_{yy}(r,s) = T^{-1}\sum_{t=1}^{T}\mathbf{y}(t-r)\mathbf{y}(t-s)' \tag{3.3.3}$$

and recall that

$$\hat{\mathbf{\Gamma}}_{yy}(r-s) = T^{-1}\sum_{t=1}^{T-(r-s)}\mathbf{y}(t)\mathbf{y}(t+r-s)'$$

where from (3.3.2), (3.3.3) is to be interpreted as containing only positively indexed terms of $\mathbf{y}(t)$. However, the difference between (3.3.3) and $\hat{\mathbf{\Gamma}}_{yy}(r-s)$ contains a finite number of terms. By the finiteness of the fourth moment of $\mathbf{y}(t)$ each of these terms is $o(T^{-\frac{1}{2}})$ *a.s* using arguments similar to those employed in Hannan and Deistler (1988, pp. 336-337), and therefore negligible for large T. Thus, it follows that

$$\mathbf{G}_{yy}(r,s) = \hat{\mathbf{\Gamma}}_{yy}(r-s) + o(T^{-\frac{1}{2}}) \quad a.s. \tag{3.3.4}$$

But from Theorem (3.2.3), $\hat{\mathbf{\Gamma}}_{yy}(r-s) = \mathbf{\Gamma}_{yy}(r-s) + O(Q_T)$ and therefore (3.3.4) can be rewritten as $\mathbf{G}_{yy}(r,s) = \mathbf{\Gamma}_{yy}(r-s) + O(Q_T) + o(T^{-\frac{1}{2}})$ *a.s*. The result in (i) follows straightforwardly by noting that the $\max(O(Q_T) + o(T^{-\frac{1}{2}})) = O(Q_T)$.

To establish the second part of the lemma, first observe that $\mathbf{G}_{\varepsilon\varepsilon}(r,s) = \hat{\mathbf{\Gamma}}_{\varepsilon\varepsilon}(r-s) + o(T^{-\frac{1}{2}})$ in much the same way as in (i). Now consider

$$
\begin{aligned}
T^{-1}\sum_{t=1}^{T}\hat{\varepsilon}_h(t-r)\hat{\varepsilon}_h(t-s)' &= T^{-1}\sum_{t=1}^{T}\{\varepsilon(t-r)+(\hat{\varepsilon}_h(t-r)-\varepsilon(t-r))\} \\
&\quad \times \{(\varepsilon(t-s)'+(\hat{\varepsilon}_h(t-s)-\varepsilon(t-s))'\} \\
&= T^{-1}\sum_{t=1}^{T}\varepsilon(t-r)\varepsilon(t-s)' \\
&\quad + T^{-1}\sum_{t=1}^{T}\varepsilon(t-r)(\hat{\varepsilon}_h(t-s)-\varepsilon(t-s))' \\
&\quad + T^{-1}\sum_{t=1}^{T}(\hat{\varepsilon}_h(t-r)-\varepsilon(t-r))\varepsilon(t-s)' \\
&\quad + T^{-1}\sum_{t=1}^{T}(\hat{\varepsilon}_h(t-r)-\varepsilon(t-r)) \\
&\quad \times (\hat{\varepsilon}_h(t-s)-\varepsilon(t-s))'. \qquad (3.3.5)
\end{aligned}
$$

Employing a result in Hannan and Kavalieris (1984, p.550) we find that the remaining terms on the right-hand side of (3.3.5) after the first can be shown to give a contribution that is $O(hQ_T^2)$. We can therefore conclude that $T^{-1}\sum_t \hat{\varepsilon}_h(t-r)\hat{\varepsilon}_h(t-s)' = \delta_{r,s}\mathbf{\Sigma} + O(hQ_T^2)$ $a.s$ since the first term in (3.3.5) for fixed r, s converges to $\delta_{r,s}\mathbf{\Sigma}$ $a.s.$

To conclude the proof of the lemma, it remains to show that

$$\mathbf{G}_{y\varepsilon}(s,r) = \mathbf{\Gamma}_{y\varepsilon}(r-s) + O(Q_T) \quad a.s.$$

To see that this is true let us rewrite $\mathbf{G}_{y\varepsilon}(s,r)$ as $T^{-1}\sum_{t=1}^{T}\mathbf{y}(t-r)\hat{\varepsilon}_h(t-s)'$ and observe that this quantity differs from $\hat{\mathbf{\Gamma}}_{y\varepsilon}(r-s)$ by a number of terms that are, uniformly in r, s, $o(T^{-\frac{1}{2}})$ $a.s.$ Now consider

$$
\begin{aligned}
T^{-1}\sum_{t=1}^{T}\mathbf{y}(t-r)\hat{\varepsilon}_h(t-s)' &= T^{-1}\sum_{t=1}^{T}\mathbf{y}(t-r)\{\hat{\varepsilon}_h(t-s)-\varepsilon(t-s)\}' \\
&\quad + T^{-1}\sum_{t=1}^{T}\mathbf{y}(t-r)\varepsilon(t-s)'. \qquad (3.3.6)
\end{aligned}
$$

For the first term in the right-hand side of (3.3.6) we obtain, using (3.2.1) and (3.2.4),

$$\sum_{j=0}^{h}T^{-1}\sum_{t=1}^{T}\mathbf{y}(t-r)\mathbf{y}(t-s-j)'(\hat{\mathbf{\Phi}}_h(j)-\mathbf{\Phi}_0(j))' - \sum_{j=h+1}^{\infty}T^{-1}\sum_{t=1}^{T}\mathbf{y}(t-r)\mathbf{y}(t-s-j)'\mathbf{\Phi}_0(j)'$$

so that for $\ell(T) \leq h \leq H_T$, the first term in the above expression is $O(Q_T)$ $a.s$ by Lemma (3.2.1) and egordicity. The other term is $o(1)$ $a.s$ since $\mathcal{E}\|\mathbf{y}(t-r)\mathbf{y}(t-s-k)'\| \leq c < \infty$ for some constant c and by virtue of (3.2.9). Hence,

$$\mathbf{G}_{y\varepsilon}(r,s) = \mathbf{\Gamma}_{y\varepsilon}(r-s) + O(Q_T)\ a.s$$

as required since $\mathcal{E}\{T^{-1}\sum_{t=1}^{T}\mathbf{y}(t-r)\varepsilon(t-s)'\} = \mathbf{\Gamma}_{y\varepsilon}(r-s)$. □

In preparation for what is to follow, consider $\hat{X}_{\nu,T} = T^{-1}\sum_{t=1}^{T}\hat{\mathbf{R}}_\nu(t)\hat{\mathbf{R}}_\nu(t)'$ and $\hat{U}_{\nu,T} = T^{-1}\sum_t \hat{\mathbf{R}}_\nu(t)\mathbf{y}(t)$ and note that, by construction, $\hat{X}_{\nu,T}$ contains six unique matrix blocks induced by the mean squares and cross-products of the variables contained in $\hat{\mathbf{R}}_\nu(t)$, that is, $(\mathbf{Y}_\nu(t)' : \hat{\mathbf{F}}_\nu(t)' : \hat{\mathbf{E}}_\nu(t)')'$. Thus, $\hat{X}_{\nu,T}$ contains components of the form

$$\begin{aligned} &T^{-1}\sum_t \mathbf{y}(t-j)\mathbf{y}(t-k)', \qquad T^{-1}\sum_t \mathbf{y}(t-j)\hat{\varepsilon}_h(t-k)', \\ &T^{-1}\sum_t \hat{\varepsilon}_h(t-j)\hat{\varepsilon}_h(t-k)', \qquad T^{-1}\sum_t (\mathbf{y}(t)-\hat{\varepsilon}_h(t))\mathbf{y}(t-j)',\ etc. \\ &\qquad\qquad\qquad\qquad\qquad\qquad for\ j,k = 1,\ldots,p. \end{aligned} \tag{3.3.7}$$

By direct application of Lemma (3.3.1) the first term in (3.3.8) is $\mathbf{\Gamma}_{yy}(k-j) + O(Q_T)$. Analogous results can be obtained for the remaining quantities in (3.3.8). Now let $\mathbf{R}_\nu(t)$ be defined as is $\hat{\mathbf{R}}_\nu(t)$ except that $\hat{\varepsilon}_h(t)$ is replaced by $\varepsilon(t)$ so that

$$\mathbf{R}_\nu(t) = \mathbf{S}_\nu \begin{bmatrix} -\zeta_p(z)\otimes\mathbf{y}(t)\otimes\mathbf{I}_v \\ -(\mathbf{y}(t)-\varepsilon(t))\otimes\mathbf{I}_v \\ \zeta_p(z)\otimes\varepsilon(t)\otimes\mathbf{I}_v \end{bmatrix} = \mathbf{S}_\nu[\mathbf{x}(t)\otimes\mathbf{I}_v]$$

where $\mathbf{S}_\nu = diag(\mathbf{S}_{\alpha(\nu)} : \mathbf{S}_{\beta(\nu)} : \mathbf{S}_{\lambda(\nu)})$ and $\mathbf{x}(t) = (-\zeta_p(z)'\otimes\mathbf{y}(t)' : -(\mathbf{y}(t)-\varepsilon(t))' : \zeta_p(z)'\otimes\varepsilon(t)')'$. As a consequence of Lemma (3.3.1), the replacement of $\hat{\varepsilon}_h(t)$ by $\varepsilon(t)$ in $\hat{\mathbf{R}}_\nu(t)$ is asymptotically negligible. Thus, we have

$$\begin{aligned} \lim_{T\to\infty} T^{-1}\sum_t \mathbf{R}_\nu(t)\mathbf{R}_\nu(t)' &= \lim_{T\to\infty} T^{-1}\sum_t \mathbf{S}_\nu(\mathbf{x}(t)\otimes\mathbf{I}_v)(\mathbf{x}(t)'\otimes\mathbf{I}_v)\mathbf{S}_\nu' \\ &= \lim_{T\to\infty} T^{-1}\sum_t \mathbf{S}_\nu(\mathbf{x}(t)\mathbf{x}(t)'\otimes\mathbf{I}_v)\mathbf{S}_\nu' \\ &= \mathcal{E}\{\mathbf{S}_\nu(\mathbf{x}(t)\mathbf{x}(t)'\otimes\mathbf{I}_v)\mathbf{S}_\nu'\} \\ &= \mathbf{S}_\nu(\mathbf{X}_\nu\otimes\mathbf{I}_v)\mathbf{S}_\nu' \end{aligned} \tag{3.3.8}$$

where $\mathbf{X}_\nu = \mathcal{E}\{\mathbf{x}(t)\mathbf{x}(t)'\}$ and is partitioned conformally with the partition induced by $\hat{X}_{\nu,T}$. A completely analogous result can be obtained for the vector $T^{-1}\sum_t \mathbf{R}_\nu(t)\mathbf{y}(t)$ as

$$\lim_{T\to\infty} T^{-1}\sum_t \mathbf{S}_\nu\{\mathbf{x}(t)\mathbf{y}(t)\otimes\mathbf{I}_v\} = \mathbf{S}_\nu(\mathbf{U}_\nu\otimes\mathbf{I}_v) \tag{3.3.9}$$

where $\mathbf{U}_\nu = \mathcal{E}\{\mathbf{x}(t)\mathbf{y}(t)\}$. The results in (3.3.9) and (3.3.10) are summarized in the following lemma. See Poskitt (1992, Lemma (3.3)) for related results.

Lemma (3.3.2):

Assume the conditions of Lemma (3.3.1) hold. Then

$$\begin{aligned} T^{-1}\sum_t \hat{\mathbf{R}}_\nu(t)\hat{\mathbf{R}}_\nu(t)' &= \mathbf{S}_\nu(\mathbf{X}_\nu\otimes\mathbf{I}_v)\mathbf{S}_\nu' + O(Q_T) \\ T^{-1}\sum_t \hat{\mathbf{R}}_\nu(t)\mathbf{y}(t) &= \mathbf{S}_\nu(\mathbf{U}_\nu\otimes\mathbf{I}_v) + O(Q_T). \end{aligned}$$

□

Some aspects of the behaviour of the least squares estimator are described in the following lemma.

Lemma (3.3.3):

Assume the conditions of Lemma (3.3.1) hold and let $[\mathbf{A}_0(z) : \mathbf{M}_0(z)]$ *denote the true (reversed) echelon form for model (3.2.1) with Kronecker indices* n_i, $i = 1,\ldots,v$. *Then* $\hat{\theta}_{LS} = [\mathbf{A}_0(z) : \mathbf{M}_0(z)] + O(Q_T), a.s,$ *provides a strongly consistent estimator of* θ_0 $([\mathbf{A}_0(z) : \mathbf{M}_0(z)])$.

Proof: Under the assumption that $X_{\nu,T}$ is positive definite, the unique solution of the normal equations (3.2.18) is given by

$$\hat{\theta}_{LS} = [T^{-1}\sum_t \hat{\mathbf{R}}_\nu(t)\hat{\mathbf{R}}_\nu(t)']^{-1}[T^{-1}\sum_t \hat{\mathbf{R}}_\nu(t)\mathbf{y}(t)].$$

By egordicity and Lemma (3.3.2),

$$\lim_{T\to\infty}\hat{\theta}_{LS} = \lim_{T\to\infty}\{T^{-1}\sum_t \hat{\mathbf{R}}_\nu(t)\hat{\mathbf{R}}_\nu(t)')^{-1}\}\{T^{-1}\sum_t \hat{\mathbf{R}}_\nu(t)\mathbf{y}(t)\}$$

$$
\begin{aligned}
&= \{\lim_{T\to\infty} T^{-1}\sum_t \hat{\mathbf{R}}_\nu(t)\hat{\mathbf{R}}_\nu(t)'\}^{-1}\{\lim_{T\to\infty} T^{-1}\sum_t \hat{\mathbf{R}}_\nu(t)\mathbf{y}(t)\} \\
&= \mathcal{E}\{\mathbf{R}_\nu(t)\mathbf{R}_\nu(t)'\}^{-1}\mathcal{E}\{\mathbf{R}_\nu(t)\mathbf{y}(t)\} + O(Q_T) \\
&= \{\mathbf{S}_\nu(\mathbf{X}_\nu \otimes \mathbf{I}_v)\mathbf{S}_\nu'\}^{-1}\{\mathbf{S}_\nu(\mathbf{U}_\nu \otimes \mathbf{I}_v)\} + O(Q_T) \\
&= \theta_0 + O(Q_T)\ a.s.
\end{aligned}
$$

Hence, $\hat{\theta}_{LS}$ is a strongly consistent estimator of θ_0. □

Finally, we may remark that it is necessary to impose conditions (1.1.4), (1.1.5) and (1.2.8) on model (3.2.1) in order to prevent the variance-covariance matrix $\mathbf{X}_\nu$ from being singular. With these conditions we note that except on very pathological situations, a unique solution, $\hat{\theta}_{LS}$, is likely to exist since $\mathbf{X}_\nu$ is positive definite for $\hat{\theta}_{LS} \in N_\delta(\theta_0)$ where $N_\delta(\theta_0) = \{\theta : \|\theta - \theta_0\| < \delta\}$ is some neighbourhood of θ_0 for $\delta > 0$. This conjecture is supported by the results in Kohn (1978) and can be inferred from the work of Dunsmuir and Hannan (1976). In the light of Lemma (3.3.3), an obvious choice for $\theta_G^{(0)}$, the starting value for the Gauss-Newton iteration scheme presented in Section (2.3) of Chapter 2, is the least squares estimator, $\hat{\theta}_{LS}$.

3.4 The Central Limit Theorem

As stated in Lemma (3.3.3), the least squares estimator, $\hat{\theta}_{LS}$, provides a strongly consistent estimate of θ_0. Thus, $\hat{\theta}_{LS} \to \theta_0$ $a.s$ and by implication, $\hat{\theta}_{LS}$ enters some neighbourhood of θ_0. Our primary concern in this section is to derive an asymptotic normality result for $\hat{\theta}_{LS}$. Although the asymptotic properties of the least squares estimator have been investigated in various contexts see, for example, Saikonnen (1986), Hannan and McDougall (1988), and Brockwell and Davis (1991) in the scalar time series models, we have not been able to find such analysis when consideration is given to vector ARMA models in their appropriate echelon canonical forms. It must, however, be mentioned that Spliid (1983) presented a heuristic proof for the case when $\mathbf{A}(0) = \mathbf{M}(0) = \mathbf{I}_v$ but no consideration is given to the

identifiability conditions (Chapter 1 (Section (1.2))) that must be imposed. We now state the main result of this section.

Theorem (3.4.1):

Under the same conditions as Lemmas (3.2.4) and (3.3.3), the vector $\sqrt{T}(\hat{\theta}_{LS} - \theta_0)$ *has an asymptotic normal distribution with zero mean vector and variance-covariance matrix* $\mathbf{\Lambda}_{LS}$ *where*

$$\mathbf{\Lambda}_{LS} = \{\mathbf{S}_\nu(\mathbf{X}_\nu \otimes \mathbf{I}_v)\mathbf{S}_\nu'\}^{-1}\{\mathbf{S}_\nu\mathbf{D}_0(\mathbf{X}_\nu \otimes \mathbf{\Sigma}_0)\mathbf{D}_0'\mathbf{S}_\nu'\}\{\mathbf{S}_\nu(\mathbf{X}_\nu \otimes \mathbf{I}_v)\mathbf{S}_\nu'\}^{-1}$$

and $\mathbf{X}_\nu = \mathcal{E}(\mathbf{x}(t)\mathbf{x}(t)')$ *with* $\mathbf{D}_0$ *being explicitly defined below.*

Proof: By definition, the least squares estimator is obtained as $\hat{\theta}_{LS} = \hat{X}_{\nu,T}^{-1}\hat{U}_{\nu,T}$. Then consider

$$\hat{\theta}_{LS} - \theta_0 = \hat{X}_{\nu,T}^{-1}[\hat{U}_{\nu,T} - \hat{X}_{\nu,T}\theta_0]. \tag{3.4.1}$$

Premultiplying both sides of (3.4.1) by $\hat{X}_{\nu,T}$, we obtain

$$\begin{aligned}
\hat{X}_{\nu,T}(\hat{\theta}_{LS} - \theta_0) &= T^{-1}\sum_{t=1}^{T}\hat{\mathbf{R}}_\nu(t)\,[\mathbf{y}(t) - \hat{\mathbf{R}}_\nu(t)'\theta_0] \\
&= T^{-1}\sum_{t=1}^{T}\hat{\mathbf{R}}_\nu(t)[(\mathbf{y}(t) - \mathbf{R}_\nu(t)'\theta_0) - (\hat{\mathbf{R}}_\nu(t) - \mathbf{R}_\nu(t))'\theta_0] \\
&= T^{-1}\sum_{t=1}^{T}\hat{\mathbf{R}}_\nu(t)[\varepsilon(t) + (\mathbf{R}_\nu(t) - \hat{\mathbf{R}}_\nu(t))'\theta_0] \\
&= T^{-1}\sum_{t=1}^{T}\hat{\mathbf{R}}_\nu(t)\{\varepsilon(t) + (\mathbf{M}_0(z) - \mathbf{I}_v)(\varepsilon(t) - \hat{\varepsilon}_h(t))\}
\end{aligned}$$

where, by definition,

$$\begin{aligned}
(\mathbf{R}_\nu(t) - \hat{\mathbf{R}}_\nu(t))'\theta_0 &= [-\mathbf{Y}_\nu(t) : -\mathbf{F}_\nu(t) : \mathbf{E}_\nu(t) - (-\mathbf{Y}_\nu(t) : -\hat{\mathbf{F}}_\nu(t) : \hat{\mathbf{E}}_\nu(t))]\theta_0 \\
&= [(\mathbf{0} \otimes \mathbf{I}_v)\mathbf{S}_{\alpha(\nu)}' : \{(\varepsilon(t) - \hat{\varepsilon}_h(t))' \otimes \mathbf{I}_v\}\mathbf{S}_{\beta(\nu)}' : \{\zeta_p(z)' \\
&\qquad \otimes (\varepsilon(t) - \hat{\varepsilon}_h(t))' \otimes \mathbf{I}_v\}\ \mathbf{S}_{\lambda(\nu)}'](\alpha_0' : \beta_0' : \lambda_0')' \\
&= \{(\varepsilon(t) - \hat{\varepsilon}_h(t))' \otimes \mathbf{I}_v\}\mathbf{S}_{\beta(\nu)}'\beta_0 + \{\zeta_p(z)' \otimes (\varepsilon(t) - \hat{\varepsilon}_h(t))' \\
&\qquad \otimes \mathbf{I}_v\}\mathbf{S}_{\lambda(\nu)}'\lambda_0 \\
&= vec\{(\mathbf{M}_0(0) - \mathbf{I}_v)(\varepsilon(t) - \hat{\varepsilon}_h(t))\}
\end{aligned}$$

$$+ \; vec\{(\mathbf{M}_0(z) - \mathbf{M}_0(0))(\varepsilon(t) - \hat{\varepsilon}_h(t))\}$$

$$= \; (\mathbf{M}_0(z) - \mathbf{I}_v)(\varepsilon(t) - \hat{\varepsilon}_h(t)),$$

the symbol $\mathbf{0}$ is used in the above expression to denote a $(1 \times pv)$ null vector. Thus,

$$\hat{X}_{\nu,T}\{\sqrt{T}(\hat{\theta}_{LS} - \theta_0)\} = T^{-\frac{1}{2}} \sum_t \hat{\mathbf{R}}_\nu(t)[\varepsilon(t) + (\mathbf{M}_0(z) - \mathbf{I}_v)(\varepsilon(t) - \hat{\varepsilon}_h(t))] \quad (3.4.2)$$

and the right-hand side of (3.4.2) can be decomposed into three constituent parts, namely:

$$-T^{-\frac{1}{2}} \sum_t \mathbf{Y}_\nu(t)'\{\varepsilon(t) + (\mathbf{M}_0(z) - \mathbf{I}_v)(\varepsilon(t) - \hat{\varepsilon}_h(t))\}, \quad (3.4.3)$$

$$-T^{-\frac{1}{2}} \sum_t \hat{\mathbf{F}}_\nu(t)'\{\varepsilon(t) + (\mathbf{M}_0(z) - \mathbf{I}_v)(\varepsilon(t) - \hat{\varepsilon}_h(t))\} \quad (3.4.4)$$

and

$$T^{-\frac{1}{2}} \sum_t \hat{\mathbf{E}}_\nu(t)'\{\varepsilon(t) + (\mathbf{M}_0(z) - \mathbf{I}_v)(\varepsilon(t) - \hat{\varepsilon}_h(t))\}. \quad (3.4.5)$$

Consider (3.4.3) and observe that it can be rewritten as

$$\mathbf{S}_{\alpha(\nu)}[-T^{-\frac{1}{2}} \sum_t (\zeta_p(z) \otimes \mathbf{y}(t) \otimes \mathbf{I}_v)\{\varepsilon(t) + (\mathbf{M}_0(z) - \mathbf{I}_v)(\varepsilon(t) - \hat{\varepsilon}_h(t))\}]. \quad (3.4.6)$$

Note that the innovation sequence is expressible in terms of the process $\mathbf{y}(t)$ as $\varepsilon(t) = \sum_{j\geq 0} \mathbf{\Phi}_0(j)\mathbf{y}(t-j)$ where $\mathbf{\Phi}_0(z) = \mathbf{K}_0^{-1}(z)$. Hence, ignoring the selection matrix for the moment and setting $\hat{\mathbf{\Phi}}_h(j) = 0$, $j \geq h+1$, the second term in (3.4.6) can be reexpressed as

$$-T^{-\frac{1}{2}} \sum_t \{(\mathbf{y}(t-1)', \ldots, \mathbf{y}(t-p)')' \otimes (\mathbf{M}_0(z) - \mathbf{I}_v) \sum_{j\geq 0} (\mathbf{\Phi}_0(j) - \hat{\mathbf{\Phi}}_h(j))\mathbf{y}(t-j)\} \quad (3.4.7)$$

since $(\zeta_p(z) \otimes \mathbf{y}(t)) = (\mathbf{y}(t-1)', \ldots, \mathbf{y}(t-p)')'$ and $(\mathbf{M}_0(z) - \mathbf{I}_v)(\varepsilon(t) - \hat{\varepsilon}_h(t)) = 1 \otimes (\mathbf{M}_0(z) - \mathbf{I}_v) \sum_{j\geq 0} (\mathbf{\Phi}_0(j) - \hat{\mathbf{\Phi}}_h(j))\mathbf{y}(t-j)$. Clearly (3.4.7) has p blocks, each of v^2 elements and its typical k^{th} block can be written as

$$-T^{-\frac{1}{2}} \sum_t \{(y_1(t-k), \ldots, y_v(t-k))' \otimes (\mathbf{M}_0(z) - \mathbf{I}_v) \sum_{j\geq 0} (\mathbf{\Phi}_0(j) - \hat{\mathbf{\Phi}}_h(j))\mathbf{y}(t-j)\},$$
$$k = 1, \ldots, p, \quad (3.4.8)$$

where the i^{th} component of this block, which consists of v elements, is of the form

$$\begin{aligned} -T^{-\frac{1}{2}} \sum_t \{\sum_{r=0}^{p}(\mathbf{M}_0(r) - \delta_{0r}\mathbf{I}_v)\sum_{j\geq 0}(\mathbf{\Phi}_0(j) - \hat{\mathbf{\Phi}}_h(j))\mathbf{y}(t-r-j)y_i(t-k)\} \\ = -T^{\frac{1}{2}}\sum_{r=0}^{p}(\mathbf{M}_0(r) - \delta_{0r}\mathbf{I}_v)\{\sum_{j\geq 0}(\mathbf{\Phi}_0(j) - \hat{\mathbf{\Phi}}_h(j))g_i(r+j,k)\}, \\ i = 1,\ldots,v \end{aligned} \tag{3.4.9}$$

where

$$\begin{aligned} g_i(r+j,k) &= T^{-1}\sum_t \mathbf{y}(t-r-j)y_i(t-k) \\ &= (g_{1i}(r+j,k),\ldots,g_{vi}(r+j,k))' \end{aligned}$$

Hence, for all $i = 1,\ldots,v$, (3.4.9) gives a $(v^2 \times 1)$ matrix

$$-T^{\frac{1}{2}}\begin{pmatrix} \sum_{r=0}^{p}(\mathbf{M}_0(r) - \delta_{0r}\mathbf{I}_v)\sum_{j\geq 0}(\mathbf{\Phi}_0(j) - \hat{\mathbf{\Phi}}_h(j))g_1(r+j,k) \\ \sum_{r=0}^{p}(\mathbf{M}_0(r) - \delta_{0r}\mathbf{I}_v)\sum_{j\geq 0}(\mathbf{\Phi}_0(j) - \hat{\mathbf{\Phi}}_h(j))g_2(r+j,k) \\ \vdots \qquad\qquad \vdots \qquad\qquad \vdots \\ \sum_{r=0}^{p}(\mathbf{M}_0(r) - \delta_{0r})\mathbf{I}_v\sum_{j\geq 0}(\mathbf{\Phi}_0(j) - \hat{\mathbf{\Phi}}_h(j))g_v(r+j,k) \end{pmatrix}$$

which, as readily verified, can be written more compactly as

$$\begin{aligned} [-T^{\frac{1}{2}}\{\mathbf{I}_v \otimes \sum_{r=0}^{p}(\mathbf{M}_0(r) - \delta_{0r}\mathbf{I}_v)\sum_{j\geq 0}(\mathbf{\Phi}_0(j) - \hat{\mathbf{\Phi}}_h(j))\}vec\,\mathbf{G}(r+j,k)], \\ k = 1,2,\ldots,p \end{aligned} \tag{3.4.10}$$

where

$$\mathbf{G}(r+j,k) = (g_1(r+j,k),\ldots,g_v(r+j,k)).$$

Hence, on application of the rule $vec\,\mathbf{ABC} = (\mathbf{C}' \otimes \mathbf{A})vec\,\mathbf{B}$, expression (3.4.10) can be rewritten as

$$\begin{aligned} -T^{\frac{1}{2}}vec\{\sum_{r=0}^{p}(\mathbf{M}_0(r) - \delta_{0r}\mathbf{I}_v)[\sum_{j=0}^{h}(\mathbf{\Phi}_0(j) - \hat{\mathbf{\Phi}}_h(j))\mathbf{G}(r+j,k) \\ + \sum_{j=h+1}^{\infty}\mathbf{\Phi}_0(j)\mathbf{G}(r+j,k)]\}, k = 1,\ldots,p. \end{aligned} \tag{3.4.11}$$

Putting $\mathbf{\Phi}_0(j) - \hat{\mathbf{\Phi}}_h(j) = \mathbf{\Phi}_h(j) - \hat{\mathbf{\Phi}}_h(j) + \mathbf{\Phi}_0(j) - \mathbf{\Phi}_h(j)$, the first term of (3.4.11) may be written as

$$vec\{-T^{\frac{1}{2}} \sum_{r=0}^{p} (\mathbf{M}_0(r) - \delta_{0r}\mathbf{I}_v) \sum_{j=0}^{h} [(\mathbf{\Phi}_h(j) - \hat{\mathbf{\Phi}}_h(j))\mathbf{G}(r+j,k) + (\mathbf{\Phi}_0(j) - \mathbf{\Phi}_h(j))\,\mathbf{G}(r+j,k)]\} \tag{3.4.12}$$

where the quantities $\mathbf{\Phi}_h(j)$ are the coefficients of the best linear predictor of $\mathbf{y}(t)$ from $\mathbf{y}(t-j)$, $j = 1, \ldots, h$. Since by stationarity, $\|\mathbf{\Gamma}_{yy}(r-k+j)\| \leq \|\mathbf{\Gamma}_{yy}(0)\|$, then it follows that $\|\mathbf{\Gamma}_{yy}(r-k+j)\|$ is uniformly bounded and employing Theorem (3.2.3) it can be shown that $\|\mathbf{G}(r+j,k)\|$ is similarly bounded. Combining the second term of (3.4.11) with (3.4.12) and employing the multivariate generalization of Baxter's inequality (Theorem (3.2.1)) expression (3.4.10) is thereby reduced to

$$vec\{-T^{\frac{1}{2}} \sum_{r=0}^{p} (\mathbf{M}_0(r) - \delta_{0r}\mathbf{I}_v)[\sum_{j=0}^{h} (\mathbf{\Phi}_h(j) - \hat{\mathbf{\Phi}}_h(j))\mathbf{G}(r+j,k)]\} \tag{3.4.13}$$

plus a term that is bounded by

$$T^{\frac{1}{2}} \cdot const \cdot \sum_{j \geq h+1}^{\infty} \|\mathbf{\Phi}_0(j)\|. \tag{3.4.14}$$

A partial fraction expansion of $\{\det \mathbf{M}(z)\}^{-1}$ shows, however, that $\|\mathbf{\Phi}_0(j)\| \leq c_0 \rho_0^j$ for some constant c_0 and it follows that (3.4.14) is dominated by $const \cdot \rho_0^h$. For large T we require $h \geq b \log T$ for $b > \frac{1}{-2\log \rho}$ and indeed, we may set h equal to the integer part of $b \log T$, b sufficiently large, such that $\rho_0^h = exp(h \log \rho_0) = exp\{-d \log T\}$. We therefore conclude that (3.4.14) is $o(1)$ $a.s$ as long as $d > \frac{1}{2}$. Now setting $\mathbf{\Phi}_h = (\mathbf{\Phi}_h(1), \ldots, \mathbf{\Phi}_h(h))$ and replacing $\mathbf{G}(r+j,k)$ by $T^{-1}\sum_t \mathbf{y}(t-r-j)\mathbf{y}(t-k)'$, we may rewrite (3.4.13) as

$$vec\{-\sum_{r=0}^{p} (\mathbf{M}_0(r) - \delta_{0r}\mathbf{I}_v)[T^{-\frac{1}{2}} \sum_{t} (\hat{\mathbf{\Phi}}_h - \mathbf{\Phi}_h)(\mathcal{Y}_h(t-r)\mathbf{y}(t-k)')]\} \tag{3.4.15}$$

where $\mathcal{Y}_h(t) = (\zeta_h(z) \otimes \mathbf{y}(t))$ and using an obvious notation, $\zeta_h(z) = (z, \ldots, z^h)'$. Define $\mathbf{\Gamma}_h = \mathcal{E}\{\mathcal{Y}_h(t)\mathcal{Y}_h(t)'\}$, $\mathbf{\Xi}_h = \mathcal{E}\{\mathcal{Y}_h(t)\mathbf{y}(t)'\}$, $\mathbf{\Delta}_h(r,k) = \mathcal{E}[\mathcal{Y}_h(t-r)\mathbf{y}(t-k)'] = (\mathbf{\Gamma}_{yy}(u+1)', \ldots, \mathbf{\Gamma}_{yy}(u+h)')'$, $u = r-k$ and note that $\mathbf{\Phi}_h$ satisfies the equation

$-\mathbf{\Phi}_h\mathbf{\Gamma}_h = \mathbf{\Xi}_h'$.

In order to proceed, we will require the following lemma.

Lemma (3.4.1):

Assume $\mathbf{\Phi}_h$ satisfies an equation of the form $-\mathbf{\Phi}_h\mathbf{\Gamma}_h = \mathbf{\Xi}_h'$. Then

$$\hat{\mathbf{\Phi}}_h - \mathbf{\Phi}_h = \{-T^{-1}\sum_t \varepsilon(t)\mathcal{Y}_h(t)'\}\hat{\mathbf{\Gamma}}_h^{-1} + o_p(1)$$

where, by construction, $\hat{\mathbf{\Gamma}}_h = T^{-1}\sum_t \mathcal{Y}_h(t)\mathcal{Y}_h(t)'$ is not a Toeplitz matrix.

Proof: Consider $-(\hat{\mathbf{\Phi}}_h - \mathbf{\Phi}_h)\hat{\mathbf{\Gamma}}_h = \hat{\mathbf{\Xi}}_h' + \mathbf{\Phi}_h\hat{\mathbf{\Gamma}}_h$. Then

$$\begin{aligned}
-(\hat{\mathbf{\Phi}}_h - \mathbf{\Phi}_h) &= \hat{\mathbf{\Xi}}_h'\hat{\mathbf{\Gamma}}_h^{-1} + \mathbf{\Phi}_h = \hat{\mathbf{\Xi}}_h'\hat{\mathbf{\Gamma}}_h^{-1} + \mathbf{\Phi}_h\hat{\mathbf{\Gamma}}_h\hat{\mathbf{\Gamma}}_h^{-1} \\
&= (\hat{\mathbf{\Xi}}_h' + \mathbf{\Phi}_h\hat{\mathbf{\Gamma}}_h)\hat{\mathbf{\Gamma}}_h^{-1} \\
&= T^{-1}[\sum_t\{\mathbf{y}(t)\mathcal{Y}_h(t)' + \mathbf{\Phi}_h\sum_t \mathcal{Y}_h(t)\mathcal{Y}_h(t)'\}]\hat{\mathbf{\Gamma}}_h^{-1} \\
&= \{T^{-1}\sum_t(\mathbf{y}(t) + \mathbf{\Phi}_h\mathcal{Y}_h(t))\,\mathcal{Y}_h(t)'\}\hat{\mathbf{\Gamma}}_h^{-1} \\
&= \{T^{-1}\sum_t(\varepsilon(t) + \varepsilon_h(t) - \varepsilon(t))\mathcal{Y}_h(t)'\}\hat{\mathbf{\Gamma}}_h^{-1} \\
&= \{T^{-1}\sum_t \varepsilon(t)\mathcal{Y}_h(t)'\}\hat{\mathbf{\Gamma}}_h^{-1} + \{T^{-1}\sum_t(\varepsilon_h(t) - \varepsilon(t))\mathcal{Y}_h(t)'\}\hat{\mathbf{\Gamma}}_h^{-1}
\end{aligned} \tag{3.4.16}$$

The second term on the right of (3.4.16) can be shown to be $o_p(1)$ using Lemma (3.2.4) and the fact that $\|\hat{\mathbf{\Gamma}}_h^{-1}\|$ is bounded by virtue of Lemma (3.2.3). Thus,

$$\hat{\mathbf{\Phi}}_h - \mathbf{\Phi}_h = \{-T^{-1}\sum_t \varepsilon(t)\mathcal{Y}_h(t)'\}\hat{\mathbf{\Gamma}}_h^{-1} + o_p(1).$$

□

Substituting the last expression into (3.4.15) and expanding we find that (3.4.10) can now be written as

$$\begin{aligned}
&vec\ \{T^{-\frac{1}{2}}\sum_{r=0}^{p}(\mathbf{M}_0(r) - \delta_{0r}\mathbf{I}_v)(T^{-1}\sum_t \varepsilon(t)\mathcal{Y}_h(t)')\hat{\mathbf{\Gamma}}_h^{-1}\sum_t \mathcal{Y}_h(t-r)\mathbf{y}(t-k)'\} + o_p(1) \\
&= vec\{T^{-\frac{1}{2}}\sum_{r=0}^{p}(\mathbf{M}_0(r) - \delta_{0r}\mathbf{I}_v)[\sum_t \varepsilon(t)\mathcal{Y}_h(t)'(\hat{\mathbf{\Gamma}}_h^{-1}\hat{\Delta}_h(r,k))]\} + o_p(1)
\end{aligned} \tag{3.4.17}$$

where $\hat{\mathbf{\Delta}}_h(r,k) = T^{-1}\sum_t \mathcal{Y}_h(t-r)\mathbf{y}(t-k)' = (\mathbf{G}(r+1,k)',\ldots,\mathbf{G}(r+h,k)')'$. We now replace $\hat{\mathbf{\Gamma}}_h^{-1}\hat{\mathbf{\Delta}}_h(r,k)$ in (3.4.17) by $\mathbf{\Gamma}_h^{-1}\mathbf{\Delta}_h(r-k)$ and show that such a replacement makes no difference asymptotically. Such a replacement also allows further simplifications to be made. Hence, we state the following lemmas, the second parts of which are required for subsequent derivations.

Lemma (3.4.2):

Assume condition A holds. Then for $h \leq (\log T)^\tau$, $\tau < \infty$, *uniformly in* h,

(i) $\|\hat{\mathbf{\Gamma}}_h^{-1}\hat{\mathbf{\Delta}}_h(r,k) - \mathbf{\Gamma}_h^{-1}\mathbf{\Delta}_h(r-k)\| = O(h^2 Q_T)$ *and*

(ii) $\|\hat{\mathbf{\Gamma}}_h^{-1}\hat{\gamma}_h(r,k) - \mathbf{\Gamma}_h^{-1}\gamma_h(r-k)\| = O(h^2 Q_T)$

where $\gamma_h(r-k) = \mathcal{E}(\mathcal{Y}_h(t-r)\varepsilon(t-k)')$ *and* $\hat{\gamma}_h(r,k) = T^{-1}\sum_t \mathcal{Y}_h(t-r)\varepsilon(t-k)'$, $r = 0,1,\ldots,p;\quad k = 1,\ldots,p.$

Proof: Consider

$$\begin{aligned}
\hat{\mathbf{\Gamma}}_h^{-1}\hat{\mathbf{\Delta}}_h(r,k) - \mathbf{\Gamma}_h^{-1}\mathbf{\Delta}_h(r-k) &= (\hat{\mathbf{\Gamma}}_h^{-1}\hat{\mathbf{\Delta}}_h(r,k) - \hat{\mathbf{\Gamma}}_h^{-1}\mathbf{\Delta}_h(r-k)) \\
&\quad + (\hat{\mathbf{\Gamma}}_h^{-1}\mathbf{\Delta}_h(r-k) - \mathbf{\Gamma}_h^{-1}\mathbf{\Delta}_h(r-k)) \\
&= \hat{\mathbf{\Gamma}}_h^{-1}(\hat{\mathbf{\Delta}}_h(r-k) - \mathbf{\Delta}_h(r-k)) \\
&\quad + (\hat{\mathbf{\Gamma}}_h^{-1}\mathbf{\Gamma}_h\mathbf{\Gamma}_h^{-1} - \hat{\mathbf{\Gamma}}_h^{-1}\hat{\mathbf{\Gamma}}_h\mathbf{\Gamma}_h^{-1})\mathbf{\Delta}_h(r-k) \\
&= \hat{\mathbf{\Gamma}}_h^{-1}\{\hat{\mathbf{\Delta}}_h(r,k) - \mathbf{\Delta}_h(r-k)\} \\
&\quad + \{\hat{\mathbf{\Gamma}}_h^{-1}(\mathbf{\Gamma}_h - \hat{\mathbf{\Gamma}}_h)\mathbf{\Gamma}_h^{-1}\}\mathbf{\Delta}_h(r-k). \quad (3.4.18)
\end{aligned}$$

As already indicated $\|\hat{\mathbf{\Gamma}}_h^{-1}\|$ is bounded and, uniformly in r and k, $\|\mathbf{G}(r,k) - \mathbf{\Gamma}(r-k)\| = O(Q_T)$ (Lemma (3.3.1 (i))) where $\mathbf{G}(r,k)$ and $\mathbf{\Gamma}_{yy}(r-k)$ respectively denote $(v \times v)$ subblocks of $\hat{\mathbf{\Gamma}}_h$ and $\mathbf{\Gamma}_h$. It follows from this result that $\|\hat{\mathbf{\Delta}}_h(r,k) - \mathbf{\Delta}_h(r-k)\|$ and $\|\hat{\mathbf{\Gamma}}_h - \mathbf{\Gamma}_h\|$ are respectively of $O(h^2 Q_T)$ since each of the normed quantities is composed of elements of the form $\mathbf{G}(r,k) - \mathbf{\Gamma}_{yy}(r-k)$. It is immediate that (3.4.18) is $O(h^2 Q_T)$ and hence the conclusion in (i). Replacing $\mathbf{\Delta}_h(r-k)$ by $\gamma_h(r-k)$ in (3.4.18) and using arguments similar to those just employed we are also led to the

statement given in (ii). □

Lemma (3.4.3):

Assume Condition A holds. Then if $h \to \infty$ *with* T *such that* $h/T \to 0$ *then*

(i) $T^{-\frac{1}{2}} \sum_t \varepsilon(t)\mathcal{Y}_h(t)'(\mathbf{\Gamma}_h^{-1}\mathbf{\Delta}_h(r-k)) = T^{-\frac{1}{2}} \sum_t \varepsilon(t)\mathbf{y}(t+r-k|t-1)' + o_p(1)$

(ii) $T^{-\frac{1}{2}} \sum_t \varepsilon(t)\mathcal{Y}_h(t)'\mathbf{\Gamma}_h^{-1}\gamma_h(r-k) = T^{-\frac{1}{2}} \sum_t \varepsilon(t)\varepsilon(t+r-k \mid t-1)' + o_p(1)$

where $\mathbf{y}(t+r-k \mid t-1) = \mathcal{E}\{\mathbf{y}(t+r-k \mid \mathcal{F}_{t-1}\}$ *and* $\varepsilon(t+r-k \mid t-1) = \mathcal{E}\{\varepsilon(t+r-k \mid \mathcal{F}_{t-1}\}, \quad r = 0, 1, \ldots, p; \quad k = 1, \ldots, p.$

Proof: Set $\ell = r - k$ and consider (i). Let $\mathbf{y}_h(t+\ell \mid t-1) = -\sum_{j=1}^{h} \mathbf{\Phi}_{\ell,h}(j)\mathbf{y}(t+\ell-j)$ denote the best linear predictor of $\mathbf{y}(t+\ell)$ based on $\mathcal{Y}_h(t)$ and similarly, let $\mathbf{y}(t+\ell \mid t-1) = -\sum_{j=1}^{\infty} \mathbf{\Phi}_{0,\ell}(j)\mathbf{y}(t+\ell-j)$ denote the projection of $\mathbf{y}(t+\ell)$ onto the space spanned by $\mathbf{y}(t-j), j = 1, \ldots, \infty$. First suppose $\ell < 0$ and note that $|\ell| < h$. Since $\mathbf{y}(t+\ell)$ belongs to the space spanned by $\mathcal{Y}_h(t)$, then it is clear that $\mathbf{y}_h(t+\ell \mid t-1) = \mathbf{y}(t+\ell)$ because $\mathbf{y}_h(t+\ell \mid t-1) = \mathcal{Y}_h(t)'\mathbf{\Gamma}_h^{-1}\mathbf{\Delta}_h$ denotes the projection of $\mathbf{y}(t+\ell)$ onto this space. Similarly, it is also clear that $\mathbf{y}(t+\ell \mid t-1) = \mathbf{y}(t+\ell)$. It then follows that $\mathbf{y}_h(t+\ell \mid t-1) - \mathbf{y}(t+\ell \mid t-1)$ is zero. Hence, (i) holds for $\ell < 0$. We now consider the case $\ell \geq 0$. Let the prediction errors associated with $\mathbf{y}_h(t+\ell \mid t-1)$ and $\mathbf{y}(t+\ell \mid t-1)$ be defined as $\mathbf{e}_h(t+\ell) = \mathbf{y}(t+\ell) - \mathbf{y}_h(t+\ell \mid t-1)$ and $\mathbf{e}(t+\ell) = \mathbf{y}(t+\ell) - \mathbf{y}(t+\ell \mid t-1)$ respectively. We will show that $\|T^{-\frac{1}{2}} \sum_t \varepsilon(t)(\mathbf{e}_h(t+\ell) - \mathbf{e}(t+\ell))'\| \to 0$ in probability. Set $\mathbf{U}_t = \sum_t \varepsilon(t)\eta_{t,\ell}'$ where $\eta_{t,\ell} = (\mathbf{e}_h(t+\ell) - \mathbf{e}(t+\ell))$. The error process $\varepsilon(t)$ is, however, uncorrelated with the lagged values of $\mathbf{y}(t)$ and hence $\varepsilon(t)$ and $\eta_{t,\ell}$ are uncorrelated. Then

$$\begin{aligned}
\mathcal{E}[\|\mathbf{U}_{t,\ell}\|] &= \mathcal{E}[\|T^{-\frac{1}{2}} \sum_{t=1}^{T} \varepsilon(t)\eta_{t,\ell}'\|] \\
&\leq T^{\frac{1}{2}}\{\mathcal{E}[\|\varepsilon(t)\|]^2 \cdot \mathcal{E}[\|\eta_{t,\ell}'\|]^2\}^{\frac{1}{2}} \\
&= T^{\frac{1}{2}}\{tr(\Sigma)\}^{\frac{1}{2}} \cdot \{\mathcal{E}\| \sum_{j=1}^{\infty} (\mathbf{\Phi}_{\ell,h}(j) - \mathbf{\Phi}_{0,\ell}(j))\mathbf{y}(t+\ell-j)\|^2\}^{\frac{1}{2}}
\end{aligned}$$

$$\begin{aligned}
&\leq T^{\frac{1}{2}}\{tr(\Sigma)\}^{\frac{1}{2}} \cdot \{(\sum_{i=1}^{\infty}\sum_{j=1}^{\infty} \|(\mathbf{\Phi}_{\ell,h}(i) - \mathbf{\Phi}_{0,\ell}(i)\| \cdot \|\mathbf{\Phi}_{\ell,h}(j) - \mathbf{\Phi}_{0,\ell}(j)\| \\
&\qquad \cdot \|\mathbf{\Gamma}_{yy}(i-j)\|)^2\}^{\frac{1}{2}} \\
&\leq T^{\frac{1}{2}}\{tr(\Sigma)\}^{\frac{1}{2}} \cdot \|\mathbf{\Gamma}_{yy}(0)\| \{(\sum_{j=1}^{\infty} \|\mathbf{\Phi}_{\ell,h}(j) - \mathbf{\Phi}_{0,\ell}(j)\|)^2\}^{\frac{1}{2}} \\
&\leq T^{\frac{1}{2}} \cdot const \cdot \sum_{j=h+1}^{\infty} \|\mathbf{\Phi}_{0,\ell}(j)\|
\end{aligned}$$

noting that the penultimate line follows from its predecessor via the fact that the $\|\mathbf{\Gamma}_{yy}(i-j)\|$ are uniformly bounded and the last line follows from Baxter's inequality (Theorem 3.2.1). Again, from Findley (1992), it can be shown that $\sum_{j=h+1}^{\infty} \|\mathbf{\Phi}_{0,\ell}(j)\| = O(\rho_0^h)$, $0 < \rho_0 < 1$, which is $o(1)$ $a.s$ by the same argument that provides a bound for (3.4.14). Hence, (i) also holds for $\ell \geq 0$.

In order to establish the second part of the lemma, first observe that $\mathbf{\Gamma}_h^{-1}\gamma_h(\ell)$ determines the regression parameters of the projection of $\varepsilon(t+\ell)$ onto $\mathcal{Y}_h(t)$, $\varepsilon_h(t+\ell \mid t-1)$. For all $h \geq 1$ these coefficients provide the solution of the Wiener-Hopf equations

$$\int_{-\pi}^{\pi} e^{ik\omega}[e^{ih\omega} - \mathbf{\Psi}_h(e^{i\omega})\mathbf{K}_0(e^{i\omega})]\mathbf{\Sigma}_0\mathbf{K}_0(e^{i\omega})^* \, d\omega \;=\; 0, \quad k = 1, \ldots, h,$$

where

$$\mathbf{\Psi}_h(z) = \sum_{j=0}^{h} \mathbf{\Psi}_h(j)\, z^j, \quad \mathbf{\Psi}_h(0) = \mathbf{I}_v,$$

denotes the operator composed from the regression coefficients of the projection. The application of Baxter's inequality to these equations yields a proof of (ii) that runs parallel to that of (i). In particular; when $\ell < 0$ we find that $\|T^{-\frac{1}{2}}\sum_t \varepsilon(t)\{\varepsilon_h(t+\ell \mid t-1) - \varepsilon(t+\ell \mid t-1)\}'\|$ converges in probability to zero, as above; when $\ell > 0$, $\varepsilon_h(t+\ell \mid t-1) = \varepsilon(t+\ell \mid t-1) = 0$ because $\varepsilon(t+\ell)$ is orthogonal to $\mathcal{Y}_h(t)$ for all $h \geq 1$ and (ii) holds trivially. □

Placing Lemmas 3.4.2(i) and 3.4.3(i) together we finally arrive at the following expression for (3.4.10),

$$\begin{aligned}
vec\{T^{-\frac{1}{2}} \sum_t \sum_{r=0}^{p} (\mathbf{M}_0(r) - \delta_{0r}\mathbf{I}_v)\, \varepsilon(t)\, \mathbf{y}(t+r-k|t-1)'\}, \\
+ \; o_p(1), \quad k = 1, \ldots, p. \qquad (3.4.19)
\end{aligned}$$

Thus, the combination of (3.4.19) with the first term of (3.4.6) gives an equivalent expression for (3.4.3) as

$$
\begin{aligned}
& vec\{T^{-\frac{1}{2}}\sum_t(\sum_{r=0}^{p}\mathbf{M}_0(r)\varepsilon(t)\mathbf{y}(t+r-k|t-1)')\} \;+\; o_p(1) \quad k=1,\ldots,p \\
&= T^{-\frac{1}{2}}\sum_t\{\sum_{r=0}^{p}\mathbf{y}(t+r-k\mid t-1)\otimes\mathbf{M}_0(r)\}vec\{\varepsilon(t)\} \;+\; o_p(1), \quad k=1,\ldots,p \\
&= T^{-\frac{1}{2}}\sum_t\{\sum_{r=0}^{p}(\mathbf{y}(t+r-k\mid t-1)\otimes\mathbf{M}_0(r)\}\,\varepsilon(t) + \quad o_p(1), \qquad (3.4.20) \\
&\qquad\qquad k=1,\ldots,p
\end{aligned}
$$

and the last line follows from the fact that a column vector is its own vectorization.

We now examine expression (3.4.5),

$$
\begin{aligned}
& T^{-\frac{1}{2}}\sum_t\hat{\mathbf{E}}_\nu(t)'\{\varepsilon(t)+(\mathbf{M}_0(z)-\mathbf{I}_v)(\varepsilon(t)-\hat{\varepsilon}_h(t))\} = \\
& \quad T^{-\frac{1}{2}}\sum_t\mathbf{E}_\nu(t)'\{\varepsilon(t)+(\mathbf{M}_0(z)-\mathbf{I}_v)(\varepsilon(t)-\hat{\varepsilon}_h(t))\} \\
& \quad - T^{-\frac{1}{2}}\sum_t(\mathbf{E}_\nu(t)-\hat{\mathbf{E}}_\nu(t))'\{\varepsilon(t)+(\mathbf{M}_0(z)-\mathbf{I}_v)(\varepsilon(t)-\hat{\varepsilon}_h(t))\}. \qquad (3.4.21)
\end{aligned}
$$

Neglect the selection matrix $\mathbf{S}_{\lambda(\nu)}$ for the moment and consider the k^{th} block of the second term on the right of (3.4.21),

$$
\begin{aligned}
& -T^{-\frac{1}{2}}\sum_t\{(\varepsilon(t-k)-\hat{\varepsilon}_h(t-k))\otimes\varepsilon(t) \\
& \quad -T^{-\frac{1}{2}}\sum_t(\varepsilon(t-k)-\hat{\varepsilon}_h(t-k))\otimes(\mathbf{M}_0(z)-\mathbf{I}_v)(\varepsilon(t)-\hat{\varepsilon}_h(t))\}, \\
& \qquad k=1,\ldots,p. \qquad (3.4.22)
\end{aligned}
$$

By direct application of Lemma (3.2.4), it is clear that the i^{th} component of (3.4.22) is either $O(h/T^{\frac{1}{2}})$ or $o_p(h/T^{\frac{1}{2}})$. In either case setting $h = O((\log T)^\tau)$, $\tau < \infty$ gives $o_p(1)$. Hence, (3.4.21) now becomes

$$
\begin{aligned}
& T^{-\frac{1}{2}}\sum_t\mathbf{E}_\nu(t)'\{\varepsilon(t)+(\mathbf{M}_0(z)-\mathbf{I}_v)(\varepsilon(t)-\hat{\varepsilon}_h(t))\}+o_p(1) \\
& \quad = \; T^{-\frac{1}{2}}\sum_t\mathbf{S}_{\lambda(\nu)}(\varepsilon(t-1)'\otimes\mathbf{I}_v,\ldots,\varepsilon(t-p)'\otimes\mathbf{I}_v)'\{\varepsilon(t) \\
& \qquad\qquad + \; (\mathbf{M}_0(z)-\mathbf{I}_v)(\varepsilon(t)-\hat{\varepsilon}_h(t))\}+\; o_p(1) \qquad (3.4.23)
\end{aligned}
$$

As before, omit the selection matrix $\mathbf{S}_{\lambda(\nu)}$ and examine the k^{th} block of the second term on the right-hand side,

$$T^{-\frac{1}{2}}\sum_t \{\varepsilon(t-k)\otimes(\mathbf{M}_0(z)-\mathbf{I}_v)(\varepsilon(t)-\hat{\varepsilon}_h(t))\} = T^{-\frac{1}{2}}\sum_t\{(\varepsilon_1(t-k),\ldots,\varepsilon_v(t-k))' \otimes [(\mathbf{M}_0(z)-\mathbf{I}_v)\sum_{j\geq 0}(\mathbf{\Phi}_0(j)-\mathbf{\Phi}_h(j))\mathbf{y}(t-j)]\}, k=1,\ldots,p, \quad (3.4.24)$$

and consider its i^{th} component,

$$T^{-\frac{1}{2}}\sum_t\{\sum_{r=0}^{p}(\mathbf{M}_0(r)-\delta_{0r}\mathbf{I}_v)\sum_{j\geq 1}(\mathbf{\Phi}_0(j)-\hat{\mathbf{\Phi}}_h(j))\mathbf{y}(t-r-j)\varepsilon_i(t-k)\},\ i=1,\ldots,v$$

$$= -T^{-\frac{1}{2}}\sum_t\{\sum_{r=0}^{p}(\mathbf{M}_0(r)-\delta_{0r}\mathbf{I}_v)\sum_{j=1}^{h}(\hat{\mathbf{\Phi}}_h(j)-\mathbf{\Phi}_h(j))\mathbf{y}(t-r-j)\varepsilon_i(t-k)\}$$

$$+ T^{-\frac{1}{2}}\sum_t\{\sum_{r=1}^{p}(\mathbf{M}_0(r)-\delta_{0r}\mathbf{I}_v)\sum_{j=1}^{h}(\mathbf{\Phi}_0(j)-\mathbf{\Phi}_h(j))\mathbf{y}(t-r-j)\varepsilon_i(t-k)$$

$$+ T^{-\frac{1}{2}}\sum_t\{\sum_{r=1}^{p}(\mathbf{M}_0(r)-\delta_{0r}\mathbf{I}_v)\sum_{j=h+1}^{\infty}\mathbf{\Phi}_0(j)\mathbf{y}(t-r-j)\varepsilon_i(t-k)\}. \quad (3.4.25)$$

By employing arguments similar to those that lead to (3.4.14), noting that $T^{-1}\sum_t \mathbf{y}(t-r-j)\varepsilon_i(t-k)$ are uniformly bounded, we find once again that the last two terms are $o_p(1)$. Thus (3.4.25) reduces to

$$-T^{-\frac{1}{2}}\sum_t\{\sum_{r=0}^{p}(\mathbf{M}_0(r)-\delta_{0r}\mathbf{I}_v)\sum_{j=1}^{h}(\hat{\mathbf{\Phi}}_h(j)-\mathbf{\Phi}_h(j))\mathbf{y}(t-r-j)\varepsilon_i(t-k)\}+o_p(1)$$

$$= T^{-\frac{1}{2}}\sum_t\{\sum_{r=0}^{p}(\mathbf{M}_0(r)-\delta_{0r}\mathbf{I}_v)(\mathbf{\Phi}_h-\hat{\mathbf{\Phi}}_h)\mathcal{Y}_h(t-r)\varepsilon_i(t-k)\} + o_p(1). \quad (3.4.26)$$

where the right-hand side expression follows from Lemma (3.4.1). Considering (3.4.26) for all i leads to an expression of the form

$$vec\{-T^{\frac{1}{2}}\sum_{r=0}^{p}(\mathbf{M}_0(r)-\delta_{0r}\mathbf{I}_v)(\mathbf{\Phi}_h-\hat{\mathbf{\Phi}}_h)\hat{\gamma}_h(r,k)\}+o_p(1),\ k=1,\ldots,p \quad (3.4.27)$$

which by employing arguments analogous to those used in moving from (3.4.15) to (3.4.17) can be rewritten as

$$vec[T^{-\frac{1}{2}}\sum_{r=0}^{p}(\mathbf{M}_0(r)-\delta_{0r}\mathbf{I}_v)\{\sum_t \varepsilon(t)\mathcal{Y}_h(t)'\hat{\mathbf{\Gamma}}_h^{-1}\hat{\gamma}_h(r,k)\}]+o_p(1), \quad k=1,2,\ldots,p. \quad (3.4.28)$$

By Lemma (3.4.2ii), $\hat{\mathbf{\Gamma}}_h^{-1}\hat{\gamma}_h(r-k)$ can be replaced by $\mathbf{\Gamma}_h^{-1}\gamma_h(r,k)$. Hence, (3.4.28) becomes

$$vec\{T^{-\frac{1}{2}}\sum_{r=0}^{p}(\mathbf{M}_0(r)-\delta_{0r}\,\mathbf{I}_v)\{\sum_t \varepsilon(t)\mathcal{Y}_h(t)'\mathbf{\Gamma}_h^{-1}\gamma_h(r-k)]\}, k=1,\ldots,p. \quad (3.4.29)$$

By Lemma(3.4.3ii), it is clear that (3.4.29) reduces to

$$vec\{T^{-\frac{1}{2}}\sum_{r=0}^{p}(\mathbf{M}_0(r)-\delta_{0r}\mathbf{I}_v)\sum_t \varepsilon(t)\varepsilon(t+r-k|t-1)'\}+o_p(1),$$
$$k=1,\ldots,p. \quad (3.4.30)$$

The combination of (3.4.30) with the first term of (3.4.23) gives an equivalent expression for (3.4.5) as

$$\begin{aligned} & vec\{T^{-\frac{1}{2}}\sum_t[\sum_{r=0}^{p}\mathbf{M}_0(r)\varepsilon(t)\varepsilon(t+r-k|t-1)']\}+o_p(1) \\ = \; & T^{-\frac{1}{2}}\sum_t\{\sum_{r=0}^{p}(\varepsilon(t+r-k\mid t-1)\otimes\mathbf{M}_0(r)\}\,\varepsilon(t) \\ & + o_p(1),\ k=1,\ldots,p. \end{aligned} \quad (3.4.31)$$

It now remains for us to consider (3.4.4),

$$\begin{aligned} & -T^{-\frac{1}{2}}\sum_t\hat{\mathbf{F}}_\nu(t)'\{\varepsilon(t)+(\mathbf{M}_0(z)-\mathbf{I}_v)(\varepsilon(t)-\hat{\varepsilon}_h(t))\} \\ = \; & -T^{-\frac{1}{2}}\sum_t[\mathbf{S}_{\beta(\nu)}\{(\mathbf{y}(t)-\hat{\varepsilon}_h(t))\otimes\mathbf{I}_v)\}\{\varepsilon(t) \\ & + (\mathbf{M}_0(z)-\mathbf{I}_v)(\varepsilon(t)-\hat{\varepsilon}_h(t))\}]. \end{aligned} \quad (3.4.32)$$

Neglecting the selection matrix $\mathbf{S}_{\beta(\nu)}$ once again and proceeding as previously, we may replace $\hat{\varepsilon}_h(t)$ by $\varepsilon(t)$ in $\hat{\mathbf{F}}_\nu(t)$ to obtain an equivalent expression to (3.4.32) as

$$-T^{-\frac{1}{2}}\sum_t\{(\mathbf{y}(t)-\varepsilon(t))\otimes\varepsilon(t)\}-T^{-\frac{1}{2}}\sum_t\{(\mathbf{y}(t)-\varepsilon(t))\otimes(\mathbf{M}_0(z)-\mathbf{I}_v)(\varepsilon(t)-\hat{\varepsilon}_h(t))\} \quad (3.4.33)$$

plus a term equivalent to (3.4.22) for $k=0$. The same line of argument leading from (3.4.21) to (3.4.23) can therefore be employed to show that this term is $o_p(1)$. The second term in (3.4.33) can be expressed as

$$\begin{aligned} & -T^{-\frac{1}{2}}\sum_t\mathbf{y}(t)\otimes(\mathbf{M}_0(z)-\mathbf{I}_v)(\varepsilon(t)-\hat{\varepsilon}_h(t)) \\ & + \; T^{-\frac{1}{2}}\sum_t\varepsilon(t)\otimes(\mathbf{M}_0(z)-\mathbf{I}_v)(\varepsilon(t)-\hat{\varepsilon}_h(t)). \end{aligned} \quad (3.4.34)$$

Repeating the steps employed in moving from (3.4.7) to (3.4.19) we find that the first term of (3.4.34) becomes

$$vec\{-T^{-\frac{1}{2}}\sum_{t}[\sum_{r=0}^{p}(\mathbf{M}_0(r)-\mathbf{I}_v)\varepsilon(t)\mathbf{y}(t+r|t-1)']\}+o_p(1). \tag{3.4.35}$$

Similarly, by adopting the method of proof that leads to (3.4.30), the second term in (3.4.34) gives

$$vec\{T^{-\frac{1}{2}}\sum_{t}[\sum_{r=0}^{p}(\mathbf{M}_0(r)-\mathbf{I}_v)\varepsilon(t)\varepsilon(t+r|t-1)']\}+o_p(1)=o_p(1) \tag{3.4.36}$$

since $\varepsilon(t+r \mid t-1)=0,\ r\geq 0$. Combining the terms in (3.4.35) and (3.4.36) with the first term of (3.4.33) we obtain

$$-T^{-\frac{1}{2}}\sum_{t}\{\sum_{r=0}^{p}(\mathbf{y}(t+r \mid t-1)'\otimes\mathbf{M}_0(r))\}\{\varepsilon(t)\}+o_p(1). \tag{3.4.37}$$

Collecting the alternative expressions for (3.4.3) - (3.4.5) together and reinstating the selection matrix $\mathbf{S}_\nu$ we have

$$\hat{\mathbf{X}}_{\nu,T}\{\sqrt{T}(\theta_{LS}-\theta_0)\} \;=\; \mathbf{S}_\nu\{T^{-\frac{1}{2}}\sum_{t}\mathbf{Z}(t)\,\varepsilon(t)\}+o_p(1) \tag{3.4.38}$$

where

$$\mathbf{Z}(t)=\sum_{r=0}^{p}\{\mathbf{x}(t+r \mid t-1)\otimes\mathbf{M}_0(r)\}$$

and

$$\begin{aligned}\mathbf{x}(t+r|t-1)' \;=\; &\{-\mathbf{y}(t+r-1|t-1)',\ldots,-\mathbf{y}(t+r-p|t-1)' : -\mathbf{y}(t+r|t-1)' :\\ &\varepsilon(t+r-1|t-1)',\ldots,\varepsilon(t+r-p|t-1)'\}.\end{aligned}$$

Furthermore, the vector of predictors, $\mathbf{x}(t+r|t-1)$, can be recursively up-dated and we may rewrite $\mathbf{Z}(t)$ as

$$\mathbf{Z}(t)=\sum_{r=0}^{p}\mathbf{V}^r\mathbf{x}(t)\otimes\mathbf{M}_0(r)$$

where $\mathbf{V}$ denotes the matrix of coefficients associated with the first non-trivial linear predictor $\mathbf{x}(t+1|t-1)=\mathbf{V}\mathbf{x}(t)$. The explicit form of the $(2p+1)v\times(2p+1)v$ matrix

$\mathbf{V}$ is given at the end of this section. It is now a relatively straightforward matter to verify that the sufficient conditions for asymptotic normality given in Kohn (1978, Theorem 5), for example, are satisfied by $T^{-\frac{1}{2}}\sum_t \mathbf{Z}(t)\varepsilon(t)$. Since $\|\hat{X}_{\nu,T} - \mathbf{S}_\nu(X_\nu \otimes \mathbf{I}_v)\mathbf{S}_\nu'\| = O(Q_T)$, via Lemma (3.3.1), it follows from (3.4.38) that $T^{\frac{1}{2}}(\hat{\theta}_{LS} - \theta_0)$ is asymptotically normal with zero mean vector and variance-covariance matrix

$$\{\mathbf{S}_\nu(\mathbf{X}_\nu \otimes \mathbf{I}_v)\mathbf{S}_\nu'\}^{-1}\mathbf{C}_0\{\mathbf{S}_\nu(\mathbf{X}_\nu \otimes \mathbf{I}_v)\mathbf{S}_\nu'\}^{-1} \tag{3.4.39}$$

where $\mathbf{C}_0 = \mathbf{S}_\nu\mathcal{E}\{[\mathbf{Z}(t)\varepsilon(t)][\mathbf{Z}(t)\varepsilon(t)]'\}\mathbf{S}_\nu'$. On application of the rule $\mathbf{AB} \otimes \mathbf{CD} = (\mathbf{A} \otimes \mathbf{C})(\mathbf{B} \otimes \mathbf{D})$, we found that $\mathbf{Z}(t)$ can be written in the form

$$\begin{aligned}\mathbf{Z}(t) &= \sum_{r=0}^{p}(\mathbf{V}^r \otimes \mathbf{M}_0(r))(\mathbf{x}(t) \otimes \mathbf{I}_v) \\ &= \mathbf{D}_0(\mathbf{x}(t) \otimes \mathbf{I}_v)\end{aligned}$$

where

$$\mathbf{D}_0 = \sum_{r=0}^{p}(\mathbf{V}^r \otimes \mathbf{M}_0(r)). \tag{3.4.40}$$

Hence,

$$\begin{aligned}\mathbf{C}_0 &= \mathbf{S}_\nu\mathbf{D}_0\mathcal{E}[\mathbf{x}(t) \otimes \varepsilon(t))(\mathbf{x}(t) \otimes \varepsilon(t))']\mathbf{D}_0'\mathbf{S}_\nu' \\ &= \mathbf{S}_\nu\mathbf{D}_0\mathcal{E}[\mathbf{x}(t)\mathbf{x}(t)' \otimes \varepsilon(t)\varepsilon(t)']\mathbf{D}_0'\mathbf{S}_\nu' \\ &= \mathbf{S}_\nu\mathbf{D}_0(\mathbf{X}_\nu \otimes \mathbf{\Sigma}_0)\mathbf{D}_0'\mathbf{S}_\nu'\end{aligned}$$

and the proof of Theorem (3.4.1) is thereby completed. □

For technical reasons we have omitted the explicit construction of $\mathbf{V}$ in the proof presented above. This will now be given. Recall that $\mathbf{y}(t+r|t-1)$ and $\varepsilon(t+r|t-1)$ are the best linear predictors, respectively, of $\mathbf{y}(t+r)$ and $\varepsilon(t+r)$ from $\mathbf{y}(s)$ $(s<t), r = 0, 1, \ldots, p$. Using the basic model (3.2.1) the linear predictor of $\mathbf{y}(t+\ell)$ can be defined recursively via

$$\begin{aligned}\mathbf{y}(t+\ell|t-1) &= -\sum_{j=1}^{p}\tilde{\mathbf{A}}_0(j)\mathbf{y}(t+\ell-j|t-1) + \sum_{j=1}^{p}\tilde{\mathbf{M}}_0(j)\varepsilon(t+\ell-j|t-1), \\ &\qquad \ell \geq 0\end{aligned} \tag{3.4.41}$$

where $\tilde{\mathbf{A}}_0(j) = \mathbf{A}_0^{-1}(0)\mathbf{A}_0(j)$ and $\tilde{\mathbf{M}}_0(j) = \mathbf{M}_0^{-1}(0)\mathbf{M}_0(j), \mathbf{M}_0(0) = \mathbf{A}_0(0)$. As noted earlier, $\varepsilon(t+r|t-1) = 0$ for $r \geq 0$. When $r < 0$, $\varepsilon(t+r \mid t-1) = \varepsilon(t+r)$. First observe that $\mathbf{V}^0 = \mathbf{I}_{(2p+1)v}$ and $\mathbf{x}(t \mid t-1) = \mathbf{x}(t)$. Now consider $r = 1$. We have $\mathbf{x}(t+1|t-1) = (\mathbf{y}(t|t-1)', \ldots, \mathbf{y}(t-p+1)' : \mathbf{y}(t+1|t-1)' : \varepsilon(t|t-1)',$ $\varepsilon(t-1)', \ldots, \varepsilon(t-p+1)')'$. Making use of (3.4.41) leads to an expression of the form $\mathbf{x}(t+1|t-1) = \mathbf{V}\mathbf{x}(t)$ where $\mathbf{V}$ is given as

$$\mathbf{V} = \begin{bmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} & \mathbf{V}_{13} \\ \mathbf{V}_{21} & \mathbf{V}_{22} & \mathbf{V}_{23} \\ \mathbf{V}_{31} & \mathbf{V}_{32} & \mathbf{V}_{33} \end{bmatrix} \tag{3.4.42}$$

wherein

$$\mathbf{V}_{11} = \begin{bmatrix} -\tilde{\mathbf{A}}_0(1) & \ldots & -\tilde{\mathbf{A}}_0(p-1) & -\tilde{\mathbf{A}}_0(p) \\ & I_{(p-1)v} & & \mathbf{0} \end{bmatrix},$$

$$\mathbf{V}_{13} = \begin{bmatrix} \tilde{\mathbf{M}}_0(1) & \tilde{\mathbf{M}}_0(2) & \ldots & \tilde{\mathbf{M}}_0(p-1) & \tilde{\mathbf{M}}_0(p) \\ & & \mathbf{0} & & \end{bmatrix},$$

$$\mathbf{V}_{21} = \begin{bmatrix} \tilde{\mathbf{A}}_0^2(1) - \tilde{\mathbf{A}}_0(2) & \tilde{\mathbf{A}}_0(1)\tilde{\mathbf{A}}_0(2) - \tilde{\mathbf{A}}_0(3) & \ldots & \tilde{\mathbf{A}}_0(1)\tilde{\mathbf{A}}_0(p) \end{bmatrix},$$

$$\mathbf{V}_{23} = \begin{bmatrix} -\tilde{\mathbf{A}}_0(1)\tilde{\mathbf{M}}_0(1) + \tilde{\mathbf{M}}_0(2) & -\tilde{\mathbf{A}}_0(1)\tilde{\mathbf{M}}_0(2) + \tilde{\mathbf{M}}_0(3) & \ldots & -\tilde{\mathbf{A}}_0(1)\tilde{\mathbf{M}}_0(p) \end{bmatrix},$$

$$\mathbf{V}_{33} = \begin{bmatrix} \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{I}_{(p-1)v} & \ldots & \mathbf{0} \end{bmatrix}$$

and the remaining components of matrix $\mathbf{V}$ are null matrices of appropriate dimension. Extending this consideration to other values of r $(2 \leq r \leq p)$ we find, after direct manipulations similar to those leading from (3.4.41) to (3.4.42), that $\mathbf{x}(t+r|t-1) = \mathbf{V}\,\mathbf{x}(t+r-1|t-1)$, which ultimately results in the formula $\mathbf{x}(t+r|t-1) = \mathbf{V}^r\mathbf{x}(t), \mathbf{V}^r = \mathbf{V}\mathbf{V}^{r-1}, r = 1, \ldots, p.$

Finally, the asymptotic normality result for $\hat{\theta}_{LS}$ established in this section is obtained assuming that n_{0i}, $i = 1, \ldots, v$ are known *a priori*. As pointed out in

Section (3.1), however, in practice the Kronecker indices usually have to be estimated from the observed realization of the generating process. Nevertheless, the theory attains some theoretical completeness since the central limit theorem will apply to the system parameter estimates corresponding to the Kronecker indices identified by a consistent procedure as is discussed in Chapter 1 (Section (1.3)). However, it should be borne in mind that the Kronecker indices could be mis-specified and in such situations some modifications of the arguments employed in the theoretical developments will have to be clearly made. For some discussion of this and other related issues see, for example, Poskitt (1990).

Appendix B

B.1 Algorithms

This section provides some discussion of the algorithms that are commonly used to estimate regression equations (3.2.3).

Given that the autocovariances of the process $\mathbf{y}(t)$, $\mathbf{\Gamma}_{yy}(r) = \mathbf{\Gamma}_{yy}(-r)'$, $r = 0, \pm 1, \ldots$, are as defined in Section (3.2), one obvious method by which to obtain the estimates of the predictor coefficients, $\mathbf{\Phi}_h(j)$, $j = 1, \ldots, h$ in (3.2.4) is via multivariate Yule-Walker equations

$$\sum_{j=0}^{h} \mathbf{\Phi}_h(j)\mathbf{\Gamma}_{yy}(j-k) = \delta_{0k}\mathbf{\Sigma}_h, \quad k = 0, 1, \ldots, h \tag{B.1.1}$$

where δ_{0k} denotes Kronecker's delta. If we replace the autocovariances $\mathbf{\Gamma}_{yy}(r)$, $r = 0, 1, \ldots, h$, appearing in (B.1.1) by the corresponding sample covariances $\hat{\mathbf{\Gamma}}_{yy}(r)$, we obtain a set of equations for the so-called Yule-Walker estimators $\hat{\mathbf{\Phi}}_h$ of $\mathbf{\Phi}_h$,

$$\hat{\mathbf{\Phi}}_h = \mathbf{G}_h^{-1}\hat{\Upsilon}_h \tag{B.1.2}$$

where $\mathbf{G}_h = [\hat{\mathbf{\Gamma}}_{yy}(j-k)]$, $j, k = 1, \ldots, h$ and $\hat{\Upsilon}_h = (\hat{\mathbf{\Gamma}}_{yy}(1)', \ldots, \hat{\mathbf{\Gamma}}_{yy}(h)')'$. The procedure just described is Toeplitz, that is, $\mathbf{\Gamma}_h$ and $\mathbf{G}_h$ have the same block down any diagonal of blocks. Implicit in the direct calculation of Yule-Walker estimates via (B.1.2) is the inversion of a $hv \times hv$ matrix $\mathbf{G}_h$ and this, of course, can be avoided by the use of a recursion. Suppose when $\mathbf{y}(t)$ is a scalar (*i.e.*, $v = 1$) we let $\hat{\phi}_h(j)$,

$\hat{\rho}(j)$, $j = 0, 1, \ldots, h$ and $\hat{\sigma}_h^2$ denote the scalar equivalents of $\hat{\mathbf{\Phi}}_h(j)$, $\hat{\mathbf{\Gamma}}_{yy}(j)$ and $\hat{\mathbf{\Sigma}}_h$, respectively. The Yule-Walker estimates, $\hat{\phi}_h$ and $\hat{\sigma}_h^2$ can be obtained using the following algorithm which we refer to as Levinson-Durbin recursion. This algorithm was originally due to Levinson (1947) and rediscovered in the statistical context by Durbin (1960).

B3.2.1 [Levinson-Durbin Algorithm]

Suppose $\hat{\rho}(r), r = 0, 1, \ldots,$ are available. Initialize the recursion with $\hat{\phi}_h(0) = 1 \; \forall \; h$, $\hat{\sigma}_0^2 = \hat{\rho}(0)$. Then the parameters, $\phi_h(j), j = 1, \ldots, h$; σ_h^2 can be recursively calculated as:

$$\begin{aligned}
\hat{\phi}_h(h) &= -\sum_{j=0}^{h-1} \hat{\phi}_{h-1}(j)\hat{\rho}(h-j)/\hat{\sigma}_{h-1}^2, h = 1, 2, \ldots, T-1 \\
\hat{\phi}_h(j) &= \hat{\phi}_{h-1}(j) + \hat{\phi}_h(h)\hat{\phi}_{h-1}(h-j), j = 1, \ldots, h-1, \\
\hat{\sigma}_h^2 &= \hat{\sigma}_{h-1}^2\{1 - \hat{\phi}_h(h)^2\}.
\end{aligned}$$

Note that the $\hat{\phi}_h(h)$ are estimates of the partial autocorrelations (also called reflection coefficient in engineering terminology). If the underlying stochastic process generating the data is a purely autoregressive process, $\hat{\phi}_h(h)$ in the convention of Box and Jenkins (1976), are extremely valuable first for deciding on the appropriateness of an autoregressive model and then for choosing an appropriate estimate of order, h, for the model to be fitted.

In the vector case ($v > 1$), $\hat{\mathbf{\Phi}}_h(j), j = 1, \ldots, h$, can also be recursively generated using the Levinson-Whittle algorithm. This algorithm is a multivariate generalization of Levinson-Durbin recursion and is due to Whittle (1963). Unlike the univariate algorithm (B3.2.1), the multivariate version requires the simultaneous solution of two sets of equations, one arising from the calculation of the forward predictor and the other in the calculation of the backward predictor. Thus, in the algorithm that follows, we let $\bar{\mathbf{\Phi}}_h$ denote the coefficients associated with the *backward* regression of

$\mathbf{y}$(t-h-1) on $\mathbf{y}(t-j), j = 1, \ldots, h$ in the model

$$\sum_{j=0}^{h} \bar{\mathbf{\Phi}}_h(j)\mathbf{y}(t-h-1+j) = \bar{\varepsilon}_h(t), \ \bar{\mathbf{\Phi}}_h(0) = \mathbf{I}_v \qquad \text{(B.1.3)}$$

and $\bar{\mathbf{\Sigma}}_h$ is the corresponding residual mean square. Similarly, $\hat{\mathbf{\Phi}}_h$ in (3.2.4) shall be referred to as the coefficient in the *forward* regression of $\mathbf{y}$(t) on $\mathbf{y}(t-j), j = 1, \ldots, h$.

B3.2.2 [Levinson-Whittle Algorithm]

We have $\hat{\mathbf{\Gamma}}_{yy}(j), j = 0, 1, \ldots$ as input. Initialize the recursion with $\hat{\mathbf{\Phi}}_h(0) = \bar{\mathbf{\Phi}}_h(0) = \mathbf{I}_v$, $\hat{\mathbf{\Sigma}}_0 = \bar{\mathbf{\Sigma}}_0 = \hat{\mathbf{\Gamma}}_{yy}(0)$.

At the h^{th} recursion compute

$$\begin{aligned}
\mathbf{\Psi}_h &= \sum_{j=0}^{h} \hat{\mathbf{\Phi}}_h(j)\hat{\mathbf{\Gamma}}_{yy}(j-h-1) = \sum_{j=0}^{h} \bar{\mathbf{\Phi}}_h(j)\hat{\mathbf{\Gamma}}_{yy}(h+1-j), \\
\hat{\mathbf{\Phi}}_h(h) &= -\mathbf{\Psi}_{h-1}\bar{\mathbf{\Sigma}}_{h-1}^{-1}, \quad \bar{\mathbf{\Phi}}_h(h) = -\mathbf{\Psi}'_{h-1}\hat{\mathbf{\Sigma}}_{h-1}^{-1} \\
\hat{\mathbf{\Phi}}_h(j) &= \hat{\mathbf{\Phi}}_{h-1}(j) + \hat{\mathbf{\Phi}}_h(h)\bar{\mathbf{\Phi}}_{h-1}(h-j), \ j = 1, \ldots, h-1, \\
\bar{\mathbf{\Phi}}_h(j) &= \bar{\mathbf{\Phi}}_{h-1}(j) + \bar{\mathbf{\Phi}}_h(h)\hat{\mathbf{\Phi}}_{h-1}(h-j), \ j = 1, \ldots, h-1, \\
\hat{\mathbf{\Sigma}}_h &= \{\mathbf{I}_v - \hat{\mathbf{\Phi}}_h(h)\bar{\mathbf{\Phi}}_h(h)\}\hat{\mathbf{\Sigma}}_{h-1}, \\
\bar{\mathbf{\Sigma}}_h &= \{\mathbf{I}_v - \bar{\mathbf{\Phi}}_h(h)\hat{\mathbf{\Phi}}_h(h)\}\bar{\mathbf{\Sigma}}_{h-1}.
\end{aligned}$$

The Toeplitz symmetry inherent in these calculations means that the estimates $\hat{\mathbf{\Phi}}_h(z) = \sum_{j=0}^{h} \hat{\mathbf{\Phi}}_h(j)z^j$, $\hat{\mathbf{\Phi}}_h(0) = \mathbf{I}_v$ and $\bar{\mathbf{\Phi}}_h(z) = \sum_{j=0}^{h} \bar{\mathbf{\Phi}}_h(j)z^j$, $\bar{\mathbf{\Phi}}_h(0) = \mathbf{I}_v$ will be stable in the sense that the $\det\{\hat{\mathbf{\Phi}}_h(z) : \bar{\mathbf{\Phi}}_h(z)\}$ are guaranteed to have their zeros outside the unit circle (Hannan, 1970, p.332). The numerical stability of the algorithm (B3.2.1), for example, has been shown to be comparable to the stability of the Cholesky or Q-R algorithm both in terms of fixed or floating-point arithmetic (Cybenko (1980)). It is often the case that when the true model is a finite order autoregression, the estimates of $\mathbf{\Phi}_h(j), j = 1, \ldots, h$ in (3.2.3) give rise to asymptotically efficient parameter estimates. Though $\hat{\mathbf{\Phi}}_h(z)$ is stable and the associated

estimates are asymptotically efficient, it is now relatively known that Toeplitz procedures have serious disadvantages (Makhoul, (1981)). For moderate sample size, $\hat{\mathbf{\Phi}}_h(z)$ can be severely biased particularly when the data generating mechanism admits an autoregressive representation with zeros close to the unit circle (Tjostheim and Paulsen, 1983; Pukkila, 1988). The bias is, of course, due to a fiction inherent in the Toeplitz procedure (B.1.2), namely that $\mathbf{y}(t) = 0, t \leq 0$ or $t > T$. Therefore, some modifications of the Toeplitz procedure have been suggested (see Makhoul (1981), Dahlhaus (1983) and Paulsen and Tjostheim (1985) for some discussions). Such bias is not present in the least squares estimator, which we shall denote by use of a *tildé* but $\tilde{\mathbf{\Phi}}_h(z) = \sum_{j=0}^{h} \tilde{\mathbf{\Phi}}_h(j) z^j$ cannot be guaranteed to be stable. The instability, of course, can be easily checked and rectified, if needs be, via either an iterative or closed-form procedure. We leave the details of these procedures till Chapter 6.

To obtain least squares estimators we regard (3.2.3) as a straightforward regression where the regressor variables are given by $\mathcal{Y}_h(t) = (\zeta_h(z) \otimes \mathbf{y}(t))$. Thus, for (3.2.3) the Least Squares Estimate for θ is obtained as in the algorithm that follow.

B3.2.3 [Least Squares Estimate - AR Model]

Input observed realization $\mathbf{y}(t), t = 1, \ldots, T$ as data.

Compute $\tilde{\mathbf{\Phi}}_h = -\hat{\mathbf{\Gamma}}_h^{-1} \hat{\mathbf{\Xi}}_h$ where

$$\begin{aligned} \hat{\mathbf{\Gamma}}_h &= T^{-1} \sum_{t=h+1}^{T} \mathcal{Y}_h(t) \mathcal{Y}_h(t)', \quad \hat{\mathbf{\Xi}}_h = T^{-1} \sum_{t=h+1}^{T} \mathcal{Y}_h(t) \mathbf{y}(t)', \\ \tilde{\mathbf{\Phi}}_h &= (\tilde{\mathbf{\Phi}}_h(1)', \ldots, \tilde{\mathbf{\Phi}}_h(h)')'. \end{aligned} \tag{B.1.4}$$

For technical reasons, only positively indexed terms of $\mathbf{y}$(t-j) are included in the sums of squares and cross-products in (B.1.4) and consequently, $\hat{\mathbf{\Gamma}}_h$ does not possess a Toeplitz structure. Efficient numerical methods for the solution of least squares problems of this type are, of course, readily available see, for example, Miller (1991)

and the references therein.

If it is considered that a stable generating function, $\mathbf{\Phi}(z)$, is desirable, an estimator that shares the advantages of both the Levinson-Whittle and Least Squares procedures may be obtained using a procedure due to Morf, Veira and Kailath (1978). This procedure is based on the characterization of the autocovariance function of a process by means of a sequence of matrix partial multivariate generalization of the maximum entropy method due to Burg (1975). For $v = 1$, ϕ_h can be obtained using the following algorithm, which is referred to as the Burg's algorithm.

B3.2.4 [Burg Algorithm]

Input a finite data record generated by a zero mean stationary process $\mathbf{y}(t)$.

Initialize the recursion with

$$\begin{aligned} \hat{b}_h(0) &= 0, \quad (h \geq 0) \\ \hat{e}_0(t) &= \hat{b}_0(t) = y(t), t = 1, \ldots, T; \\ \hat{\phi}_h(0) = 1; \qquad & y(t) = 0, \ (t \leq 0, \ t > T). \end{aligned}$$

At the $(h+1)^{th}$ stage compute

$$\begin{aligned} \hat{\phi}_{h+1}(h+1) &= \{-2\sum_{h+1}^{T} \hat{e}_h(t)\hat{b}_h(t-1)\}/\{\sum_{h+1}^{T} \hat{e}_h(t)^2 + \sum_{h+1}^{T} \hat{b}_h(t-1)^2\}, \ h \geq 0, \\ \hat{e}_h(t) &= \hat{e}_{h-1}(t) + \hat{\phi}_h(h)\hat{b}_{h-1}(t-1), \\ \hat{b}_h(t) &= \hat{b}_{h-1}(t-1) + \hat{\phi}_h(h)\hat{e}_{h-1}(t), \\ \hat{\phi}_{h+1}(j) &= \hat{\phi}_h(j) + \hat{\phi}_{h+1}(h+1)\hat{\phi}_h(h+1-j), \ j = 1, \ldots, h; \ h \geq 0. \end{aligned}$$

Essentially, Burg's algorithm is based on least squares approach but instead of minimizing the *forward* prediction error variance, as was done in the case of least squares, Algorithm B3.2.4 seeks to select the partial correlations (or reflection coefficients) in such a way that the average variance obtained by running the prediction error filter both forward and backward over the data span $(1 \leq t \leq T)$ is minimized. It is

evident that working entirely within the data span the source of bias associated with the Toeplitz calculation is eliminated. Of course, an estimated covariance function can be produced as a byproduct of the algorithm. It should be mentioned, however, that the algorithm B3.2.4 does not generalize directly to vector case since the forward and backward autoregression matrices are not the same as in the univariate situation. Also, the forward and backward one-step prediction error covariance matrices are different, although they have the same determinant (Akaike, 1973). These and other related issues concerning the implementation of the algorithm B3.2.4 in the multivariate case are well known (See e.g. Jones (1978)).

It is perhaps worth pointing out that the observed differences in the estimates obtained from the Algorithms B3.2.2 to B3.2.4 are only of a finite sample phenomenon. Asymptotically, these differences become negligible and these estimators can be shown to have the same limiting distribution (Hannan, 1970; Hannan and Deistler, 1988). It follows thus that the choice made by the applied worker in moderate to large sample may be governed by questions of availability and convenience.

Chapter 4

Asymptotic Relative Efficiency Of Least Squares Estimators

4.1 Motivation

When dealing with time series data it is often the practice to estimate the parameters associated with the underlying process via the Maximum (Gaussian) likelihood method. Since the Maximum likelihood procedures are iterative and computationally rather expensive, it seems prudent to commence the estimation scheme with a strongly consistent estimator as was mentioned in Chapter 2. This will not only decrease the possibility of non-convergence, but also reduces the number of iterations that are needed to achieve the selected convergence criterion. In the light of Lemma (3.3.3), the least squares estimator could provide appropriate initial values for the computation of the Gaussian estimates once an appropriate model, or collection of specifications deemed worthy of further investigation, has been delineated. This estimator may even have strong appeal where the Maximum (Gaussian) likelihood procedure seems too costly or computationally prohibitive (Saikonnen (1986)). However, there is a trade-off between the relative simplicity and ease of implementation of ordinary least squares regression and the increased variability of the least squares estimator over the Gaussian estimator.

The inefficiency of least squares estimators relative to the Maximum (Gaussian) likelihood estimator in various context is, of course, well known (Bloomfield and Watson (1975)) in the ordinary regression models. In the scalar time series models, McDougall (1988) provides some discussion of the comparative performance of various regression procedures, (Spliid (1983)); lagged regression procedures (Stoica et al (1985)); Pseudo linear regression (Solo (1978)), in relation to a fully efficient procedure (e.g. Hannan-Rissanen (1982)). We have been unable, however, to find a detailed analysis in the literature of the case of vector time series using ARMA echelon canonical forms. Following Poskitt and Salau (1993b) we can now make a very precise statement about the relative magnitude of the cost and benefits of using these alternative estimators, which is described in part in the following sections.

4.2 Asymptotic Comparison Of Estimators

Our purpose in the present section is to investigate the asymptotic efficiency of least squares estimator, $\hat{\theta}_{LS}$, in relation to the Gaussian estimator, $\hat{\theta}_G$. The asymptotic properties of the Gaussian (Maximum likelihood) estimator provides a lower bound to which the properties of alternative estimators can be compared. Thus we have the following definition.

Definition: An estimator is said to be efficient if it has the same limiting distribution as that of an equivalent Gaussian estimator constructed on Gaussian assumptions. As noted earlier, the Gaussian assumptions will not be maintained and for our purpose Condition A will suffice. Similarly, all the standard assumptions of Chapter 1 are preserved throughout our discussions in this chapter.

The performance of the least squares estimators may therefore be judged, in part, by their efficiency. This aspect of the procedure will now be discussed. For ease of reference and discussion, we begin by recapitulating some notation and definitions from Chapters 2 and 3. In Chapter 2, $\mathbf{\Lambda}_G$, the variance-covariance matrix of θ_G, was presented in the frequency domain as

$$\mathbf{\Lambda}_G = (\mathbf{S}_\nu \mathbf{\Omega} \mathbf{S}_\nu')^{-1} \tag{4.2.1}$$

where

$$\mathbf{\Omega} = (2\pi)^{-1} \int_{-\pi}^{\pi} (\mathbf{P}(e^{i\omega})\, \mathbf{P}(e^{i\omega})^* \otimes \{\mathbf{M}_0(e^{i\omega}) \mathbf{\Sigma}_0 \mathbf{M}_0(e^{i\omega})^*\}'^{-1})\, d\omega$$

and

$$\mathbf{P}(z) = \begin{bmatrix} -\zeta_p(z) \otimes \mathbf{K}_0(z) \mathbf{\Sigma}_0^{\frac{1}{2}} \\ (\mathbf{K}_0(z) - \delta_{0j} \mathbf{I}_v) \mathbf{\Sigma}_0^{\frac{1}{2}} \\ \zeta_p(z) \otimes \mathbf{\Sigma}_0^{\frac{1}{2}} \end{bmatrix}.$$

Here again, an asterisk has been used to denote a complex conjugate transpose of the indicated argument. Also, the subscript zero is used here to emphasize an evaluation at the true parameter values, the notational convention we shall persist with in the remaining sections of this chapter.

Similarly, the variance-covariance matrix of the least squares estimator, $\mathbf{\Lambda}_{LS}$, was obtained in Chapter 3 as a matrix quadratic form involving the expectation of the outer product of $\mathbf{x}(t)$ or Gramians as

$$\begin{aligned}\mathbf{\Lambda}_{LS} &= \{\mathbf{S}_\nu \mathcal{E}(\mathbf{x}(t)\mathbf{x}(t)' \otimes \mathbf{I}_v)\mathbf{S}_\nu'\}^{-1}\{\mathbf{S}_\nu \mathbf{D}_0(\mathcal{E}(\mathbf{x}(t)\mathbf{x}(t)') \otimes \mathbf{\Sigma}_0)\mathbf{D}_0'\mathbf{S}_\nu'\} \\ &\quad \times \{\mathbf{S}_\nu \mathcal{E}(\mathbf{x}(t)\mathbf{x}(t)' \otimes \mathbf{I}_v)\mathbf{S}_\nu'\}^{-1} \\ &= \{\mathbf{S}_\nu(\mathbf{X}_\nu \otimes \mathbf{I}_v)\mathbf{S}_\nu'\}^{-1}\{\mathbf{S}_\nu \mathbf{D}_0(\mathbf{X}_\nu \otimes \mathbf{\Sigma}_0)\mathbf{D}_0'\mathbf{S}_\nu'\} \\ &\quad \times \{\mathbf{S}_\nu(\mathbf{X}_\nu \otimes \mathbf{I}_v)\mathbf{S}_\nu'\}^{-1} \qquad (4.2.2)\end{aligned}$$

where

$$\mathbf{x}(t) = (-\zeta_p(z)' \otimes \mathbf{y}(t)' : -(\mathbf{y}(t) - \varepsilon(t))' : \zeta_p(z)' \otimes \varepsilon(t)')'.$$

The matrices $\mathbf{X}_\nu$ and $\mathbf{D}_0$ are as defined in (3.3.9) and (3.4.40), respectively.

We now wish to compare the asymptotic performance of $\hat{\theta}_{LS}$ to that of $\hat{\theta}_G$ via a comparison of the corresponding variance-covariance matrices (4.2.1) and (4.2.2), respectively. To facilitate this comparison we will first reconcile the two modes of representation using the standard isometric-isomorphism between the time and frequency domains. Define

$$\mathbf{w}_0(t) = \mathbf{x}(t) \otimes \mathbf{M}_0^{-1}(z)\mathbf{\Sigma}_0^{-\frac{1}{2}} \qquad (4.2.3)$$

and

$$\mathbf{n}_0(t) = \mathbf{x}(t) \otimes \mathbf{\Sigma}_0^{\frac{1}{2}} \qquad (4.2.4)$$

where $\mathbf{\Sigma}_0^{\frac{1}{2}}$ is the unique Cholesky lower triangular factor of $\mathbf{\Sigma}_0$. Then using (4.2.3) it is clear that $\mathbf{\Lambda}_G$ can be rewritten as

$$\begin{aligned}\mathbf{\Lambda}_G^{-1} &= \mathbf{S}_\nu \mathcal{E}\{\mathbf{w}_0(t)\mathbf{w}_0(t)'\}\mathbf{S}_\nu' \\ &= \mathbf{S}_\nu \mathcal{E}\{(\mathbf{x}(t) \otimes \mathbf{M}_0^{-1}(z)\mathbf{\Sigma}_0^{-\frac{1}{2}})(\mathbf{x}(t) \otimes \mathbf{M}_0^{-1}(z)\mathbf{\Sigma}_0^{-\frac{1}{2}})'\}\mathbf{S}_\nu'.\end{aligned}$$

Observe that when (1.1.4) holds $\mathbf{y}(t)$ can be written via Wold's decomposition as $\mathbf{y}(t) = \mathbf{A}^{-1}(z)\mathbf{M}(z)\varepsilon(t) = \mathbf{K}(z)\varepsilon(t)$ where the generating polynomial $\mathbf{K}(z)$ is

defined by (1.1.7) with $\mathbf{K}(0) = \mathbf{I}_v$. Then it follows that $\mathbf{x}(t)$ has the form

$$\mathbf{x}(t) = \{-\zeta_p(z)' \otimes (\mathbf{K}(z)\varepsilon(t))' : -(\mathbf{K}(z) - \mathbf{I}_v)\varepsilon(t))' : \zeta_p(z)' \otimes \varepsilon(t)'\}'.$$

Thus, given that $\mathbf{X}_\nu = \mathcal{E}(\mathbf{x}(t)\mathbf{x}(t)')$ it can be readily checked that the matrix $\mathbf{X}_\nu$ can be written in the frequency domain as $(2\pi)^{-1}\int_{-\pi}^{\pi}\mathbf{P}(e^{i\omega})\mathbf{P}(e^{i\omega})^* d\,\omega$. Also, we have the relationship

$$(\mathbf{X}_\nu \otimes \mathbf{\Sigma}_0) = \mathcal{E}(\mathbf{n}_0(t)\mathbf{n}_0(t)') = (2\pi)^{-1}\int_{-\pi}^{\pi}(\mathbf{P}(e^{i\omega})\mathbf{P}(e^{i\omega})^* \otimes \mathbf{\Sigma}_0)\, d\omega.$$

In what follows, we will require an explicit expression for $\mathcal{E}(\mathbf{D}_0\mathbf{n}_0(t)\mathbf{w}_0(t)')$. Thus, we have

$$\begin{aligned}
\mathcal{E}(\mathbf{D}_0\mathbf{n}_0(t)\mathbf{w}_0(t)') &= \sum_{r=0}^{p}(\mathbf{V}^r \otimes \mathbf{M}_0(r))\mathcal{E}(\mathbf{n}_0(t)\mathbf{w}_0(t)') \\
&= \sum_{r=0}^{p}(\mathbf{V}^r \otimes \mathbf{M}_0(r))\mathcal{E}\{(\mathbf{x}(t) \otimes \mathbf{\Sigma}_0^{\frac{1}{2}})(\mathbf{x}(t)' \otimes \mathbf{\Sigma}_0^{'-\frac{1}{2}}\mathbf{M}_0^{-1}(z)')\} \\
&= \frac{1}{2\pi}\int_{-\pi}^{\pi}\sum_{r=0}^{p}(\mathbf{V}^r \otimes \mathbf{M}_0(r))\{\mathbf{P}(e^{i\omega})\mathbf{P}(e^{i\omega})^* \\
&\qquad \otimes \mathbf{M}_0^{-1}(e^{i\omega})'\}\, d\omega \qquad (4.2.5)
\end{aligned}$$

since $\mathbf{X}_\nu = \mathcal{E}\{\mathbf{x}(t)\mathbf{x}(t)'\} = (2\pi)^{-1}\int_{-\pi}^{\pi}\mathbf{P}(e^{i\omega})\mathbf{P}(e^{i\omega})^*\, d\omega$.

We now establish the following result.

Theorem (4.2.1):

The variance-covariance matrices of $\hat{\theta}_{LS}$ and $\hat{\theta}_G$ satisfy the inequality $\mathbf{\Lambda}_{LS} \geq \mathbf{\Lambda}_G$ where the inequality relation has the interpretation that $\mathbf{\Lambda}_{LS} - \mathbf{\Lambda}_G$ is positive semi-definite.

Proof: The inequality stated in the theorem, using (4.2.1) and (4.2.2), is equivalent to

$$\mathbf{S}_\nu\mathbf{\Omega}\mathbf{S}_\nu' \geq \{\mathbf{S}_\nu(\mathbf{X}_\nu \otimes \mathbf{I}_v)\mathbf{S}_\nu'\}\{\mathbf{S}_\nu\mathbf{D}_0(\mathbf{X}_\nu \otimes \mathbf{\Sigma}_0)\mathbf{D}_0'\mathbf{S}_\nu'\}^{-1}\{\mathbf{S}_\nu(\mathbf{X}_\nu \otimes \mathbf{I}_v)\mathbf{S}_\nu'\}. \qquad (4.2.6)$$

To establish the inequality relation in (4.2.6) set $\mathbf{N}_0 = \mathcal{E}(\mathbf{n}_0(t)\mathbf{w}_0(t)')$ and consider

the process

$$\begin{bmatrix} \mathbf{S}_\nu \mathbf{D}_0 \mathbf{n}_0(t) \\ \mathbf{S}_\nu \mathbf{w}_0(t) \end{bmatrix} = \begin{bmatrix} \mathbf{S}_\nu \mathbf{D}_0 (\mathbf{x}(t) \otimes \boldsymbol{\Sigma}_0^{\frac{1}{2}}) \\ \mathbf{S}_\nu (\mathbf{x}(t) \otimes \mathbf{M}_0'(z)^{-1} \boldsymbol{\Sigma}_0^{-\frac{1}{2}}) \end{bmatrix}. \tag{4.2.7}$$

Since

$$\begin{aligned} \mathbf{S}_\nu \mathcal{E}\{\mathbf{D}_0(\mathbf{x}(t) \otimes \boldsymbol{\Sigma}_0^{\frac{1}{2}})(\mathbf{x}(t)' \otimes \boldsymbol{\Sigma}_0^{-\frac{1}{2}'} \mathbf{M}_0(z)^{-1})\} \mathbf{S}_\nu' &= \mathbf{S}_\nu \mathbf{D}_0 \mathcal{E}(\mathbf{n}_0(t) \mathbf{w}_0(t)') \mathbf{S}_\nu' \\ &= \mathbf{S}_\nu \mathbf{D}_0 \mathbf{N}_0 \mathbf{S}_\nu' \end{aligned}$$

it is clear that the variance-covariance matrix (or Gramian) of the process (4.2.7) is

$$\begin{bmatrix} \mathbf{S}_\nu \mathbf{D}_0 (\mathbf{X}_\nu \otimes \boldsymbol{\Sigma}_0) \mathbf{D}_0' \mathbf{S}_\nu' & \mathbf{S}_\nu \mathbf{D}_0 \mathbf{N}_0 \mathbf{S}_\nu' \\ (\mathbf{S}_\nu \mathbf{D}_0 \mathbf{N}_0 \mathbf{S}_\nu')' & \mathbf{S}_\nu \boldsymbol{\Omega} \mathbf{S}_\nu' \end{bmatrix}. \tag{4.2.8}$$

The matrix in (4.2.8) is, of course, positive semi-definite. Now consider the difference $(\mathbf{X}_\nu \otimes \mathbf{I}_v) - \mathbf{D}_0 \mathbf{N}_0$ and substituting for $\mathbf{D}_0 \mathbf{N}_0$ using (4.2.5) we have

$$\begin{aligned} (\mathbf{X}_\nu \otimes \mathbf{I}_v - \mathbf{D}_0 \mathbf{N}_0) &= \frac{1}{2\pi} \int_{-\pi}^{\pi} (\mathbf{P}(e^{i\omega}) \mathbf{P}(e^{i\omega})^* \otimes \mathbf{I}_v) d\,\omega - \frac{1}{2\pi} \int_{-\pi}^{\pi} \sum_{r=0}^{p} (\mathbf{V}^r \otimes \mathbf{M}_0(r)) \\ &\quad \times \{\mathbf{P}(e^{i\omega}) \mathbf{P}(e^{i\omega})^* \otimes \mathbf{M}_0(e^{i\omega})'^{-1}\} d\,\omega \\ &= \frac{1}{2\pi} \int_{-\pi}^{\pi} \{(\mathbf{I} \otimes \mathbf{M}_0(e^{i\omega})') - \sum_{r=0}^{p} (\mathbf{V}^r \otimes \mathbf{M}_0(r))\} \\ &\quad \times \{\mathbf{P}(e^{i\omega}) \mathbf{P}(e^{i\omega})^* \otimes \mathbf{M}_0(e^{i\omega})'^{-1}\} d\,\omega \\ &= \sum_{r=0}^{p} \frac{1}{2\pi} \int_{-\pi}^{\pi} \{(\mathbf{I} e^{i\omega r} - \mathbf{V}^r) \otimes \mathbf{I}_v\} \times \{\mathbf{P}(e^{i\omega}) \mathbf{P}(e^{i\omega})^* \\ &\quad \otimes \mathbf{M}_0(r) \mathbf{M}_0(e^{i\omega})'^{-1}\} d\,\omega \qquad (4.2.9) \\ &= \sum_{r=0}^{p} \mathcal{E}[\{(\mathbf{x}(t+r) - \mathbf{x}(t+r|t-1) \otimes \mathbf{M}_0(r)\}(\mathbf{x}(t)) \otimes \mathbf{M}_0'(z)^{-1})'] \\ &= \sum_{r=0}^{p} \mathcal{E}\{\varepsilon(t+r) \otimes \mathbf{M}_0(r)\}\{(\mathbf{x}(t) \otimes \mathbf{M}_0'(z)^{-1})'\} \end{aligned}$$

where $\varepsilon(t+r) = \mathbf{x}(t+r) - \mathbf{x}(t+r|t-1)$ and we have employed the rule $a\,\mathbf{A} \otimes \mathbf{B} = \mathbf{A} \otimes a\,\mathbf{B}$ in (4.2.9) with a being a scalar quantity, and $\mathbf{A}$ and $\mathbf{B}$ are matrices. Since by construction $\varepsilon(t+r)$ is orthogonal to the space spanned by $\mathbf{x}(t)$ and, indeed, $\mathbf{x}(t)$ is a Markov process, it follows that $\mathcal{E}[\{\mathbf{x}(t+r) - \mathbf{x}(t+r \mid t-1)\}\mathbf{x}(t-j)'] = 0$,

$j \geq 0,\ r = 0, \ldots, p$. We are therefore led to the conclusion that $(\mathbf{X}_\nu \otimes \mathbf{I}_\upsilon) = \mathbf{D}_0\mathbf{N}_0$. Then, it follows that the equality

$$\begin{bmatrix} \mathbf{S}_\nu\mathbf{D}_0(\mathbf{X}_\nu \otimes \boldsymbol{\Sigma}_0)\mathbf{D}_0'\mathbf{S}_\nu' & \mathbf{S}_\nu\mathbf{D}_0\mathbf{N}_0\mathbf{S}_\nu' \\ (\mathbf{S}_\nu\mathbf{D}_0\mathbf{N}_0\mathbf{S}_\nu')' & \mathbf{S}_\nu\boldsymbol{\Omega}\mathbf{S}_\nu' \end{bmatrix} = \begin{bmatrix} \mathbf{S}_\nu\mathbf{D}_0(\mathbf{X}_\nu \otimes \boldsymbol{\Sigma}_0)\mathbf{D}_0'\mathbf{S}_\nu' & \mathbf{S}_\nu(\mathbf{X}_\nu \otimes \mathbf{I}_\upsilon)\mathbf{S}_\nu' \\ (\mathbf{S}_\nu(\mathbf{X}_\nu \otimes \mathbf{I}_\upsilon)\mathbf{S}_\nu')' & \mathbf{S}_\nu\boldsymbol{\Omega}\mathbf{S}_\nu' \end{bmatrix}$$

holds and hence,

$$\begin{aligned} \det \begin{bmatrix} \mathbf{S}_\nu\mathbf{D}_0(\mathbf{X}_\nu \otimes \boldsymbol{\Sigma}_0)\mathbf{D}_0'\mathbf{S}_\nu' & \mathbf{S}_\nu\mathbf{D}_0\mathbf{N}_0\mathbf{S}_\nu' \\ (\mathbf{S}_\nu\mathbf{D}_0\mathbf{N}_0\mathbf{S}_\nu')' & \mathbf{S}_\nu\boldsymbol{\Omega}\mathbf{S}_\nu' \end{bmatrix} &= \det\{\mathbf{S}_\nu\mathbf{D}_0(\mathbf{X}_\nu \otimes \boldsymbol{\Sigma}_0)\mathbf{D}_0'\mathbf{S}_\nu'\} \\ \times\ \det\{(\mathbf{S}_\nu\boldsymbol{\Omega}\mathbf{S}_\nu') - (\mathbf{S}_\nu(\mathbf{X}_\nu \otimes \mathbf{I}_\upsilon)\mathbf{S}_\nu')(\mathbf{S}_\nu\mathbf{D}_0(\mathbf{X}_\nu \otimes \boldsymbol{\Sigma}_0)\mathbf{D}_0'\mathbf{S}_\nu')^{-1} \\ (\mathbf{S}_\nu(\mathbf{X}_\nu \otimes \mathbf{I}_\upsilon)\mathbf{S}_\nu')\} \ &\geq 0. \end{aligned} \tag{4.2.10}$$

An application of basic algebra to (4.2.10) confirms that the inequality relation in (4.2.6) obtains, as required. □

Theorem (4.2.1) tells us that use of the least squares estimator could result in a loss of efficiency but it does not, as it stands, provide a simple summary measure of the relative performance of the estimators $\hat{\theta}_{LS}$ and $\hat{\theta}_G$. Following what may be regarded as standard practice we will henceforth judge the relative efficiency of the two estimators by reference to the ratio of the determinants of their variance-covariance matrices. Thus, the asymptotic relative efficiency of $\hat{\theta}_{LS}$ with respect to $\hat{\theta}_G$ is given by the measure

$$\begin{aligned} e_0(\hat{\theta}_{LS}, \hat{\theta}_G) &= \{\det \boldsymbol{\Lambda}_G / \det \boldsymbol{\Lambda}_{LS}\} \\ &= \{\det(\mathbf{S}_\nu(\mathbf{X}_\nu \otimes \mathbf{I}_\upsilon)\mathbf{S}_\nu')\}^2\{\det(\mathbf{S}_\nu\boldsymbol{\Omega}\mathbf{S}_\nu') \times \det(\mathbf{S}_\nu(\mathbf{X}_\nu \otimes \boldsymbol{\Sigma}_0)\mathbf{S}_\nu')\}^{-1}. \end{aligned}$$

From Theorem (4.2.1) it is obvious that $0 \leq e_0(\hat{\theta}_{LS}, \hat{\theta}_G) \leq 1$. The choice itself is motivated by the notion that the volume of the concentration ellipsoids obtained from the asymptotic Gaussian distribution are proportional to the value of the determinant of the variance-covariance matrix so $e_0(\hat{\theta}_{LS}, \hat{\theta}_G)$ provides a natural measure of the relative precision of the two estimators, see Cramer(1977, Section 22.7). The

values of $\det \mathbf{\Lambda}_{LS}$ and $\det \mathbf{\Lambda}_G$ are referred to as generalised variances. Of course, generalised variances have been employed in the context of seemingly unrelated regression equations, (see, Zellner (1962), Chun-tu (1991)), to examine the relative merits of ordinary and Aitken generalised least squares, to which the following result is intimately related.

Proposition (4.2.1):

If $\mathbf{M}_0(z) = \mathbf{I}_v$ *then* $\mathrm{e}_0(\hat{\theta}_{LS}, \hat{\theta}_G) = 1$ *when either*

(i) the Kronecker indices are all equal, i.e. $n_1 = n_2 = \ldots = n_v$, *or*

(ii) the Kronecker indices, n_i, $i = 1, \ldots, v$ *are not equal but* $\mathbf{\Sigma}_0 = diag(\sigma_{11}, \ldots, \sigma_{vv})$.

Proof: If $n_1 = n_2 = \ldots = n_v$ then $\mathbf{A}_0(0) = \mathbf{M}_0(0) = \mathbf{I}_v$ and the corresponding selection matrix, $\mathbf{S}_\nu$, is of form

$$\mathbf{S}_\nu = \begin{pmatrix} \mathbf{I}_{pv^2} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \mathbf{I}_{pv^2} \end{pmatrix}. \tag{4.2.11}$$

Using an obvious notation, let $\mathbf{X}_\nu$ be partitioned into submatrices as

$$\mathbf{X}_\nu = \begin{pmatrix} X_{\alpha\alpha} & X_{\alpha\beta} & X_{\alpha\lambda} \\ X_{\beta\alpha} & X_{\beta\beta} & X_{\beta\lambda} \\ X_{\lambda\alpha} & X_{\lambda\beta} & X_{\lambda\lambda} \end{pmatrix}. \tag{4.2.12}$$

From (4.2.1), (4.2.11), (4.2.12) and the assumption that $\mathbf{M}_0(z) = \mathbf{I}_v$ we find that

$$\begin{aligned} \mathbf{\Lambda}_G^{-1} &= \mathbf{S}_\nu(\mathbf{X}_\nu \otimes \mathbf{\Sigma}_0^{-1})\mathbf{S}_\nu' \\ &= \mathbf{S}_\nu \begin{pmatrix} X_{\alpha\alpha} \otimes \mathbf{\Sigma}_0^{-1} & X_{\alpha\beta} \otimes \mathbf{\Sigma}_0^{-1} & X_{\alpha\lambda} \otimes \mathbf{\Sigma}_0^{-1} \\ X_{\beta\alpha} \otimes \mathbf{\Sigma}_0^{-1} & X_{\beta\beta} \otimes \mathbf{\Sigma}_0^{-1} & X_{\beta\lambda} \otimes \mathbf{\Sigma}_0^{-1} \\ X_{\lambda\alpha} \otimes \mathbf{\Sigma}_0^{-1} & X_{\lambda\beta} \otimes \mathbf{\Sigma}_0^{-1} & X_{\lambda\lambda} \otimes \mathbf{\Sigma}_0^{-1} \end{pmatrix} \mathbf{S}_\nu' \\ &= \begin{pmatrix} X_{\alpha\alpha} \otimes \mathbf{\Sigma}_0^{-1} & X_{\alpha\lambda} \otimes \mathbf{\Sigma}_0^{-1} \\ X_{\lambda\alpha} \otimes \mathbf{\Sigma}_0^{-1} & X_{\lambda\lambda} \otimes \mathbf{\Sigma}_0^{-1} \end{pmatrix}. \end{aligned}$$

Hence, noting that $\det(\mathbf{A} \otimes \mathbf{B}) = (\det \mathbf{A})^m \cdot (\det \mathbf{B})^n$ for any $n \times n$ and $m \times m$ matrices $\mathbf{A}$ and $\mathbf{B}$, respectively, we have

$$\begin{aligned}
\det \mathbf{\Lambda}_G &= \det(\chi \otimes \mathbf{\Sigma}_0^{-1})^{-1} \\
&= \det\{(\chi \otimes \mathbf{I}_v)(\mathbf{I}_{d(\nu)} \otimes \mathbf{\Sigma}_0^{-1})\}^{-1} \\
&= \det\{(\chi \otimes \mathbf{I}_v)\}^{-1}\{\det(\mathbf{I}_{d(\nu)}\}^v \cdot \{\det(\mathbf{\Sigma}_0^{-1})\}^{-d(\nu)} \\
&= \det(\chi \otimes \mathbf{I}_v)^{-1} \times (\det \mathbf{\Sigma}_0)^{d(\nu)}
\end{aligned}$$

where

$$\chi = \begin{pmatrix} X_{\alpha\alpha} & X_{\alpha\lambda} \\ X_{\lambda\alpha} & X_{\lambda\lambda} \end{pmatrix}$$

and $d(\nu)$ is used, as in the previous chapters, to denote the total number of freely varying parameters in the matrix pair $[\mathbf{A}(z)\colon \mathbf{M}(z)]$ or in the vector θ. Similarly,

$$\begin{aligned}
\mathbf{\Lambda}_{LS} &= \{\mathbf{S}_\nu(\mathbf{X}_\nu \otimes \mathbf{I}_v)\mathbf{S}_\nu'\}^{-1}\{\mathbf{S}_\nu\mathbf{D}_0(\mathbf{X}_\nu \otimes \mathbf{\Sigma}_0)\mathbf{D}_0'\mathbf{S}_\nu'\}\{\mathbf{S}_\nu(\mathbf{X}_\nu \otimes \mathbf{I}_v)\mathbf{S}_\nu'\}^{-1} \\
&= \{\chi \otimes \mathbf{I}_v\}^{-1}(\chi \otimes \mathbf{\Sigma}_0)\{\chi \otimes \mathbf{I}_v\}^{-1},
\end{aligned}$$

which follows by direct application of (4.2.11) and (4.2.12). It should be noted that the assumption $\mathbf{M}_0(z) = \mathbf{I}_v$ implies that $\mathbf{D}_0 = \mathbf{I}_{(2p+1)v^2}$ in the above expression. Thus,

$$\begin{aligned}
\det \mathbf{\Lambda}_{LS} &= \{\det(\chi \otimes \mathbf{I}_v)\}^{-2} \det(\chi \otimes \mathbf{I}_v)(\det \mathbf{\Sigma}_0)^{d(\nu)} \\
&= \det(\chi \otimes \mathbf{I}_v)^{-1}(\det \mathbf{\Sigma}_0)^{d(\nu)}
\end{aligned}$$

since $\chi \otimes \mathbf{\Sigma}_0 = (\chi \otimes \mathbf{I}_v)\,(\mathbf{I}_{d(\nu)} \otimes \mathbf{\Sigma}_0)$. Hence, the relative efficiency measure now becomes

$$\begin{aligned}
e_0(\hat{\theta}_{LS}, \hat{\theta}_G) &= \frac{\det \mathbf{\Lambda}_G}{\det \mathbf{\Lambda}_{LS}} \\
&= \frac{(\det \mathbf{\Sigma}_0)^{d(\nu)}}{\det(\chi \otimes \mathbf{I}_v)} \times \frac{\det(\chi \otimes \mathbf{I}_v)}{(\det \mathbf{\Sigma}_0)^{d(\nu)}} \\
&= 1.
\end{aligned}$$

To establish the second part of the proposition we proceed as follows. Let $\mathbf{X}_\nu$ be partitioned as in (4.2.12). Then

$$\begin{aligned}\boldsymbol{\Lambda}_G^{-1} &= \mathbf{S}_\nu(\mathbf{X}_\nu \otimes \boldsymbol{\Sigma}_0^{-1})\mathbf{S}_\nu' \\ &= \begin{pmatrix} \mathbf{S}_{\alpha(\nu)}(X_{\alpha\alpha} \otimes \boldsymbol{\Sigma}_0^{-1})\mathbf{S}_{\alpha(\nu)}' & \mathbf{S}_{\alpha(\nu)}(X_{\alpha\beta} \otimes \boldsymbol{\Sigma}_0^{-1})\mathbf{S}_{\beta(\nu)}' & \mathbf{S}_{\alpha(\nu)}(X_{\alpha\lambda} \otimes \boldsymbol{\Sigma}_0^{-1})\mathbf{S}_{\lambda(\nu)}' \\ \mathbf{S}_{\beta(\nu)}(X_{\beta\lambda} \otimes \boldsymbol{\Sigma}_0^{-1})\mathbf{S}_{\alpha(\nu)}' & \mathbf{S}_{\beta(\nu)}(X_{\beta\beta} \otimes \boldsymbol{\Sigma}_0^{-1})\mathbf{S}_{\beta(\nu)}' & \mathbf{S}_{\beta(\nu)}(X_{\beta\lambda} \otimes \boldsymbol{\Sigma}_0^{-1})\mathbf{S}_{\lambda(\nu)}' \\ \mathbf{S}_{\lambda(\nu)}(X_{\lambda\alpha} \otimes \boldsymbol{\Sigma}_0^{-1})\mathbf{S}_{\alpha(\nu)}' & \mathbf{S}_{\lambda(\nu)}(X_{\lambda\beta} \otimes \boldsymbol{\Sigma}_0^{-1})\mathbf{S}_{\beta(\nu)}' & \mathbf{S}_{\lambda(\nu)}(X_{\lambda\lambda} \otimes \boldsymbol{\Sigma}_0^{-1})\mathbf{S}_{\lambda(\nu)}' \end{pmatrix}\end{aligned}$$

since $\mathbf{S}_\nu = diag(\mathbf{S}_{\alpha(\nu)} : \mathbf{S}_{\beta(\nu)} : \mathbf{S}_{\lambda(\nu)}$. Some straightforward but somewhat extensive matrix manipulations that exploit the fact that $\boldsymbol{\Sigma}_0$ is presumed to be diagonal suggests that the determinant of $\boldsymbol{\Lambda}_G$ can be written as

$$\det\{\mathbf{S}_\nu(\mathbf{X}_\nu \otimes \boldsymbol{\Sigma}_0^{-1})\mathbf{S}_\nu'\}^{-1} = \{(\det \tilde{\mathbf{X}}_\nu)\}^{-1}\{\Pi_{i=1}^v(\sigma_{ii})^{\alpha_i}\}^{-1}$$

where $\tilde{\mathbf{X}}_\nu = \mathbf{S}_\nu(\mathbf{X}_\nu \otimes \mathbf{I}_v)\mathbf{S}_\nu'$ and $\sum_{i=1}^v \alpha_i = d(\nu)$. Consequently,

$$\begin{aligned}\det \boldsymbol{\Lambda}_G &= \det\{\mathbf{S}_\nu(\mathbf{X}_\nu \otimes \boldsymbol{\Sigma}_0^{-1})\mathbf{S}_\nu'\}^{-1} \\ &= \{(\det \tilde{\mathbf{X}}_\nu\}^{-1}\{\Pi_{i=1}^v(\sigma^{ii})^{\alpha_i}\}\end{aligned}$$

where $\sigma^{ii} = (\sigma_{ii})^{-1}$ for $i = 1, \ldots, v$. Hence, the relative efficiency is

$$\begin{aligned}e_0(\hat{\theta}_{LS}, \hat{\theta}_G) = \frac{\det \boldsymbol{\Lambda}_G}{\det \boldsymbol{\Lambda}_{LS}} &= \frac{\{(\det \tilde{\mathbf{X}}_\nu)\}^{-1}\{\Pi_{i=1}^v(\sigma^{ii})^{\alpha_i}\}}{\det\{\mathbf{S}_\nu(\mathbf{X}_\nu \otimes \mathbf{I}_v)\mathbf{S}_\nu'\}^{-2}\det\{\mathbf{S}_\nu(\mathbf{X}_\nu \otimes \boldsymbol{\Sigma}_0)\mathbf{S}_\nu'\}} \\ &= \frac{\{(\det \tilde{\mathbf{X}}_\nu)\}^{-1}\Pi_{i=1}^v(\sigma_{ii})^{\alpha_i}}{\{\det \tilde{\mathbf{X}}_\nu\}^{-2}\det\{\tilde{\mathbf{X}}_\nu\}\{\Pi_{i=1}^v(\sigma_{ii})^{\alpha_i}\}} \\ &= 1.\end{aligned}$$

□

To conclude this section, we make some few remarks. Though Proposition (4.2.1) dwelt on some specific cases when the relative efficiency is 1, it is interesting to note that the results do not hold in a situation where the Kronecker indices are as in (*ii*) but $\boldsymbol{\Sigma}_0$ is not a diagonal matrix. Once again the proof is straightforward and can be shown in an analogous manner to the proof of Proposition (4.2.1 ii).

To gain further understanding of the relative performance of the least squares estimator to the Gaussian one, it will be of interest to investigate the extent to which the above results carry-over to cases where the moving average operator, $\mathbf{M}_0(z)$, is chosen differently from an identity matrix. This issue will be addressed in the subsequent sections. An important part of this investigation, of course, will be through illustration via some numerical examples. Such an analysis should provide sufficient insight for understanding the asymptotic efficiency of least squares estimators. Before proceeding with the numerical considerations, it behoves on us to specify how the variance-covariance matrices $\mathbf{\Lambda}_G$ and $\mathbf{\Lambda}_{GLS}$ are to be evaluated and this is considered next.

4.3 Some Computational Aspects

In order to use $e_0(\hat{\theta}_{LS}, \hat{\theta}_G)$ to obtain relevant information about the relative efficiency of least squares estimators consideration has to be given first to the numerical evaluation of $\mathbf{\Lambda}_{LS}$ and $\mathbf{\Lambda}_G$. Although the frequency domain representation is useful as a tool to facilitate theoretical derivations and simplify interpretation, from the point of view of numerical evaluation it is more convenient to express the elements of matrices $\mathbf{\Lambda}_{LS}$ and $\mathbf{\Lambda}_G$ in terms of their time domain equivalents. For example, the elements of $\mathbf{X}_\nu$ are obtained from the coefficients of powers of z in the generating functions $\mathbf{\Gamma}_{yy}(z)$, $\mathbf{\Gamma}_{y\varepsilon}(z)$ and $\mathbf{\Gamma}_{\varepsilon\varepsilon}(z)$ where, for any two jointly stationary processes, say $\eta(t)$ and $\epsilon(t)$ with cross-covariance $\mathbf{\Gamma}_{\eta\epsilon}(r)$, $r = 0, \pm 1, \pm 2, \ldots$, we shall write

$$\mathbf{\Gamma}_{\eta\epsilon}(z) = \sum_{r=-\infty}^{\infty} \mathbf{\Gamma}_{\eta\epsilon}(r) z^r$$

for the covariance generating function. We will also recall that the cross-covariances can be recovered from the corresponding spectral densities through the inverse Fourier transforms. In particular,

$$\mathbf{\Gamma}_{0,yy}(r) = \mathcal{E}(\mathbf{y}(t)\mathbf{y}(t+r)')$$

$$= \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{ir\omega}(\mathbf{K}_0(e^{i\omega})\mathbf{\Sigma}_0\mathbf{K}_0(e^{i\omega})^*)\, d\,\omega.$$

since $\mathbf{y}(t) = \sum_{j\geq 0}\mathbf{K}(j)\varepsilon(t-j)$. The last expression can be equivalently written as

$$\mathbf{\Gamma}_{0,yy}(r) = \sum_{j=0}^{\infty}\mathbf{K}_0(j)\mathbf{\Sigma}_0\mathbf{K}_0(j+r)', \quad \mathbf{\Gamma}_{0,yy}(-r)' = \mathbf{\Gamma}_{0,yy}(r), \quad r \geq 0. \tag{4.3.1}$$

It follows from (4.3.1) that $\mathbf{\Gamma}_{0,yy}(r)$ corresponds to the coefficient of z^r in the power series expansion $\mathbf{P}_1(z)\mathbf{P}_1(z^{-1})'$ where $\mathbf{P}_1(z) = \mathbf{K}_0(z)\mathbf{\Sigma}_0^{\frac{1}{2}}$ and $\mathbf{K}_0(z) = \sum_{j\geq 0}\mathbf{K}_0(j)z^j$. Let $\mathbf{P}_2 = \mathbf{\Sigma}_0^{\frac{1}{2}}$. Similar relationships involving $\mathbf{P}_1(z)$ and $\mathbf{P}_2$ can be established for $\mathbf{\Gamma}_{y\varepsilon}(z)$ and $\mathbf{\Gamma}_{\varepsilon\varepsilon}(z)$, i.e. $\mathbf{\Gamma}_{y\varepsilon}(z) = \mathbf{P}_1(z)\mathbf{P}_2'$ and $\mathbf{\Gamma}_{\varepsilon\varepsilon}(z) = \mathbf{P}_2\mathbf{P}_2'$. Hence, we can obtain the $(2p+1)v \times (2p+1)v$ matrix $\mathbf{X}_\nu$ as follows. As in (4.2.12), we define

$$\mathbf{X}_\nu = \begin{bmatrix} X_{\alpha\alpha} & X_{\alpha\beta} & X_{\alpha\lambda} \\ X'_{\alpha\beta} & X_{\beta\beta} & X_{\beta\lambda} \\ X'_{\alpha\lambda} & X'_{\beta\lambda} & X_{\lambda\lambda} \end{bmatrix}$$

where $X_{\alpha\alpha}$ is $pv \times pv$, $X_{\alpha\beta}$ is $pv \times v$, $X_{\alpha\lambda}$ is $pv \times pv$, $X_{\beta\beta}$ is $v \times v$, $X_{\beta\lambda}$ is $v \times pv$, and $X_{\lambda\lambda}$ is $pv \times pv$ and the remaining subblocks are obtained by transposition as indicated. Then the submatrices of the above partition are determined as follows:

$(X_{\alpha\alpha})_{jk}$ is the coefficient of z^{k-j} in the expansion $\mathbf{P}_1(z)\mathbf{P}_1(z^{-1})'$,

$(X_{\alpha\lambda})_{jk}$ is the coefficient of z^{k-j} in the expansion $\mathbf{P}_1(z)\mathbf{P}_2'$ and

$(X_{\lambda\lambda})_{jk}$ is the coefficient of z^{k-j} in the expansion $\mathbf{P}_2\mathbf{P}_2'$.

The subscript jk denotes the elements in the $(j,k)^{th}$, $j,k = 1,\ldots,p$, $v \times v$ subblock of the indicated matrix;

$(X_{\alpha\beta})_{j1}$ is the coefficient of z^j in the expansion $\mathbf{P}_1(z)\{\mathbf{P}_1(z^{-1}) - \mathbf{P}_2\}'$, $j = 1,\ldots,p$, and

$(X_{\beta\lambda})_{1j}$ is the coefficient of z^j in the expansion $\mathbf{P}_1(z)\mathbf{P}_2'$, $j = 1,\ldots,p$.

The $v \times v$ matrix $X_{\beta\beta}$ is given by the coefficient of z^0 in the expansion $(\mathbf{P}_1(z) - \mathbf{P}_2)(\mathbf{P}_1(z) - \mathbf{P}_2)'$.

In order to determine $\mathbf{P}_1(z)$ it is necessary to evaluate the impulse response coefficients $\mathbf{K}_0(j)$, $j = 1,2,\ldots$ of the transfer function $\mathbf{K}_0(z) = \mathbf{A}_0(z)^{-1}\mathbf{M}_0(z)$.

Recall that $\mathbf{A}_0^{-1}(z) = Adj\ \mathbf{A}_0(z)/(\det \mathbf{A}_0(z))$, so we may write $\mathbf{K}_0(z)$ as $\{Adj\ \mathbf{A}_0(z)\mathbf{M}_0(z)\}/\det \mathbf{A}_0(z)$. The matrix in the curly bracket is a polynomial and hence is easily evaluated and the convolution of each entry with $1/\det \mathbf{A}_0(z)$ can be carried out in a straightforward fashion using the Euclidean division algorithm. It is the division by $\det \mathbf{A}_0(z)$, of course, that generates a power series for $\mathbf{K}_0(z)$. In order to accommodate this computationally we have chosen to truncate the expansion after 612 terms. For the cases we have considered this coincides with a point in the expansion where the elements of the impulse response coefficient are less than 10^{-32}. Thus infinite expansions of the type presented in (4.3.1) are replaced by corresponding finite summations and the effect of the truncation will be of the same order of magnitude. Given $\mathbf{X}_\nu$, constructed in the manner just described, the computation of $(\mathbf{X}_\nu \otimes \mathbf{I}_v)$ and $(\mathbf{X}_\nu \otimes \mathbf{\Sigma}_0)$ is straightforward and $\mathbf{\Lambda}_{LS}$ can be evaluated as a product of matrices as given in (4.2.2) once the appropriate rows and columns have been selected.

The construction of $\mathbf{\Lambda}_G$ can be carried out in a manner similar to $\mathbf{X}_\nu$. By a slight abuse of notation, let us redefine $\mathbf{P}_1(z) = \mathbf{K}_0(z)\mathbf{\Sigma}_0^{\frac{1}{2}} \otimes \mathbf{M}_0'(z)^{-1}\mathbf{\Sigma}_0^{-\frac{1}{2}}$ and set $\mathbf{P}_2(z) = \mathbf{\Sigma}_0^{\frac{1}{2}} \otimes \mathbf{M}_0'(z)^{-1}\mathbf{\Sigma}_0^{-\frac{1}{2}}$. We obtain the matrix $\mathbf{\Lambda}_G$ by using the partitioned form

$$\mathbf{\Lambda}_G^{-1} = \mathbf{S}_\nu \begin{pmatrix} \Omega_{\alpha\alpha} & \Omega_{\alpha\beta} & \Omega_{\alpha\lambda} \\ \Omega_{\alpha\beta}' & \Omega_{\beta\beta} & \Omega_{\beta\lambda} \\ \Omega_{\alpha\lambda}' & \Omega_{\beta\lambda}' & \Omega_{\lambda\lambda} \end{pmatrix} \mathbf{S}_\nu'$$

where the submatrices are obtained from appropriate combinations of the coefficients of powers of z in the expansions $\mathbf{P}_1(z)\mathbf{P}_1(z^{-1})'$, $\mathbf{P}_1(z)\mathbf{P}_2(z^{-1})'$ and $\mathbf{P}_2(z)\mathbf{P}_2(z^{-1})'$.

It should be mentioned, however, that some algorithms are readily available for generating the theoretical auto-covariances of an ARMA process see, for example, Ansley (1980), Shea (1987), Pate and Davies (1988) and Mittnik (1990). Though these algorithms would provide exact solution, they involve solving a large system of linear equations which make their practical applications somewhat cumbersome than the approach we have adopted here. In fact our method will yield comparable

results in terms of numerical accuracy if the truncation of the infinite sum in (4.3.1) is appropriately carried out.

We next consider the construction of the matrix $\mathbf{D}_0$ which is associated with the evaluation of $\mathbf{\Lambda}_{LS}$. In particular, the focus of interest will be on the evaluation of the matrix $\mathbf{V}^r$, $r = 1, \ldots, p$. We may recall that the matrix $\mathbf{V}^r$ can be recursively generated using $\mathbf{V}^r = \mathbf{V}^{r-1}\mathbf{V}, r = 1, \ldots, p$ where, of course, $\mathbf{V}^0 = \mathbf{I}_{(2p+1)v}$. Examination of the structure of $\mathbf{V}^r$ reveals that its computations can be carried out in a fairly efficient manner without the need to use the above recursive formula. To show this let $\mathbf{V}^r$, $r = 1, \ldots, p$, be partitioned as

$$\mathbf{V}^r = \begin{pmatrix} \mathbf{V}_{11,r} & \mathbf{V}_{12,r} & \mathbf{V}_{13,r} \\ \mathbf{V}_{21,r} & \mathbf{V}_{22,r} & \mathbf{V}_{23,r} \\ \mathbf{V}_{31,r} & \mathbf{V}_{32,r} & \mathbf{V}_{33,r} \end{pmatrix} \tag{4.3.2}$$

where $\mathbf{V}_{11,r}$ is $pv \times pv$, $\mathbf{V}_{12,r}$ is $pv \times v$, $\mathbf{V}_{13,r}$ is $pv \times pv$, $\mathbf{V}_{21,r}$ is $v \times pv$, $\mathbf{V}_{22,r}$ is $v \times v$, $\mathbf{V}_{23,r}$ is $v \times pv$, $\mathbf{V}_{31,r}$ is $pv \times pv$, $\mathbf{V}_{32,r}$ is $pv \times v$, and $\mathbf{V}_{33,r}$ is $pv \times pv$. For ease of exposition, it may be helpful to denote the $(i,j)^{th}$ $v \times v$ subblock of $\mathbf{V}_{11,r}$ by $\mathbf{V}_{11,r}(i,j)$ and the remaining submatrices in (4.3.2) are assumed to be defined in an analogous manner. We now turn to the explicit construction of submatrices in (4.3.2). It is particularly revealing that these submatrices lend themselves to a set of recursions. To motivate discussion in this direction consider the generating functions

$$\mathbf{\Psi}_{(\ell)}(z) = -\sum_{j=1}^{\ell} \tilde{\mathbf{A}}_0(j)\mathbf{\Psi}_{(\ell-j)}(z) + \mathbf{m}_{(\ell)}(z), \ \ell = 1, \ldots, r \tag{4.3.3}$$

and

$$\tilde{\mathbf{\Phi}}_{(\ell)}(z) = -\sum_{j=1}^{\ell} \tilde{\mathbf{A}}_0(j)\tilde{\mathbf{\Phi}}_{(\ell-j)}(z) - \mathbf{a}_{(\ell)}(z), \ \ell = 1, \ldots, r \tag{4.3.4}$$

for $r = 1, \ldots, p$ where

$$\tilde{\mathbf{A}}_0(j) = \mathbf{A}_0(0)^{-1}\mathbf{A}_0(j), \quad \tilde{\mathbf{M}}_0(j) = \mathbf{A}_0(0)^{-1}\mathbf{M}_0(j),$$

$$\mathbf{m}_{(\ell)}(z) = z^{-\ell} \sum_{j=\ell+1}^{p} \tilde{\mathbf{M}}_0(j) z^j, \quad \mathbf{m}_{(\ell)}(z) \equiv \mathbf{0} \text{ for } \ell + 1 > p,$$

$$\mathbf{\Psi}_{(0)}(z) = \sum_{j=0}^{p} \mathbf{\Psi}_{(0)}(j) z^j = \sum_{j=0}^{p} \tilde{\mathbf{M}}_0(j) z^j, \quad \mathbf{\Psi}_{(0)}(j) = \tilde{\mathbf{M}}_0(j) \text{ for all } j,$$

$$\mathbf{a}_{(\ell)}(z) = z^{-\ell} \sum_{j=\ell+1}^{p} \tilde{\mathbf{A}}_0(j) z^j, \quad \mathbf{a}_{(\ell)}(z) \equiv \mathbf{0}, \quad \ell + 1 > p,$$

and

$$\tilde{\mathbf{\Phi}}_{(0)}(z) = \sum_{j=0}^{p} \tilde{\mathbf{\Phi}}_{(0)}(j) z^j = -\sum_{j=0}^{p} \tilde{\mathbf{A}}_0(j) z^j, \quad \tilde{\mathbf{\Phi}}_{(0)}(j) = -\tilde{\mathbf{A}}_0(j) \text{ for all } j.$$

wherein, the subscript (0) is used to denote the values at which the above recursions are initiated. The elements of submatrices contained in $\mathbf{V}^r$ can now be formed via (4.3.3) and (4.3.4) in the following way:

$$\mathbf{V}_{11,r}(i,j) = \tilde{\mathbf{\Phi}}_{(r-i)}(j), \quad j = 1, \ldots, p; \; i = 1, \ldots, r$$

$$\mathbf{V}_{21,r}(1,j) = \tilde{\mathbf{\Phi}}_{(\ell)}(j), \quad j = 1, \ldots, p; \; \ell = 1, \ldots, r$$

and for $i = r+1, \ldots, p; \; j = 1, \ldots, p,$

$$\mathbf{V}_{11,r}(i,j) = \begin{cases} \mathbf{I}_v & \text{if } i - j = 1 \\ \mathbf{0} & \text{otherwise.} \end{cases}$$

In the same fashion,

$$\mathbf{V}_{13,r}(i,j) = \begin{cases} \mathbf{\Psi}_{(r-i)}(j), & j = 1, \ldots, p; \; i = 1, \ldots, r \\ \mathbf{0} & j = 1, \ldots, p; \quad \text{i= r+1, ..., p} \end{cases}$$

$$\mathbf{V}_{23,r}(1,j) = \mathbf{\Psi}_{(\ell)}(j), \quad j = 1, \ldots, p; \quad \ell = 1, \ldots, r,$$

and

$$\mathbf{V}_{33,r} = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{I}_{v(p-r)} & \mathbf{0} \end{pmatrix}.$$

The remaining submatrices in (4.3.2) are null.

For $r = 2, \ldots, p$ and $p > 1$, somewhat straightforward manipulations reveal that further computational savings can be made in the construction of $\mathbf{V}^r$. To be specific, we found that it may not be necessary to construct the submatrices $\mathbf{V}_{11,r}$, $\mathbf{V}_{13,r}$ and $\mathbf{V}_{33,r}$ via the generating functions (4.3.3) and (4.3.4). The submatrix $\mathbf{V}_{33,r}$ along with $\mathbf{V}_{12,r}$, $\mathbf{V}_{22,r}$ and $\mathbf{V}_{13,r}$ are found to be null. The construction of $\mathbf{V}_{11,r}$ and $\mathbf{V}_{13,r}$ is now accomplished via the following steps. Set $r = 2$.

Step 1 : Copy the first $(p-1)v$ rows of $\mathbf{V}_{11,r-1}$ and $\mathbf{V}_{13,r-1}$ into the last $(p-1)v$ rows of $\mathbf{V}_{11,r}$ and $\mathbf{V}_{13,r}$, respectively.

Step 2: Replace the first $v \times pv$ rows of $\mathbf{V}_{11,r}$ and $\mathbf{V}_{13,r}$ with $\mathbf{V}_{21,r-1}$ and $\mathbf{V}_{23,r-1}$, respectively.

Step 3: Construct $\mathbf{V}_{21,r}$ and $\mathbf{V}_{23,r}$ via the generating functions (4.3.3) and (4.3.4), respectively.

Step 4: Repeat the steps 1 - 3 for r =3, ..., p.

The evaluation of matrix $\mathbf{D}_0$ is thus achieved by conducting the matrix manipulations $\sum_{r=0}^{p} \mathbf{M}_0(r) \otimes \mathbf{V}^r$. In the computation of matrix $\mathbf{D}_0$, however, considerable savings in storage space can be achieved by performing the calculations in (4.3.3) and (4.3.4) in conjunction with those in steps (1-4) in a loop on r and accumulating results for $\mathbf{D}_0$ as one goes along. With this device only one storage location will be required for matrix $\mathbf{V}^r$ as opposed to $p+1$ storage locations that would have been used.

4.4 Numerical Illustrations

The theoretical discussion of the previous sections does not provide a guide as to the relative magnitude of the loss that may be incurred from using $\hat{\theta}_{LS}$ rather than $\hat{\theta}_G$. The purpose of the present section is to provide further insight into this

issue via the use of some numerical examples. We wish, therefore, to calculate the value of $e_0(\hat{\theta}_{LS}, \hat{\theta}_G)$ in a number of different circumstances.

A suite of programs for conducting the calculations of $\mathbf{\Lambda}_{LS}$ and $\mathbf{\Lambda}_G$ have been written in FORTRAN 77 and implemented on a SUN SPARC Station 1. The accuracy of the programs was checked in various special cases. As an example, consider a bivariate ARMA process with Kronecker indices, $n_1 = 1, \quad n_2 = 2$,

$$\mathbf{A}_0(z) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \begin{pmatrix} 1.336 & -1.776 \\ 0.156 & -0.673 \end{pmatrix} z + \begin{pmatrix} 0 & 0 \\ -0.498 & -0.018 \end{pmatrix} z^2,$$

$$\mathbf{M}_0(z) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

and innovation variance-covariance matrix $\mathbf{\Sigma}_0 = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}$ given by $\sigma_{11} = \sigma_{22} = 1.0$ and $\sigma_{12} = 0.0$. A value of 80169.96, correct to two decimal places, was obtained for both $\det \mathbf{\Lambda}_{LS}$ and $\det \mathbf{\Lambda}_G$. Clearly the relative efficiency measure, $e_0(\hat{\theta}_{LS}, \hat{\theta}_G)$, is one and is in accord with the theoretical results obtained in Section (4.2). We now turn to demonstrate by means of an example the case where the efficiency is less than unity. As an illustration let $\mathbf{A}_0(z)$ and $\mathbf{M}_0(z)$ be as defined in the last example and consider $\mathbf{\Sigma}_0$ for $\sigma_{11} = \sigma_{22} = 1$ and $\sigma_{12} = 0.2, 0.4, 0.6, 0.8$. We obtained results which are summarized in the following table where we have taken $\rho = \sigma_{12}$. It is clear from Table (4.1) that substantial loss in efficiency could occur under the

Table 4.1: Results for $e_0(\hat{\theta}_{LS}, \hat{\theta}_G)$ when $\mathbf{\Sigma}_0$ is not necessarily diagonal.

ρ	0.0	0.2	0.4	0.6	0.8
$e_0(\hat{\theta}_{LS}, \hat{\theta}_G)$	1.0	0.89	0.60	0.26	0.06

hypothesis of Proposition (4.2.1 (ii)) if the variance-covariance matrix $\mathbf{\Sigma}_0$ is not a

diagonal matrix.

We will now generalize to the situation where $\mathbf{M}_0(z)$ is different from an identity matrix, $\mathbf{I}_v$. As a motivation, consider the scalar ARMA process,

$$y(t) - \phi y(t-1) = \epsilon(t) - \varphi\epsilon(t-1), \quad \phi \neq \varphi, \quad |\phi| < 1, \ |\varphi| < 1,$$

where $\epsilon(t)$ is an innovation sequence with mean zero and variance σ^2. For simplicity we write the generating polynomials as $\phi(z) = 1 - \phi z$ and $\varphi(z) = 1 - \varphi z$ where, by definition, the zeros of $\phi(z)$ and $\varphi(z)$ are assumed to be outside the unit circle. In this special case, after some straightforward algbera, we find that

$$\mathbf{\Lambda}_{LS} = \frac{\sigma^2}{(\phi-\varphi)^2}\begin{pmatrix} (1-\phi\varphi)^2 & (1-\phi\varphi)(1-\phi^2) \\ (1-\phi\varphi)(1-\phi^2) & (1-\phi^2)(1+\varphi^2-2\phi\varphi) \end{pmatrix}. \quad (4.4.1)$$

According to Walker (1964), and Box and Jenkins (1976, p. 246), the corresponding asymptotic variance-covariance matrix of the Gaussian (Maximum likelihood) estimator is given by

$$\mathbf{\Lambda}_G = \frac{\sigma^2}{(\phi-\varphi)^2}\begin{pmatrix} (1-\phi^2)(1-\phi\varphi)^2 & (1-\phi^2)(1-\varphi^2)(1-\phi\varphi) \\ (1-\phi^2)(1-\varphi^2)(1-\phi\varphi) & (1-\phi\varphi)^2(1-\varphi^2) \end{pmatrix}. \quad (4.4.2)$$

Note that we have neglected the scaling factor T in (4.4.1) and (4.4.2), and such an exlusion, as easily checked, is inconsequential to the accuracy of $e_0(\hat{\theta}_{LS}, \hat{\theta}_G)$. Thus, the relative efficiency measure is obtained as $e_0(\hat{\theta}_{LS}, \hat{\theta}_G) = 1 - \varphi^2$. This result tells us that the relative efficiency decreases towards the invertibility boundary. This pattern of behaviour seems to be typical of the scalar ARMA models and has been alluded to elsewhere in the literature see, for example, Brockwell and Davis (1991, p. 253).

In the vector case, however, our results show that this condition is necessary but not sufficient for high efficiency. As an illustration, consider a bivariate ARMA process with Kronecker indices $n_1 = 1 \ \ n_2 = 2$,

$$\mathbf{A}_0(z) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \begin{pmatrix} -0.437 & 1.5604 \\ -0.062 & 1.1 \end{pmatrix} z + \begin{pmatrix} 0 & 0 \\ 0.613 & -0.22 \end{pmatrix} z^2,$$

$$\mathbf{M}_0(z) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \begin{pmatrix} 0.001 & 0.003 \\ 0.008 & 0.005 \end{pmatrix} z + \begin{pmatrix} 0 & 0 \\ 0.004 & 0.002 \end{pmatrix} z^2$$

and the innovation variance-covariance matrix given by $\sigma_{11} = \sigma_{22} = 1.0$ and $\sigma_{12} = 0.2$. A figure of just over 0.85 was obtained for the efficiency measure of the least squares estimator relative to the Gaussian estimator for this process. In view of the size of the coefficients in $\mathbf{M}_0(z)$ such an outcome might have been anticipated. Let us, therefore, change the magnitude of coefficients in the moving average operator to

$$\mathbf{M}_0(z) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \begin{pmatrix} 0.316 & -0.055 \\ 0.1807 & -0.31 \end{pmatrix} z + \begin{pmatrix} 0 & 0 \\ -0.5173 & 0.09 \end{pmatrix} z^2$$

whilst holding all other values of the parameters of the process fixed. We now record an efficiency value $e_0(\hat{\theta}_{LS}, \hat{\theta}_G)$ of a little under 0.26. What is, perhaps, surprising is that for both processes, the roots of the moving average operator, taken to two decimal places, are 200.00 and $-166.67 \pm 22.22i$. This is in contrast to what occurs in the univariate time series case where the relative efficiency has a direct relationship with the distance of the roots of $\mathbf{M}_0(z)$ from the unit circle. It is possible to construct moving average operators with roots farther away from the unit circle where the relative efficiency can be less than, say, 20% of its maximum. However, it is still evidently true that when the coefficients on the moving average operator are very small, as in the case above, the roots of the operator will also be well outside the unit circle and $e_0(\hat{\theta}_{LS}, \hat{\theta})$ close to unity.

Another feature which is, perhaps, worth mentioning concerns the invariance of the efficiency measure, $e_0(\hat{\theta}_{LS}, \hat{\theta}_G)$, to σ^2 in the univariate case. This issue was investigated in the vector case and it was shown in Salau (1992) that it is usually the case for the bivariate processes we considered that the structure of the variance-covariance matrix $\mathbf{\Sigma}_0$ does indeed influence the efficiency measure. As an illustration

consider a bivariate ARMA process with Kronecker indices $n_1 = 1$, $n_2 = 2$,

$$\mathbf{A}_0(z) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \begin{pmatrix} 1.336 & -0.371 \\ 2.065 & -0.673 \end{pmatrix} z + \begin{pmatrix} 0 & 0 \\ -0.498 & 0.018 \end{pmatrix} z^2$$

$$\mathbf{M}_0(z) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \begin{pmatrix} 0.001 & 0.003 \\ 0.008 & 0.005 \end{pmatrix} z + \begin{pmatrix} 0 & 0 \\ 0.004 & 0.002 \end{pmatrix} z^2$$

with $\sigma_{11} = \sigma_{22} = 1.0$, $\sigma_{12} = 0.0$ and $e(\hat{\theta}_{LS}, \hat{\theta}_G) = 1.0$. Set $\sigma_{11} = \sigma_{22} = 1.0$ and $\rho = \sigma_{12}$. Keeping the magnitude of the coefficients in $[\mathbf{A}_0(z) : \mathbf{M}_0(z)]$ unchanged, we will now assess the response of $e_0(\hat{\theta}_{LS}, \hat{\theta}_G)$ to different choice of ρ $(0 \leq \rho < 1)$. We now summarize our findings for this particular example in the following table.

Table 4.2: Results for $e_0(\hat{\theta}_{LS}, \hat{\theta}_G)$ for different choice of $\mathbf{\Sigma}_0$.

ρ	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$e_0(\hat{\theta}_{LS}, \hat{\theta}_G)$	0.99	0.97	0.88	0.75	0.59	0.42	0.26	0.13	0.05	0.01

It is seen from Table (4.2) that the efficiency measure $e_0(\hat{\theta}_{LS}, \hat{\theta}_G)$ is not invariant to the structure of $\mathbf{\Sigma}_0$ and in fact, the value of $e_0(\hat{\theta}_{LS}, \hat{\theta}_G)$ decreases rapidly to zero as $\mathbf{\Sigma}_0$ approaches singularity. Our view is not that one should use the structure of $\mathbf{\Sigma}_0$ in isolation to judge the efficiency of least squares estimator, but rather that it should be used in conjunction with an examination of $e(\hat{\theta}_{LS}, \hat{\theta}_G)$. However, the result is only suggestive since there is no true ARMA process and moreover, in practice, $\mathbf{\Sigma}_0$ will have to be estimated using the estimates of θ_0.

The above results imply that the pattern of behaviour of $e_0(\hat{\theta}_{LS}, \hat{\theta}_G)$ as a function of θ_0 in the vector case is not easily discernible. They do, however, convey a definite message for the practitioner. Although there are situations where the least squares estimator will perform well there are also subsets of the parameter space where the loss of efficiency can be considerable. Since in practice we will rarely, if

ever, know which situation we are dealing with it would seem advisable to evaluate $\hat{\theta}_G$. This will involve implementing Gauss-Newton type iterative calculations that can be initiated at $\hat{\theta}_{LS}$ as is discussed in Chapter 2 (Section (2.3)). Finally, we stress that all results in this Chapter have been asymptotic. This means that they will only apply when the sample size has become large. We do not, however, know how large it has to be. This caution must be kept in mind when the results of this Chapter are applied to practical cases.

Chapter 5

Generalized Least Squares Procedure for ARMA Estimation

5.1 Background and Chapter structure

Since the work of Box and Jenkins (1970) there has been a renewed and continuing interest in improved methods for estimating the parameters of scalar linear time series models see Hannan and Rissanen (1982) for example. Parallel developments in the vector case, where there are many inputs and outputs at each time point, have been widely studied. The problem of parameter estimation of such models has been considered by a number of authors, for example, Hannan (1970), Wilson (1973), Nicholls (1976), Anderson (1980), Jenkins and Alavi (1981), Spliid (1983), Hannan and Kavalieris (1984) and Shea (1987), amongst others. The estimation procedures proposed are generally based on optimization of the likelihood function constructed as if the data were from a Gaussian process or some approximation to that. In particular, emphasis has been placed on the use of exact or approximate maximum likelihood method to obtain the fully efficient (maximum likelihood) estimate. As was shown in Chapter 2, the computation of the Maximum likelihood estimator typically requires iterative procédures such as the Newton-Raphson or Gauss-Newton method which are based on the calculation of the gradient vector and the Hessian matrix of the log likelihood function, and specialized computations. Our interest in this problem arose, however, from the recent work of Koreisha and Pukkila (1990) who proposed a generalized least squares (GLS) procedure for estimating the parameters of ARMA models when the observable sequence is scalar. These authors have conducted simulation studies indicating the superiority of the GLS technique over other estimation procedures for the cases they examined. This result therefore raises the practical question of how best to estimate ARMA models in finite samples and the theoretical question of the nature of the differences between the alternative estimation procedures.

One of the main thrusts of this chapter is to examine the relationship between the GLS and Gaussian estimation schemes vis-a-vis the development of the theory for

the GLS procedure in the vector case. The latter has not been so readily forthcoming which may not be surprising in view of the inherently greater complexity of these models. Computationally the problem is much greater and, as this account will show, the theory is much more difficult.

We are concerned with structures generating the vector $\mathbf{y}(t)$ of v components of the form

$$\sum_{j=0}^{p} \mathbf{A}(j)\mathbf{y}(t-j) = \sum_{j=0}^{p} \mathbf{M}(j)\varepsilon(t-j) \tag{5.1.1}$$

where $\{\varepsilon(t)\}$ is the unobservable disturbance sequence satisfying (1.1.2) or, more generally, condition A and $p = \max(n_1, \ldots, n_v)$. Here $\mathbf{y}(t)$ is observed and is assumed to be stationary. As part of the generality sought for model (5.1.1), the matrix pair $[\mathbf{A}(z) : \mathbf{M}(z)]$ will be assumed to be in (reversed) echelon canonical form thus possessing the unique properties (1.2.8). Also, the standard assumptions of Chapter 1 are assumed to hold throughout this chapter. We introduce θ as a general coefficient parameter vector, so that, for example, $\theta = (\alpha' : \beta' : \lambda')'$ where α, β and λ are as defined in Chapter 2 (Section (2.2)). For later convenience, d_α, d_β and d_λ will be used, as before, to specify the dimension of vectors α, β and λ, respectively. We will also write θ_{GLS} for the GLS estimator of θ_0 (true value). The symbols hat and $\mathcal{E}$ will be used to signify an estimate of the indicated quantity and expectation operator, respectively.

The rest of this chapter is organized as follows. In the following section a detailed description of the GLS scheme is presented. The discussion emphasizes certain details that are exploited in the subsequent sections. Some theoretical considerations concerning the relationship between the GLS and the Gaussian estimation schemes are provided and the asymptotic convergence of the GLS estimator to the Gaussian estimator is established. This is followed in Section (5.4) by a detailed description of the alternative numerical implementation for the GLS procedure. We conclude the chapter with some simulation results which illustrate the different aspects of the procedure.

5.2 The Generalized Least Squares Scheme

Having observed $\mathbf{y}(t) = (y_1(t), \ldots, y_v(t))'$, $t = 1, \ldots, T$ assumed to be generated by a member of the class of ARMA processes (5.1.1), the GLS procedure is implemented using the following three stage estimation scheme which is a multivariate generalization of the method proposed by Koreisha and Pukkila (1990).

Stage 1: As an initiation of the procedure, an autoregression of order h is used to obtain $\hat{\varepsilon}_h(t)$ from

$$\mathbf{y}(t) = -\sum_{j=1}^{h} \hat{\mathbf{\Phi}}_h(j)\mathbf{y}(t-j) + \hat{\varepsilon}_h(t), \quad \hat{\mathbf{\Phi}}_h(0) = \mathbf{I}_v, \ \mathbf{y}(t) = 0 \ \ (t \leq 0) \tag{5.2.1}$$

and, for later convenience, define $\hat{\mathbf{\Sigma}}_h = T^{-1}\sum_{t=1}^{T} \hat{\varepsilon}_h(t)\hat{\varepsilon}_h(t)'$. In essence, the purpose of this stage is to provide estimates of the innovation sequence $\{\varepsilon(t)\}$, $t = 1, \ldots, T$, using the observed values, $\mathbf{y}(t), t = 1, \ldots, T$. In the light of assertion made in Hannan et al (1989) that there is no theory that unequivocally states which criterion must always be used to determine the order of an autoregressive approximation the choice of h is problematic. Pukkila et al (1990) have provided some basis for fixing h at $T^{\frac{1}{2}}$ through extensive simulation studies in small sample situations. The solution of (5.2.1) can then be obtained via a multivariate generalization of the Burg's algorithm (B3.2.4) and the virtues for using this algorithm have already been discussed in Appendix B. However, in order, asymptotically, to analyse this procedure we will suppose that h is chosen as a function of T such that $h \leq (\log T)^{\tau}$, $1 < \tau < \infty$. This limitation in the growth of h allows us to appeal to results on convergence rates, which are uniform in h, that are needed for the asymptotic results we obtain later. We will return to this issue in Section (5.5).

In order to simplify exposition in what that follows it will be assumed that the Kronecker indices, $\nu = \{n_1, \ldots, n_v\}$, are known *a priori* or have been consistently identified using one of the available procedures as is discussed in Chapter 1 (Section (1.3)). To facilitate the presentation of the second stage, let us substitute $\hat{\varepsilon}_h(t)$ for

$\varepsilon(t)$ in (5.1.1) to obtain the regression model

$$\mathbf{y}(t) - \hat{\varepsilon}_h(t) = -(\mathbf{A}(z) - \mathbf{I}_v)\mathbf{y}(t) + (\mathbf{M}(z) - \mathbf{I}_v)\hat{\varepsilon}_h(t) + \mathbf{e}(t) \tag{5.2.2}$$

where $\{\mathbf{e}(t)\}$ is assumed to be a sequence of error terms, and $\mathbf{A}(z) = \sum_{j=0}^{p} \mathbf{A}(j)z^j$ and $\mathbf{M}(z) = \sum_{j=0}^{p} \mathbf{M}(j)z^j$, as before, denote the generating functions. Now, applying the rule $vec\mathbf{ABC} = (\mathbf{C}' \otimes \mathbf{A})vec\mathbf{B}$ to both sides of equation (5.2.2), using arguments similar to those leading to (3.2.16), results in the following equation

$$\begin{aligned} \mathbf{y}(t) - \hat{\varepsilon}_h(t) \;=\; & \{-\zeta_p(z)' \otimes \mathbf{y}(t)' \otimes \mathbf{I}_v\}\mathbf{S}'_{\alpha(\nu)}\alpha - \{(\mathbf{y}(t) - \hat{\varepsilon}_h(t))' \otimes \mathbf{I}_v\}\mathbf{S}'_{\beta(\nu)}\beta \\ & + \{\zeta_p(z)' \otimes \hat{\varepsilon}_h(t)' \otimes \mathbf{I}_v\}\mathbf{S}'_{\lambda(\nu)}\lambda + \mathbf{e}(t) \end{aligned}$$

which can be succintly written as

$$\mathbf{y}(t) - \hat{\varepsilon}_h(t) = \mathbf{Y}_\nu(t)\alpha + \hat{\mathbf{F}}_\nu(t)\beta + \hat{\mathbf{E}}_\nu(t)\lambda + \mathbf{e}(t)$$

where $\mathbf{Y}_\nu(t)$, $\hat{\mathbf{F}}_\nu(t)$ and $\hat{\mathbf{E}}_\nu(t)$ are as defined in Chapter 3 (Section (3.2.2)) and we will also let $\hat{\mathbf{R}}_\nu(t)' = (\mathbf{Y}_\nu(t) : \hat{\mathbf{F}}_\nu(t) : \hat{\mathbf{E}}_\nu(t))$ defines the regressor variables. These constructions therefore lead to stage 2.

Stage 2: Regress $\mathbf{y}(t) - \hat{\varepsilon}_h(t)$ on the regressor variables $\hat{\mathbf{R}}_\nu(t)' = (\mathbf{Y}_\nu(t) : \hat{\mathbf{F}}_\nu(t) : \hat{\mathbf{E}}_\nu(t)), t = 1, \ldots, T$, to obtain $\hat{\theta}_{LS}$ $(\hat{\mathbf{A}}_{LS}(j) : \hat{\mathbf{M}}_{LS}(j))$, the least squares estimate of $\theta = (\alpha' : \beta' : \lambda')'$, which minimizes the criterion function $\sum_t \|\mathbf{y}(t) - \hat{\varepsilon}_h(t) - \hat{\mathbf{R}}_\nu(t)'\theta\|^2$ with respect to the freely varying coefficients in $[\mathbf{A}(z) : \mathbf{M}(z)]$.

In practice, however, $\hat{\varepsilon}_h(t)$ may exhibit a behaviour that is not intuitively compatible with either assumption (1.1.2) or condition A. This phenomenon may occur when T is small or h is not sufficiently large to guarantee that $\hat{\varepsilon}_h(t)$ approximate the true innovation $\{\varepsilon(t)\}$ in a manner made explicit in Lemma (3.2.4). In such a situation, $\hat{\varepsilon}_h(t)$ will contain random variations due to estimation errors and consequently, $\hat{\theta}_{LS} = (\hat{\mathbf{A}}_{LS}(j) : \hat{\mathbf{M}}_{LS}(j))$ may be biased and, of course, inefficient. In this view, and by way of motivation, the innovation sequence may be decomposed into two components as

$$\varepsilon(t) = \hat{\varepsilon}_h(t) + \mathbf{e}(t). \tag{5.2.3}$$

Hence, substituting the right-hand side of (5.2.3) for $\varepsilon(t)$ in (5.1.1) results in a regression equation

$$\mathbf{y}(t) - \hat{\varepsilon}_h(t) = \hat{\mathbf{R}}_\nu(t)'\theta + \epsilon(t) \tag{5.2.4}$$

where $\epsilon(t) = \sum_{j=0}^{p} \mathbf{M}(j)\mathbf{e}(t-j),\ t = 1, \ldots, T.$

If we let $\gamma_{\epsilon\epsilon}(k)$ be the $v \times v$ matrix autocovariance of $\epsilon(t)$ at lag k then

$$\gamma_{\epsilon\epsilon}(k) = \gamma_{\epsilon\epsilon}(-k)' = \sum_{i=0}^{p}\sum_{j=0}^{p} \mathbf{M}(i)\mathbf{\Sigma}_e(i, j+k)\mathbf{M}(j)',$$

where $\mathbf{\Sigma}_e(i,j) = \lim_{T\to\infty} T^{-1}\sum_{t=1}^{T} \mathbf{e}(t-i)\mathbf{e}(t-j)'$ for $i, j \geq 0$. From the results presented in Lemma (3.2.4), however, $T^{-1}\sum_{t=1}^{T}\{\varepsilon(t-i) - \hat{\varepsilon}_h(t-i)\}\{\varepsilon(t-j) - \hat{\varepsilon}_h(t-j)\}'$, ignoring the terms involving $o_p(1)$, can be approximated by $\frac{hv}{T}\delta_{i,j}\mathbf{\Sigma}$ to sufficient accuracy for large T. Hence, $\epsilon(t)$, as in econometrics literature, can be treated as moving average residuals since $\gamma_{\epsilon\epsilon}(k)$ is proportional to

$$\begin{cases} \sum_{j=0}^{p-k} \mathbf{M}(j)\mathbf{\Sigma}\mathbf{M}(j+k)' & ,(0 \leq k \leq p) \\ \mathbf{0} & ,(\ k > p). \end{cases} \tag{5.2.5}$$

Thus, the efficient estimation of θ in (5.2.4) involves the use of the generalized least squares technique. To motivate this and the subsequent step we let

$$\begin{aligned} \epsilon &= [\epsilon(p+1)', \epsilon(p+2)', \ldots, \epsilon(T)']' \\ &= [\epsilon(t)]_{t=p+1}^{T} \end{aligned}$$

denote the vector formed by $\mathcal{T} = T - p$ consecutive observations on $\epsilon(t)$ from $t = p+1$ to T. Then set $\mathbf{\Omega}_{GLS} = \mathcal{E}(\epsilon\epsilon')$ where the script letter $\mathcal{E}$, as earlier indicated, signifies the expectation operator and also note that $\mathbf{\Omega}_{GLS}$ is a $\mathcal{T}v \times \mathcal{T}v$ symmetric block-Toeplitz matrix whose first row of $v \times v$ blocks is $\gamma_{\epsilon\epsilon}(0), \gamma_{\epsilon\epsilon}(1), \ldots, \gamma_{\epsilon\epsilon}(\mathcal{T}-1)$. In practice, the elements of $\mathbf{\Omega}_{GLS}$ are rarely known and will have to be estimated from the observed quantities. This matrix can be consistently estimated by replacing $\mathbf{M}(j)$ by $\hat{\mathbf{M}}_{LS}(j)$ and $\mathbf{\Sigma}$ by $\hat{\mathbf{\Sigma}}_h$ in (5.2.5). Denote the estimate of $\mathbf{\Omega}_{GLS}$ by $\hat{\mathbf{\Omega}}_{GLS}$ and since $\hat{\gamma}(k) = 0$ for $k > p$, then it is easily verified that $\hat{\mathbf{\Omega}}_{GLS}$ has a block band

structure. If we set

$$\mathbf{X}_{GLS} = [\mathbf{R}_\nu(t)']_{t=p+1}^{T} = \mathbf{X}\,\mathbf{S}'_\nu,$$

and

$$\mathbf{Y}_{GLS} = [\mathbf{y}(t) - \hat{\varepsilon}_h(t)]_{t=p+1}^{T}$$

where

$$\mathbf{X} = [-\zeta_p(z)' \otimes \mathbf{y}(t)' \otimes \mathbf{I}_v : -(\mathbf{y}(t) - \hat{\varepsilon}_h(t))' \otimes \mathbf{I}_v : \zeta_p(z)' \otimes \hat{\varepsilon}_h(t)' \otimes \mathbf{I}_v]_{t=p+1}^{T},$$

and

$$\mathbf{S}'_\nu = diag(\mathbf{S}'_{\alpha(\nu)} : \mathbf{S}'_{\beta(\nu)} : \mathbf{S}'_{\lambda(\nu)}),$$

then we have

Stage 3: Evaluate the GLS estimator as

$$\hat{\theta}_{GLS} = (\mathbf{X}'_{GLS}\hat{\mathbf{\Omega}}^{-1}_{GLS}\mathbf{X}_{GLS})^{-1}(\mathbf{X}'_{GLS}\hat{\mathbf{\Omega}}^{-1}_{GLS}\mathbf{Y}_{GLS}). \tag{5.2.6}$$

This completes the third stage of the scheme but one problem remains, however, in connection with (5.2.6) and that concerns the evaluation of the variance-covariance matrix, $\hat{\mathbf{\Omega}}_{GLS}$. The explicit construction of $\hat{\mathbf{\Omega}}_{GLS}$ in a manner just described may not be convenient from the computational point of view and it therefore seems of practical interest to know whether the computation of $\hat{\mathbf{\Omega}}_{GLS}$ can be accomplished in a fairly simple manner. Taking advantage of the problem structure, we find that $\hat{\Omega}_{GLS}$ can be re-expressed as

$$\hat{\mathbf{\Omega}}_{GLS} = \hat{\mathbf{C}}_{GLS}(\mathbf{I}_T \otimes \hat{\mathbf{\Sigma}}_h)\hat{\mathbf{C}}'_{GLS} \tag{5.2.7}$$

where $\mathbf{I}_T$ is a $T \times T$ identity matrix and $\mathbf{C}_{GLS}$ is a $\mathcal{T}v \times Tv$ full row rank matrix which has a block band structure. The first diagonal block of the band has a $v \times v$ matrix $\mathbf{M}(p)$ as its elements, $\mathbf{M}(p-j+1)$ $(j = 2, \ldots, p+1)$ down the subsequent

j^{th} diagonal block and zeros elsewhere. Thus, the matrix $\mathbf{C}_{GLS}$ has an explicit form

$$\mathbf{C}_{GLS} = \begin{bmatrix} \mathbf{M}(p) & \dots & \mathbf{M}(0) & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{M}(p) & \dots & \mathbf{M}(0) & \dots & \mathbf{0} \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \mathbf{0} & \dots & \mathbf{0} & \mathbf{M}(p) & \dots & \mathbf{M}(0) \end{bmatrix}. \tag{5.2.8}$$

In concluding this section, we need to emphasize that the exclusion of the constant factor $\frac{hv}{T}$ from the formula defining $\gamma_{\epsilon\epsilon}(k)$ in (5.2.5), as can be observed from (5.2.6), is of no consequence to the accuracy of $\hat{\theta}_{GLS}$. Also, it is worthwhile to note that when the observed sequence $\mathbf{y}(t)$ is scalar, $\hat{\sigma}_h^2$ (a univariate analogue of $\hat{\mathbf{\Sigma}}_h$) does not contribute to the expression (5.2.6) but such does not seem to hold in the vector case. Finally, there are virtues and vices of the above procedure that might be pointed out. Using the GLS procedure at stage 3 it is only the covariance matrix, $\mathbf{\Omega}_{GLS}$, of the process $\epsilon(t)$ that needs to be formed for $\mathbf{M}(j) = \hat{\mathbf{M}}_{LS}(j)$, the moving average parameters from stage 2, and this matrix is subsequently updated using $\hat{\mathbf{M}}(j)$ corresponding to $\hat{\theta}_{GLS}^{(i)}$, $i = 1, 2, \dots$, if (5.2.6) were to be iterated, as it should be in small sample situations. This matrix is not altered when the zeros of $\det \hat{\mathbf{M}}(z)$ are *flipped*, that is, a zero z_j say, is replaced by $\bar{z}_j^{-1}$ if $|z_j| < 1$. This obviates the necessity of flipping zeros that are inside $|z| = 1$ which cannot be avoided in a filtering procedure such as the Gaussian estimation scheme (see Chapter 2, Section (2.3)). On the other hand, the GLS procedure uses the estimated innovation sequence from stage 1 throughout and there will be little change from an iteration of this technique since to effect that change one would need improved estimates of the innovation sequence. However, as $T \to \infty$, the GLS technique will give efficient estimates but the computational effort required to obtain $\hat{\theta}_{GLS}$ will be very severe since the evaluation of $\hat{\theta}$ via (5.2.6) involves the inversion of a $\mathcal{T}v \times \mathcal{T}v$ matrix $\hat{\mathbf{\Omega}}_{GLS}$. A method which circumvents this huge matrix inversion is proposed in Section (5.4) but, before discussing this technique, we shall first dwell on some theoretical issues.

5.3 Theoretical Considerations

In this section we investigate the interconnections between the GLS and Gaussian estimation schemes. The logical question of how GLS procedure performs in relation to the Gaussian estimation scheme can be answered in several ways. One way is to examine the effects of different choices within the algorithm on the parameter estimates, asymptotically or in finite samples. From this, we shall be able to make a comparative assessment of the performance of the two procedures. A useful device for carrying out this evaluation is via simulation studies. With this, the effect of the initial conditions and the choice of preliminary estimates, which seem to differ for both the GLS and the Gaussian estimation schemes, can also be examined. However, a serious limitation of simulation is that it may be of limited value or inconclusive. In other words, it may be difficult to tell whether the conclusion based on simulation results has universal implications or is simply a product of the particular set of data used. Thus, to obtain results of more general validity, it is necessary to analyze the asymptotic properties of the estimates obtained from the algorithms. The convergence properties of the estimates and their limiting distributions will give an indication of the performance of the algorithms.

To carry out a valid and useful assessment of the GLS algorithm, it must be emphasized that the third stage of the procedure works on the presumption that the error process, $\mathbf{e}(t)$, is uncorrelated with constant variance-covariance matrix. Hence, the covariance matrix of the errors in (5.2.4), as noted in the last section, is *known* up to a scalar factor, and hence the GLS procedure may be used to get more efficient estimates. It is to be observed, of course, that using $\mathbf{y}(t) - \hat{\varepsilon}_h(t)$ as a regressand in place of $\mathbf{y}(t)$ employed in the Gaussian case, may improve the parameter estimates in finite samples and it is quite possible that it could also have some second order effects which may prove useful in small samples. It can be argued via Lemma (3.2.4) that $\hat{\varepsilon}_h(t)$ will provide a good approximation to $\varepsilon(t)$ for sufficiently large T

and this, in turn, tends to suggest that the subtraction of $\hat{\varepsilon}_h(t)$ from $\mathbf{y}(t)$ to form the regressand may have no asymptotic effect and hence (5.2.4) can be rewritten as

$$\mathbf{y}(t) = \hat{\mathbf{R}}_\nu(t)'\theta + \varepsilon(t) + o_p(\sqrt{\frac{hv}{T}}). \tag{5.3.1}$$

Thus, as $T \to \infty$ (5.3.1) will reduce to the form

$$\mathbf{y}(t) = \mathbf{R}_\nu(t)'\theta + \varepsilon(t) \tag{5.3.2}$$

where $\mathbf{R}_\nu(t)$ is $\hat{\mathbf{R}}_\nu(t)$ with $\varepsilon(t)$ replacing $\hat{\varepsilon}_h(t)$. In this regard, we should attain asymptotic efficiency since working with (5.3.2) is equivalent to estimating (2.1.1). The arguments put forward here point to the fact that the GLS procedure is essentially the same as the Gaussian estimation scheme, the difference being due to some effects resulting from different initiations, which based on previous discussions, will disappear asymptotically but may be important in practice. This shows the heuristic nature of the arguments we have used which needs to be supported by further analysis.

From the foregoing discussion it is evident that there is an inherent close relationship between the GLS and Gaussian estimation schemes and this interconnection will now be formally established in the following subsections.

5.3.1 Equivalent interpretation

Consider (5.2.7) and observe that

$$\hat{\mathbf{\Omega}}_{GLS}^{-1} = \hat{\mathbf{C}}_{GLS}'^{\dagger}(\mathbf{I}_T \otimes \hat{\mathbf{\Sigma}}_h^{-1})\hat{\mathbf{C}}_{GLS}^{\dagger} \tag{5.3.3}$$

where $\hat{\mathbf{C}}_{GLS}^{\dagger} = \hat{\mathbf{C}}_{GLS}'(\hat{\mathbf{C}}_{GLS}\hat{\mathbf{C}}_{GLS}')^{-1}$ is the Moore-Penrose generalized inverse of $\mathbf{C}_{GLS}$, a matrix of full row rank $\mathcal{T}v$. For a detailed discussion of the property of generalized inverses see, for example, Farell (1985, p. 132). Now substituting (5.3.3) into (5.2.6) gives an expression of the form

$$\begin{aligned}\hat{\theta}_{GLS} &= \{\mathbf{X}_{GLS}'\hat{\mathbf{C}}_{GLS}'^{\dagger}(\mathbf{I}_T \otimes \hat{\mathbf{\Sigma}}_h^{-1})\hat{\mathbf{C}}_{GLS}^{\dagger}\mathbf{X}_{GLS}\}^{-1}\{\mathbf{X}_{GLS}'\hat{\mathbf{C}}_{GLS}'^{\dagger}(\mathbf{I}_T \otimes \hat{\mathbf{\Sigma}}_h^{-1})\hat{\mathbf{C}}_{GLS}^{\dagger}\mathbf{Y}_{GLS}\} \\ &= \{\mathbf{Z}_{GLS}'(\mathbf{I}_T \otimes \hat{\mathbf{\Sigma}}_h^{-1})\mathbf{Z}_{GLS}\}^{-1}\{\mathbf{Z}_{GLS}'(\mathbf{I}_T \otimes \hat{\mathbf{\Sigma}}_h^{-1})\mathbf{W}_{GLS}\}\end{aligned} \tag{5.3.4}$$

where $\mathbf{Z}_{GLS} = \hat{\mathbf{C}}^{\dagger}_{GLS}\mathbf{X}_{GLS}$ and $\mathbf{W}_{GLS} = \hat{\mathbf{C}}^{\dagger}_{GLS}\mathbf{Y}_{GLS}$. Note that the calculation of $\mathbf{Z}_{GLS}$ and $\mathbf{W}_{GLS}$ amounts to obtaining the solutions of $\hat{\mathbf{C}}_{GLS}\mathbf{Z}_{GLS} = \mathbf{X}_{GLS}$ and $\hat{\mathbf{C}}_{GLS}\mathbf{W}_{GLS} = \mathbf{Y}_{GLS}$, respectively. These equations tell us that the components of $\mathbf{Z}_{GLS}$ and $\mathbf{W}_{GLS}$ can be recursively generated as those of $\mathbf{Z}_G$ and $\mathbf{W}_G$ (defined in Chapter 2, Subsection (2.3.1)) from the elements of quasi-derivative processes $\mathbf{d}(t),\ t = 1,\ldots,T$ given by

$$\begin{aligned}\sum_{j=0}^{p}\hat{\mathbf{M}}_{LS}(j)\mathbf{d}(t-j) &= (-\zeta_p(z)'\otimes\mathbf{y}(t)'\otimes\mathbf{I}_v : -(\mathbf{y}(t)-\hat{\varepsilon}_h(t))'\otimes\mathbf{I}_v \\ &\quad : \zeta_p(z)'\otimes\hat{\varepsilon}_h(t)'\otimes\mathbf{I}_v),\ t = p+1,\ldots,T \qquad (5.3.5)\end{aligned}$$

where the implied initial values $\mathbf{d}(1),\ldots,\mathbf{d}(p)$ are given by the first pv rows of $\hat{\mathbf{C}}^{\dagger}_{GLS}\mathbf{X}$. Consequently we see that (5.3.4) and (5.3.5) provide a representation of the GLS estimator that parallels the construction of the Gaussian estimator via a multivariate regression involving the derivative processes determined in (2.3.9).

From the results presented in this subsection we find that although the GLS and Gaussian estimation procedures are derived from different view points they admit common interpretation. Their apparent differences arise from the choice of initial estimates, treatment of implied or explicit initial conditions and use of effective sample size. That these apparent differences can be of significance in terms of small, finite sample performance has already been attested to in the literature see, Koreisha and Pukkila (1990). Given these differences, however, it is desirable to know if the two procedures will always produce different results as the sample size T tends to infinity. As we will establish in the next subsection, the two estimators are in fact asymptotically equivalent in the sense that they will yield the same estimates, almost surely, as sample size increases.

5.3.2 Asymptotic Convergence

In this section we wish to establish the asymptotic convergence of the GLS estimator to the Gaussian estimator. As part of the preliminary investigation we

shall require the following lemmas which allow simplifications to be made in the subsequent derivations. To state the lemmas succintly it is helpful to introduce the following notation. Define

$$\tilde{\mathbf{X}}_{GLS} = \begin{pmatrix} \mathbf{0} \\ \dots \\ \mathbf{X}_{GLS} \end{pmatrix}, \quad \tilde{\mathbf{Y}}_{GLS} = \begin{pmatrix} \mathbf{0} \\ \dots \\ \mathbf{Y}_{GLS} \end{pmatrix}$$

and

$$\tilde{\mathbf{\Omega}}_{GLS} = \begin{pmatrix} \mathbf{I}_{pv} & \vdots & \mathbf{0} \\ \dots & \cdot & \dots \\ \mathbf{0} & \vdots & \mathbf{\Omega}_{GLS} \end{pmatrix}$$

so that (5.2.6) may be rewritten as

$$\hat{\theta}_{GLS} = (\tilde{\mathbf{X}}'_{GLS}\hat{\tilde{\mathbf{\Omega}}}^{-1}_{GLS}\tilde{\mathbf{X}}_{GLS})^{-1}(\tilde{\mathbf{X}}'_{GLS}\hat{\tilde{\mathbf{\Omega}}}^{-1}_{GLS}\tilde{\mathbf{Y}}_{GLS}) \tag{5.3.6}$$

where $\tilde{\mathbf{X}}_{GLS}$, $\tilde{\mathbf{Y}}_{GLS}$ and $\tilde{\mathbf{\Omega}}_{GLS}$ are matrices of dimensions $Tv \times d(\nu)$, $Tv \times 1$ and $Tv \times Tv$, respectively. Put $\bar{\mathbf{Y}}_G = [\mathbf{y}(t) - \hat{\mathbf{e}}_\theta(t)]_{t=1}^T$ and define

$$\bar{\theta}_G = (\mathbf{X}'_G\hat{\mathbf{\Omega}}_G^{-1}\mathbf{X}_G)^{-1}(\mathbf{X}'_G\hat{\mathbf{\Omega}}_G^{-1}\bar{\mathbf{Y}}_G). \tag{5.3.7}$$

Also, let $\|\cdot\|$ denote the Euclidean matrix norm. We can now state the required lemmas.

Lemma (5.3.1)

Under the current regularity conditions

$$\lim_{T\to\infty} T^{-1}\{\tilde{\mathbf{X}}'_{GLS}\hat{\tilde{\mathbf{\Omega}}}^{-1}_{GLS}\tilde{\mathbf{X}}_{GLS} : \mathbf{X}'_G\hat{\mathbf{\Omega}}_G^{-1}\mathbf{X}_G\} = (\mathbf{Q}_1 : \mathbf{Q}_2)$$

exists where $\mathbf{Q}_1$ *and* $\mathbf{Q}_2$ *are non-singular matrices.*

Proof: This result follows immediately from egordicity of $\{\mathbf{y}(t)\}$ and $\{\varepsilon(t)\}$ using a mode of argument employed in Fuller (1976, chapt. 9). □

Lemma (5.3.2)

If $\bar{\theta}_G$ and $\hat{\theta}_G$ are as defined in (5.3.7) *and* (2.3.13), *respectively, then* $\|\hat{\theta}_G - \bar{\theta}_G\| = O(Q_T)$ *a.s*

Proof: First observe that

$$\begin{aligned}
\lim_{T\to\infty}\|\hat{\theta}_G - \bar{\theta}_G\| &= \lim_{T\to\infty}\|(\frac{\mathbf{X}_G'\hat{\mathbf{\Omega}}_G^{-1}\mathbf{X}_G}{T})^{-1}(\frac{\mathbf{X}_G'\hat{\mathbf{\Omega}}_G^{-1}\mathbf{Y}_G}{T}) - (\frac{\mathbf{X}_G'\hat{\mathbf{\Omega}}_G^{-1}\mathbf{X}_G}{T})^{-1}(\frac{\mathbf{X}_G'\hat{\mathbf{\Omega}}_G^{-1}\bar{\mathbf{Y}}_G}{T})\| \\
&= \lim_{T\to\infty}\|(\frac{\mathbf{X}_G'\hat{\mathbf{\Omega}}_G^{-1}\mathbf{X}_G}{T})^{-1}\{\frac{\mathbf{X}_G'\hat{\mathbf{\Omega}}_G^{-1}}{T}(\mathbf{Y}_G - \bar{\mathbf{Y}}_G)\}\| \\
&= \lim_{T\to\infty}\|(\frac{\mathbf{X}_G'\hat{\mathbf{\Omega}}_G^{-1}\mathbf{X}_G}{T})^{-1}\{\frac{\mathbf{X}_G'\hat{\mathbf{\Omega}}_G}{T}[\mathbf{y}(t) - \hat{\mathbf{e}}_\theta(t)]_{t=1}^T + \hat{\mathbf{C}}_G[\hat{\mathbf{e}}_\theta(t)]_{t=1}^T \\
&\quad - [\mathbf{y}(t) - \hat{\mathbf{e}}_\theta(t)]_{t=1}^T\}\| \\
&\leq \lim_{T\to\infty}\|(\frac{\mathbf{X}_G'\hat{\mathbf{\Omega}}_G^{-1}\mathbf{X}_G}{T})^{-1}\| \cdot \|\{\frac{\mathbf{X}_G'\hat{\mathbf{\Omega}}_G^{-1}}{T}(\hat{\mathbf{C}}_G[\hat{\mathbf{e}}_\theta(t)]_{t=1}^T)\}\| \\
&= \|\mathbf{Q}_1^{-1}\| \cdot \lim_{T\to\infty}\|\frac{\mathbf{X}_G'\hat{\mathbf{C}}_G'^{-1}}{T}(\mathbf{I}_T \otimes \hat{\mathbf{\Sigma}}^{-1})[\hat{\mathbf{e}}_\theta(t)]_{t=1}^T\| \\
&= \|\mathbf{Q}_1^{-1}\| \cdot \lim_{T\to\infty}\|\frac{\mathbf{Z}_G'}{T}(\mathbf{I}_T \otimes \hat{\mathbf{\Sigma}})[\hat{\mathbf{e}}_\theta(t)]_{t=1}^T\| \\
&= \|\mathbf{Q}_1^{-1}\| \cdot \lim_{T\to\infty}\|T^{-1}\sum_{t=1}^T \frac{\partial\hat{\mathbf{e}}_\theta(t)'}{\partial\theta}\hat{\mathbf{\Sigma}}^{-1}\hat{\mathbf{e}}_\theta(t)\|
\end{aligned}$$

where we have written $\hat{\mathbf{\Sigma}}$ for $\hat{\mathbf{\Sigma}}_T(\hat{\theta}_G)$ for convenience, a shorthand that will be maintained in the sequel and $\|\mathbf{Q}_1^{-1}\|$ follows from Lemma (5.3.1).

Now consider the second term of the product in the last expression. This gives the limit of the norm of the Gaussian normal equations evaluated at $\theta_G^{(0)}$, which will converge to zero via the consistency of the least squares estimates. Formally, substituting for $\frac{\partial\hat{\mathbf{e}}_\theta(t)'}{\partial\theta}$ using (2.3.8) we notice that its typical elements are of the form

$$\begin{aligned}
&T^{-1}\sum_t \hat{\eta}(t-u)'\hat{\mathbf{\Sigma}}^{-1}\hat{\mathbf{e}}_\theta(t), \qquad T^{-1}\sum_t(\hat{\eta}(t) - \hat{\xi}(t))'\hat{\mathbf{\Sigma}}^{-1}\hat{\mathbf{e}}_\theta(t), \quad \text{and} \\
&T^{-1}\sum_t \hat{\xi}(t-u)'\hat{\mathbf{\Sigma}}^{-1}\hat{\mathbf{e}}_\theta(t), \qquad u = 1,\ldots,p. \qquad (5.3.8)
\end{aligned}$$

where the range of summation is from $t = 1$ to T. The first term of (5.3.8) can be

rewritten as

$$T^{-1}\sum_t \hat{\eta}(t-u)'\hat{\mathbf{\Sigma}}^{-1}\{\varepsilon(t) + (\hat{\mathbf{e}}_\theta(t)-\varepsilon(t))\} = T^{-1}\sum_t \hat{\eta}(t-u)'\hat{\mathbf{\Sigma}}^{-1}\varepsilon(t) + T^{-1}\sum_t \hat{\eta}(t-u)'\hat{\mathbf{\Sigma}}^{-1}(\hat{\mathbf{e}}_\theta(t)-\varepsilon(t)). \quad (5.3.9)$$

Since $\eta(t-u)$ is orthogonal to $\varepsilon(t)$ by construction then by the result in Hannan and Kavalieris (1984, p. 557) the first quantity in (5.3.9) is $O(Q_T)$. Employing the rule $vec(\mathbf{ABC}) = (\mathbf{C}' \otimes \mathbf{A})vec\mathbf{B}$ and noting that $\varepsilon(t) = vec\{\varepsilon(t)\}$ the second term on the right handside of (5.3.9) can be rewritten using (2.3.7) as

$$\begin{aligned} T^{-1}\sum_t \{\mathbf{y}(t-u) \otimes \hat{\mathbf{M}}(z)^{-1}\hat{\mathbf{\Sigma}}^{-1}\}vec\{\hat{\mathbf{e}}_\theta(t)-\varepsilon(t)\} \\ = T^{-1}\sum_t \{\sum_{j\geq 0}(\mathbf{y}(t-u-j)\otimes \hat{\mathbf{L}}(j))\}vec\{\hat{\mathbf{e}}_\theta(t)-\varepsilon(t)\} \\ = T^{-1}\sum_t vec\{\sum_{j\geq 0}\hat{\mathbf{L}}(j)(\hat{\mathbf{e}}_\theta(t)-\varepsilon(t))\mathbf{y}(t-u-j)'\} \quad (5.3.10) \end{aligned}$$

where $\mathbf{L}(j)$ denotes the j^{th} coefficient of $\mathbf{M}(z)^{-1}\mathbf{\Sigma}^{-1}$. Since $\varepsilon(t)$ is expressible in terms of the process $\mathbf{y}(t)$ as $\varepsilon(t) = \sum_{j\geq 0}\mathbf{\Phi}_0(j)\mathbf{y}(t-j)$ then we have

$$T^{-1}\sum_t vec\{\sum_{j\geq 0}\hat{\mathbf{L}}(j)(\hat{\mathbf{e}}_\theta(t) - \varepsilon(t))\mathbf{y}(t-u-j)'\} = vec\{\sum_{j\geq 0}\hat{\mathbf{L}}(j)(\sum_{k=0}^{b_T}(\hat{\mathbf{\Phi}}(k)-\mathbf{\Phi}_0(k)))\times T^{-1}\sum_{t=1}^{T}\mathbf{y}(t-k)\mathbf{y}(t-u-j)'\} \quad (5.3.11)$$

plus a term that is bounded by

$$const\cdot \sum_{k\geq b_T+1}\|\mathbf{\Phi}_0(k)\| \quad (5.3.12)$$

where $\hat{\mathbf{\Phi}}(k)$ is presumed to be zero for $k > b_T$ and $b_T \leq (\log T)^\tau$, $\tau < \infty$. By miniphase assumption (2.2.3), $\mathbf{\Phi}_0(k)$ and $\mathbf{L}(j)$ converge to zero at a geometric rate, that is, $\|\mathbf{L}(j)\|$ and $\|\mathbf{\Phi}_0(j)\|$ approach zero faster than $\kappa\rho_0^j$ as $j \to \infty$ for some constant κ and $0 < \rho_0 < 1$ and hence (5.3.12) can be shown to be $o(T^{-s})$ for some $s > 0$. Hence (5.3.10) reduces to

$$vec\{\sum_{j=0}^{\infty}\hat{\mathbf{L}}(j)\sum_{k=0}^{b_T}(\hat{\mathbf{\Phi}}(k)-\mathbf{\Phi}_0(k))(\mathcal{E}\{\mathbf{y}(t-k)\mathbf{y}(t-u-j)'\}+O(Q_T))\}+o(T^{-s}) = O(Q_T) \quad (5.3.13)$$

where the right-hand side quantity follows from Theorem (3.2.3) combined with the fact that the $\sum_{j\geq 0}\|\mathbf{L}(j)\| < \infty$, $\max_k \|(\hat{\mathbf{\Phi}}(k) - \mathbf{\Phi}_0(k))\| = O(Q_T)$ *a.s* and that $\sum_{k=1}^{b_T}\|\mathcal{E}\{\mathbf{y}(t-j)\mathbf{y}(t-k)'\}\| \leq c < \infty$ for some constant c. Similar arguments to those used above show that the other terms in (5.3.8) are $O(Q_T)$. The result of the lemma follows by noting that $\|\mathbf{Q}^{-1}\| = O(1)$. □

Lemma (5.3.3):

Let $\tilde{\mathbf{X}} = (\mathbf{x}(1)', \ldots, \mathbf{x}(T)')'$ *and* $\tilde{\mathbf{Y}} = (\mathbf{y}(1)', \ldots, \mathbf{y}(T)')'$ *be any two column vectors of observations on the stationary and egordic stochastic processes* $\mathbf{x}(t)$ *and* $\mathbf{y}(t)$. *If* $\mathbf{\Omega}_G$ *and* $\tilde{\mathbf{\Omega}}_{GLS}$ *are the variance-covariance matrices associated with the Gaussian and GLS procedures, respectively, and* $\mathbf{x}(t)$ *and* $\mathbf{y}(t)$ *have finite fourth moment then*

$$\|\frac{\tilde{\mathbf{X}}'}{T}(\mathbf{\Omega}_G - \tilde{\mathbf{\Omega}}_{GLS})\tilde{\mathbf{Y}}\| = o(T^{-\frac{1}{2}}).$$

Proof: First note, by construction,

$$\mathbf{\Omega}_G - \tilde{\mathbf{\Omega}}_{GLS} = \begin{pmatrix} \mathbf{G} - \mathbf{I}_{pv} & \vdots & \mathbf{U} \\ \ldots & \cdot & \ldots \\ \mathbf{U}' & \vdots & \mathbf{0} \end{pmatrix}$$

where, employing (2.3.12) and (5.2.8),

$$\begin{aligned} \mathbf{G} &= \begin{pmatrix} \mathbf{M}(0) & \ldots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{M}(p-1) & \ldots & \mathbf{M}(0) \end{pmatrix} \cdot (\mathbf{I}_p \otimes \mathbf{\Sigma}) \cdot \begin{pmatrix} \mathbf{M}(0)' & \ldots & \mathbf{M}(p-1)' \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \ldots & \mathbf{M}(0)' \end{pmatrix} \\ &= \begin{pmatrix} g_{\epsilon\epsilon}(1,1) & \ldots & g_{\epsilon\epsilon}(1,p) \\ \vdots & \ddots & \vdots \\ g_{\epsilon\epsilon}(p,1)' & \ldots & g_{\epsilon\epsilon}(p,p) \end{pmatrix}, \end{aligned}$$

$$g_{\epsilon\epsilon}(i,j) = g_{\epsilon\epsilon}(j,i)' = \sum_{r=0}^{\min(i,j)-1} \mathbf{M}(r)\mathbf{\Sigma}\mathbf{M}(j - min(i,j) + r)', \; i \leq j = 1, \ldots, p,$$

and

$$\mathbf{U} = \begin{pmatrix} \gamma_{\epsilon\epsilon}(p) & \mathbf{0} & \dots & \dots & \dots & \dots & \mathbf{0} \\ \gamma_{\epsilon\epsilon}(p-1) & \gamma_{\epsilon\epsilon}(p) & \mathbf{0} & \dots & \dots & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \gamma_{\epsilon\epsilon}(1) & \dots & \dots & \gamma_{\epsilon\epsilon}(p) & \mathbf{0} & \dots & \mathbf{0} \end{pmatrix}$$

are $pv \times pv$ and $pv \times \mathcal{T}v$, respectively. Now consider, noting that $\|g_{\epsilon\epsilon}(\cdot,\cdot)\| \leq \|\gamma_{\epsilon\epsilon}(0)\|$,

$$\begin{aligned} \|\frac{\tilde{\mathbf{X}}'}{T}(\mathbf{\Omega}_G - \tilde{\mathbf{\Omega}}_{GLS})\tilde{\mathbf{Y}}\| &= \|\frac{\tilde{\mathbf{X}}'}{T}\begin{pmatrix} \mathbf{G} - \mathbf{I}_{pv} & \vdots & \mathbf{U} \\ \dots & \cdot & \dots \\ \mathbf{U}' & \vdots & \mathbf{0} \end{pmatrix}\tilde{\mathbf{Y}}\| \\ &\leq |T^{-1}\sum_{i=1}^{p}\mathbf{x}(i)'\mathbf{y}(i)| + |T^{-1}\sum_{i=1}^{p}\sum_{j=1}^{p}\mathbf{x}(i)'g_{\epsilon\epsilon}(i,j)\mathbf{y}(j)| \\ &\quad + |T^{-1}\sum_{i=1}^{p}\sum_{j=1}^{i}\mathbf{x}(i)'\gamma_{\epsilon\epsilon}(p+j-i)\mathbf{y}(j)| \\ &\quad + |T^{-1}\sum_{j=1}^{p}\sum_{i=1}^{j}\mathbf{x}(i)'\gamma_{\epsilon\epsilon}(p+i-j)\mathbf{y}(j)| \\ &\leq T^{-1}p\max_{1\leq i\leq p}|\mathbf{x}(i)'\mathbf{y}(i)| \\ &\quad + T^{-1}\sum_{i=1}^{p}\sum_{j=1}^{p}\|\gamma_{\epsilon\epsilon}(0)\|\cdot\max_{1\leq i,j\leq p}|\mathbf{x}(i)'\mathbf{y}(j)| \\ &\quad + T^{-1}\sum_{i=1}^{p}\sum_{j=1}^{i}\|\gamma_{\epsilon\epsilon}(0)\|\cdot\max_{1\leq i,j\leq p}|\mathbf{x}(i)'\mathbf{y}(j)| \\ &\quad + T^{-1}\sum_{j=1}^{p}\sum_{i=1}^{j}\|\gamma_{\epsilon\epsilon}(0)\|\cdot\max_{1\leq i,j\leq p}|\mathbf{x}(i)'\mathbf{y}(j)| \\ &= T^{-1}\cdot const\cdot p^2 o(T^{\frac{1}{2}}) \\ &= o(T^{-\frac{1}{2}}) \end{aligned}$$

where the penultimate line follows from its predecessor using a result in Fuller (1976, p.180), noting that $\max_{1\leq i,j\leq p}|\mathbf{x}(i)'\mathbf{y}(j)| = o(T^{\frac{1}{2}})$ by the fourth moment property.

□

Thus we state the main theorem of this section.

Theorem (5.3.1)

Assume that the conditions of Lemmas (5.3.1), (5.3.2) and (5.3.3) hold. If $\hat{\theta}_G$ *and* $\hat{\theta}_{GLS}$ *are the Gaussian and GLS estimators defined by (2.3.13) and (5.3.6), respectively, then* $\|(\hat{\theta}_{GLS} - \hat{\theta}_G)\| = O(Q_T)$.

Proof: By Lemma (5.3.2) we may substitute $\bar{\theta}_G$ for $\hat{\theta}_G$ since $\|\hat{\theta}_{GLS} - \hat{\theta}_G\| \leq \|\theta_{GLS} - \bar{\theta}_G\| + \|\bar{\theta}_G - \hat{\theta}_G\|$. Also, as an immediate consequence of Theorem (2.4.1) and the strong consistency of $\hat{\Omega}_{GLS}$, we may also replace $\hat{\Omega}_G$ and $\hat{\tilde{\Omega}}_{GLS}$ by Ω_G and $\tilde{\Omega}_{GLS}$, respectively, thereby introducing errors that are $O(Q_T)$. Hence, from (5.2.6) and (5.3.7), and for notational simplicity in the derivations, writing $\tilde{\mathbf{X}}_{GLS}$ as $\mathbf{X}_1$, $\tilde{\Omega}_{GLS}$ as Ω_1 and $\tilde{\mathbf{Y}}_{GLS}$ as $\mathbf{Y}_1$ and substituting in a similar way the number 2 for the subscript G, neglecting the $O(Q_T)$ term, we have

$$\begin{aligned}
\|(\hat{\theta}_{GLS} - \bar{\theta}_G)\| &= \|(\mathbf{X}_1'\Omega_1^{-1}\mathbf{X}_1)^{-1}(\mathbf{X}_1'\Omega_1^{-1}\mathbf{Y}_1) - (\mathbf{X}_2'\Omega_2^{-1}\mathbf{X}_2)^{-1}(\mathbf{X}_2'\Omega_2^{-1}\mathbf{Y}_2)\| \\
&= \|(\mathbf{X}_1'\Omega_1^{-1}\mathbf{X}_1)^{-1}(\mathbf{X}_1'\Omega_1^{-1}\mathbf{Y}_1) - (\mathbf{X}_1'\Omega_1^{-1}\mathbf{X}_1)^{-1}(\mathbf{X}_2\Omega_2^{-1}\mathbf{Y}_2) \\
&\quad + (\mathbf{X}_1'\Omega_1^{-1}\mathbf{X}_1)^{-1}(\mathbf{X}_2'\Omega_2^{-1}\mathbf{Y}_2) - (\mathbf{X}_2'\Omega_2^{-1}\mathbf{X}_2)^{-1}(\mathbf{X}_2'\Omega_2^{-1}\mathbf{Y}_2)\| \\
&\leq \|(\mathbf{X}_1'\Omega_1^{-1}\mathbf{X}_1)^{-1}\|\|(\mathbf{X}_1'\Omega_1^{-1}\mathbf{Y}_1) - (\mathbf{X}_2'\Omega_2^{-1}\mathbf{Y}_2)\| \\
&\quad + \|(\mathbf{X}_1'\Omega_1^{-1}\mathbf{X}_1)^{-1} - (\mathbf{X}_2'\Omega_2^{-1}\mathbf{X}_2)^{-1}\| \cdot \|\mathbf{X}_2'\Omega_2^{-1}\mathbf{Y}_2\| \\
&\leq \|(\frac{\mathbf{X}_1'\Omega_1^{-1}\mathbf{X}_1}{T})^{-1}\| \cdot \|(\frac{\mathbf{X}_1'\Omega_1^{-1}\mathbf{Y}_1}{T}) - (\frac{\mathbf{X}_2'\Omega_2^{-1}\mathbf{Y}_2}{T})\| \\
&\quad + \|(\frac{\mathbf{X}_1'\Omega_1^{-1}\mathbf{X}_1}{T})^{-1} - (\frac{\mathbf{X}_2'\Omega_2^{-1}\mathbf{X}_2}{T})^{-1}\| \cdot \|\frac{\mathbf{X}_2'\Omega_2^{-1}\mathbf{Y}_2}{T}\|. \quad (5.3.14)
\end{aligned}$$

Consider the first term in the right-hand side of (5.3.14) and look at

$$\begin{aligned}
\|(\frac{\mathbf{X}_1'\Omega_1^{-1}\mathbf{Y}_1}{T}) - (\frac{\mathbf{X}_2'\Omega_2^{-1}\mathbf{Y}_2}{T})\| &= \|(\frac{\mathbf{X}_1'\Omega_1^{-1}\mathbf{Y}_1}{T}) - (\frac{\mathbf{X}_1'\Omega_1^{-1}\mathbf{Y}_2}{T}) + (\frac{\mathbf{X}_1'\Omega_1^{-1}\mathbf{Y}_2}{T}) \\
&\quad - (\frac{\mathbf{X}_1'\Omega_2\mathbf{Y}_2}{T}) + (\frac{\mathbf{X}_1'\Omega_2^{-1}\mathbf{Y}_2}{T}) - (\frac{\mathbf{X}_2'\Omega_2^{-1}\mathbf{Y}_2}{T})\| \\
&\leq \|\frac{\mathbf{X}_1'\Omega_1^{-1}}{T}(\mathbf{Y}_1 - \mathbf{Y}_2)\| + \|\frac{\mathbf{X}_1'}{T}(\Omega_1^{-1} - \Omega_2^{-1})\mathbf{Y}_2\| \\
&\quad + \|(\mathbf{X}_1' - \mathbf{X}_2')\frac{\Omega_2\mathbf{Y}_2}{T}\|
\end{aligned}$$

$$= \|(\frac{\mathbf{X}_1'\mathbf{\Omega}_1^{-1}}{T}[\hat{\mathbf{e}}_\theta(t) - \hat{\varepsilon}_h(t)]_{t=1}^T)\| + \|\frac{\mathbf{X}_1'}{T}(\mathbf{\Omega}_1^{-1} - \mathbf{\Omega}_2^{-1})\mathbf{Y}_2\|$$

$$+ \|(\mathbf{X}_1' - \mathbf{X}_2')\frac{\mathbf{\Omega}_2^{-1}\mathbf{Y}_2}{T}\|. \qquad (5.3.15)$$

Notice that the first term on the right-hand side of (5.3.15) contains, excluding a scaling factor of $O(1)$, quantities of the form $T^{-1}\sum_t \mathbf{y}(t-j)'\{\hat{\mathbf{e}}_\theta(t) - \hat{\varepsilon}_h(t)\} = T^{-1}\sum_t\{\sum_{k=1}^v(\hat{e}_{k,\theta}(t) - \hat{\varepsilon}_{k,h}(t))y_k(t-j)\}$ and $T^{-1}\sum_t \hat{\varepsilon}_h(t-j)'\{\hat{\mathbf{e}}_\theta(t) - \hat{\varepsilon}_h(t)\} = T^{-1}\sum_t\{\sum_{k=1}^v \hat{\varepsilon}_{k,h}(t-j)(\hat{e}_{k,\theta}(t) - \hat{\varepsilon}_{k,h}(t-j))\}$ whose typical elements are

$$T^{-1}\sum_t y_k(t-j)\{\hat{e}_{k,\theta}(t) - \hat{\varepsilon}_{k,h}(t)\} \qquad (5.3.16)$$

and

$$T^{-1}\sum_t \hat{\varepsilon}_{k,h}(t-j)\{\hat{e}_{k,\theta}(t) - \hat{\varepsilon}_{k,h}(t)\}, \qquad (5.3.17)$$

respectively. Now it is clear that (5.3.16) can be rewritten as

$$\begin{aligned} T^{-1}\sum_t y_k(t-j)\{\hat{e}_{k,\theta}(t) - \hat{\varepsilon}_{k,h}(t)\} &= T^{-1}\sum_t y_k(t-j)\hat{e}_{k,\theta}(t) - T^{-1}\sum_t y_k(t-j)\hat{\varepsilon}_{k,h}(t) \\ &= T^{-1}\sum_t y_k(t-j)(\hat{e}_{k,\theta}(t) - \varepsilon_k(t)) \\ &\quad - T^{-1}\sum_t y_k(t-j)\{\hat{\varepsilon}_{k,h}(t) - \varepsilon_k(t)\}. \qquad (5.3.18) \end{aligned}$$

Employing arguments that parallel those leading to the conclusion that follows (5.3.13), we observe that the first term in (5.3.18) is $O(Q_T)$. Similar manipulations via arguments involving Baxter's inequality (Theorem (3.2.1)) and Lemma (3.2.1) show that the second term in (5.3.18) is $O(Q_T)$. To see this observe that $\varepsilon_k(t) = \sum_{s=1}^v \phi_{ks,0}(z)y_s(t)$ and that $\hat{\varepsilon}_{k,h}(t) - \varepsilon_k(t) = \sum_{s=1}^v(\hat{\phi}_{ks,h}(z) - \phi_{ks,0}(z))y_s(t)$. We can now rewrite $T^{-1}\sum_t y_k(t-j)(\hat{\varepsilon}_{k,h}(t) - \varepsilon_k(t))$ as

$$\sum_{s=1}^v\{T^{-1}\sum_t[\sum_{u=1}^h(\hat{\phi}_{ks,h}(u) - \phi_{ks,0}(u))y_s(t-u)y_k(t-j)]\} + o(1) \qquad (5.3.19)$$

since $\mathcal{E}|(y_k(t-j)y_k(t-u)| = c < \infty$, $\hat{\phi}_{ks,h}(u) \equiv 0$ for $u > h$ and $\sum_{u>h}|\phi_{ks,0}(u)|$ is dominated by $const.\rho^h$. Given that the $\max_{1\le u\le h}(\hat{\phi}_{ks,h}(u) - \phi_{ks,0}(u))$ is $O(Q_T)$, $u \le h$, $h \le (\log T)^\tau$, $\tau < \infty$, it is apparent that (5.3.19) is itself $O(Q_T)$. By slight

modification of the arguments just employed, the quantity (5.3.17) can also be shown to be $O(Q_T)$.

Let us now turn to the second term in (5.3.15). We have

$$\|\frac{\mathbf{X}_1'}{T}(\mathbf{\Omega}_1^{-1} - \mathbf{\Omega}_2^{-1})\mathbf{Y}_2\| = \|\frac{\mathbf{X}_1'\mathbf{\Omega}_2^{-1}}{T}(\mathbf{\Omega}_2 - \mathbf{\Omega}_1)\mathbf{\Omega}_1^{-1}\mathbf{Y}_2\|$$

which, by direct application of Lemma (5.3.3), is $o(T^{-\frac{1}{2}})$ since, by construction, $\mathbf{\Omega}_1$ and $\mathbf{\Omega}_2$ are invertible matrices.

Finally, a typical element of the last term in (5.3.15), excluding once again the scaling factor of $O(1)$, is of the form

$$\begin{aligned} T^{-1}\sum_t(\hat{\varepsilon}_{h,k}(t-j) &- \hat{e}_{k,\theta}(t-j))(y_i(t) - \hat{e}_{i,\theta}(t)) \\ &= T^{-1}\sum_t(\hat{\varepsilon}_{h,k}(t-j) - \hat{e}_{k,\theta}(t-j))y_i(t) \\ &\quad - T^{-1}\sum_t(\hat{\varepsilon}_{h,k}(t-j) - \hat{e}_{k,\theta}(t-j))\hat{e}_{i,\theta}(t). \quad (5.3.20) \end{aligned}$$

Both the first and second terms in (5.3.20) can be shown to be $O(Q_T)$ employing arguments analogous to those used in providing such a bound for (5.3.16) and (5.3.17).

Since $(\frac{\mathbf{X}_2'\mathbf{\Omega}_2^{-1}\mathbf{Y}_2}{T})$ is $O(1)$ and that each component of $(\frac{\mathbf{X}_1'\mathbf{\Omega}_1^{-1}\mathbf{X}_1}{T})$ - $(\frac{\mathbf{X}_2'\mathbf{\Omega}_2^{-1}\mathbf{X}_2}{T})$, using arguments similar to those leading to the conclusion that follows (5.3.15), can be shown to be $O(Q_T)$ it follows that the second term in (5.3.14) is $O(Q_T)$. Hence the theorem is proved. □

It is important to stress that Theorem (5.3.1) carries with it the implication that the GLS estimator, $\hat{\theta}_{GLS}$, yields the same value of the estimates as the Gaussian estimator, $\hat{\theta}_G$, when the sample size is sufficiently large. With regard to the limiting distribution we find, after adopting derivations that parallel those employed to establish the result in Theorem (5.3.1), that $\sqrt{T}\|\hat{\theta}_{GLS} - \hat{\theta}_G\| = o_p(1)$. This result tells us that $\hat{\theta}_{GLS}$ is asymptotically efficient in the sense that the asymptotic covariance matrix in its limiting distribution is that given by the inverse of the information matrix in the Gaussian case. The discussion of the last paragraph is summarized

immediately below.

Corollary (5.3.1)

If $\mathbf{y}(t)$ *is an ARMA process of the form (5.1.1) satisfying condition A, (1.1.5) and that the conditions of Theorem (5.3.1) hold then* $\sqrt{T}(\hat{\theta}_{GLS} - \theta_0)$ *has an asymptotic normal distribution with zero mean vector and variance-covariance matrix,* $\mathbf{\Lambda}_G = \{\mathbf{S}_\nu \mathcal{E}(\mathbf{w}_0(t)\mathbf{w}_0(t)')\mathbf{S}_\nu'\}^{-1}$ *where* $\mathbf{w}_0(t) = \mathbf{x}(t) \otimes \mathbf{M}_0(z)^{-1}\mathbf{\Sigma}_0^{-\frac{1}{2}}$ *and* $\mathbf{x}(t) = (-\zeta_p(z)' \otimes \mathbf{y}(t)' : -(\mathbf{y}(t) - \varepsilon(t))' : \zeta_p(z)' \otimes \varepsilon(t)')'$.

Proof: See Chapter 2 (Section (2.4)) for an explicit derivation of $\mathbf{\Lambda}_G$. □

Although the results of this section provide us with information about the asymptotic convergence of the GLS estimator to the Gaussian estimator, they do not provide any hints about how large T has to be for the results to be applicable. Therefore, to get some insight into the practical convergence rate, transient behaviour, and finite-sample properties of the estimators (GLS and Gaussian), the analysis presented above has been complemented by some simulation studies. These are presented in Section (5.5) but before discussing these results it is necessary to consider the numerical implementation of the GLS scheme and this is discussed next.

5.4 Numerical Implementation

This section discusses the computational aspects of the GLS procedure in relation to model (5.1.1). From our discussion in the previous sections the expression (5.2.6) seems to be the most useful formulation for evaluating $\hat{\theta}_{GLS}$ but, as the sample size increases, this may not be attractive from the computational point of view. The difficulties with the direct use of (5.2.6) to evaluate $\hat{\theta}_{GLS}$ arise from the construction and inversion of $\hat{\mathbf{\Omega}}_{GLS}$. Although some algorithms that employ the Cholesky decomposition are readily available for conducting the inversion of the

general symmetric Toeplitz matrices see, for example, Ansley (1979), Golub and Van Loan (1983) and Kailath et al (1986), there are problems with their use. For some discussion of the numerical difficulties associated with such method see Golub and Van-Loan (1989, Chap. 4). These practical difficulties will obviously present themselves directly if any attempt is made to evaluate $\hat{\theta}_{GLS}$ from the basic definition (5.2.6). Similar concern about inverting large Toeplitz (band) matrices in the case $v = 1$ has been expressed elsewhere in the literature see, for example, Zinde-Walsh (1988) but these problems seem somewhat understated by Koreisha and Pukkila (1990), who simply recommend the use of Cholesky factorization of $\mathbf{\Omega}_{GLS}$. We now aim at removing these sources of difficulty by exploiting the theoretical developments of Subsection (5.3.1) to provide simple numerical procedures for evaluating $\hat{\theta}_{GLS}$. Though the expressions presented in that subsection are satisfactory in terms of abstract theoretical analysis, they do not lend themselves directly to numerical implementation since the determination of the initial values and the succeeding iterates involves, once again, a similar matrix inversion and hence the basic problem is not yet removed. Nevertheless, by means of alternative algebraic manipulations it is possible to show that the evaluation of $\hat{\theta}_{GLS}$ can be carried out using a combination of specialized calculations that avoid the construction and direct inversion of $\hat{\mathbf{\Omega}}_{GLS}$.

In order to facilitate discussion, consider a factorization of the form $\mathbf{\Omega}_{GLS} = \mathbf{\Lambda}\mathbf{\Lambda}'$. As a byproduct of the theoretical analysis of the previous section, an obvious choice for $\mathbf{\Lambda}$ using (5.2.7) is

$$\mathbf{\Lambda} = \mathbf{C}_{GLS}(\mathbf{I}_T \otimes \mathbf{\Sigma}_h^{\frac{1}{2}}) \tag{5.4.1}$$

and $\mathbf{\Sigma}_h^{\frac{1}{2}}$ is a unique Cholesky factor of $\mathbf{\Sigma}_h$. The advantage of (5.4.1) is that it provides an alternative factorization for $\mathbf{\Omega}_{GLS}$ which avoids the direct evaluation of $\hat{\mathbf{\Omega}}_{GLS}^{-1}$ and/or its Cholesky factor as recommended by Koreisha and Pukkila (op. cit.). The construction of $\mathbf{\Omega}_{GLS}$ as an appropriate product of quantities of the form $\mathbf{C}_{GLS}(\mathbf{I}_T \otimes \mathbf{\Sigma}_h^{\frac{1}{2}})$ leads to the suggestion of several simplifications that can be made

to obtain $\hat{\theta}_{GLS}$. In particular, it allows the evaluation of $\hat{\theta}_{GLS}$ to be carried out as a combination of specialized computations and a sequence of recursions. With such a device it is practical and feasible to implement the GLS algorithm for any desired length of data.

Let us now turn to the explicit construction of $\mathbf{\Omega}_{GLS}$ via (5.4.1). Now partition $\mathbf{\Lambda}$ into two submatrices as $\mathbf{\Lambda} = (\bar{\mathbf{D}} : \bar{\mathbf{C}})$ where $\bar{\mathbf{D}}$ and $\bar{\mathbf{C}}$ are $\mathcal{T}v \times pv$ and $\mathcal{T}v \times \mathcal{T}v$ matrices, respectively. In particular,

$$\bar{\mathbf{D}} = \begin{bmatrix} \mathbf{M}(p) & \mathbf{M}(p-1) & \dots & \mathbf{M}(1) \\ \mathbf{0} & \mathbf{M}(p) & \dots & \mathbf{M}(2) \\ \vdots & \mathbf{0} & \ddots & \vdots \\ \mathbf{0} & \dots & \mathbf{0} & \mathbf{M}(p) \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix} (\mathbf{I}_p \otimes \mathbf{\Sigma}_h^{\frac{1}{2}})$$

and $\bar{\mathbf{C}} = \mathcal{C}(\mathbf{I}_{\mathcal{T}} \otimes \mathbf{\Sigma}_h^{\frac{1}{2}})$ wherein the $\mathcal{T}v \times \mathcal{T}v$ matrix $\mathcal{C}$ is of the form

$$\mathcal{C} = \begin{bmatrix} \mathbf{M}(0) & \mathbf{0} & \dots & \mathbf{0} & \dots & \dots & \mathbf{0} \\ \mathbf{M}(1) & \mathbf{M}(0) & \dots & \mathbf{0} & \dots & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \ddots & \ddots & & \vdots \\ \mathbf{M}(p) & \mathbf{M}(p-1) & \dots & \mathbf{M}(0) & \dots & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{M}(p) & \dots & \mathbf{M}(1) & \mathbf{M}(0) & \dots & \mathbf{0} \\ \vdots & \mathbf{0} & \ddots & \ddots & \ddots & \ddots & \vdots \\ \mathbf{0} & \dots & \mathbf{0} & \mathbf{M}(p) & \dots & \dots & \mathbf{M}(0) \end{bmatrix}. \tag{5.4.2}$$

Substituting for $\mathbf{\Lambda}$ in $\mathbf{\Omega}_{GLS}$ it becomes a simple exercise to verify that

$$\begin{aligned} \mathbf{\Omega}_{GLS} &= \bar{\mathbf{D}}\bar{\mathbf{D}}' + \bar{\mathbf{C}}\bar{\mathbf{C}}' \\ &= \bar{\mathbf{D}}\bar{\mathbf{D}}' + \mathbf{\Omega} \end{aligned} \tag{5.4.3}$$

where $\mathbf{\Omega} = \bar{\mathbf{C}}\bar{\mathbf{C}}'$. Thus, substituting for $\hat{\mathbf{\Omega}}_{GLS}$ using (5.4.3) we notice that the basic formula (5.2.6) can be rewritten as

$$\hat{\theta}_{GLS} = \{\mathbf{X}'_{GLS}(\hat{\mathbf{\Omega}} + \hat{\bar{\mathbf{D}}}\hat{\bar{\mathbf{D}}}')^{-1}\mathbf{X}_{GLS}\}^{-1}\{\mathbf{X}'_{GLS}(\hat{\mathbf{\Omega}} + \hat{\bar{\mathbf{D}}}\hat{\bar{\mathbf{D}}}')^{-1}\mathbf{Y}_{GLS}\}. \tag{5.4.4}$$

As has been noted earlier, expression (5.4.4) is not, however, well suited for computation as it stands since it involves an inversion of a $\mathcal{T}v \times \mathcal{T}v$ matrix. The direct inversion of matrix $(\hat{\boldsymbol{\Omega}} + \hat{\bar{\mathbf{D}}}\hat{\bar{\mathbf{D}}}')$ can thus be avoided by employing a result in matrix algebra which is now stated in the following lemma.

Lemma (5.4.1) *Let* $\mathbf{A}$, $\mathbf{B}$ *and* $\mathbf{C}$ *be matrices of compatible dimensions, so that* $\mathbf{A}^{-1}$, *the product* $\mathbf{BC}$ *and the sum* $\mathbf{A}+\mathbf{BC}$ *exist. Then*

$$(\mathbf{A}+\mathbf{BC})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}(\mathbf{I}+\mathbf{CA}^{-1}\mathbf{B})^{-1}\mathbf{CA}^{-1} \tag{5.4.5}$$

where $\mathbf{I}$ *is an identity matrix of an appropriate dimension.*

Proof: For convenience, set $\mathbf{Q} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}(\mathbf{CA}^{-1}\mathbf{B}+\mathbf{I})^{-1}\mathbf{CA}^{-1}$ and post multiply $\mathbf{Q}$ by $\mathbf{A}+\mathbf{BC}$ to obtain

$$\begin{aligned}
\mathbf{Q}\{\mathbf{A}+\mathbf{BC}\} &= \mathbf{I}+\mathbf{A}^{-1}\mathbf{BC} - \mathbf{A}^{-1}\mathbf{B}(\mathbf{CA}^{-1}\mathbf{C}+\mathbf{I})^{-1}\mathbf{C} - \mathbf{A}^{-1}\mathbf{B}(\mathbf{CA}^{-1}\mathbf{B}+\mathbf{I})^{-1} \\
&\qquad \times \mathbf{CA}^{-1}\mathbf{BC} \\
&= \mathbf{I}+\mathbf{A}^{-1}\mathbf{B}(\mathbf{CA}^{-1}\mathbf{B}+\mathbf{I})^{-1}\{(\mathbf{CA}^{-1}\mathbf{B}+\mathbf{I})\mathbf{C} - \mathbf{C} - \mathbf{CA}^{-1}\mathbf{BC}\} \\
&= \mathbf{I}+\mathbf{A}^{-1}\mathbf{B}(\mathbf{CA}^{-1}\mathbf{B}+\mathbf{I})^{-1}\{\mathbf{CA}^{-1}\mathbf{BC}+\mathbf{C} - \mathbf{C} - \mathbf{CA}^{-1}\mathbf{BC}\} \\
&= \mathbf{I}+\mathbf{A}^{-1}\mathbf{B}(\mathbf{CA}^{-1}\mathbf{B}+\mathbf{I})^{-1}\{0\} = \mathbf{I}
\end{aligned}$$

which establishes (5.4.5). □

Remark: *Lemma (5.4.1) is but a special case of the matrix inversion lemma. The current proof, as it stands, uses arguments parallel to that of Henderson and Searle (1981) and is presented because it emphasizes the role played by the lemma in the subsequent derivations.*

Now set $\mathbf{A} = \boldsymbol{\Omega}$, $\mathbf{B} = \bar{\mathbf{D}}$, $\mathbf{C} = \bar{\mathbf{D}}'$ and apply Lemma (5.4.1) we have

$$(\boldsymbol{\Omega} + \bar{\mathbf{D}}\bar{\mathbf{D}}')^{-1} = \boldsymbol{\Omega}^{-1} - \boldsymbol{\Omega}^{-1}\bar{\mathbf{D}}(\mathbf{I}_{pv} + \bar{\mathbf{D}}'\boldsymbol{\Omega}^{-1}\bar{\mathbf{D}})^{-1}\bar{\mathbf{D}}'\boldsymbol{\Omega}^{-1}. \tag{5.4.6}$$

The condition of the lemma is thus satisfied since, by construction, $\boldsymbol{\Omega}$ is positive definite, and therefore non-singular.

It is sometimes convenient to find inverses of matrices in terms of submatrices and as such, the expression in the right-hand side of (5.4.6) can be further simplified somewhat by rewriting $\bar{\mathbf{D}}\boldsymbol{\Omega}^{-1}\bar{\mathbf{D}}$ in a partitioned form. Assume for the moment that $\boldsymbol{\Omega}$ has been constructed and partitioned into four submatrices as

$$\boldsymbol{\Omega} = \begin{pmatrix} \boldsymbol{\Omega}_{11} & \vdots & \boldsymbol{\Omega}_{12} \\ \dots & \cdot & \dots \\ \boldsymbol{\Omega}_{21} & \vdots & \boldsymbol{\Omega}_{22} \end{pmatrix}, \quad \boldsymbol{\Omega}_{21} = (\boldsymbol{\Omega}_{12})'$$

such that $\boldsymbol{\Omega}_{11}$ is a $pv \times pv$ matrix and applying similar reasoning to matrix $\bar{\mathbf{D}}$ gives $\bar{\mathbf{D}} = (\mathbf{D}' : \mathbf{O})'$ where $\mathbf{D}$ is a $(pv \times pv)$ upper submatrix of $\bar{\mathbf{D}}$,

$$\mathbf{D} = \begin{pmatrix} \mathbf{M}(p) & \mathbf{M}(p-1) & \dots & \dots & \mathbf{M}(1) \\ \mathbf{0} & \mathbf{M}(p) & \dots & \dots & \mathbf{M}(2) \\ \mathbf{0} & \mathbf{0} & \mathbf{M}(p) & \dots & \mathbf{M}(3) \\ \vdots & \vdots & \ddots & & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{M}(p) \end{pmatrix} (\mathbf{I}_p \otimes \boldsymbol{\Sigma}_h^{\frac{1}{2}}),$$

and $\mathbf{O}$ denotes a $(\mathcal{T}-p)v \times pv$ null matrix. Thus, following Graybill (1969, p.165), it can be readily shown that

$$\boldsymbol{\Omega}^{-1} = \begin{pmatrix} \boldsymbol{\Omega}^{11} & \vdots & \boldsymbol{\Omega}^{12} \\ \dots & \cdot & \dots \\ \boldsymbol{\Omega}^{21} & \vdots & \boldsymbol{\Omega}^{22} \end{pmatrix}, \quad \boldsymbol{\Omega}^{21} = (\boldsymbol{\Omega}^{12})' \tag{5.4.7}$$

where $\boldsymbol{\Omega}^{11} = (\boldsymbol{\Omega}_{11} - \boldsymbol{\Omega}_{12}\boldsymbol{\Omega}_{22}^{-1}\boldsymbol{\Omega}_{21})^{-1}$, $\boldsymbol{\Omega}^{22} = (\boldsymbol{\Omega}_{22} - \boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12})^{-1}$ and $\boldsymbol{\Omega}^{12} = -\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12}(\boldsymbol{\Omega}_{22} - \boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12})^{-1}$. Hence, employing (5.4.7) in $(\mathbf{I}_{pv} + \bar{\mathbf{D}}'\boldsymbol{\Omega}^{-1}\bar{\mathbf{D}})$ gives

$$\mathbf{I}_{pv} + (\mathbf{D}' : \mathbf{O}) \cdot \begin{pmatrix} \boldsymbol{\Omega}^{11} & \vdots & \boldsymbol{\Omega}^{12} \\ \dots & \cdot & \dots \\ \boldsymbol{\Omega}^{12} & \vdots & \boldsymbol{\Omega}^{22} \end{pmatrix} \cdot \begin{pmatrix} \mathbf{D} \\ \dots \\ \mathbf{O} \end{pmatrix} = \mathbf{I}_{pv} + \mathbf{D}'\boldsymbol{\Omega}^{11}\mathbf{D}. \tag{5.4.8}$$

We are therefore led to the conclusion that $(\mathbf{I}_{pv} + \bar{\mathbf{D}}'\boldsymbol{\Omega}^{-1}\bar{\mathbf{D}})^{-1} = (\mathbf{I}_{pv} + \mathbf{D}'\boldsymbol{\Omega}^{11}\mathbf{D})^{-1}$ and so, (5.4.6) may be rewritten as

$$(\boldsymbol{\Omega} + \bar{\mathbf{D}}\bar{\mathbf{D}}')^{-1} = \boldsymbol{\Omega}^{-1} - \boldsymbol{\Omega}^{-1}\bar{\mathbf{D}}\mathbf{H}\bar{\mathbf{D}}'\boldsymbol{\Omega}^{-1} \tag{5.4.9}$$

where $\mathbf{H} = (\mathbf{I}_{pv} + \mathbf{D}'\boldsymbol{\Omega}^{11}\mathbf{D})^{-1}$.

Let us now return to (5.4.4) and first consider

$$\begin{aligned}\mathbf{X}'_{GLS}(\hat{\boldsymbol{\Omega}} + \hat{\mathbf{D}}\hat{\mathbf{D}}')^{-1}\mathbf{X}_{GLS} &= \mathbf{X}'_{GLS}\{\hat{\boldsymbol{\Omega}}^{-1} - \hat{\boldsymbol{\Omega}}^{-1}\hat{\mathbf{D}}\hat{\mathbf{H}}\hat{\mathbf{D}}'\hat{\boldsymbol{\Omega}}^{-1}\}\mathbf{X}_{GLS} \\ &= \mathbf{X}'_{GLS}\hat{\boldsymbol{\Omega}}^{-1}\mathbf{X}_{GLS} - \mathbf{X}'_{GLS}\hat{\boldsymbol{\Omega}}^{-1}\hat{\mathbf{D}}\hat{\mathbf{H}}\hat{\mathbf{D}}'\hat{\boldsymbol{\Omega}}^{-1}\mathbf{X}_{GLS}\end{aligned} \tag{5.4.10}$$

where $(\hat{\boldsymbol{\Omega}} + \hat{\mathbf{D}}\hat{\mathbf{D}}')$ has been replaced by the expression on the right-hand side of (5.4.9). Since $\hat{\boldsymbol{\Omega}}^{-1} = \hat{\mathbf{C}}'^{-1}\hat{\mathbf{C}}^{-1}$ then it follows that (5.4.10) has the form

$$\begin{aligned}\mathbf{X}'_{GLS}(\hat{\boldsymbol{\Omega}} + \hat{\mathbf{D}}\hat{\mathbf{D}}')^{-1}\mathbf{X}_{GLS} &= \mathbf{X}'_{GLS}(\hat{\mathbf{C}}'^{-1}\hat{\mathbf{C}}^{-1})\mathbf{X}_{GLS} - \mathbf{X}'_{GLS}(\hat{\mathbf{C}}'^{-1}\hat{\mathbf{C}}^{-1}\hat{\mathbf{D}}\hat{\mathbf{H}}\hat{\mathbf{D}}' \\ &\quad \times \hat{\mathbf{C}}'^{-1}\hat{\mathbf{C}}^{-1})\mathbf{X}_{GLS} \\ &= (\hat{\mathbf{C}}^{-1}\mathbf{X}_{GLS})'(\hat{\mathbf{C}}^{-1}\mathbf{X}_{GLS}) - (\hat{\mathbf{C}}^{-1}\mathbf{X}_{GLS})' \\ &\quad \times (\hat{\mathbf{C}}^{-1}\hat{\mathbf{D}}\hat{\mathbf{H}}\hat{\mathbf{D}}'\hat{\mathbf{C}}'^{-1})(\hat{\mathbf{C}}^{-1}\mathbf{X}_{GLS}) \\ &= \mathbf{Z}'\mathbf{Z} - \mathbf{Z}'\hat{\mathbf{C}}^{-1}(\hat{\mathbf{D}}\hat{\mathbf{H}}\hat{\mathbf{D}}')\hat{\mathbf{C}}'^{-1}\mathbf{Z} \\ &= \mathbf{Z}'\mathbf{Z} - \mathbf{W}'(\hat{\mathbf{D}}\hat{\mathbf{H}}\hat{\mathbf{D}}')\mathbf{W}\end{aligned} \tag{5.4.11}$$

where $\mathbf{Z} = \hat{\mathbf{C}}^{-1}\mathbf{X}_{GLS}$ and $\mathbf{W} = (\hat{\mathbf{C}}')^{-1}\mathbf{Z}$. It has been found that the elements of matrix $\mathbf{W}$ beyond the $(pv)^{th}$ row do not contribute to the matrix product $\mathbf{W}'(\hat{\mathbf{D}}\hat{\mathbf{H}}\hat{\mathbf{D}}')\mathbf{W}$ in (5.4.11). Suppose $\mathbf{W}$ is partitioned after the $(pv)^{th}$ row so that $\mathbf{W}' = (\mathbf{W}'_1 : \mathbf{W}'_2)$. Then $\mathbf{W}'(\hat{\mathbf{D}}\hat{\mathbf{H}}\hat{\mathbf{D}}')\mathbf{W}$ can be further simplified as $\mathbf{W}'_1\hat{\mathbf{D}}\hat{\mathbf{H}}\hat{\mathbf{D}}'\mathbf{W}_1$ since $\hat{\mathbf{D}}' = (\hat{\mathbf{D}}' : \mathbf{O})$. Hence, (5.4.11) becomes

$$\mathbf{Z}'\mathbf{Z} - \mathbf{W}'_1\hat{\mathbf{D}}\hat{\mathbf{H}}\hat{\mathbf{D}}'\mathbf{W}_1. \tag{5.4.12}$$

It thus remains to obtain an equivalent expression for $(\mathbf{X}'_{GLS}\hat{\boldsymbol{\Omega}}^{-1}_{GLS}\mathbf{Y}_{GLS})$. For by explicit calculation using Lemma (5.4.1),

$$\begin{aligned}(\mathbf{X}'_{GLS}\hat{\boldsymbol{\Omega}}^{-1}_{GLS}\mathbf{Y}_{GLS}) &= \mathbf{X}'_{GLS}(\hat{\boldsymbol{\Omega}} + \hat{\mathbf{D}}\hat{\mathbf{D}}')^{-1}\mathbf{Y}_{GLS} \\ &= \mathbf{X}'_{GLS}\{\hat{\boldsymbol{\Omega}}^{-1} - \hat{\boldsymbol{\Omega}}^{-1}\hat{\mathbf{D}}\hat{\mathbf{H}}\hat{\mathbf{D}}'\hat{\boldsymbol{\Omega}}^{-1}\}\mathbf{Y}_{GLS} \\ &= \mathbf{X}'_{GLS}\hat{\boldsymbol{\Omega}}^{-1}\mathbf{Y}_{GLS} - \mathbf{X}'_{GLS}\hat{\boldsymbol{\Omega}}^{-1}\hat{\mathbf{D}}\hat{\mathbf{H}}\hat{\mathbf{D}}'\hat{\boldsymbol{\Omega}}^{-1}\mathbf{Y}_{GLS}\end{aligned}$$

$$
\begin{aligned}
&= \mathbf{X}'_{GLS}\hat{\bar{\mathbf{C}}}^{'-1}\hat{\bar{\mathbf{C}}}^{-1}\mathbf{Y}_{GLS} - \mathbf{X}'_{GLS}\hat{\bar{\mathbf{C}}}^{'-1}\hat{\bar{\mathbf{C}}}^{-1}\hat{\mathbf{D}}\hat{\mathbf{H}}\hat{\mathbf{D}}' \\
&\quad \times \hat{\bar{\mathbf{C}}}^{'-1}\hat{\bar{\mathbf{C}}}^{-1}\mathbf{Y}_{GLS} \\
&= \mathbf{Z}'\mathbf{U} - \mathbf{W}'(\hat{\mathbf{D}}\hat{\mathbf{H}}\hat{\mathbf{D}}')\mathbf{V} \qquad (5.4.13)
\end{aligned}
$$

where $\mathbf{U} = \hat{\bar{\mathbf{C}}}^{-1}\mathbf{Y}_{GLS}$ and $\mathbf{V} = (\hat{\bar{\mathbf{C}}}^{'-1}\mathbf{U})$. Partitioning $\mathbf{V}$ as $\mathbf{W}$ so that $\mathbf{V}' = (\mathbf{V}'_1 : \mathbf{V}'_2)$. Then it is easily verified that (5.4.13) reduces to $\mathbf{Z}'\mathbf{U} - \mathbf{W}'_1\hat{\mathbf{D}}\hat{\mathbf{H}}\hat{\mathbf{D}}'\mathbf{V}_1$. Thus, an equivalent expression for the GLS estimator is

$$
\hat{\theta}_{GLS} = \{\mathbf{Z}'\mathbf{Z} - \mathbf{W}'_1(\hat{\mathbf{D}}\hat{\mathbf{H}}\hat{\mathbf{D}}')\mathbf{W}_1\}^{-1}\{\mathbf{Z}'\mathbf{U} - \mathbf{W}'_1(\hat{\mathbf{D}}\hat{\mathbf{H}}\hat{\mathbf{D}}')\mathbf{V}_1\}. \qquad (5.4.14)
$$

Examining (5.4.14), there seems to exist some features of the proposed technique that have not been fully exploited. In particular, it is revealing that the elements of matrices $\mathbf{Z}$ and $\mathbf{W}$ and those of vectors $\mathbf{U}$ and $\mathbf{V}$ can, respectively, be generated via a sequence of recursions. To show this we proceed as follows. Putting $\mathbf{Z} = \bar{\mathbf{C}}^{-1}\mathbf{X}_{GLS}$ amounts to obtaining the solution of

$$
\bar{\mathbf{C}}\mathbf{Z} = \mathbf{X}_{GLS} \qquad (5.4.15)
$$

for $\mathbf{Z}$. Now define $\mathcal{M}(z) = \sum_{j=0}^{p}\mathcal{M}(j)z^j$ where $\mathcal{M}(j) = \mathbf{M}(j)\boldsymbol{\Sigma}_h^{\frac{1}{2}}$, $j = 0, 1, \ldots, p$ and let $(\mathbf{X}(1)', \ldots, \mathbf{X}(\mathcal{T})')'$ and $(\mathbf{Z}(1)', \ldots, \mathbf{Z}(\mathcal{T})')'$ denote partitions of $\mathbf{X}_{GLS}$ and $\mathbf{Z}$ where $\mathbf{X}(t)\mathbf{S}_\nu$ and $\mathbf{Z}(t)$, $t = 1, \ldots, \mathcal{T}$, represent the t^{th} block of v rows, respectively. Substituting $\hat{\bar{\mathbf{C}}}$ for $\bar{\mathbf{C}}$ in (5.4.15) we find that $\mathbf{Z}(t)$ satisfies the forward difference equations

$$
\begin{aligned}
\sum_{j=0}^{p}\hat{\mathcal{M}}(j)\mathbf{Z}(t-j) &= \mathbf{X}(t),\ t = 1, \ldots, \mathcal{T}, \\
&= (-\zeta_p(z)' \otimes \mathbf{y}(p+t)' \otimes \mathbf{I}_v : -(\mathbf{y}(t+p)' - \hat{\varepsilon}_h(p+t)') \otimes \mathbf{I}_v \\
&\qquad : \zeta_p(z)' \otimes \hat{\varepsilon}_h(p+t)' \otimes \mathbf{I}_v)\mathbf{S}'_\nu \qquad (5.4.16)
\end{aligned}
$$

subject to $\mathbf{Z}(t) = 0$, $t \leq 0$. Also, from $\mathbf{W} = \hat{\bar{\mathbf{C}}}^{'-1}\mathbf{Z}$ it follows that $\hat{\bar{\mathbf{C}}}'\mathbf{W} = \mathbf{Z}$. Substituting for $\hat{\bar{\mathbf{C}}}'$ and letting $(\mathbf{W}(1)', \ldots, \mathbf{W}(\mathcal{T})')'$ define a partition analogous to that of $\mathbf{X}_{GLS}$ and $\mathbf{Z}$, it can be easily verified that $\mathbf{W}$ satisfies a set of backward

recursions on

$$\sum_{j=0}^{p} \hat{\mathcal{M}}(j)\mathbf{W}(t+j) = \mathbf{Z}(t), \quad t = \mathcal{T}, \ldots, 1 \tag{5.4.17}$$

with $\mathbf{W}(t) = 0$ for $t > \mathcal{T}$. The components of $\mathbf{U}$ and $\mathbf{V}$ can be generated in an analogous manner using

$$\sum_{j=0}^{p} \hat{\mathcal{M}}(j)\mathbf{U}(t-j) = \mathbf{y}(p+t) - \hat{\varepsilon}_h(p+t), \; t = 1, \ldots, \mathcal{T}, \tag{5.4.18}$$

$\mathbf{U}(t) = 0, \; t \leq 0$, and

$$\sum_{j=0}^{p} \hat{\mathcal{M}}(j)\mathbf{V}(t+j) = \mathbf{U}(t), \; t = \mathcal{T}, \ldots, 1, \tag{5.4.19}$$

$\mathbf{V}(t) = 0, \; t > \mathcal{T}$.

To complete the evaluation of $\hat{\theta}_{GLS}$ via (5.4.14) it now only remains for us to determine $\hat{\mathbf{H}}$. This involves calculating $\hat{\mathbf{\Omega}}^{-1} = (\hat{\bar{\mathbf{C}}}\hat{\bar{\mathbf{C}}}')^{-1}$. It is often the case to evaluate $\hat{\bar{\mathbf{C}}}^{-1}$ using numerical algorithms but, as has previously been indicated, the computations involved as T increases become increasingly laborious even for a high speed computer. In effect the estimates of $\hat{\theta}_{GLS}$ may suffer from computational round-off errors. The rounding error may appear slight, or even inconsequential, when considering a matrix of low order but, could sometimes accumulates to be very serious in lengthy calculations such as those involved in the inversion of large, possibly sparse and ill-conditioned, block Toeplitz (band) matrices. This is all the more difficult because situations in which such errors will be serious cannot be specified with high degrees of certainty, and a mathematical discussion of the cumulative effects of rounding errors is very involved in anything but simple situations. See, for example, Stewart (1973) for some discussion. It is therefore important to have an appreciation for the possible consequences of rounding errors and possibly guard against them in calculations of this nature. Thus we need some device to avoid problems arising from the direct inversion of matrix $\hat{\mathbf{\Omega}}^{-1}$.

Fortunately, the elements of $\hat{\mathbf{\Omega}}^{-1}$ can be computed indirectly by observing that $\bar{\mathbf{C}}^{-1} = (\mathbf{I}_T \otimes \mathbf{\Sigma}_h^{-\frac{1}{2}})\mathcal{C}^{-1}$ and the inverse of $\mathcal{C}$ is easily obtained from the coefficient

matrices in $\hat{\mathbf{M}}(z)^{-1} = adj\hat{\mathbf{M}}(z)/\det\hat{\mathbf{M}}(z)$. The matrix $adj\hat{\mathbf{M}}(z)$ is a polynomial where adj stands for the adjoint or adjugate of the indicated matrix. The convolution of each entry of $adj\hat{\mathbf{M}}(z)$ with $1/\{\det\hat{\mathbf{M}}(z)\}$ can be carried out in a straightforward fashion using the Euclidean division algorithm. Now set $\mu(z) = adj\hat{\mathbf{M}}(z)/\{\det\hat{\mathbf{M}}(z)\} = \sum_{j\geq 0}\mu(j)z^j$ where $\mu(j)$, $j = 0, 1, \ldots$ are $v \times v$ matrices. A useful result derives from the fact that the elements of $\hat{\mathcal{C}}^{-1}$ occur as the coefficients of $\mu(z)$ and in particular the successive block of columns of $\mathcal{C}^{-1}$ contain the coefficient matrices $(\mu(0)', \ldots, \mu(\mathcal{T}-i)')'$ with i denoting the i^{th} column. Thus, the $\mathcal{T}v \times \mathcal{T}v$ matrix $\mathcal{C}^{-1}$ is given by

$$\mathcal{C}^{-1} = \begin{bmatrix} \mu(0) & \mathbf{0} & \ldots & \ldots & \mathbf{0} \\ \mu(1) & \mu(0) & \mathbf{0} & \ldots & \vdots \\ \mu(2) & \mu(1) & \mu(0) & \ldots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \mu(\mathcal{T}-1) & \mu(\mathcal{T}-2) & \ldots & \ldots & \mu(0) \end{bmatrix}.$$

Given $\mathcal{C}^{-1}$, constructed in a manner just described, the computation of $\hat{\bar{\mathbf{C}}}^{-1}$ is straightforward and can be evaluated as a product of matrices $(\mathbf{I}_{\mathcal{T}} \otimes \hat{\mathbf{\Sigma}}_h^{-\frac{1}{2}})$ and $\hat{\mathcal{C}}^{-1}$. Hence the computation of $\hat{\mathbf{\Omega}}^{-1}$ now becomes elementary. Moreover, since $\bar{\mathbf{D}} = (\mathbf{D}' : \mathbf{O})'$ and $\mathbf{H} = (\mathbf{I}_{pv} + \mathbf{D}'\mathbf{\Omega}^{11}\mathbf{D})^{-1}$ where, as indicated, $\mathbf{\Omega}^{11}$ is the $pv \times pv$ submatrix in the top-left hand corner of $\mathbf{\Omega}^{-1}$ obtained subsequent to partitioning after the first pv rows and columns then $\mathbf{\Omega}^{11}$ can be evaluated without the need to construct $\mathbf{\Omega}^{-1}$. Thus, the computation of $\hat{\mathbf{\Omega}}^{11}$ is accomplished by multiplying the first pv rows of $\hat{\bar{\mathbf{C}}}^{\prime-1}$ by the first pv columns in $\hat{\bar{\mathbf{C}}}^{-1}$.

The computational appeal of the method proposed here lies in the fact that the direct inversion of a block band Toeplitz matrix $\hat{\mathbf{\Omega}}_{GLS}$ associated with the basic formula (5.2.6) has been completely avoided. Instead, the evaluation of $\hat{\theta}_{GLS}$ is achieved via a specialized computations and a sequence of recursions. This seems to provide a convenient way of implementing the GLS scheme replacing $O((\mathcal{T}v)^2)$ floating point operations by $O(\mathcal{T}v^2)$ thereby bringing about considerable computational

savings when the sample size T is anything other than a very small number.

In the implementation of the proposed scheme, however, we have noted one potential problem. In practice, it will often be necessary to modify some aspects of this algorithm in order for it to operate successfully. In particular, the recursions in (5.4.16) – (5.4.19) like those of (2.3.9) require $\hat{\mathbf{M}}(z)$ to be stable. Obviously, if $\hat{\mathbf{M}}(z)$ is unstable, the quantities $\mathbf{Z}(t)$, $\mathbf{W}(t)$, $\mathbf{U}(t)$ and $\mathbf{V}(t)$ will grow unboundedly with t and numerical overflow would cause the program to abort. Procedures for checking the stability of an operator $\hat{\mathbf{M}}(z)$ and determining its stable equivalent are clearly required and that form the subject-matter of Chapter 6.

5.5 Some Empirical Evidence

Some simulation experiments based on scalar ARMA processes have been conducted to illustrate the asymptotic theory developed in the previous sections. In particular, the convergence of the GLS estimator to the Gaussian estimator is investigated. Some discussions are also provided in relation to the accuracy of the alternative method of implementation proposed in Section (5.4). The illustration is confined to the scalar ARMA case both to avoid the increase in complexity and because the results give a general indication. For all the scalar ARMA processes employed in the current investigation, the data was generated using random number generator (G05EGF) contained in the Numerical Algorithms Group (NAG) library. The routine generates realizations from a given ARMA structure and initializes the series to a stationary position using a procedure due to Wilson (1979). For each of the processes used the sample size T was determined via a Fibonacci sequence, $T_j = T_{j-1} + T_{j-2}$, $j = 3, \ldots, 6$, where the first two values in the sequence were set at $T_1 = 75$ and $T_2 = 150$. Thus, the range of T values considered in our investigation are $T = 75, 150, 225, 375, 600$ and 975. The number of replications performed for each T is given by the relation $R = \lfloor (100,000/T) \rfloor$ where $\lfloor \cdot \rfloor$ is used to denote

Table 5.1: Scalar ARMA Processes employed in simulation

Notation	ARMA structures	Parameter values
P_1	MA (1)	$\theta = 0.3$
P_2	MA(1)	$\theta = 0.9$
P_3	MA(2)	$\theta_1 = 0.3,\ \theta_2 = -0.2$
P_4	MA(2)	$\theta_1 = 1.42,\ \theta_2 = -0.73$
P_5	ARMA(1, 1)	$\phi_1 = 0.8,\ \theta_1 = -0.7$
P_6	ARMA(1, 1)	$\phi_1 = 0.5,\ \theta_1 = -0.5$
P_7	ARMA(1, 2)	$\phi_1 = -0.8,\ \theta_1 = 0.3, \theta_2 = -0.2$
P_8	ARMA(1, 2)	$\phi_1 = 0.6,\ \theta_1 = 1.0, \theta_2 = -0.64$

the integer part of the indicated argument. For each sample size, the computations of the parameter estimates were carried out using a suite of programs designed to perform both the GLS and Gaussian estimation algorithms as presented in Section (5.4) and Chapter 2 (Section (2.3)), respectively. These programs, supplemented by the least squares routine (G05EGF) in the NAG library, were written in FORTRAN 77 and implemented on a SUN SPARC Station 1 utilizing double precision for all real and complex values. For $\hat{\theta}_{GLS}$ the order of the autoregressive approximation used at the first stage was fixed at $h = T^{\frac{1}{2}}$, as suggested by Pukkila et al (1990), and this is in close accord with our theoretical requirements since $T^{\frac{1}{2}} \leq (\log T)^{1.8}$ for the range of T values that we consider.

For ease of discussion of simulation results, we first introduce in Table (5.1) the ARMA structures employed in the investigation arranged according to the number of parameters they possess with the smallest first. It is also worth mentioning that some of the ARMA processes employed by Koreisha and Pukkila (1990) were included among the structures presented in Table (5.1).

To assess the computational accuracy of the technique proposed in Section (5.4) a comparison is made with the method of implementation suggested by Koreisha and

Pukkila (op. cit) which we henceforth refer to as the conventional method. To ensure homogeneous results the same set of data were used to evaluate $\hat{\theta}_{GLS}$ for each of the two methods. The processes $P2$, $P6$ and $P8$ are used for illustration and sample sizes 50 and 100 are considered. The empirical results obtained are summarized in Tables (5.2) and (5.3). In these tables, however, the symbols CGLS and AGLS have been used to denote the conventional method and the technique proposed in Section (5.4), respectively. Table (5.2) contains the parameter estimates, standard errors (σ), mean squared errors (MSE) and bias (B) over all invertible cases from 200 replications on a sample size of 50 and we also use IFL to indicate the number of times $\hat{\mathbf{M}}(z)$ was unstable (non-invertible). Here the flipping operation simply consists of replacing the zeros, z_j say, that are inside the unit circle by $\bar{z}_j^{-1}$ before the next iteration. Table (5.3) is similarly defined but for $T = 100$. The results presented in Tables (5.2) and (5.3) were constructed for the first, third and fifth iteration and the i^{th} *iteration* is denoted by iter i. It is obvious from the results in Tables (5.2) and (5.3) that both CGLS and AGLS produce identical estimates, in most cases, to the third order of magnitude but the greater accuracy of AGLS is evident. These tables also show the frequency at which the non-invertible estimates occur using both methods of implementation. What is perhaps surprising is that there are some cases when unstable estimates occur using CGLS but these seem to be virtually absent when AGLS is used. However, this does not mean that AGLS can completely avoid the stability problem earlier mentioned but it is only a clear indication that AGLS can generate unstable estimates less often than the CGLS. These obvious advantages of using AGLS and the difficulties associated with the use of CGLS lead to the suggestion that AGLS should be employed when consideration is given to the GLS procedure in the estimation of ARMA models.

As a measure of convergence of the GLS estimator to the Gaussian estimator, the average squared difference between the two estimators $C_R = \frac{1}{R}\sum_{i=1}^{R}\{\sum_{j=1}^{p+q}(\hat{\theta}_{ij,G} - \hat{\theta}_{ij,GLS})^2\}$ is compared with the criterion measure, $C_T = T^{-1}$. Convergence is taken

Table 5.2: Comparison of results for CGLS and AGLS over 200 replications for a sample of size 50

Simulated model	Parameters	CGLS			AGLS		
		iter 1	iter 3	iter 5	iter 1	iter 3	iter 5
P2	$\hat{\theta}$	0.7851	0.7929	0.7930	0.7856	0.7930	0.7934
	$\sigma(\hat{\theta})$	0.0888	0.0866	0.0866	0.0831	0.0805	0.0805
	B	0.1149	0.1071	0.1070	0.1144	0.1070	0.1065
	IFL	1	3	3	0	0	0
P6	$\hat{\phi}$	0.4470	0.4423	0.4421	0.4471	0.4425	0.4422
	$\sigma(\hat{\phi})$	0.1723	0.1742	0.1742	0.1721	0.1733	0.1733
	B	0.0530	0.0577	0.0579	0.0529	0.0575	0.0578
	$\hat{\theta}$	-0.4762	-0.4859	-0.4866	-0.4759	-0.4866	-0.4864
	$\sigma(\hat{\theta})$	0.1632	0.1644	0.1645	0.1625	0.1643	0.1641
	B	-0.0238	-0.0141	-0.0134	-0.0241	-0.0133	-0.0136
	IFL	0	0	0	0	0	0
P8	$\hat{\phi}$	0.2222	0.2775	0.2838	0.2227	0.2794	0.2864
	$\sigma(\hat{\phi})$	0.3914	0.4170	0.4285	0.3915	0.4177	0.4296
	B	0.3778	0.3225	0.3162	0.3772	0.3206	0.3136
	$\hat{\theta}_1$	0.6341	0.6865	0.6922	0.6342	0.6881	0.6945
	$\sigma(\hat{\theta}_1)$	0.3664	0.3966	0.4113	0.3668	0.3992	0.4144
	B	0.3659	0.3135	0.3078	0.3658	0.3119	0.3055
	$\hat{\theta}_2$	-0.4399	-0.4681	-0.4707	-0.4399	-0.4691	-0.4720
	$\sigma(\hat{\theta}_2)$	0.2201	0.2256	0.2284	0.2198	0.2257	0.2286
	B	-0.2001	-0.1719	-0.1693	-0.2001	-0.1709	-0.1680
	IFL	0	0	0	0	0	0

Table 5.3: Comparison of results for CGLS and AGLS over 200 replications for a sample of size 100

Simulated model	Parameters	CGLS			AGLS		
		iter 1	iter 3	iter 5	iter 1	iter 3	iter 5
P2	$\hat{\theta}$	0.8387	0.8443	0.8446	0.8383	0.8445	0.8448
	$\sigma(\hat{\theta})$	0.0499	0.0480	0.0486	0.0475	0.0465	0.0466
	B	0.0613	0.0557	0.0554	0.0617	0.0555	0.0552
	IFL	0	0	1	0	0	0
P6	$\hat{\phi}$	0.4706	0.4657	0.4654	0.4706	0.4657	0.4655
	$\sigma(\hat{\phi})$	0.1089	0.1093	0.1092	0.1086	0.1089	0.1089
	B	0.0294	0.0343	0.0346	0.0293	0.0343	0.0345
	$\hat{\theta}$	-0.4917	-0.5001	-0.5005	-0.4916	-0.5001	-0.5005
	$\sigma(\hat{\theta})$	0.1115	0.1114	0.1114	0.1115	0.1113	0.1113
	B	-0.0083	0.0001	0.0005	-0.0084	0.0001	0.0005
	IFL	0	0	0	0	0	0
P8	$\hat{\phi}$	0.4126	0.5025	0.5162	0.4138	0.5049	0.5188
	$\sigma(\hat{\phi})$	0.2423	0.2119	0.2012	0.2419	0.2112	0.2002
	B	0.1874	0.0975	0.0838	0.1862	0.0951	0.0812
	$\hat{\theta}_1$	0.8363	0.9192	0.9321	0.8375	0.9216	0.9347
	$\sigma(\hat{\theta}_1)$	0.2185	0.1848	0.1743	0.2186	0.1850	0.1743
	B	0.1637	0.0808	0.0678	0.1625	0.0784	0.0653
	$\hat{\theta}_2$	-0.5485	-0.5869	-0.5921	-0.5491	-0.5879	-0.5931
	$\sigma(\hat{\theta}_2)$	0.1244	0.0957	0.0897	0.1245	0.0957	0.0897
	B	-0.0915	-0.0531	-0.0479	-0.0909	-0.0521	-0.0468
	IFL	0	0	0	0	0	0

to have occurred if $C_R \leq C_T$. The choice of C_T in itself is guided by the results of Subsection (5.3.2). In Table (5.4), the values of C_T and C_R are presented for processes $P_1 - P_8$ over the whole range of T values given above. Since the accuracy of the individual parameter estimates would be better described by the mean squared error, this was also constructed and included in Table (5.4) for both the GLS and Gaussian estimators. For aesthetic reasons, we have organized the ARMA structures, $P_1 - P_8$, into Table (5.4) according to the sample size at which convergence appears to have occurred and the values of C_R were not recorded beyond that sample size.

The results presented in Table (5.4) are based on the parameter estimates from the first iteration. Our choice is guided by the fact that in moderate to large samples the parameter estimates from the two procedures, in most cases, seem not to show much alteration with iterations. From Table (5.4), we see that convergence has taken place for most of the processes considered by the time a sample size of 150, or less, has been reached. This reflects the general situation in all the simulations experiments we performed. There are, of course, cases where convergence has not actually occurred by our criterion until T is at least 375. See for example processes P_4 and P_8. Some insight into the mechanism giving rise to such behaviour can be gained by examining the mean squared error of the two estimators. When convergence does not take place for $T \leq 150$ it is usually the case that the performance of the GLS estimates is only slowly improving with sample size. Conversely, the performance of the Gaussian estimator, which tends to be poor in small samples, improves very rapidly as the value of T increases. As an illustration, we show in Table (5.5) the mean squared errors plus the standard errors and biases of the estimates for the mixed process P_8 over the whole range of T values. The results in Table (5.5) reflect the situation where a substantial sample size ($T > 150$) might be required before the two estimators converge. Though the GLS estimator has superior small sample performance this does not seem to carry over into the large sample situation, here the

Table 5.4: Convergence results for the GLS and Gaussian estimators.

	T	75	150	225	375	600	975
	R	1333	666	444	266	166	102
	C_T	0.0133	0.0067	0.0044	0.0027	0.0017	0.001
	C_R	0.0075					
P_2	MSE(θ_{GLS})	0.0108					
	MSE($\hat{\theta}_G$)	0.0109					
	C_R	0.0040					
P_3	MSE($\hat{\theta}_{GLS}$)	0.0344					
	MSE($\hat{\theta}_G$)	0.0342					
	C_R	0.0092					
P_5	MSE($\hat{\theta}_{GLS}$)	0.0194					
	MSE($\hat{\theta}_G$)	0.0325					
	C_R	0.0086					
P_7	MSE($\hat{\theta}_{GLS}$)	0.0537					
	MSE($\hat{\theta}_G$)	0.0540					
	C_R	0.0159	0.0005				
P_1	MSE($\hat{\theta}_{GLS}$)	0.0168	0.0074				
	MSE($\hat{\theta}_G$)	0.0166	0.0072				
	C_R	0.2592	0.0015				
P_6	MSE($\hat{\theta}_{GLS}$)	0.0370	0.0179				
	MSE($\hat{\theta}_G$)	0.0412	0.0180				
	C_R	0.0365	0.0128	0.0102	0.0015		
P_4	MSE($\hat{\theta}_{GLS}$)	0.0372	0.0108	0.0052	0.0028		
	MSE($\hat{\theta}_G$)	0.1108	0.0409	0.0218	0.0060		
	C_R	0.2988	0.0306	0.0175	0.0055	0.0016	
P_8	MSE($\hat{\theta}_{GLS}$)	0.2940	0.0966	0.0395	0.0183	0.0081	
	MSE($\hat{\theta}_G$)	0.2377	0.1222	0.0454	0.0157	0.0071	

Table 5.5: Summary of simulation results for process P_8 example

Parameters	T = 75; R = 1333		T = 150; R = 666		T = 225; R = 444	
	GLS	G	GLS	G	GLS	G
$\hat{\phi}$	0.369	0.508	0.479	0.601	0.531	0.600
$\sigma(\hat{\phi})$	0.289	0.308	0.180	0.247	0.119	0.149
B	0.231	0.092	0.121	-0.001	0.069	0.000
$\hat{\theta}_1$	0.790	0.926	0.896	1.000	0.945	0.999
$\sigma(\hat{\theta}_1)$	0.270	0.302	0.164	0.229	0.110	0.137
B	0.210	0.074	0.104	0.000	0.055	0.001
$\hat{\theta}_2$	-0.517	-0.558	-0.587	-0.612	-0.612	-0.619
$\sigma(\hat{\theta}_2)$	0.157	0.176	0.094	0.089	0.068	0.062
B	-0.123	-0.082	-0.053	-0.028	-0.028	-0.021
Parameters	T = 375;	R = 266	T = 600;	R = 166	T = 975;	R = 102
	GLS	G	GLS	G	GLS	G
$\hat{\phi}$	0.558	0.601	0.580	0.604	0.587	0.600
$\sigma(\hat{\phi})$	0.087	0.088	0.059	0.057	0.042	0.040
B	0.042	-0.001	0.020	-0.004	0.013	0.000
$\hat{\theta}_1$	0.968	1.000	0.987	1.000	0.992	1.000
$\sigma(\hat{\theta}_1)$	0.074	0.075	0.053	0.053	0.038	0.035
B	0.032	0.000	0.013	0.000	0.008	0.000
$\hat{\theta}_2$	-0.624	-0.628	-0.632	-0.634	-0.636	-0.636
$\sigma(\hat{\theta}_2)$	0.048	0.045	0.034	0.033	0.027	0.026
B	-0.016	- 0.012	-0.008	-0.006	-0.004	-0.004

Gaussian estimator seems to perform rather well in comparison with its competitor. For large absolute values of the coefficients, in particular, the GLS estimates are often seriously biased with relatively large mean squared error.

In summary, we have shown that the GLS and Gaussian estimation procedures have equivalent representations. It is also shown that modifications can be made to allow the GLS scheme to be implemented via a sequence of recursions and specialized computations, thus permitting estimation of parameters to be carried out for any desired length of data. Extensive simulations have also been conducted for various model structures and the empirical behaviours of the GLS estimator have been compared with that of the Gaussian estimator. Although the scope of the simulation experiments reported here might be limited, our results strongly suggest that the sample size does not have to be extremely large for the GLS and Gaussian estimators to converge and the gain in performance of the GLS estimator reported in Koreisha and Pukkila (1990) appears to be, very much, a small sample phenomenon.

Chapter 6

Stability Consideration In Estimation

6.1 Motivation

The problem addressed in this chapter is motivated by a consideration of the Gauss-Newton iterative scheme to estimate the parameters of an autoregressive moving average model,

$$\sum_{j=0}^{p} \mathbf{A}(j)\mathbf{y}(t-j) = \sum_{j=0}^{p} \mathbf{M}(j)\varepsilon(t-j), \quad \mathbf{A}(0) = \mathbf{M}(0) \tag{6.1.1}$$

where the disturbance sequence $\{\varepsilon(t)\}$ satisfies assumption (1.1.2) or more generally, Condition A. As in Chapter 1, $\mathbf{y}(t)$ is the (observed) v-dimensional output process and $(\mathbf{A}(j),\ \mathbf{M}(j)) \in \mathcal{R}^{v\times v}$ for all j are parameter matrices. The standard assumptions of Chapter 1 are assumed throughout here. In particular, the matrix pair $[\mathbf{A}(z) : \mathbf{M}(z)]$ are assumed to be left co-prime and in (reversed) echelon form (cf properties (1.2.8)) where the generating functions, $\mathbf{A}(z)$ and $\mathbf{M}(z)$, are as defined in (1.1.3) and satisfy the assumptions (1.1.4) and (2.2.3). The imposition of the structure (1.2.8) on model (6.1.1), however, extends the class of models considered in Poskitt and Salau (1993a) in dealing with the issue addressed in this chapter. Such an extension is found necessary because the basic problem considered here will be manifest whichever canonical form is employed.

As a way of introducing the problem let us once again consider the iterative scheme,

$$\hat{\theta}_G^{(j+1)} = \hat{\theta}_G^{(j)} - (\sum_{t=1}^{T} \frac{\partial \mathbf{e}_\theta(t)'}{\partial\theta}\boldsymbol{\Sigma}(\theta_G)^{-1}\frac{\partial \mathbf{e}_\theta(t)}{\partial\theta'})^{-1}(\sum_{t=1}^{T} \frac{\partial \mathbf{e}_\theta(t)'}{\partial\theta}\boldsymbol{\Sigma}(\theta_G)^{-1}\mathbf{e}_\theta(t))|_{\theta=\hat{\theta}_G^{(j)}},$$
$$j = 0, 1, 2, \ldots, \tag{6.1.2}$$

presented in Chapter 2 (Section (2.3)) where θ is a vector of freely varying parameters in (6.1.1) and is as defined in the previous chapters. $\theta_G^{(0)}$ is some consistent estimator given prior to the commencement of the iterative scheme (6.1.2). In the light of Lemma (3.3.3), one possible choice for $\theta_G^{(0)}$ is the least squares estimator, $\hat{\theta}_{LS}$.

Now interpreting $\hat{\theta}_G^{(j+1)} - \hat{\theta}_G^{(j)}$ as the coefficient in a regression of $\mathbf{e}_\theta(t)$ on $-\frac{\partial \mathbf{e}_\theta(t)}{\partial\theta'}$

then it is only necessary to evaluate $\mathbf{e}_\theta(t)$ and $\frac{\partial \mathbf{e}_\theta(t)}{\partial \theta'}$ at $\hat{\theta}_G^{(0)}$ in order to complete a simple calculation yielding a fully efficient estimate. The residuals $\mathbf{e}_\theta(t)$ are generated via

$$\begin{aligned} \mathbf{e}_\theta(t) &= \sum_{j=0}^{t-1} \mathbf{\Phi}(j)\mathbf{y}(t-j), \\ \mathbf{\Phi}(z) &= \mathbf{M}(z)^{-1}\mathbf{A}(z) = \sum_{j=0}^{\infty} \mathbf{\Phi}(j)z^j \end{aligned} \tag{6.1.3}$$

where the subscript θ is used once again to denote the dependence of the residuals on the parameter vector θ. Similarly, the subscript T will be used in the sequel to indicate the dependence of the indicated argument on the sample size T. It needs be mentioned once again that the effect of starting up the process with $(\mathbf{y}(t), \mathbf{e}_\theta(t)) = 0,\ t \leq 0$ is quite easily seen in (6.1.3) namely, the infinite order of pure vector autoregressive representation is truncated at lag $t-1$ for $\mathbf{y}(t)$. In our earlier results (Chapter 2, Section (2.3)) we obtained an explicit expression for $\frac{\partial \mathbf{e}_\theta(t)}{\partial \theta'}$ as

$$\frac{\partial \mathbf{e}_\theta(t)}{\partial \theta'} = (\eta(t-1), \ldots, \eta(t-p) : \eta(t) - \zeta(t) : \zeta(t-1), \ldots, \zeta(t-p))\mathbf{S}_\nu'$$

where the $v \times v^2$ matrix sequences $\eta(t)$ and $\zeta(t)$ are defined recursively via

$$\sum_{j=0}^{p} \mathbf{M}(j)\eta(t-j) = (\mathbf{y}(t)' \otimes \mathbf{I}_v)$$

and

$$\sum_{j=0}^{p} \mathbf{M}(j)\zeta(t-j) = -(\mathbf{e}_\theta(t)' \otimes \mathbf{I}_v) \tag{6.1.4}$$

with $\eta(t) = \zeta(t) = 0,\ t \leq 0$ and $\mathbf{M}(z)$ denotes the moving average component of $\theta_G^{(j)},\ j \geq 0$.

From expressions (6.1.3) - (6.1.4) it is clear that unless the iterates, and $\hat{\theta}_G^{(0)} = \hat{\theta}_{LS}$ in particular, are such that the associated moving-average operator can be guaranteed to be miniphase and invertible the calculations may break down due to numerical overflow. Of course, in practical situation random flunctuations may cause the estimates to occur outside the stability and invertibiliy region even though the

true parameters satisfy the conditions required for stability. This problem seems to be a fatal flaw in the use of iterative scheme (6.1.2) or the recursions in (5.4.16) - (5.4.19) and has been alluded to elsewhere in the literature, see Hannan and Deistler (1988, p.300). As is shown in Chapter 5 the possibility of generating non-stationary and/or non-invertible operators appears to be a common feature in the estimation of model (6.1.1). An example of this phenomenon in the vector case is given in Hannan and Kavalieris (1984, pp.528-529) when fitting an echelon canonical structure with Kronecker indices $n_1 = 1$ and $n_2 = 2$ to the well known mink-muskrat data of the Hudson Bar Company, Jones (1914). This points to an obvious fact that the stability problem appears to trouble all existing estimation algorithms where the only exception could be a procedure due to Walker (1962) which seems difficult to extend to the multivariate case.

However, some procedures are readily available in both the statistical and engineering literature for overcoming this problem see, for example, Davis (1963), Wilson (1972), Robinson (1983), and Hannan and Kavalieris (op. cit.). Although these procedures provide some framework for solving the stability problem, there are some problems with their use and these follow from the following theorem.

Theorem (6.1.1):

Let $\mathcal{W}(z) = \sum_{j=0}^{n} \mathcal{W}(j) z^j$ *be a* $v \times v$ *matrix polynomial with real coefficients of full rank on* $|z| = 1$ *but otherwise arbitrary. If*

$$\mathcal{G}(z) = \mathcal{W}(z)\mathcal{W}(z^{-1})'$$

then there exists an equivalent polynomial, denoted by $\mathcal{W}_\dagger(z)$, *such that* $\mathcal{G}(z) = \mathcal{W}_\dagger(z)\mathcal{W}_\dagger(z^{-1})'$ *and* $\mathcal{W}_\dagger(z)$ *has real coefficients and* $\det \mathcal{W}_\dagger(z) \neq 0, |z| \leq 1$. *Moreover,* $\mathcal{W}_\dagger(z)$ *is unique up to post multiplication by an arbitrary* $v \times v$ *real orthogonal matrices and can be uniquely determined by imposing the requirement that* $\mathcal{W}_\dagger(0)$ *is lower triangular with positive diagonal elements.*

Proof: This theorem is but a special case of the spectral factorization result associated with Wold's decomposition. See, for example, Rosanov (1967). □

The factorization property in Theorem (6.1.1) is often established by reference to abstract arguments, via the theories and ideas of stochastic processes for instance, and these do not readily lend themselves to practical implementation. To obviate this problem Wilson (1972) has suggested the use of an iterative method to obtain $\mathcal{W}_\dagger(z)$ which is implemented by solving a large system of linear equations or using numerical integration to obtain successive sets of coefficients. The convergence of the iterates to $\mathcal{W}_\dagger(z)$ is stated to be quadratic in nature but precise information on the sensitivity of the procedure to variations in convergence criteria or choice of numerical integration technique is not given. An alternative algebraic approach, which, like Wilson's, works explicitly with the spectrum $\mathcal{G}(z)$ is presented in Robinson (1983, chap. 5). The approach taken by Robinson is to break down $\mathcal{G}(z)$ into a product of linear factors and to group these so as to ensure that $\det \mathcal{W}_\dagger(z) \neq 0, \mid z \mid \leq 1$. The derivations and algorithm, however, require that the degree of $\det \mathcal{W}(z)$ equal nv and that the zeros are distinct and in the context of echelon canonical form these conditions cannot be guaranteed. We may, however, avoid factorizing $\mathcal{G}(z)$ both because of the labour involved and, more importantly because we may not be sure that $\mathcal{G}(z)$ will factor. In fact Tuan (1984) provided an example in the univariate case where $\mathcal{G}(z)$ does not factor. Of course, this may arise in situations where the model is mis-specified or in the presence of sample variability. A different method altogether was introduced in Hannan and Kavalieris (op. cit.) to obtain $\mathcal{W}_\dagger(z)$. Their approach appears attractive because of its directness and simplicity but, like Wilson's procedure, is iterative and moreover, it is not easy to see how to generate $\mathcal{W}_\dagger(z)$ such that a zero, $|\xi| < 1$ say, will be replaced by $|\bar{\xi}^{-1}|$. In other words, there is no guarantee that the relationship $\mathcal{W}(z)\mathcal{W}(z^{-1})' = \mathcal{W}_\dagger(z)\mathcal{W}_\dagger(z^{-1})'$ will hold. Nevertheless, the procedure has some virtues and we defer the discussion of these until

we have presented our procedure.

For reasons given above we now seek an algebraic procedure that will evaluate $\mathcal{W}_\dagger(z)$ for an arbitrary operator $\mathcal{W}(z)$. A feature of the method advocated here is that it is not necessary to explicitly construct $\mathcal{G}(z)$ for, although the technique relies implicitly on the spectral factorization theorem, it is motivated by the following easily established consequence:

Corollary (6.1.1):

Given $\mathcal{G}(z)$ as in Theorem (6.1.1), any rational spectral factor $\tilde{\mathcal{W}}(z)$ such that $\mathcal{G}(z) = \tilde{\mathcal{W}}(z)\tilde{\mathcal{W}}(z^{-1})^*$, where the asterisk denotes the complex conjugate transpose, is given by $\tilde{\mathcal{W}}(z) = \mathcal{W}(z)\mathbf{E}(z)$ where $\mathbf{E}(z)$ is a $v \times v$ rational matrix such that $\mathbf{E}(z)\mathbf{E}(z^{-1})^* = I$ on $\mid z \mid = 1$. □

Using the result of Corollary (6.1.1) has the merit of leading to a direct, closed-form solution to the problem that is relatively straightforward to implement numerically. As a preliminary step to the determination of $\mathcal{W}_\dagger(z)$, a simple method of assessing the stability properties of $\mathcal{W}(z)$ is, therefore, required. This forms the focus of the next section. Before continuing, note that we will henceforth refer to the requirement that $\det \mathcal{W}(z) \neq 0, \mid z \mid < 1$, as merely the stability condition and will drop explicit reference to the equivalent qualification miniphase and invertible, except where the context makes the use of the latter more natural.

6.2 Tests for Stability

In this section the problem is considered of checking whether the condition $\det \mathcal{W}(z) \neq 0, \mid z \mid \leq 1$, is violated. For ease of exposition, we commence by presenting a result from the algebra of matrix polynomials.

Theorem (6.2.1):

Every non-singular $v \times v$ polynomial $\mathcal{W}(z)$ can be transformed by left multiplication

by a unimodular matrix $\mathbf{U}(z)$ *to a unique matrix* $\mathbf{H}(z)$*, called the Hermite normal form, with the following properties:*

(i) $\mathbf{H}(z)$ *is lower triangular*

(ii) $h_{ii}(z)$ $(i = 1, \ldots, v)$ *are monic polynomials.*

(iii) $\delta\{h_{ji}\} < \delta\{h_{ii}\}$, $j \neq i$, *where* δ *denotes the degree of the polynomial indicated.*

(iv) $\delta\{\det \mathbf{H}(z)\} = \delta\{\det \mathcal{W}(z)\}$ *and the zeros of* $\det\{\mathcal{W}(z)\}$ *and* $\det\{\mathbf{H}(z)\}$ *coincide.*

Proof: Using some elementary row transformations properties (i) - (iii) can be readily established. For a detailed exposition of the required manipulations see, for example, Wolovich (1974, chap. 2) and MacDuffee (1956, p.78). Property (iv) follows from the fact that by construction $\mathbf{H}(z) = \mathbf{U}(z)\mathcal{W}(z)$ and hence

$$\det\{\mathbf{H}(z)\} = c \cdot \det \mathcal{W}(z)$$

where c is a non-zero constant since $\mathbf{U}(z)$ is unimodular. The result is now immediate, see Birkhoff and Maclane (1977, chapter 3, Theorem 4). □

It is, perhaps, worth pointing out that the reduction of a polynomial matrix to its Hermite normal form is a fairly efficient numerical procedure which is effected via the Euclidean division algorithm. Note that

$$\begin{aligned} h_{ii}(z) &= z^{n_i} + \sum_{r=0}^{n_i-1} h_{ii}(r)z^r, \quad i = 1, \ldots, v; \\ h_{ij}(z) &= \sum_{r=0}^{n_{ji}} h_{ij}(r)z^r, \quad 0 < i < j = 2, \ldots, v;\ n_{ji} < n_i, \end{aligned}$$

and if $h_{ii}(z)$ is unity all other entries in the ith column are zero.

Now suppose $\mathcal{W}(z)$ is a polynomial matrix of interest and that $\mathbf{H}(z)$ is its Hermite Normal form. From Theorem (6.2.1) it is clear that the stability of $\mathcal{W}(z)$ can be assessed by examining the diagonal elements of $\mathbf{H}(z)$ without having to construct

the determinantal equation. The location of the zeros of $h_{ii}(z)$, $i = 1, \ldots, v$, relative to the unit circle can also be ascertained without explicitly determining their values using the following lemmas.

Lemma (6.2.1) [Product rule]:

Let

$$p(z) = p_0 + p_1 z + \cdots + p_n z^n; \ \{p_n \neq 0\} \tag{6.2.1}$$

$$= p_n \prod_{i=1}^{n} (z - \xi_i) \tag{6.2.2}$$

where $\xi_i, i = 1, \ldots, n$ *are the zeros of* $p(z)$. *If* $| \xi_i | > 1$ *for all* $i = 1, \ldots, n$, *then* $| \frac{p_0}{p_n} | > 1$.

Proof: The result is obvious by equating (6.2.1) and (6.2.2). □

Lemma (6.2.1) implies that a necessary condition for all of the $h_{ii}(z), i = 1, \ldots, v$, to be stable is that $| h_{ii}(0) | > 1$ and this provides a simple first check for stability: failure to meet this condition implies that there is at least one zero of $h_{ii}(z)$ that lies within the unit disc. If $| h_{ii}(0) | > 1$ then the Schur-Cohn criterion provides necessary and sufficient conditions that can be readily evaluated to determine if $h_{ii}(z) = 0$ for some $| z | < 1$.

Lemma (6.2.2) [Schur-Cohn]:

Let $p(z)$ *be defined as in Lemma (6.2.1) and set*

$$a_{m-1}(j) = a_m(0)a_m(j) - a_m(m-j)a_m(m) \tag{6.2.3}$$

$j = 0, \ldots, m-1$, $m = n, \ldots, 1$ *with* $a_n(j) = p_j, j = 0, \ldots, n$. *Then* $p(z) \neq 0$, $| z | \leq 1$, *if and only if* $a_i(0) > 0, i = n-1, \ldots, 0$.

Proof: Let

$$p_m(z) = \sum_{j=0}^{m} a_m(j) z^j$$

and suppose that $a_m(0) > 0$ and $p_m(z)$ is stable. From the recursive formula (6.2.3)

$$a_{m-1}(0) = (a_m(0))^2 - (a_m(m))^2$$

and since $p_m(z)$ is assumed to be stable Lemma (6.2.1) implies that $\mid \frac{a_m(0)}{a_m(m)} \mid > 1$ and hence that $a_{m-1}(0) > 0$.

Now put

$$p_m(z) = a_m(m) \prod_{i=1}^{m} (z - \xi_i)$$

with $\mid \xi_i \mid > 1,\ i = 1, \ldots, m$ and write the corresponding reverse polynomial as

$$\begin{aligned} p_m^R(z) &= z^m p_m(z^{-1}) \\ &= a_m(m) \prod_{i=1}^{m} (1 - \xi_i z) \end{aligned}$$

Then

$$\mid p_m(z) \mid \geq \mid p_m^R(z) \mid,\ \mid z \mid \leq 1, \tag{6.2.4}$$

since the ratio

$$\frac{p_m(z)}{p_m^R(z)} = \prod_{i=1}^{n} \frac{(z - \xi_i)}{(1 - \xi_i z)}$$

is bounded below in modulus by unity for $\mid z \mid \leq 1$. From (6.2.3) we observe that

$$p_{m-1}(z) = a_m(0) p_m(z) - a_m(m) p_m^R(z) \tag{6.2.5}$$

and employing this expression in conjuction with (6.2.3) and (6.2.4) we find that

$$\begin{aligned} \mid p_{m-1}(z) \mid &\geq \mid a_m(0) \mid \cdot \{ \mid p_m(z) \mid - \mid \frac{a_m(m)}{a_m(0)} \mid \cdot \mid p_m^R(z) \mid \} \\ &> 0,\ \mid z \mid \leq 1, \end{aligned}$$

implying that $p_{m-1}(z)$ is stable. Hence we have shown that $a_m(0) > 0$ and $p_m(z)$ stable imply that $a_{m-1}(0) > 0$ and $\mid p_{m-1}(z) \mid \neq 0,\ \mid z \mid \leq 1$, and application of the principle of finite induction establishes necessity. Conversely, assume that $a_i(0) > 0,\ i = n-1, \ldots, 0$. It is clear that $p_1(z)$ must be stable because $a_0(0) > 0$. We need to verify, therefore, that $a_{m-1}(0) > 0$ and $\mid p_{m-1}(z) \mid \neq 0,\ \mid z \mid \leq 1$, imply that $p_m(z)$ is stable in order to complete the reverse induction. From (6.2.5) it follows that

$$z\, p_{m-1}^R(z) = a_m(0) p_m^R(z) - a_m(m) p_m(z) \tag{6.2.6}$$

and expressing $p_m(z)$ in terms of $p_{m-1}(z)$ and $p^R_{m-1}(z)$ using (6.2.5) and (6.2.6) gives

$$p_m(z) = \frac{a_m(0)}{a_{m-1}(0)}\left[p_{m-1}(z) + \frac{a_m(m)}{a_m(0)} z p^R_{m-1}(z)\right].$$

The inequality (6.2.4) and, (6.2.5) for m = m-1, now imply that $\mid p_m(z) \mid > 0$ for $\mid z \mid \leq 1$, that is, $p_m(z)$ has no zeros within the unit disc, as required. □

We should perhaps mention that the Schur-Cohn result is, of course, well known and has been discussed in some detail in both the engineering literature, Barnett (1972), and in the context of complex analysis, Henrici (1974). The current proof is presented because it emphasizes the role played by the product $\prod \frac{(z-\xi_i)}{(1-\xi_i z)}$, which is itself intimately related to Corollary (6.1.1). Matrix functions of the type $\mathbf{E}(z)$ are referred to as extended-unitary and are associated with the invariant subspaces of the Hardy space, H_2, of functions analytic in $|z| < 1$ and square integrable on $|z| = 1$, which are characterized by inner functions. The function-theoretic structure of such functions is known and, ignoring singular components, is such that their determinant is a product $c \cdot b(z)$ where c is a complex constant of unit modulus and b(z) is a Blaschke function. For a detailed exposition of the above, see Helson (1964). Blaschke functions are products of the form

$$b(z) = z^k \prod_{j=1}^{\infty} \frac{(\tau_j - z)}{(1 - \bar{\tau}_j z)} \cdot \frac{\bar{\tau}_j}{\mid \tau_j \mid}$$

where k is zero or a positive integer and the τ_j are complex numbers satisfying $0 < \mid \tau_j \mid < 1$ and $\sum_{j=1}^{\infty}(1 - \mid \tau_j \mid) < \infty$. It is this partial characterization that forms the technical background to the reflection scheme outlined in the next section.

It is important to note, however, that it is only when $h_{ii}(z)$ is found unstable by one or other of these tests that it will be necessary to evaluate the zeros of $h_{ii}(z)$. Should $h_{ii}(z)$ be found to be unstable by one or other of these tests then, and only then, it will be required to evaluate the zeros of $h_{ii}(z)$ as a preliminary step to the reflection process. Although the computation of the zeros is not of direct concern here it is clear that the accuracy with which the zeros of $h_{ii}(z)$ can be evaluated will

impinge on the numerical accuracy of the technique. Standard algorithms designed to extract the zeros of quadratic and higher degree polynomials with a high level of precision are readily available, Conte (1965) and Hildebrand (1974), but the difficulty in obtaining these zeros accurately increases rapidly with the polynomial order. This point highlights another feature of our procedure. Although it is possible to work directly with det $\mathcal{W}(z)$, since det $\mathcal{W}(z)$ can be readily calculated without having to determine $\mathbf{H}(z)$, as in Robinson (1983, Chapter 4) for example, if $\mathcal{W}(z)$ is found to be unstable then all the zeros of det $\mathcal{W}(z)$ will require evaluation. With the current approach it is only necessary to calculate the zeros of those factors $h_{ii}(z)$ of det $\mathcal{W}(z) = \prod_{i=1}^{v} h_{ii}(z)$ that lead to instability. This suggests that the current method should yield reasonably good numerical accuracy and moreover working with $\mathbf{H}(z)$ rather than $\mathcal{W}(z)$ has additional advantages as we shall see in the following section.

6.3 The Proposed Procedure

For ease of presentation our discussion in this section has been organized into two subsections dealing with the actual reflection scheme used to generate $\mathcal{W}_\dagger(z)$ from $\mathcal{W}(z)$ should the stability of $\mathcal{W}(z)$ be deemed inappropriate and the algorithmic summary of the procedure, respectively. The latter is presented in a user-ready format so as to provide some guide for the practitioner as to the numerical implementation of the scheme.

6.3.1 Reflection Scheme

Now suppose that the ith diagonal element of $\mathbf{H}(z)$ is unstable and that its zeros $\xi_{i,j},\ j = 1, \ldots, n_i$ are known.

Factorizing $h_{ii}(z)$ we obtain

$$\begin{aligned} h_{ii}(z) &= \prod_{j=1}^{n_i}(z-\xi_{i,j}) \\ &= \prod_{j=1}^{n_i}(z-\xi_{i(j)}) \end{aligned}$$

where $\xi_{i(j)}$ denote the zeros of $h_{ii}(z)$ ordered such that $\mid \xi_{i(j)} \mid \leq \mid \xi_{i(j+1)} \mid$. Presume also that $\mid \xi_{i(j)} \mid < 1,\ j = 1,\ldots,r_i \leq n_i$ and, for ease of exposition, that the $h_{ii}(z),\ i = 1,\ldots,v$, are coprime.

Since $h_{ii}(z)$ is divisible by $(z-\xi_{i(j)})$ it follows, in particular, that $\det \mathbf{H}(z) = 0$ for $z = \xi_{i(1)}$ and that $\mathbf{H}(\xi_{i(1)})$ is singular. Now assume that we can find a constant unitary matrix $\mathbf{V}_{\xi_{i(1)}}$ such that $\mathbf{H}(\xi_{i(1)})\,\mathbf{V}_{\xi_{i(1)}}$ is a matrix whose ith column is identically zero. By construction the ith column of $\mathbf{H}(z)\,\mathbf{V}_{\xi_{i(1)}}$ is divisible by $(z-\xi_{i(1)})$. Set

$$\mathbf{D}_{\xi_{i(1)}}(z) = diag(1,\ldots,1,d_{\xi_{i(1)}}(z),1,\ldots,1)$$

where

$$d_{\xi_{i(1)}}(z) = \frac{(1-\bar{\xi}_{i(1)}z)}{(z-\xi_{i(1)})} \tag{6.3.1}$$

occupies the ith location on the main diagonal. It is now easily verified that the determinant of $\mathbf{H}_{i(1)}(z) = \mathbf{H}(z)\,\mathbf{V}_{\xi_{i(1)}}\,\mathbf{D}_{\xi_{i(1)}}(z)$ equals $constant \cdot \det \mathbf{H}(z) \cdot d_{\xi_{i(1)}}(z)$ and that $\mathbf{E}_{i(1)}(z) = \mathbf{V}_{\xi_{i(1)}}\mathbf{D}_{\xi_{i(1)}}(z)$ is extended unitary. Thus, $\mathbf{H}_{i(1)}(z)$ is equivalent to $\mathbf{H}(z)$ but the zero $\xi_{i(1)}$ has been reflected through the unit circle and replaced by $\bar{\xi}_{i(1)}^{-1} = \frac{\xi_{i(1)}}{|\xi_{i(1)}|^2}$.

To construct $\mathbf{V}_{\xi_{i(1)}}$ consider the singular value decomposition $\mathbf{H}(\xi_{i(1)})\mathbf{P} = \mathbf{Q}\mathbf{\Lambda}$ where the columns of $\mathbf{P}$ are the orthonormal eigenvectors of the Hermitian product $\mathbf{H}(\xi_{i(1)})^*\mathbf{H}(\xi_{i(1)})$, $\mathbf{\Lambda}^2 = diag(\lambda_1^2,\ldots,\lambda_v^2)$ contains the associated non-negative eigenvalues, and $\mathbf{Q} = \mathbf{H}(\xi_{i(1)})\,\mathbf{P}\,\mathbf{\Lambda}^-$ where $\mathbf{\Lambda}^- = diag(\frac{1}{\lambda_1},\ldots,\frac{1}{\lambda_v}),\ \lambda_i \geq 0$ and $\frac{1}{\lambda_i} = 0$ if $\lambda_i = 0,\ i = 1,\ldots,v$. It is easy to check that the matrix $\mathbf{P}$ satisfies the above requirement for $\mathbf{V}_{\xi_{i(1)}}$ and hence $\mathbf{V}_{\xi_{i(1)}}$ is obtained by forming the matrix of

orthonormalized eigenvectors of $\mathbf{H}(\xi_{i(1)})^*\mathbf{H}(\xi_{i(1)})$. This can be achieved, of course, using standard routines such as MATLAB and F02AXF in the Numerical Algorithm Group (NAG) library. One important feature of $\mathbf{V}_{\xi_{i(j)}}$, $j = 1, \ldots, n_i$, to note is that for any two of its eigenvectors, V_s and V_k say, we have $V_s^* V_k = 0$, $s \neq k$, where, as before, the asterisk indicates the complex conjugate combined with transposition.

At this stage an operator $\mathcal{W}_{i(1)}(z)$, given by

$$\begin{aligned}\mathcal{W}_{i(1)}(z) &= \mathbf{U}(z)^{-1}\mathbf{H}_{i(1)}(z) \\ &= \mathbf{U}(z)^{-1}\mathbf{H}(z)\mathbf{V}_{\xi_{i(1)}}\mathbf{D}_{\xi_{i(1)}}(z) \\ &= \mathcal{W}(z)\mathbf{V}_{\xi_{i(1)}}\mathbf{D}_{\xi_{i(1)}}(z),\end{aligned}$$

has implicitly been generated which is such that $\mathcal{W}_{i(1)}(z)$ provides an alternative factorization of $\mathcal{W}(z)\mathcal{W}(z^{-1})'$ but differs from $\mathcal{W}(z)$ in that the zero $\bar{\xi}_{i(1)}^{-1}$ has been exchanged for $\xi_{i(1)}$. Moreover, since by construction the i^{th} column of $\mathbf{H}(z)\mathbf{V}_{\xi_{i(1)}}$ is divisible by $(z - \xi_{i(1)})$ then the same must be true of $\mathcal{W}(z)\mathbf{V}_{\xi_{i(1)}}$ because $\mathcal{W}(\xi_{i(1)}) = \mathbf{U}(\xi_{i(1)})^{-1}\mathbf{H}(\xi_{i(1)})$. Regarding each column of $\mathcal{W}(z)\mathbf{V}_{\xi_{i(1)}}$ as a linear combination of the columns of $\mathcal{W}(z)$ with weights given by the corresponding column of $\mathbf{V}_{\xi_{i(1)}}$ it is apparent that not only does $\delta(\det \mathcal{W}_{i(1)}(z)) = \delta(\det \mathcal{W}(z))$ but the row degrees of $\mathcal{W}(z)$ and $\mathcal{W}_{i(1)}(z)$ are also the same.

By successive applications of the above procedure the remaining zeros of $h_{ii}(z)$, $\xi_{i(2)}, \ldots, \xi_{i(r_i)}$, that lie inside the unit circle can be reflected through the unit circle, the zero with the smallest modulus being changed to its complex conjugate reciprocal at each repetition.

It is worthy of note that if the $h_{ii}(z)$ are not coprime, the above procedure is straightforward to modify. If $h_{rr}(z)$ and $h_{ss}(z)$ have the common factor $(z-\xi)$, $|\xi| < 1$, for example, we can employ the same technique to determine a unitary matrix $\mathbf{V}_\xi$ such that the r and s columns of $\mathbf{H}(\xi)\mathbf{V}_\xi$ are null. Then $\mathbf{H}(z)\mathbf{V}_\xi\mathbf{D}_\xi(z)$, where $\mathbf{D}_\xi(z) = diag(1, \ldots, 1, d_\xi(z), 1, \ldots, 1, d_\xi(z), 1, \ldots, 1)$ with $d_\xi(z)$ in both the r and s leading diagonal positions, will be such that ξ will be reflected through the unit

circle in both locations simultaneously.

Once all the zeros of det $\mathcal{W}(z)$ that lie inside the unit circle have been modified in the manner just described we are left with a polynomial $\tilde{\mathcal{W}}(z) = \mathcal{W}(z)\,\mathbf{E}(z)$ where $\mathbf{E}(z)$ is an extended-unitary matrix equal to a product of terms of the form $\mathbf{V}_\xi \mathbf{D}_\xi(z)$ considered above. From the Corollary (6.1.1) it is clear that $\mathcal{W}(z)\,\mathcal{W}(z^{-1})' = \tilde{\mathcal{W}}(z)\,\tilde{\mathcal{W}}(z^{-1})^*$ where $\tilde{\mathcal{W}}(z)$ has the same degree as $\mathcal{W}(z)$. Moreover, det $\tilde{\mathcal{W}}(z) \neq 0, \quad \mid z \mid < 1$, by construction. It follows that $\tilde{\mathcal{W}}(z)\mathbf{U} = \mathcal{W}_\dagger(z)$ for some constant unitary matrix $\mathbf{U}$. In particular, $\tilde{\mathcal{W}}(0)\mathbf{U} = \mathcal{W}_\dagger(0)$ which implies that $\tilde{\mathcal{W}}(0)\,\tilde{\mathcal{W}}(0)^* = \mathcal{W}_\dagger(0)\mathcal{W}_\dagger(0)'$ is real valued. Setting $(\tilde{\mathcal{W}}(0)\tilde{\mathcal{W}}(0)^*)^{\frac{1}{2}}$ equal to the unique lower triangular Cholesky factor of the product it is easily checked that $\tilde{\mathbf{U}} = \tilde{\mathcal{W}}(0)^{-1}(\tilde{\mathcal{W}}(0)\tilde{\mathcal{W}}(0)^*)^{\frac{1}{2}}$ is unitary. Thus we can generate a polynomial matrix $\tilde{\mathcal{W}}(z)\tilde{\mathbf{U}}$ such that $\tilde{\mathcal{W}}(0)\tilde{\mathbf{U}}$ is real valued, lower triangular with positive diagonal elements and equating coefficients in $\mathcal{G}(z) = \tilde{\mathcal{W}}(z)\tilde{\mathbf{U}}\tilde{\mathbf{U}}^*\tilde{\mathcal{W}}(z^{-1})^*$ we can readily deduce from the Hermitian symmetry of $\mathcal{G}(z)$ that all the coefficients of $\tilde{\mathcal{W}}(z)\tilde{\mathbf{U}}$ are real valued. It follows that $\tilde{\mathcal{W}}(z)\tilde{\mathbf{U}}$ equals the unique normalised stable equivalent of $\mathcal{W}(z)$, $\mathcal{W}_\dagger(z)$, and hence evaluation of $\tilde{\mathbf{U}}$ and subsequent determination of $\tilde{\mathcal{W}}(z)\tilde{\mathbf{U}}$, which we observe can be achieved using standard finite algorithms, completes the stabilization process.

6.3.2 Algorithmic Summary

In order to facilitate a clearer understanding and provide a guide as to the numerical implementation of the scheme we now summarize the procedure in the following steps.

Step 1: Convert the given polynomial matrix $\mathcal{W}(z)$ into Hermite Normal form with $h_{ii}(z), i = 1, \ldots, v$ denoting its diagonal elements.

Step 2: For $i = 1, \ldots, v$ test $h_{ii}(z)$ for stability using the product rule and/or the Schur-Cohn condition.

Step 3: If the test in step 2 fails (i. e. $h_{ii}(z)$ has some zeros inside the unit circle for some i) then evaluate the zeros of those $h_{ii}(z)$ that are unstable.

Step 4: Denote the zeros of such $h_{ii}(z)$ by $\xi_{i,j}, j = 1, \ldots, n_i$ and order the zeros in ascending order of magnitude such that $\mid \xi_{i(j)} \mid \leq \mid \xi_{i(j+1)} \mid$. Assume $\mid \xi_{i(j)} \mid < 1$ for $j = 1, \ldots, r_i \leq n_i$. Set $j = 1$.

Step 5: Construct the diagonal matrix

$$\mathbf{D}_{\xi_{i(1)}} = diag(1, \ldots, 1, d_{\xi_{i(j)}}, 1, \ldots, 1)$$

with $d_{\xi_{i(j)}}(z)$ as defined in (6.3.1).

Step 6: Form the matrix $\mathbf{V}_{\xi_{i(j)}}$ from the orthonormalized eigen-vectors associated with $\mathbf{H}(\xi_{i(j)})^*\mathbf{H}(\xi_{i(j)})$ and evaluate

$$\mathcal{W}_{i(j)}(z) = \mathcal{W}(z)\mathbf{V}_{\xi_{i(j)}}\mathbf{D}_{\xi_{i(j)}}$$

noting that $\mathcal{W}(z) = \mathbf{U}(z)^{-1}\mathbf{H}(z)$.

Repeat steps 5 and 6 for $j = 2, \ldots, r_i$ with $\mathcal{W}_{i(j-1)}$ replacing $\mathcal{W}(z)$ in step 6.

Step 7: For all i for which $h_{ii}(z)$ is unstable reiterate steps 4 through to 6 to obtain $\tilde{\mathcal{W}}(z)$.

Step 8: Compute the matrix $\tilde{\mathbf{U}} = \tilde{\mathcal{W}}(0)^{-1}(\tilde{\mathcal{W}}(0)\tilde{\mathcal{W}}(0)^*)^{\frac{1}{2}}$ and obtain $\mathcal{W}_\dagger(z)$, the stable equivalent of $\mathcal{W}(z)$, as $\mathcal{W}_\dagger(z) = \tilde{\mathcal{W}}(z)\tilde{\mathbf{U}}$.

In what follows we will illustrate this algorithm.

6.4 Illustrative Examples

In this section we now wish to demonstrate by means of numerical examples on how the procedure introduced in Section (6.3) can be applied in practical situations.

As an illustration, consider the matrix operator $\mathcal{W}(z)$ defined by

$$\mathcal{W}(z) = \begin{pmatrix} 2.25 & -3 \\ -4.75 & 6 \end{pmatrix} + \begin{pmatrix} 0 & 4 \\ -4.75 & -2 \end{pmatrix} z + \begin{pmatrix} -1 & -1 \\ 3 & -9 \end{pmatrix} z^2 + \begin{pmatrix} 0 & 0 \\ 3 & 3 \end{pmatrix} z^3 \tag{6.4.1}$$

The first step in the procedure is to generate the Hermite normal form, H(z) (cf. Theorem(6.2.1)) and this is obtained as

$$\mathbf{H}(z) = \begin{pmatrix} z^2 + \frac{1}{4} & 0 \\ 2z + 2 & z - 3 \end{pmatrix} \tag{6.4.2}$$

with the corresponding unimodular matrix

$$\mathbf{U}(z) = \begin{pmatrix} 3z^2 - 2 & z - 1 \\ 3z + 3 & 1 \end{pmatrix}. \tag{6.4.3}$$

The second step of the procedure is to test the diagonal elements of $\mathbf{H}(z)$, (i.e. $h_{11}(z)$, $h_{22}(z)$) for stability. By direct application of the product rule and/or Schur-Cohn criterion, it can be easily verified that $h_{11}(z)$ has zeros for $\mid z \mid < 1$. Hence, as required in step 3 above, the roots of $h_{11}(z)$ need to be evaluated and it is easy to check that $h_{11}(z) = z^2 + \frac{1}{4}$ has $\xi_{1,1} = \frac{1}{2}i$, $\xi_{1,2} = -\frac{1}{2}i$ as its zeros. Since these zeros are in conjugate pairs then the condition of step 4 is naturally satisfied and we are therefore led into step 5. Set $\xi_{1(1)} = \frac{i}{2}$ and construct the diagonal matrix

$$\mathbf{D}_{\xi_{1(1)}}(z) = \begin{pmatrix} \frac{1+\frac{i}{2}z}{z-\frac{i}{2}} & 0 \\ 0 & 1 \end{pmatrix}.$$

In order to complete step 6 we first compute the matrix product

$$\mathbf{H}(\xi_{1(1)})^*\mathbf{H}(\xi_{1(1)}) = \begin{pmatrix} 5.0 & -5.5 + 4i \\ -5.5 - 4i & 9.25 \end{pmatrix}.$$

Then the corresponding unitary matrix generated by the orthonormal eigenvectors of this Hermitian matrix is

$$\mathbf{V}_{\xi_{1(1)}} = \left(\frac{1}{\sqrt{2.85}}\right) \begin{pmatrix} 1.1 - 0.8i & 1.0 \\ 1.0 & -1.1 - 0.8i \end{pmatrix}$$

and it is readily verified that

$$\mathbf{H}(\frac{i}{2})\,\mathbf{V}_{\xi_{1(1)}} = \left(\frac{1}{\sqrt{2.85}}\right)\begin{pmatrix} 0.0 & 0.0 \\ 0.0 & 5.7+2.85i \end{pmatrix}.$$

As has earlier been noted, the first column of the matrix $\mathbf{H}(z)\,\mathbf{V}_{\xi_{1(1)}}$ is clearly divisible by $z-\frac{i}{2}$. Employing some basic matrix manipulations it can be further verified that

$$\mathcal{W}(\frac{i}{2})\mathbf{V}_{\xi_{1(1)}} = \left(\frac{1}{\sqrt{2.85}}\right)\begin{pmatrix} 0.0 & 7.125 \\ 0.0 & -15.675-7.8375i \end{pmatrix}$$

and this, of course, implies that the first column of the matrix $\mathcal{W}(z)\mathbf{V}_{\xi_{i(1)}}$ is also divisible by $(z-\frac{i}{2})$. Thus, the matrix operator equivalent to $\mathcal{W}(z)$ in which the zero $\xi_{1(1)}=\frac{i}{2}$ has been replaced by $\bar{\xi}^{-1}_{1(1)}=-2i$ is

$$\begin{aligned}\mathcal{W}_{1(1)}(z) &= \mathcal{W}(z)\,\mathbf{V}_{\xi_{1(1)}}\,\mathbf{D}_{\xi_{1(1)}}(z) \\ &= \begin{pmatrix} \phi_{1(1),11}(z) & \phi_{1(1),12}(z) \\ \phi_{1(1),21}(z) & \phi_{1(1),22}(z) \end{pmatrix}\end{aligned}$$

where $\phi_{1(1),11}(z)=\{(-0.4-1.05i)z^2+(-1.575+2.6i)z+(3.6-1.05i)\}/\sqrt{2.85}$;

$\phi_{1(1),12}(z)=\{(0.1+0.8i)z^2-(4.4+3.2i)z+(5.55+2.4i)\}/\sqrt{2.85}$;

$\phi_{1(1),21}(z) = \{(1.2+3.15i)z^3+(5.925-4.65i)z^2+(-5.275-3.05i)z+(-7.6+1.55i)\}/\sqrt{2.85}$ and

$\phi_{1(1),22}(z) = \{(-0.3-2.4i)z^3+(12.9+7.2i)z^2+(-2.55+1.6i)z+(-11.35-4.8i)\}/\sqrt{2.85}$.

Repeating steps 5 and 6 for $\xi_{1(2)}=-\frac{i}{2}$ leads to a stable equivalent matrix $\tilde{\mathcal{W}}(z)=\mathcal{W}(z)\,\mathbf{E}(z)$ given by

$$\frac{1}{\sqrt{6.8133}}\left[\begin{pmatrix} 0.1125-3.6i & 10.06875-4.05i \\ 1.3425+8.04i & -20.73375+7.77i \end{pmatrix}\right.$$

$$\left.+\begin{pmatrix} -2.31+1.92i & -7.845+3.96i \\ 5.2125+4.2i & -5.86875+0.45i \end{pmatrix}z\right.$$

$$+\begin{pmatrix} 0.988125+0.255i & 0.3825-0.51i \\ 3.965625-6.525i & 22.3875-10.35i \end{pmatrix} z^2$$

$$+\left.\begin{pmatrix} 0.0 & 0.0 \\ -2.964375-0.765i & -1.1475+1.53i \end{pmatrix} z^3\right]$$

where

$$\mathbf{E}(z) = \prod_{i=1}^{2} \mathbf{V}_{\xi_i}\mathbf{D}_{\xi_i}(z) = \left(\frac{1}{\sqrt{2.85}}\right)\begin{pmatrix} 1.1-0.8i & 1.0 \\ 1.0 & -1.1-0.8i \end{pmatrix}\cdot\begin{pmatrix} \frac{1+\frac{i}{2}z}{z-\frac{i}{2}} & 0 \\ 0 & 1 \end{pmatrix}$$

$$\times \left(\frac{1}{\sqrt{2.390625}}\right)\begin{pmatrix} -0.785-0.88i & 1.0 \\ 1.0 & 0.785-0.88i \end{pmatrix}\cdot\begin{pmatrix} \frac{1-\frac{i}{2}z}{z+\frac{i}{2}} & 0 \\ 0 & 1 \end{pmatrix}$$

is the product of the extended unitary matrices obtained at each successive reflection. Note that step 7 is not required in this example since $h_{22}(z)$ is stable. We should, perhaps, comment that the explicit construction of $\mathbf{E}(z)$ as presented above is given only for purposes of exposition, it is not required in practice. In the implementation of the procedure each component of $\mathbf{E}(z)$ is generated and applied successively.

As earlier pointed out, to obtain the unique real stable equivalent matrix polynomial requires a post multiplication of $\tilde{\mathcal{W}}(z)$ by a constant unitary matrix and this construction constitutes step 8. Thus, $\mathcal{W}_\dagger(z)$, to three decimal places, is

$$\mathcal{W}_\dagger(z) = \tilde{\mathcal{W}}(z)\,\tilde{\mathbf{U}} = \begin{pmatrix} 1.678 & 0.0 \\ -3.453 & 0.262 \end{pmatrix} + \begin{pmatrix} -1.312 & -0.371 \\ -0.968 & 0.88 \end{pmatrix} z$$

$$+\begin{pmatrix} 0.066 & 0.164 \\ 3.739 & 0.619 \end{pmatrix} z^2 + \begin{pmatrix} 0.0 & 0.0 \\ -0.197 & -0.492 \end{pmatrix} z^3$$

where

$$\tilde{\mathbf{U}} = \tilde{\mathcal{W}}(0)^{-1}(\tilde{\mathcal{W}}(0)\tilde{\mathcal{W}}(0)^*)^{\frac{1}{2}}$$

$$= -(20.43984375)^{-1}\begin{pmatrix} 0.20295-6.4344i & -18.002925+7.2414i \\ -17.998275-7.2402i & 0.20115-6.4368i \end{pmatrix}$$

and

$$\tilde{\mathcal{W}}(0)\,\tilde{\mathcal{W}}(0)^* = \begin{pmatrix} 130.755 & -269.024 \\ -269.024 & 556.705 \end{pmatrix}.$$

As can be easily verified, the Cholesky factor is

$$(\tilde{\mathcal{W}}(0)\,\tilde{\mathcal{W}}(0)^*)^{\frac{1}{2}} = \begin{pmatrix} 11.435 & 0.0 \\ -23.527 & 1.788 \end{pmatrix}.$$

Similarly, the corresponding determinantal equation, written to four decimal places, is obtained as

$$\det \mathcal{W}_\dagger(z) = 0.4404 - 0.1468\,z + 0.1101\,z^2 - 0.0367\,z^3.$$

Hence the zeros of $\det \mathcal{W}_\dagger(z)$, as can be easily checked, are $\xi_1 = 3.0$, $\xi_2 = 2i$ and $\xi_3 = \bar{\xi}_2$ as required.

In choosing this example we are constrained to a simple, perhaps obvious, problem that can be easily verified. Despite the comparative simplicity of this example, however, it illustrates all of the salient features of the procedure and, dealing as it does with the complex root case, covers the most complicated situation likely to be encountered. Problems that will manifest themselves in practice can only be of the same kind as that just presented, they cannot introduce anything different in principle.

Although the procedure has been illustrated using an arbitrary operator, the scheme can be readily adapted to deal with cases where $\mathcal{W}(z)$ is in echelon form. However, our choice of the example presented above is motivated by the desire to reflect the general applicability of the proposed procedure. This in itself is necessitated when observing that multivariate filtering problems and that of network synthesis in system and control engineering have both generated the requirement of a canonical factor of the spectral density in the construction of frequency response functions of the linear filter. As a further illustration, the procedure is applied to an unstable

preliminary estimate of the moving average operator,

$$\mathbf{M}_T^{(0)}(z) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \begin{pmatrix} 0.38 & -0.22 \\ 0.52 & -1.10 \end{pmatrix} z + \begin{pmatrix} 0.0 & 0.0 \\ -0.45 & 0.02 \end{pmatrix} z^2, \quad (6.4.4)$$

obtained in Hannan and Kavalieris (1984, p. 528-529) when fitting an echelon canonical structure with Kronecker indices $n_1 = 1$ and $n_2 = 2$ to the mink-muskrat data. The details of the various steps leading to the stable equivalent operator are provided in Appendix (C, p. 165).

6.5 Stability Problem In ARMA Estimation

From our discussion in previous sections it is clear that if any of the iterates $\hat{\theta}_G^{(j)}$, $j \geq 0$, are such that the associated moving average operator, denoted by $\hat{\mathbf{M}}_T^{(j)}(z)$, violates the miniphase and invertibility assumption (2.2.3) then the iterations may break down due to numerical overflow. In practice, random flunctuations in small samples can cause $\hat{\mathbf{M}}_T^{(j)}(z)$ to have zeros in $|z| < 1$ and such a situation is typical of ARMA estimation. Since the $\hat{\mathbf{M}}_T^{(j)}(z), j \geq 0$, are consistent they will, for T sufficiently large, lie within a ϑ_T neighbourhood of the true $\mathbf{M}(z)$ where ϑ_T varies inversely with $T^{\frac{1}{2}}$, viz

$$\|\hat{\mathbf{M}}_T^{(j)}(z) - \mathbf{M}_0(z)\| < \vartheta_T$$

for all $T > T_\vartheta$ where the norm is defined by reference to Parseval's relation,

$$\|\hat{\nabla}_T^{(j)}(z)\|^2 = (2\pi)^{-1}\, tr \int_{-\pi}^{\pi} \hat{\nabla}_T^{(j)}(e^{i\omega})\hat{\nabla}_T^{(j)}(e^{i\omega})^* d\omega,$$

$\hat{\nabla}_T^{(j)} = \hat{\mathbf{M}}_T^{(j)}(z) - \mathbf{M}_0(z)$. This implies, of course, that the zeros of $\det \hat{\mathbf{M}}_T^{(j)}(z)$ and $\det \mathbf{M}_0(z)$ will also be arbitrarily close. Suppose, therefore, using an obvious notation, that $\mid \hat{\xi}_{T(1)}^{(j)} - \xi_{(1)} \mid < \rho(\vartheta_T)$ and that $\mid \hat{\xi}_{T(1)}^{(j)} \mid < 1$. Since the true zero lies outside the unit circle this means that $\mid \hat{\xi}_{T(1)}^{(j)} \mid > 1 - \rho(\vartheta_T)$ and $\mid \xi_{(1)} \mid < 1 + \rho(\vartheta_T)$. This in turn intimates that $\hat{\mathbf{M}}_T^{(j)}(z)$ may not be miniphase and invertible when T

is not large enough to ensure that $\rho(\vartheta_T)$ is sufficiently small and/or the true zero is itself close to the unit circle. An example of this phenomenon was reported in Hannan and Kavalieris (op. cit.) where (6.4.4) is obtained from a sample of size, $T = 62$, and, implicitly, $|\xi_1| < 1.07$. These authors suggest that the difficulty with $\hat{\mathbf{M}}_T^{(j)}(z)$, $j \geq 0$, might be overcome by replacing this operator with

$$\hat{\mathbf{M}}_s^{(j)}(z) = \hat{\mathbf{M}}_T^{(j)}(0) + s\{\hat{\mathbf{M}}_T^{(j)}(z) - \hat{\mathbf{M}}_T^{(j)}(0)\}, \quad 0 < s < 1,$$

where s is chosen as close to one as is consistent with the requirement that $\hat{\mathbf{M}}_s^{(j)}(z)$ satisfies (2.2.3). Such an approach appears attractive because of its simplicity, but no precise guidelines are offered on how to evaluate s in practice. Of course, if the technique were to be used within the framework of Gauss-Newton iterative scheme, as it is designed to be, then s should be chosen such that the objective (likelihood) function, $\mathbf{\Sigma}_s(\hat{\theta}_G)$, is minimized where we have used the subscript s to emphasize the dependence of $\mathbf{\Sigma}(\theta_G)$ on $\mathbf{M}_s(z)$. As mentioned in Chapter 2, various methods of choosing s in this way have been suggested in the literature. In particular, Kavalieris (1984) provides a lucid account of this and for the refinements of his approach see, for example, Bard (1974, p. 110-116), and Kowalik and Osborne (1968, p. 71). The choice of s therefore raises the question of how best to implement the technique when dealing with the unstable initial estimator, $\hat{\mathbf{M}}_T^{(0)}(z)$. Since the main objective at this stage of estimation is to obtain a stable operator with which to commence the Gauss-Newton iterative scheme, it seems plausible to employ a rather relatively cheap technique to obtain $\mathbf{M}_s^{(0)}(z)$ should $\mathbf{M}_T^{(0)}(z)$ be found unstable. For this reason, we now seek an ad-hoc procedure for choosing s such that the least unstable zero in $\mathbf{M}_T^{(0)}(z)$ is replaced by its stable equivalent in $\mathbf{M}_s^{(0)}(z)$. One obvious advantage derivable from this approach is that it can be guaranteed that all other zeros of $\mathbf{M}_T^{(0)}(z)$ that are inside $|z| = 1$ are reflected through the unit circle all at once. To achieve this set $s = 1$ and evaluate $\hat{\mathbf{M}}_s^{(0)}(z)$ successively for $s = s - 0.1$, $s - 0.2$, $s - 0.3, \ldots$ until such a point when $\bar{\xi}_{(1)}^{-1}$ is bounded by the least zero in $\hat{\mathbf{M}}_{s_\ell}^{(0)}(z)$

and $\hat{\mathbf{M}}_{s_r}^{(0)}(z)$ for some values of s denoted by s_ℓ and s_r. The appropriate choice of s can now be made by searching over the grid line adjoining s_ℓ and s_r. As an illustration, this technique is applied to $\hat{\mathbf{M}}_T^{(0)}(z)$ defined by (6.4.4) in which the zeros of $\det \hat{\mathbf{M}}_T^{(0)}(z)$ are evaluated and written to two decimal places as $\xi_{(1)} = 0.94$, $\xi_{(2)} = 2.02 + 2.75i$ and $\xi_{(3)} = \bar{\xi}_{(2)}$. We find that the suitable choice of s in this case is $s = 0.88$ and this gives rise to a stable initial operator

$$\hat{\mathbf{M}}_s^{(0)}(z) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \begin{pmatrix} 0.334 & -0.194 \\ 0.458 & -0.968 \end{pmatrix} z + \begin{pmatrix} 0.0 & 0.0 \\ -0.396 & 0.018 \end{pmatrix} z^2$$

whose zeros are approximately $\xi_{(1)} = 1.07$, $\xi_{(2)} = -2.07 + 3.01i$ and $\xi_{(3)} = \bar{\xi}_{(2)}$.

At this point it might be argued that when the zeros of $\det \hat{\mathbf{M}}_T^{(0)}(z)$ are very close to the unit circle (as in this case) it may seem preferrable to choose s so that the zeros of $\hat{\mathbf{M}}_s^{(0)}(z)$ are bounded away from the unit circle. This has an additional advantage of forcing the initialization effects, which may arise from using (6.1.3), to die a little more rapidly. In this case, a value of $s < 0.88$ may be considered. In view of this consideration it may be advisable to evaluate the loss function, $\det\{T^{-1}\sum_{t=1}^{T} \mathbf{e}_\theta(t,s)\mathbf{e}_\theta(t,s)'\}$, for successive values of s so as to make sure that $\hat{\mathbf{M}}_s^{(0)}(z)$ obtained is at least a local minimum. However, we have not investigated this issue further here but we must emphasize that it is a worthwhile venture.

In direct contrast, an application of the methodology presented in Section (6.3) shows that the stable, miniphase and invertible, equivalent of $\hat{\mathbf{M}}_T^{(0)}(z)$, written to two decimal places, is

$$\hat{\mathbf{M}}_\dagger^{(0)}(z) = \begin{pmatrix} 1.0 & 0.0 \\ 0.04 & 1.13 \end{pmatrix} + \begin{pmatrix} 0.38 & -0.23 \\ 0.50 & -1.10 \end{pmatrix} z + \begin{pmatrix} 0.0 & 0.0 \\ -0.45 & 0.03 \end{pmatrix} z^2$$

with zeros of the determinantal equation obtained as $\xi_{(1)} = 1.07$, $\xi_{(2)} = -2.02 + 2.75i$ and $\xi_{(3)} = \bar{\xi}_{(2)}$, corrected to two decimal places. See Appendix C for details.

Note that although the row degrees of $\mathbf{M}_\dagger^{(0)}(z)$ are still 1 and 2 it no longer corresponds to an echelon form with Kronecker indices $k_1 = 1$ and $k_2 = 2$ because

the coefficient matrix of z^0 is lower triangular rather than an identity. This reflects a general property of the stable equivalent operator that is also apparent from our numerical example, viz: although the row degrees of the original operator will be maintained there is no guarantee that $\mathcal{W}_\dagger(z)$ will satisfy any other coefficient restrictions placed on $\mathcal{W}(z)$. In general this feature will be immaterial, however, since it is the spectral characteristics that are of importance and these will be identical. Thus, in the current context it implies that the coefficient values given in $\mathbf{M}_\dagger^{(0)}(z)$ should not be interpreted as alternative initial estimates of the unknown model parameters in the usual sense. The construction of the operator $\mathbf{M}_\dagger^{(0)}(z)$ serves as a device for implementing the first step of the Gauss-Newton iteration. Indeed, as previously observed, the parameter adjustment $\hat{\theta}_T^{(j+1)} - \hat{\theta}_T^{(j)}$ may be viewed as being derived from a multivariate generalized least squares regression of the residuals on the derivative processes and, as such, it depends only on second moment properties. Since $\hat{\mathbf{M}}_T^{(0)}(z)\hat{\mathbf{M}}_T^{(0)}(z^{-1})' = \hat{\mathbf{M}}_\dagger^{(0)}(z)\hat{\mathbf{M}}_\dagger^{(0)}(z^{-1})'$ it can be shown that $\hat{\theta}_T^{(1)}$ will be invariant with respect to which operator is employed. We would therefore suggest the use of $\hat{\mathbf{M}}_\dagger^{(0)}(z)$ in place of the initial value. This recommendation follows Hannan and Deistler (1988, p.300). They suggest the use of Wilson's (1972) algorithm, referred to above, to generate the stable equivalent, although once again guidelines on the precise numerical implementation are not given. The use of $\hat{\mathbf{M}}_\dagger^{(0)}(z)$ as an implied initial value and, in an obvious notation, $\hat{\mathbf{M}}_\dagger^{(j)}(z)$ for $j > 1$, appears to us to be justified by the following features. First, the stable equivalent can be readily evaluated in the manner described. This does not entail anything more than the direct application of a sequence of algebraic manipulations that can be implemented using the program written for generating Hermite Normal form in conjunction with standard, finite algorithms. Secondly, and perhaps more importantly, since the stable equivalent can be evaluated algebraically the procedure is void of subjective choices concerning such matters as convergence and should be independent of numerical implementation.

Finally, the purpose of this chapter has been to present a closed-form, numerically achievable procedure for solving the difficult but important problem of stable spectral factorization, and to illustrate the theory by simple but non-trivial examples. The main result is embodied in Section (6.3), and it would be extremely useful to have available a handy computer routine, which the author is currently pursuing, for implementing this fundamental algorithm. Moreover, despite the comparative computational advantages of our procedure it is of interest to contrast the performance of $\hat{\mathbf{M}}_{\dagger}(z)$ with $\hat{\mathbf{M}}_{s}(z)$ in terms of the minimization of the objective function. For conciseness, we do not deal with the topic here. If the results of this chapter succeed in provoking and stimulating further research in this direction then one of the main objectives of this thesis will have been realized.

Appendix C

C.1 An Echelon Structure Example

In this appendix we employ the reflection scheme of Section(6.3) to obtain a stable equivalent operator for $\hat{\mathbf{M}}_T^{(0)}$ defined by (6.4.4) as

$$\hat{\mathbf{M}}_T^{(0)}(z) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \begin{pmatrix} 0.38 & -0.22 \\ 0.52 & -1.10 \end{pmatrix} z + \begin{pmatrix} 0.0 & 0.0 \\ -0.45 & 0.02 \end{pmatrix} z^2.$$

Executing the step 1 of the algorithm the Hermite Normal form, $\mathbf{H}(z)$, corresponding to $\hat{\mathbf{M}}_T^{(0)}(z)$ is obtained, using a suite of programs written for the purpose, as

$$\mathbf{H}(z) = \begin{pmatrix} z^3 + 3.1028z^2 + 7.8775z - 10.9409 & 0.0 \\ -0.4155z^2 - 1.2891z - 5.0 & 1.0 \end{pmatrix}.$$

Testing the diagonal elements of $\mathbf{H}(z)$ for stability as in step 2 it was found that $h_{11}(z)$ is not zero-free for $|z| < 1$. As step 3 suggests, the zeros of $h_{11}(z)$ are evaluated using standard routine as $\xi_{1(1)} = 0.94$, $\xi_{1(2)} = -2.20 + 2.75i$ and $\xi_{1(3)} = \bar{\xi}_{1(2)}$, corrected to two decimal places. Since only one zero lies inside $|z| = 1$ we proceed directly to step 5.

Setting $\xi_{1(1)} = 0.93778768$ and, for convenience, construct the diagonal matrix $\mathbf{D}_{\xi_{1(1)}}(z)$ as

$$\mathbf{D}_{\xi_{1(1)}}(z) = \begin{pmatrix} \frac{1+\xi_{1(1)}(z)}{z-\xi_{1(1)}} & 0 \\ 0 & 1 \end{pmatrix}.$$

In preparation for the implementation of step 6, we require

$$\mathbf{H}(\xi_{1(1)}) = \begin{pmatrix} 0.0 & 0.0 \\ -6.574 & 1.0 \end{pmatrix},$$

$$\hat{\mathbf{M}}_T^{(0)}(\xi_{1(1)}) = \begin{pmatrix} 1.356 & -0.206 \\ 0.092 & -0.014 \end{pmatrix}$$

and the Hermitian matrix,

$$\mathbf{H}(\xi_{1(1)})'\mathbf{H}(\xi_{1(1)}) = \begin{pmatrix} 43.222 & -6.574 \\ -6.574 & 1.0 \end{pmatrix}.$$

Then the corresponding unitary matrix generated by the orthonormal eignevectors of this Hermitian matrix is

$$\mathbf{V}_{\xi_{1(1)}} = \begin{pmatrix} 0.150 & -0.989 \\ 0.989 & 0.150 \end{pmatrix}$$

and it is easily checked that

$$\hat{\mathbf{M}}_T^{(0)}(\xi_{1(1)})\mathbf{V}_{\xi_{1(1)}} = \begin{pmatrix} 0.0 & -1.372 \\ 0.0 & -0.093 \end{pmatrix},$$

to three decimal places. By implication, the first column of $\hat{\mathbf{M}}_T^{(0)}(z)\mathbf{V}_{\xi_{1(1)}}$ is clearly divisible by $z-\xi_{1(1)}$. Thus, the stable matrix operator, $\tilde{\mathbf{M}}_T(z) = \hat{\mathbf{M}}_T^{(0)}(z)\mathbf{V}_{\xi_{1(1)}}\mathbf{D}_{\xi_{1(1)}}(z)$, equivalent to $\hat{\mathbf{M}}_T^{(0)}(z)$ in which the zero $\xi_{1(1)} = 0.93778768$ has been replaced by $\xi_{1(1)}^{-1} = 1.066339383$ is

$$\tilde{\mathbf{M}}_T^{(0)}(z) = \begin{pmatrix} 0.171 & -0.989 \\ 1.124 & 0.150 \end{pmatrix} + \begin{pmatrix} -0.160 & -0.409 \\ -1.003 & -0.679 \end{pmatrix} z + \begin{pmatrix} 0.0 & 0.0 \\ -0.048 & 0.448 \end{pmatrix} z^2.$$

Since step 7 is not applicable in the present case we proceed to step 8. As part of the requirement for carrying out step 8 we need to compute the unitary matrix

$\tilde{\mathbf{U}}$. This is obtained as

$$\begin{aligned}\tilde{\mathbf{U}} &= \tilde{\mathbf{M}}_T^{(0)}(0)^{-1}(\tilde{\mathbf{M}}_T^{(0)}(0)\tilde{\mathbf{M}}_T^{(0)}(0)')^{\frac{1}{2}} \\ &= \begin{pmatrix} 0.170 & 0.985 \\ -0.985 & 0.170 \end{pmatrix}\end{aligned}$$

where the matrix $\tilde{\mathbf{M}}_T^{(0)}(0)\tilde{\mathbf{M}}_T^{(0)}(0)'$, written to three decimal places, is

$$\tilde{\mathbf{M}}_T^{(0)}(0)\tilde{\mathbf{M}}_T^{(0)}(0)' = \begin{pmatrix} 1.007 & 0.044 \\ 0.044 & 1.286 \end{pmatrix}$$

and has

$$(\tilde{\mathbf{M}}_T^{(0)}(0)\tilde{\mathbf{M}}_T^{(0)}(0)')^{\frac{1}{2}} = \begin{pmatrix} 1.003 & 0.0 \\ 0.043 & 1.133 \end{pmatrix}$$

as its Cholesky factor.

Thus, $\hat{\mathbf{M}}_\dagger^{(0)}(z)$, to three decimal places, is

$$\begin{aligned}\hat{\mathbf{M}}_\dagger^{(0)}(z) &= \tilde{\mathbf{M}}_T^{(0)}(z)\tilde{\mathbf{U}} = \begin{pmatrix} 1.003 & 0.0 \\ 0.043 & 1.133 \end{pmatrix} + \begin{pmatrix} 0.375 & -0.228 \\ 0.499 & -1.104 \end{pmatrix} z \\ &\quad + \begin{pmatrix} 0.0 & 0.0 \\ -0.45 & 0.029 \end{pmatrix} z^2\end{aligned}$$

and the zeros of $\det \hat{\mathbf{M}}_\dagger^{(0)}(z)$, written to three decimal places, are 1.066 and $2.02 \pm 2.754i$ as expected.

We conclude by mentioning that all the computations carried out in the foregoing were conducted in FORTRAN 77 and implemented on a SUN SPARC Station 1 utilizing double precision for all real and complex values. Also, each of the matrices presented above is written to three decimal places.

Bibliography

An, H-Z, Chen, Z-G and E. J. Hannan (1982). Autocorrelation, autoregression and autoregressive approximation, *Ann. Statist.* **10**, 926-936.

Akaike, H (1969). Fitting autoregressive models for prediction, *Ann. Inst. Statist. Math.*, **21**, 243-247.

Akaike, H (1973). Block Toeplitz matrix inversion, *SIAM J. Appl. Math.*, **24**, 234-241.

Akaike, H (1974a). A new look at the statistical model identification, *IEEE Trans. Autom. Control*, **AC-19**, 716-723.

Akaike, H (1974b). Markovian representation of stochastic processes and its application to the analysis of autoregressive moving average processes, *Ann. Inst. Statist. Math.*, **26**, 363-387.

Akaike, H (1975). Markovian representation of stochastic processes by canonical variable, *SIAM J. Control*, **13**, 162-173.

Akaike, H (1976). Canonical correlation analysis of time series and the use of an information criterion, In *system identification: Advances and case studies* (eds. R. H. Mehra and D. G. Lainiotis), 27-96, New York: Academic Press.

Anderson, T. W. (1980). Maximum likelihood estimation for vector autoregressive moving average models, In *directions in time series* (eds. D. R. Brillinger

and G. C. Tiao), 49-59.

Ansley, C. F. (1979). An algorithm for the exact likelihood of a mixed autoregressive moving average process, *Biometrika*, **66**, 59-65.

Ansley, C. F. (1980). Computation of the theoretical autocovariance function for a vector ARMA process, *J. Statist. Comput. Simul.*, **12**, 15-24.

Ansley, C. F. and P. Newbold (1980). Finite sample properties of estimators for autoregressive moving average models, *J. Econometr.*, **13**, 159-183.

Ansley, C. F. and R. Kohn (1983). Exact likelihood of vector autoregressive moving average process with missing or aggregated data, *Biometrika*, **70**, 275-278.

Aoki, M. (1987). *State space modelling of time series*, New York: Springer-Verlag.

Astrom, K. J. (1980). Maximum likelihood and prediction error methods, *Automatica*, **16**, 551-574.

Astrom, K. J. and T. Soderstrom (1974). Uniqueness of the maximum likelihood estimates of the parameters of an ARMA model, *IEEE Trans. Autom. Control,* **AC-19**, 769-773.

Bard, Y. (1974). *Non-linear parameter estimation*, New York: Academic Press.

Barnett, S. (1972). *Matrices in control theory*, New York: Van Nostrand Reinhold.

Baxter, G. (1963). A norm of inequality for a "finite section" Wienner-Hopf equation, *Ill. J. Math.*, **7**, 97-103.

Birkhoff, G. and S. Mclane (1977). *A Survey of Modern Algebra* (4^{th} *ed.)*, New York: MacMillan.

Bloomfield, P. and G. S. Watson (1975). The inefficiency of least squares, *Biometrika*, **12**, 121-128.

Box, G. E. P. and G. Jenkins (1976). *Time Series Analysis, Forecasting and Control*, San Francisco, CA: Holden-day.

Brockwell, P. J. and R. A. Davis (1991). *Time Series: Theory and Methods (2^{nd} ed.)*, New York: Springer-Verlag.

Burg, J. P. (1975). Maximum entropy spectral analysis, *Ph.D dissertation*, University of Standford, California.

Chun-tu, L. (1991). The efficiency of least squares estimator of a seemingly unrelated regression model, *Comm. Statist.-Simul. Comput.*, **20**, 919-925.

Cooper, D. M. and E. F. Wood (1982). Identifying multivariate time series models, *J. Time Ser. Anal.*, **3**, 153-164.

Conte, S. D. (1965). *Elementary numerical analysis –An algorithmic approach*, New York: McGraw-Hill.

Cramer, H. (1977). *Mathematical methods of statistics*, Princeton: Princeton University Press.

Crowther, M. J. (1976). Maximum likelihood estimation for dependent observations, *J. R. Statist. Soc. Ser.* **B 35**, 45-53.

Cybenko, G. (1980). The numerical stability of the Levinson-Durbin algorithm for Toeplitz systems of equations, *SIAM J. Sci. Statist. Comput.*, **1**, 303-319.

Dahlhaus, R. (1983). Spectral analysis with tapered data, *J. Time Ser. Anal.*, **4**, 163-176.

Dahlhaus, R. (1988). Small sample effects in time series analysis: A new asymptotic theory and a new estimate, *Ann. Stat.*, **16**, 808-841.

Davis, M. C. (1963). Factoring the spectral matrix, *IEEE Trans. Autom. Control*, **AC-8**, 296-305.

Deistler, M (1985). General structure and parameterization of ARMA and state-space systems and its relation to statistical problems, *Handbook of statistics 5* (eds. E. J. Hannan, P. R. Krishnaiah and M. M. Rao), 257-277, Amsterdam: North-Holland.

Deistler, M. and E. J. Hannan (1981). Some properties of the parameterization of ARMA systems with unknown order, *J. Multivariate Anal.*, **11**, 474-484.

Deistler, M and B. M. Potscher (1984). The behaviour of the likelihood function for ARMA models, *Adv. Appl. Prob.*, **16**, 843-865.

Dickson, B. W., Kailath, T., and M. Morf (1974). Canonical matrix fraction and state-space descriptions for deterministic and stochastic linear systems, *IEEE Trans. Autom. Control*, **AC-19**, 656-667.

Draper, N. and H. Smith (1981). *Applied regression analysis,* 2^{nd} *ed.*, New York: Wiley.

Dunsmuir, W. and E. J. Hannan (1976). Vector linear time series models, *Adv. Appl. Prob.*, **8**, 339-364.

Durbin, J. (1960). The fitting of time series models, *Rev. Int. Inst. Statist.*, **28**, 233-289.

Farell, R. H. (1985). *Multivariate calculation: Use of the continuous groups*, New York: Springer-Verlag.

Findley, D. F. (1992). Convergence of finite multistep predictors from incorrect models and its role in model selection, *Note di Matematica* – Under review.

Forney, G. D. (1975). Minimal bases of rational vector spaces with applications to multivariate linear systems, *SIAM J. Control*, **13**, 493-520.

Fuller, W. A. (1976). *Introduction to statistical time series*, New York: Wiley.

Gardner, G. A., Harvey, C., and G. D. A. Phillips (1980). An algorithm for maximum likelihood estimation of autoregressive-moving average models by means of Kalman filtering, *Appl. Statist.*, **29**, 311-322.

Golub, G. H. and C. F. Van Loan (1983). *Matrix Computations*, Baltimore, MD: Johns Hopkins University Press.

Graybill, F. A. (1969). *Introduction to matrices with applications in statistics*, Belmont, CA: Wadsworth Publishing Company.

Hannan, E. J. (1969). The identification of vector autoregressive moving average system, *Biometrika*, **56**, 223-225.

Hannan, E. J. (1970). *Multiple time series*, New York: Wiley.

Hannan, E. J. (1973). The asymptotic theory of linear time series models, *J. Appl. Prob.*, **10**, 130-145.

Hannan, E. J. (1975). Multivariate ARMA theory, In *essays in honour of Professor A. W. Phillips* (ed. M. Peston), New York: Wiley.

Hannan, E. J. (1979). The statistical theory of linear systems, In *Developments in statistics, vol. 2* (ed. P. R. Krishnaiah), 83-121, New York: Academic Press.

Hannan, E. J (1980). The estimation of the order of an ARMA process, *Ann. Statist.*, **8**, 1071-1081.

Hannan, E. J. (1981). Estimating the dimension of a linear system, *J. Multivariate Anal.*, **11**, 459-473.

Hannan, E. J. and M. Deistler (1988). *The statistical theory of linear systems*, New York: Wiley.

Hannan, E. J., Dunsmuir, W. and M. Deistler (1980). Estimation of vector ARMAX models, *J. Multivariate Anal.*, **10**, 275-295.

Hannan, E. J. and C. C. Heyde (1972). On the limit theorems for quadratic functions of discrete time series, *Ann. Math. Statist.*, 2058-2066.

Hannan, E. J. and L. Kavalieris (1983). The convergence of autocorrelations and autoregressions, *Austr. J. Statist.*, **25**, 287-297.

Hannan, E. J. and L. Kavalieris (1984). Multivariate linear time series models, *Adv. Appl. Prob.*, **16**, 492-561.

Hannan, E. J. and L. Kavalieris (1986). Regression, autoregression models, *J. Time Ser. Anal.*, **7**, 27-50.

Hannan, E. J. and A. J. McDougall (1988). Regression procedures for ARMA estimation, *J. Am. Statist. Assoc.*, **83**, 490-498. Correction (1989), **84**, 636.

Hannan, E. J., McDougall, A. M. and D. S. Poskitt (1989). Recursive estimation for autoregressions, *J. R. Statist. Soc. Ser.*, **B 51**, 217 - 233.

Hannan, E. J. and B. G. Quinn (1979). The determination of the order of an autoregression, *J. R. Statist. Soc. Ser.*, **B 41**, 190-195.

Hannan, E. J. and J. Rissanen (1982). Recursive estimation of mixed autoregressive moving average order, *Biometrika*, **69**, 8l-94. Correction (1983), **70**, 303.

Helson, H (1964). *Lectures on invariant subspaces*, New York: Academic Press.

Henderson, H. V. and S. R. Searle (1979). Vec and vech operators for matrices with some uses in Jacobians and multivariate statistics, *The Canadian J. Statist.*, **7**, 65-81.

Henderson, H. V. and S. R. Searle (1981). On deriving the inverse of a sum of matrices, *SIAM Review.*, **23**, 53-60.

Henrici, P. (1974). *Applied and computational complex analysis, vol. 1*, New York: Wiley.

Hildebrand, F. B. (1974). *Introduction to numerical analysis*, 2^{nd} ed., New York: McGraw-Hill.

Hillmer, S. C. and G. C. Tiao (1979). Likelihood function of stationary multiple autoregressive moving average models, *J. Am. Statist. Assoc.*, **74**, 652-660.

Jenkins, G. M. and A. S. Alavi (1981). Some aspects of modelling and forecasting multivariate time series. *J. Time Ser. Anal.*, **2**, 1-47.

Jones, J. W. (1914). Fur Farming in Canada, *Commission of Conservation, Ottawa*

Jones, R. H. (1975). Fitting autoregressions, *J. Am. Statist. Assoc.*, **70**, 590-592.

Jones, R. H. (1978). Multivariate autoregression estimation using residuals, In *Applied Time Series Analysis* (ed. D. F. Findley), 139-161, New York: Academic Press.

Judge, G. G., Griffiths, W. E., Hill, R. C., Lutkepohl, H. and T. C. Lee (1985). *The theory and practice of econometrics*, 2^{nd} ed., New York: Wiley.

Kailath, T (1980). *Linear systems*, Englewood Cliffs, NJ: Prentice-Hall.

Kailath, T., Bruckstein, A. and D. Morgan (1986). Fast matrix factorizations via discrete transmission lines, *Linear Algebra and its applications*, **75**, 1-25.

Kalman, R. E., Falb, P. L., and M. A. Arbib (1969). *Topics in mathematical system theory*, New York: McGraw-Hill.

Kashyap, R. L. (1970). Maximum likelihood identification of stochastic linear systems, *IEEE Trans. Autom. Control* **AC-15**, 25-34.

Kashyap, R. L. and A. R. Rao (1976). *Dynamic stochastic models from empirical data*, New York: Academic Press.

Kavalieris, L (1984). Statistical problems in linear systems, *Ph.D dissertation*, Australian National University, Canberra.

Kohn, R. (1978). Asymptotic properties of time domain Gaussian estimators, *Adv. Appl. Prob.*, **10**, 339-359.

Kohn, R. (1979). Asymptotic estimation and hypothesis testing results for vector linear time series models, *Econometrica*, **47**, 1005-1030.

Koreisha, S. and T. Pukkila (1990). A generalized least squares approach for estimation of autoregressive moving average models, *J. Time Ser. Anal.*, **11**, 139-151.

Kowalik, J. and M. R. Osborne (1968). *Methods for unconstrained optimization problems*, New York: Elsevier.

Levinson, N. (1947). The Weiner RMS (root mean square) error criterion in filter design and prediction, *J. Math. Physics*, **25**, 261-278.

Ljung, L. and T. Soderstrom (1983). *Theory and practice of recursive identification*, London: The MIT Press.

Lutkepohl, H. (1985). Comparison of criteria for estimating the order of a vector autoregressive process, *J. Time Ser. Anal.*, **6**, 35-52.

Lutkepohl, H. (1991). *Introduction to multiple time series analysis*, New York: Springer-Verlag.

MacDuffee, C. C. (1956). *The theory of matrices*, New York: Chelsea.

Magnus, J. R. and H. Neudecker (1988). *Matrix differential calculus with applications in statistics and econometrics*, Chichester: Wiley.

Makhoul, J. (1981). Lattice methods in spectral estimation, In *applied time series analysis*, vol. 2 (ed. D. F. Findley), 301-326, New York: Academic Press.

McDougall, A. J. (1988). Algorithms in time series, *Ph.D dissertation*, Australian National University, Canberra.

Mittnik, S. (1990). Computation of theoretical autocovariance matrices of multivariate autoregressive moving average time series, *J. R. Statist. Soc. Ser.* **B 52**, 151-155.

Morf, M., Viera, A. and T. Kailath (1978). Covariance characterization of partial autocorrelation matrices, *Ann. Statist.*, **6**, 643-648.

Miller, A. J. (1991). Least squares routines to supplement those of Gentleman, *Appl. Statist.*, **1**, 458-478.

Nicholls, D. F. (1976). The efficient estimation of vector linear time series models, *Biometrika*, **63**, 381-390.

Nicholls, D. F. (1977). A comparison of estimation methods for vector linear time series models, *Biometrika*, **64**, 85-90.

Nicholls, D. F. and A. D. Hall (1979). The exact likelihood function of multivariate autoregressive moving average models, *Biometrika*, **66**, 259-264.

Osborn, R. R. (1977). Exact and approximate maximum likelihood estimators for vector moving average processes, *J. R. Statist. Soc. Ser.*, **B 39**, 114-118.

Pagan, A. R. and D. F. Nicholls (1976). Exact maximum likelihood estimation of regression models with finite order moving average errors, *Rev. Econ. Stud.*, **43**, 383-387.

Pate, M. B. and N. Davies (1988). Computation of population and sample correlation and partial correlation matrices in MARMA(P,Q) time series, *Appl. Statist.*, **37**, 127-155.

Paulsen, J. and D. Tjostheim (1985). On the estimation of residual variance and order in autoregressive time series, *J. R. Statist. Soc.*, **B 47**, 216-228.

Phadke, M. S. and G. Kedem (1978). Computation of the exact likelihood function of multivariate moving average models, *Biometrika*, **65**, 511-519.

Phillips, P. C. B (1976). The iterated minimum distance estimator and the quasi maximum likelihood estimator, *Econometrica*, **44**, 449-460.

Poskitt, D. S. (1990). Estimation and structure determination of multivariate input-output systems, *J. Multivariate Anal.*, **33**, 157-182.

Poskitt, D. S (1992). Identification of echelon canonical forms for vector linear processes using least squares, *Ann. Statist.*, **20**, 195-215.

Poskitt, D. S. and M. O. Salau (1993a). Stable spectral factorization with applications to the estimation of time series models, *Comm. Statist. – Theor. Meth.*, **22**(2), 347-367.

Poskitt, D. S. and M. O. Salau (1993b). On the asymptotic relative efficiency of Gaussian and least squares estimators for vector ARMA models, *J. Multivariate Anal.* – To appear.

Poskitt, D. S. and A. R. Tremayne (1982). Diagnostics tests for multiple time series models, *Ann. Statist.*, **10**, 114-120.

Popov, V. M (1969). Some properties of the control systems with irreducible matrix transfer functions, *Lecture notes in mathematics*, 169-180, Berlin: Springer-Verlag.

Pukkila, T. M (1988). An improved estimation method for univariate autoregressive models, *J. Multivariate Anal.*, **27**, 422-433.

Pukkila, T., Koreisha, S. and J. Kallinen (1990). The identification of ARMA models, *Biometrika*, **77**, 537-548.

Reinsel, G. C. (1976). Maximum likelihood estimation of vector autoregressive moving average models, *Technical Report No. 117*, Dept. of Statistics, Carnegie-Mellon University, Pittburg.

Reinsel, G. C., S. Basu and S. F. Yap (1992). Maximum likelihood estimators in the multivariate autoregressive moving-average model from a generalized least squares viewpoint, *J. Time Ser. Anal.*, **13**, 133-145.

Rissanen, J (1974). Basis of invariants and canonical forms for linear dynamic systems, *Automatica*, **10**, 175-182.

Rissanen, J. (1976). Minimax entropy estimation of models for vector processes, In *system identification: Advances and case studies* (eds. R. K. Mehra and D. G. Lainiotis), New York: Academic Press, 97-119.

Rissanen, J. and P. E. Caines (1979). The strong consistency of maximum likelihood estimators for ARMA processes, *Ann. Statist.*, **7**, 297-315.

Robinson, E. A. (1983). *Multichannel time series analysis with digital computer programs* (2^{nd} ed.), Texas: Goose Pond Press.

Rosanov, Y. A. (1967). *Stationary random processes*, San Francisco, CA: Holden-Day.

Rosenbrock, H. H (1970). *State space and multivariable theory*, London: Nelson.

Saikonnen, P. (1986). Asymptotic properties of some preliminary estimators for autoregressive moving average time series models, *J. Time Ser. Anal.*, **7**, 133-155.

Salau, M. O. (1992). Asymptotic relative efficiency of least squares estimators for vector time series models, *A paper presented at the eleventh Australian Statistical Association conference held in Perth, Western Australia.*

Schwarz, G. (1978). Estimating the dimension of a model, *Ann. Statist.*, **6**, 461-464.

Seber, G. A. F. (1977). *Linear regression analysis*, New York: Wiley.

Shea, B. L. (1984). Maximum likelihood estimation of multivariate ARMA processes via the Kalman filter. In the *Time series analysis: Theor. Pract.* **5** (ed. O. D. Anderson), 91-101, Elsevier Science Publishers B. V., North-Holland.

Shea, B. L. (1987). Estimation of multivariate time series. *J. Time Ser. Anal.*, **8**, 95-109.

Shibata, R. (1976). Selection of the order of an autoregressive model by Akaike's information criterion, *Biometrika*, **63**, 117-126.

Shibata, R. (1980). Asymptotically efficient selection of the order of the model for estimating the parameters of a linear process, *Ann. Statist.*, **8**, 147-164.

Solo, V. (1978). Time series recursions and stochastic approximation, *Ph.D dissertation*, Australian National University, Canberra.

Solo, V. (1984). The exact likelihood for a multivariate ARMA model, *J. Multivariate Anal.*, **15**, 164-173.

Solo, V. (1986). Topics in advanced time series analysis, *Lecture notes in mathematics* (eds. G. del Pino and R. Rebolledo), New York: Springer-Verlag.

Spliid, H (1983). A fast estimation method for the vector autoregressive moving average model with exogenous variable, *J. Am. Statist. Assoc.*, **78**, 843-849.

Steinberg, H., Gasser, T. and J. Franke (1985). Fitting autoregressive processes to EEG time series: an empirical comparison of estimates of the order, *IEEE-ASSP*, **33**, 143-150.

Stewart, G. W. (1973). *An introduction to matrix computations*, New York: Academic Press.

Stoica, P., Soderstrom, T., Ahlen, A., and G. Solbrand (1985). On the convergence of pseudo-linear regression algorithms, *Int. J. Control*, **41**, 1429-1444.

Tiao, G. C. and G. E. P. Box (1981). Modelling multiple time series with applications, *J. Am. Statist. Assoc.*, **76**, 802-816.

Tiao, G. C. and R. S. Tsay (1989). Model specification in multivariate time series, *J. R. Statist. Soc. Ser.* **B 51**, 157-213.

Tjostheim, D. and J. Paulsen (1983). Bias of some commonly used time series estimates, *Biometrika*, **70**, 389-400.

Tsay, R. S. (1989a). Identifying multivariate time series models, *J. Time Ser. Anal.*, **10**, 357-372.

Tsay, R. S. (1989b). Parsimonious parameterization of vector autoregressive moving average models, *J. Bus. Econ. Statist.*, **7**, 327-341.

Tsay, R. S. (1991). Two canonical forms for vector ARMA processes, *Statistica Sinica*, **1**, 247-269.

Tuan, P. D. (1984). The estimation of parameters for autoregressive moving average models, *J. Time Ser. Anal.*, **5**, 53-68.

Walker, A. M. (1962). Large sample estimation of parameters for autoregressive processes with moving average residuals, *Biometrika*, **49**, 117-131.

Walker, A. M. (1964). Asymptotic properties of least squares estimates of parameters of the spectrum of a stationary non-deterministic time series, *J. Austr. Math. Soc.* **4**, 363-384.

Whittle, P. (1963). On the fitting of multivariate autoregressions, and the approximate canonical factorization of a spectral density matrix, *Biometrika*, **50**, 129-134.

Wilson, D. A. and A. Kumar (1982). Derivative computations for the log-likelihood function, *IEEE Trans. Autom. Control*, **AC-27**, 230-232.

Wilson, G. T. (1972). The factorization of matricial spectral densities, *SIAM J. Appl. Math.*, **23**, 420-426.

Wilson, G.T. (1973). The estimation of parameters in multivariate time series models, *J. R. Statist. Soc.*, **B 35**, 76-85.

Wilson, G. T. (1979). Some efficient computational procedures for high order ARMA models, *J. Statist. Comput. Simul.*, **8**, 301-309.

Wolovich, W. A. (1974). *Linear multivariable systems*, New York: Springer-Verlag.

Zellner, A (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias, *J. Am. Statist. Assoc.*, **57**, 348-368.

Zinde-Walsh, V (1988). Some exact formulae for autoregressive moving average processes, *Econometr. Theory*, **4**, 384-402.