STATISTICAL PROBLEMS IN LINEAR SYSTEMS

BY

LAIMONIS KAVALIERIS

A THESIS SUBMITTED TO THE

AUSTRALIAN NATIONAL UNIVERSITY

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

JUNE 1984

THE AUSTRALIAN
LIBRARY

DECLARATION

Most of the results in Chapter 3 have appeared in Hannan and Kavalieris (1983b, 1984a, 1984b) or have been accepted for publication. Except where otherwise stated, the work presented in this thesis is my own.

L. KAVALIERIS

ACKNOWLEDGEMENTS

I am grateful to my supervisor, Professor E.J. Hannan, for the assistance and encouragement he has given me throughout my studies at ANU. I would also like to thank Chris Jirasek for her patient and excellent typing of this thesis. Finally, I thank Winifred for the support she has given me during these past years of study.

SUMMARY

In recent years there has been increasing emphasis on the use of finitely parameterised linear models and particularly ARMA models, in analysing time series data. The structure theory of such linear systems has been extensively investigated by system engineers using the state space representation of linear systems that arises naturally in systems control. Chapter 1 describes those parts of the structure of linear systems that relate to identification and estimation problems considered later.

Usually the structure of the model describing a particular time series will not be known a priori; in general the data will not even be generated by a finitely parameterised model. Thus it is necessary to obtain procedures to select a model of sufficient complexity (e.g. having sufficiently many parameters) to describe those aspects of the data that are of interest to the experimenter. The complexity of the model would be expected to depend on the quality of the available data. Criteria known as AIC and BIC that are associated with the work of Akaike are important for this purpose. In Chapter 2 we survey published results concerning the consistency of these criteria. Methods for the identification of the structure of multivariate ARMAX models based on AIC and BIC but using only linear estimation procedures are investigated in Chapter 3. Theorems relating to the consistency of these identification procedures are also discussed. Chapters 4 and 5 are devoted to proving the results discussed in Chapter 3.

CONTENTS

DECLARATION		i
ACKNOWLEDGEMENTS		ii
SUMMARY		iii
INTRODUCTION		1
NOTATION		3
CHAPTER 1:	Algebraic and topological description of linear systems	
1.1	ARMAX and state space representations	4
1.2	Identification from second order statistics	9
1.3	Structure of spaces of rational transfer functions	17
1.4	Structure of $\overline{M(n)}$	30
CHAPTER 2:	Structure and parameter estimation based on maximum likelihood methods	
2.1	Introduction	39
2.2	Consistency of maximum likelihood estimates	42
2.3	Parameter estimation when the McMillan degree is known	47
2.4	Order estimation based on maximum likelihood	50
CHAPTER 3:	Estimation of ARMAX models	
3.1	Introduction	59
3.2	Recursive parameter estimation for autoregressive processes	66
3.3	Stage I	70
3.4	Stage II	81
3.5	Stage III	102
3.6	Estimation of transfer function models	116
CHAPTER 4:	Limit theorems and technical lemmas	
4.1	Uniform convergence of covariances	124
4.2	Properties of $\hat{\Gamma}_h$	128

CHAPTER 5: Proofs of theorems from Chapter 3

5.1 Stage I 140

5.2 Stage II 160

5.3 Stage III 176

5.4 Transfer function models 179

REFERENCES 188

INDEX TO DEFINITIONS 197

INTRODUCTION

Until the late 1960's the main techniques of time series analysis were the non-parametric methods based on Fourier transforms and methods based on approximation by auto-regressions. With the increased application of computers the use of finitely parameterised models, and particularly ARMA models, was popularised through the influence of Box and Jenkins (1970). Following the lead of R.E. Kalman linear systems were investigated by control engineers in terms of the state space formulation that arises naturally in this context. ARMA and ARMAX models are conveniently represented in this form. The structure theory of such linear systems has been extensively studied by these workers. In Chapter 1 we present a synthesis of parts of this structure theory that relates to ARMAX structure and the identification and estimation problems considered later.

In the identification problems we consider we are not placing restrictions on any parameter values that occur in the model, thus the structure of the model we fit is determined by the data rather than a priori knowledge. When finite parameter models are used to describe time series that are generated by more general processes, the difficulty that arises is to select a model of sufficient complexity to describe the features of interest, but also one that is commensurate with the quality (and quantity) of data available. The identification of such models leads to inference problems that cannot be solved using maximum likelihood methods; criteria introduced by Akaike (1969,1972), Rissanen (1976) are important here. These criteria make it possible to compare models of different dimensions (i.e. different

numbers of parameters) and to select an optimal model. Under the assumption that the true model is in one of the model sets considered, we collate some consistency results for these criteria in Chapter 2. The criteria alluded to above require the maximisation of a likelihood function for each of the model sets considered, which is a time consuming non-linear optimisation problem. This is a particular problem when the data is multivariate for then the number of possible model sets that may need to be considered becomes very large indeed.

The main thrust of this thesis is to study procedures based on the criteria of Akaike and Rissanen but which do not require the solution of a non-linear optimisation problem. In particular we provide algorithms that are, as much as practicable, carried out in a manner that is recursive in the dimension of the model. We construct an algorithm for the identification of an ARMAX process for which the amount of computation that is required can be bounded a priori. Other procedures are studied for the identification of transfer function models. When the data is indeed generated by an ARMAX process (or transfer function model) we show that our identification procedures are consistent. Of course consistency is only asymptotic as the sample size increases. As well as providing methods for the identification of time series models our techniques also yield consistent estimates of the model parameters that can be used to initialise an iterative procedure to obtain statistically efficient parameter estimates. In Chapter 3 algorithms and statements of the theorems pertaining to the asymptotic consistency of the identification procedure and the initial parameter estimates are presented. An

algorithm to obtain asymptotically efficient parameter estimates is also provided. A number of simulation results are given for simple models to illustrate the conclusions of the theorems for samples of moderate size. Chapters 4 and 5 are devoted to proofs of the theorems discussed in Chapter 3.

NOTATION

T	sample size
Q_T	$(\log \log T/T)^{\frac{1}{2}}$
I_s	$s \times s$ identity matrix
$A = [a_{ij}]_{i,j = 1,\ldots,n}$	an $n \times n$ matrix with a_{ij} as the element in i^{th} row and j^{th} column
(i,j) element of A	element in i^{th} row and j^{th} column of A
$\|\|A\|\|$	Euclidean norm of A, $\|\|A\|\| = (\sum\sum_{ij} a_{ij}^2)^{\frac{1}{2}}$
$x_t = O(y_t)$	$\lim \sup \|x_t/y_t\| < \infty$
$x_t = o(y_t)$	$\lim \|x_t/y_t\| = 0$

If $A = [a_{ij}]_{i,j = 1,\ldots,n}$ then

$A = O(y_t)$	$\sup_{ij} \|a_{ij}\| = O(y_t)$
$A = o(y_t)$	$\sup_{ij} \|a_{ij}\| = o(y_t)$

CHAPTER 1

ALGEBRAIC AND TOPOLOGICAL DESCRIPTION OF LINEAR SYSTEMS

§1.1 ARMAX and state space representations

In this chapter we are principally concerned with the structure of manifolds whose elements are linear systems rather than the description of any particular system. In developing the structure theory, the probabilistic properties of the time series obtained as outputs from such systems are of no importance. In fact the description introduced in §§1.3, 1.4 is the structure theory of certain spaces of matrices of rational functions. The value to statistical inference derives from the consequent parameterisation and the fact that the manifold topology is the natural topology in which to consider convergence of maximum likelihood estimates. In this chapter it is through the behaviour of the maximum likelihood estimate that the probabilistic properties play a part. For the remainder of the development in this chapter we will assume sufficient conditions have been enforced to ensure the existence of the various time series considered.

Let $y(t)$ be an observed process of s components $y_j(t)$, $j = 1,\ldots,s$ generated by a linear process of the form

$$\text{(1.1.1)} \quad \sum_0^p A(j)y(t-j) = \sum_0^q B(j)\varepsilon(t-j) + \sum_1^r D(j)z(t-j) \ ,$$

$$A(0) = B(0) = I_s \ .$$

Here $z(t)$ is another observed process of m components, often referred to as the *input* or vector of *exogenous variables*. The sequence $y(t)$ is often called the *output* or vector of *endogenous variables*. The $\varepsilon(t)$ are also s dimensional, but are unobserved, and will be always assumed to

be independent of the inputs. Many of the later results continue to hold when $z(t)$ is correlated with $\varepsilon(s)$, $s < t$, but we will not investigate that here. The process (1.1.1), with $A(0) = B(0)$ but not necessarily $A(0) = I_s$, is known as an ARMAX process (Autoregressive moving average process with exogenous variables). When inputs are absent it is called an ARMA process. Often, when $s = 1$, (1.1.1) will be described as ARMAX(p,q,r) (or ARMA(p,q)), the integers p,q,r referring to the highest lags present in the expression (1.1.1). It is always assumed, without further mention, that the $\varepsilon(t)$ are a stationary and ergodic sequence and

$$E\{\varepsilon(t)\} = 0, \; E\{\varepsilon(t)\varepsilon(s)'\} = \delta_{st}\Sigma_0$$

where Σ_0 is a positive definite matrix.

Define the lag operators

$$a(L) = \sum_0^p A(j)L^j, \; b(L) = \sum_0^q B(j)L^j, \; d(L) = \sum_1^r D(j)L^j$$

where

$$Ly(t) = y(t-1) ,$$

hence

$$a(L)y(t) = \sum_0^p A(j)y(t-j) .$$

Expressions such as this will be abbreviated to $ay(t)$ when the meaning is clear. The functions

$$a(z) = \sum_0^p A(j)z^{-j} , \; b(z) = \sum_0^q B(j)z^{-j} , \; d(z) = \sum_1^r D(j)z^{-j}$$

are assumed to satisfy

$$\begin{aligned} &\det a(z) \neq 0 \quad |z| \geqslant 1 \\ &\det b(z) \neq 0 \quad |z| \geqslant 1 . \end{aligned} \tag{1.1.2}$$

A function of the form of $a(z)$ that satisfies (1.1.2) is called stable. Then, assuming that $z(t)$ is stationary with a

finite second moment so that the series converge, (1.1.1) has a stationary solution of the form

$$
\begin{aligned}
&y(t) = u(t) + v(t) \\
&u(t) = \sum_0^\infty K(j)\varepsilon(t-j) \ , \ v(t) = \sum_1^\infty L(j)z(t-j) \ .
\end{aligned}
\tag{1.1.3}
$$

The coefficient matrices of the *transfer functions*

$$
k(z) = \sum_0^\infty K(j)z^{-j} \ , \ \ell(z) = \sum_1^\infty L(j)z^{-j}
$$

are obtained from the expression

$$
[k(z),\ell(z)] = a(z)^{-1}[b(z),d(z)] \ .
$$

Conditions (1.1.2) imply

$$
\begin{aligned}
&k(z),\ell(z) \quad \text{analytic} \quad |z| \geqslant 1 \\
&\det k(z) \neq 0 \qquad\quad |z| \geqslant 1 \ .
\end{aligned}
\tag{1.1.4}
$$

If $\det b(z)$ is permitted to have zeros on $|z| = 1$ the stationary solution (1.1.3) still exists. The $\varepsilon(t)$ are the *linear innovations*, i.e. the prediction errors of the best linear predictor of $y(t)$ from $y(t-j)$, $z(t-j)$, $j = 1,2,\ldots,\infty$.

Alternatively an ARMAX process may be represented as the *transfer function model*

$$
y(t) = \ell z(t) + \eta(t) \tag{1.1.5}
$$

where $\ell(z) = \sum_1^\infty L(j)z^{-j}$ is a rational transfer function and $\eta(t)$ is a noise sequence that is independent of the input $z(t)$. This kind of model is appealing since it is the transfer function $\ell(z)$ that is of chief interest, while the noise could be adequately described by quite a simple model. We assume the noise is generated by an ARMA process

$$
a\eta(t) = b\varepsilon(t) \tag{1.1.6}
$$

where $a(z)$, $b(z)$ and $\varepsilon(t)$ have the same meanings as before. The system (1.1.5) will be known as the *output model* and (1.1.6) the *noise model*.

A third description is the *state space form*

$$
\begin{aligned}
&x(t+1) = Fx(t) + Lz(t) + \xi(t-1)\\
&y(t) = Hx(t) + \zeta(t)\\
&E\{\xi(s)\xi(t)'\} = \delta_{st}Q,\ E\{\zeta(s)\zeta(t)'\} = \delta_{st}R\\
&E\{\xi(s)\zeta(t)'\} = \delta_{st}S
\end{aligned}
\tag{1.1.7}
$$

when all of the eigenvalues of F are less than unit modulus. The vector $x(t)$ is called the *state vector*. If for $j \geqslant 1$ we define

$$y(t+j \mid t) = \sum_0^\infty K(\ell+j)\varepsilon(t-\ell) + \sum_0^\infty L(\ell+j)z(t-\ell)$$

and

$$x(t)' = [y(t+1 \mid t)', \ldots, y(t+n \mid t)']$$

where $n \geqslant \max\{p,q,r\}$, then the output of (1.1.1) is seen to be obtained from

$$x(t+1) = \begin{bmatrix} 0 & I_s & 0 & \cdots & 0 \\ 0 & 0 & I_s & & \vdots \\ \vdots & \vdots & & & 0 \\ \vdots & \vdots & & & I_s \\ -A(n) & -A(n-1) & \cdots & & -A(1) \end{bmatrix} x(t) + \begin{bmatrix} L(1) \\ \vdots \\ \vdots \\ \vdots \\ L(n) \end{bmatrix} z(t) + \begin{bmatrix} K(1) \\ \vdots \\ \vdots \\ \vdots \\ K(n) \end{bmatrix} \varepsilon(t)$$

$$y(t) = [I_s, 0, \ldots, 0]x(t) + \varepsilon(t)\ .$$

The eigenvalues of the matrix F in this form are the zeros of $\det a(z)$ and hence less than unit modulus. The form

$$
\begin{aligned}
x(t+1) &= Fx(t) + Lz(t) + K\varepsilon(t)\\
y(t) &= Hx(t) + \varepsilon(t)
\end{aligned}
\tag{1.1.8}
$$

is known as the *prediction error form* of (1.1.7). The state space representation of a given system is not uniquely defined. The output of (1.1.8) may be written as

$$y(t) = HF^N x(t-N) + \sum_1^N HF^{j-1}L\, z(t-j) + \sum_1^N HF^{j-1}K\, \varepsilon(t-j) + \varepsilon(t)$$

thus $y(t)$ is dependent on the state vector at an initial time $t-N$ together with $z(t-j)$, $j = 1,\ldots,N$, $\varepsilon(t-j)$, $j = 0,\ldots,N$. The effect of the initial state diminishes when the initial time is further in the past because the eigenvalues of F are less than unit modulus. As $N \to \infty$ (i.e. the initial time $\to -\infty$) the output converges to a limiting form

$$y(t) = \sum_1^\infty HF^{j-1}L\, z(t-j) + \sum_1^\infty HF^{j-1}K\, \varepsilon(t-j) + \varepsilon(t) .$$

Now $y(t)$ is in the form (1.1.3) with

$$[k(z),\ell(z)] = H(z\, I - F)^{-1}[K,L] + [I_s,0] \tag{1.1.9}$$

and the state space model (1.1.8) gives rise to a representation where the transfer function is rational.

A rational function may be written in the form

$$[k(z),\ell(z)] = \tilde{a}(z)^{-1}[\tilde{b}(z),\tilde{d}(z)] \tag{1.1.10}$$

where $\tilde{a}(z)$, $\tilde{b}(z)$, $\tilde{d}(z)$ are matrices of polynomials. By $\mathrm{diag}\{z^{-\nu_i}\}$ we mean a diagonal matrix with entries $z^{-\nu_i}$ in the i^{th} place along the main diagonal. An ARMAX form may be recovered by defining

$$[a(z),b(z),d(z)] = \mathrm{diag}\{z^{-\nu_i}\}[\tilde{a}(z),\tilde{b}(z),\tilde{d}(z)]$$

yielding the form $ay(t) = b\varepsilon(t) + dz(t)$. Here ν_i is the highest degree of the polynomials in the i^{th} row of

$[\tilde{a}(z),\tilde{b}(z),\tilde{d}(z)]$. The decomposition (1.1.10) is known as the *matrix fraction description* of the transfer function $[k(z),\ell(z)]$, often abbreviated to *mfd*. Again the matrix fraction description is not unique and it is the problem of finding finite parameterisations of spaces of transfer functions in terms of their state space representations and matrix fraction descriptions that occupies §§1.3, 1.4.

§1.2 Identification from second order statistics

As we have indicated there is a certain non-uniqueness in the description of a linear system using, for example, ARMAX or state space representations. Here we decide which characteristics of a linear system can be identified from a complete knowledge of the second order statistics. Denote the spectral density of $z(t)$ by $f_z(\omega)$. Then from (1.1.3)

$$f_y(\omega) = \frac{1}{2\pi} k(e^{-i\omega}) \$_0 \, k(e^{-i\omega})^* + \ell(e^{-i\omega}) f_z(\omega) \ell(e^{-i\omega})^*$$

$$f_{yz}(\omega) = \ell(e^{-i\omega}) f_z(\omega)$$

give the spectrum of $y(t)$ and cross-spectrum of $y(t)$ and $z(t)$. As $\ell(z)$ is a rational function it is fully determined from a finite number of points and very weak conditions are required on $f_z(\omega)$ to determine $\ell(z)$ uniquely from the knowledge of $f_{yz}(\omega)$ and $f_z(\omega)$. It is enough to require

$$\int_{-\pi}^{\pi} p(e^{i\omega}) f_z(\omega) p(e^{i\omega})^* d\omega \neq 0 \tag{1.2.1}$$

for every $p(z) = \sum_0^k p(j) z^j \not\equiv 0$, $k < \infty$. The condition (1.2.1) is satisfied when $f_z(\omega)$ is continuous and $f_z(\omega) \not\equiv 0$, or when $f_z(\omega)$ is non-zero on a set of positive Lebesgue

measure. The kind of input sequence excluded here is one obtained as a solution to a homogeneous difference equation, i.e. one composed of a linear combination of finitely many exponentially damped trigonometric functions. With (1.2.1) we are able to determine

$$f_u(\omega) = \frac{1}{2\pi} k(e^{-i\omega}) \Sigma_0 k(e^{-i\omega})^* \quad . \tag{1.2.2}$$

To recover $k(z)$ and Σ_0 we have

THEOREM 1.2.1. (Rozanov, 1967) *If* $f_u(\omega)$ *is a rational spectral density matrix and* $f_u(\omega) > 0$ *for almost all* $\omega \in [-\pi,\pi]$ *, then there exists a unique factorisation of the form (1.2.2) where*

$$k(z) \text{ is analytic} \quad |z| \geqslant 1$$

$$\det k(z) \neq 0 \quad |z| > 1$$

$$k(0) = I_s \ , \quad \Sigma_0 > 0 \ .$$

As a result processes that give rise to the same transfer functions are indistinguishable (using their second order statistics) and are said to be *observationally equivalent.* Any representation of a process, whether ARMAX or state space, having transfer functions $[k(z),\ell(z)]$ is called a *realisation* of $[k(z),\ell(z)]$. The matrix fraction descriptions $p(z)[a(z),b(z),d(z)]$, where $p(z)$ is a matrix of polynomials with determinant not identically zero, are observationally equivalent. Similarly in a state space representation (1.1.7) the state vector may be transformed $x(t) \longrightarrow Tx(t)$, T invertible inducing the transformation

$$\begin{aligned} F &\longrightarrow TFT^{-1} \ , \ K \longrightarrow TK \\ L &\longrightarrow TL \quad , \ H \longrightarrow HT^{-1} \end{aligned} \tag{1.2.3}$$

which gives an observationally equivalent state space form.

The transformation (1.2.3) is often known as a *system similarity* transformation.

A degree of uniqueness for state space forms may be obtained as follows. Here we are concerned with the algebraic aspects of the description so we shall continue as if $\varepsilon(t)$ is on the same footing as $z(t)$ using the terminology of linear systems as found, for example in Kailath (1980). In (1.1.7) let F be an $n \times m$ matrix and

$$Co = [[K,L],\ F[K,L],\ldots,F^{n-1}[K,L]] \ .$$

If Co is of full rank the system is said to be *controllable*; Co is known as the controllability matrix. When Co is not of full rank a non-singular matrix T may be found so that

$$T\ Co = \begin{bmatrix} Co_1 \\ 0 \end{bmatrix}$$

where Co_1 has $n_1 < n$ rows and is of rank n_1. Let $\tilde{F}$, $\tilde{K}$, $\tilde{L}$, $\tilde{H}$ be obtained from the similarity transformation (1.3.2). Then as

$$T\ Co = [[\tilde{K},\tilde{L}],\ \tilde{F}[\tilde{K},\tilde{L}],\ldots,\tilde{F}^{n-1}[\tilde{K},\tilde{L}]]$$

we have

$$\tilde{K} = \begin{bmatrix} K_1 \\ 0 \end{bmatrix},\ \tilde{L} = \begin{bmatrix} L_1 \\ 0 \end{bmatrix}\ ,\ \tilde{F} = \begin{bmatrix} F_{11} & F_{12} \\ 0 & F_{22} \end{bmatrix}\ ,$$

$$\tilde{H} = [H_1\ H_2]\ ,\quad T\,x(t) = \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix}$$

where the partition is after n_1 rows and columns. Now $x_2(t) = F_{22}^{t+1}\ x_2(0)$ which will converge to zero as the

eigenvalues of F_{22} are less than unit modulus. Thus the effect of $x_2(t)$ is transient. In any case, the transfer functions of the system

$$\begin{aligned} x_1(t+1) &= F_{11}x_1(t) + K_1\varepsilon(t) + L_1z(t) \\ y(t) &= H_1x_1(t) + \varepsilon(t) \end{aligned} \tag{1.2.4}$$

computed from (1.1.9) are unchanged. As Co_1 is of rank n_1 so is $[[K_1,L_1],\ F_{11}[K_1,L_1],\ldots,F_{11}^{n_1-1}[K_1,L_1]]$ hence (1.2.4) is an observationally equivalent controllable system. Similarly, beginning with a controllable system (1.1.8) put

$$Ob = \begin{bmatrix} H \\ HF \\ \vdots \\ HF^{n-1} \end{bmatrix}$$

The matrix Ob is the *observability matrix*; if Ob is of full rank the system is *observable*. If Ob is not of full rank then there is a system similarity transformation T such that $Ob\ T^{-1} = [Ob,0]$ where Ob_1 has full rank n_1. Now if $\tilde{F}$, $\tilde{H}$, $\tilde{K}$, $\tilde{L}$ are defined by the system similarity they can by partitioned as

$$T\,x(t) = \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} \quad \tilde{H} = [H_1\ \ 0] \quad \tilde{F} = \begin{bmatrix} F_{11} & 0 \\ F_{21} & F_{22} \end{bmatrix} \quad \tilde{K} = \begin{bmatrix} K_1 \\ K_2 \end{bmatrix} \quad \tilde{L} = \begin{bmatrix} L_1 \\ L_2 \end{bmatrix}$$

and

$$\begin{aligned} x_1(t+1) &= F_{11}x_1(t) + K_1\varepsilon(t) + L_1z(t) \\ x_2(t+1) &= F_{21}x_1(t) + F_{22}x_2(t) + K_2\varepsilon(t) + L_2z(t) \\ y(t) &= H_1x_1(t) + \varepsilon(t) \end{aligned}$$

so that $x_2(t)$ does not influence the output $y(t)$ and

may be omitted from the state vector giving an observable system. As the original system was controllable, T Co still has full rank, the submatrix

$$Co_1 = [[K_1,L_1],\ F_{11}[K_1,L_1],\ldots,F_{11}^{n_1-1}[K_1,L_1]]$$

has rank n_1 and the transformed system remains controllable. Thus the dimension of the state vector may be reduced to obtain an observationally equivalent representation that is both controllable and observable. A state space representation is called *minimal* if the state vector has the smallest dimension amongst observationally equivalent realisations. The above construction can be extended to the following theorem, see Kailath (1980) for details.

THEOREM 1.2.2. (Kalman, 1963)

(i) *A state space representation is minimal if and only if it is both observable and controllable.*

(ii) *Any two controllable and observable realisations of* $[k(z),\ell(z)]$ *are related by a system similarity transformation.*

The following observation shows that the observable and controllable aspects are all we can hope to learn about a system from second order statistics.

THEOREM 1.2.3. (Kalman, 1963)

"Knowledge of the transfer function identifies the controllable and observable part, and this part alone of the dynamical system which generated it. Moreover this part is identified uniquely up to a system similarity."

Parallel results are obtained for the matrix fraction description of the dynamical system. For the matrix theory

involved we refer to MacDuffee (1956). Further extensions may be found in Kailath (1980, Chapter VI). A polynomial matrix $u(z)$ is *unimodular* if $\det u(z)$ is a non-zero constant. Thus $u(z)^{-1}$ is also a polynomial matrix and $u(z)$ is a unit in the ring of polynomial matrices. A polynomial matrix $a(z)$ is *left prime* if, whenever $a(z) = b(z)d(z)$, $b(z),d(z)$ are matrices of polynomials, and $b(z)$ is square matrix, then $b(z)$ is unimodular. The analogous definition of right primeness requires $d(z)$ to be square when $a(z)$ is right prime if $d(z)$ is unimodular. A characterisation of primeness is given by the following lemma.

LEMMA 1.2.4. *A matrix of polynomials* $a(z)$ *is left prime if and only if* $a(z)$ *has full row rank for every* z . *Similarly* $a(z)$ *is right prime if and only if* $a(z)$ *has full column rank for every* z .

Note that the scalars used to construct linear combinations are polynomials. We notice that from this lemma it is clear the only (left or right) prime square matrices are the unimodular ones.

Some degree of uniqueness in the matrix fraction description $[k(z),\ell(z)] = a(z)^{-1}[b(z),d(z)]$ is achieved by requiring the polynomial matrix $[a(z),b(z),d(z)]$ to be left prime. Then the observationally equivalent classes are given by

THEOREM 1.2.5. (Hannan, 1969, Popov, 1969) *Two matrix fraction descriptions* $[a,b,d]$, $[a_1,b_1,d_1]$ *where* $[a,b,d]$ *is left prime are observationally equivalent if and only if there exists a polynomial matrix* u *such that*

$$u[a,b,d] = [a_1,b_1,d_1] .$$

This leads to the following result that is closely related

to theorem 1.2.2.

THEOREM 1.2.6. *The matrix fraction description* [a(z),b(z), d(z)] *is left prime if and only if* deg{det a(z)} *is minimal among all observationally equivalent matrix fraction descriptions.*

The minimal degree of det a(z) is called the *McMillan degree* of the system (or sometimes the *order*). It is invariant under the transformation

$$[a,b,d] \longrightarrow u[a,b,d]$$

when u is unimodular.

THEOREM 1.2.7. (Popov, 1969) *The McMillan degree of a system is the same as the dimension of the state vector in a minimal state space realisation.*

A class of mfd's (or state space realisations) is said to be *identifiable* if a mfd (state space form) is uniquely determined from the transfer functions [k(z),ℓ(z)]. In the situation of interest here, 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. This is the approach taken in this study. It corresponds to a modelling problem in which we are concerned with obtaining a model which describes certain aspects of the data without trying to model the physical processes generating the data. This kind of model is sometimes referred to as a "*black box*" model. When there is a priori knowledge about the system (for example some parameters may be set to zero because of physical constraints on the system) canonical forms are less valuable and the question arises whether the a priori restrictions are sufficient to guarantee identifiability. This is referred to as *structural identifia-*

bility. One particular identified class of ARMAX models that has been used in relation to some theorems in chapter 3 is where, in (1.1.1), the maximal lag of each component $y_i(t)$, $\varepsilon_i(t)$, $z_i(t)$ is bounded by p_i, q_i, r_i. From the ARMAX form (1.1.1) construct the matrix C where the i, $i+s$, $i+2s$ columns of C are the i^{th} columns of $A(p_i)$, $B(q_i)$, $D(r_i)$ respectively.

THEOREM 1.2.8. (Hannan, 1971) *Let the maximal lags of the components* $y_i(t), \varepsilon_i(t), z_i(t)$ *occurring in the ARMAX form be bounded as above. If* C *is of full rank then the ARMAX form is identified.*

Denote the set of all transfer functions having an ARMAX representation with bounds p_i, q_i, r_i as above and where C is of full rank by $M(p_i, q_i, r_i)$. Many aspects of the structure of the $M(p_i, q_i, r_i)$ are similar to properties of the echelon submanifolds discussed in the next sections. For a discussion of this see Deistler, Dunsmuir and Hannan (1978), Deistler (1981). The spaces $M(p_i, q_i, r_i)$ are a generalisation of the natural univariate ARMAX models

$$\sum_0^p A(j)y(t-j) = \sum_0^q B(j)\varepsilon(t-j) + \sum_1^r D(j)z(t-j), \quad A(0) = B(0) = 1.$$

where at least one of $A(p)$, $B(q)$, $D(r)$ is non-zero. Then the conditions of structural identifiability in Theorem 1.2.8 are easily seen to be satisfied. We shall continue to use these univariate models referring to them as ARMAX(p,q,r) or ARMA(p,q). A number of more general conditions leading to structural identifiability are discussed in Hannan (1971). In the remainder of this thesis we shall make no use of the multivariate models $M(p_i, q_i, r_i)$ or consider questions of structural identifiability.

§1.3 Structure of Spaces of Rational Transfer functions

The algebraic and topological properties considered in the remaining sections are more easily discussed when conditions (1.1.2) are not imposed on the transfer functions $[k(z),\ell(z)]$. This still means that the coefficient matrices $K(j)$, $L(j)$ are defined from $[k(z),\ell(z)] = a(z)^{-1}[b(z),d(z)]$ but this decomposition is only valid for $|z|$ sufficiently large. A function $k(z)$ such that $\lim_{|z|\to\infty} k(z)$ is finite is called *proper*. If the limit is zero then the function is *strictly proper*. Thus in the next sections it is the structure of matrices of rational functions $[k(z),\ell(z)]$ such that $[k(z) - I_s,\ell(z)]$ is strictly proper that is studied. The consequence of (1.1.2), (1.1.4) is that we restrict ourselves to consider subsets of the otherwise simple structures obtained below. While these restrictions will not substantially alter the structure theory that is developed, they will lead to tedious complications in the statements of the results. We leave the discussion of (1.1.2), (1.1.4) to the conclusion of §1.4.

Let $M(n)$ *denote those matrices of rational functions* $[k(z),\ell(z)]$ *where* $[k(z) - I_s,\ell(z)]$ *is strictly proper and has McMillan degree* n . The topology on $M(n)$ is the relative pointwise topology T_{pt} in which a sequence $[k_n(z),\ell_n(z)]$ converges if and only if the coefficient matrices $[K_n(j),L_n(j)] \to [K(j),L(j)]$ for each $j = 1,2,\ldots$. The key result for understanding the various coordinate systems is the following theorem due to Clark (1976).

THEOREM 1.3.1. $M(n)$ *is a real analytic manifold of dimension*

$$n(2s + m) .$$

A related result has been proved in Kalman (1974), Hazewinkel and Kalman (1975). When F, K, L have the meanings in (1.1.8) and the dimension of the state vector is fixed these authors show that the quotient space of completely controllable pairs $\{F,[K,L]\}$ under system similarity transformations is a closed subset of a projective algebraic variety. Of course in their treatment, systems with stochastic inputs $\varepsilon(t)$ are not considered, but treating $\varepsilon(t)$ and $z(t)$ on the same footing, their algebraic results apply to $\{F,[K,L]\}$. However extensions to observable *and* controllable systems do not seem to be available.

As a consequence of the manifold structure of $M(n)$ there is an atlas of charts $\{U_\alpha,\phi_\alpha\}$, $\alpha \in A$ where the U_α are an open covering of $M(n)$. Each ϕ_α is an homeomorphism between U_α and an open subset of Euclidean space of dimension $n(2s + m)$ giving rise to a local coordinate system for U_α . We proceed to give an explicit construction of $\{U_\alpha,\phi_\alpha\}$ that is effectively the main part of Clark's proof of Theorem 1.4.1. To complete the proof of Theorem 1.4.1 it remains to show that ϕ_α is a homeomorphism between U_α (with the topology T_{pt}) and an open subset of $\mathbb{R}^{n(2s + m)}$ (with the natural Euclidean topology) and that for any α, $\beta \in A$ the mapping

$$\phi_\alpha \circ \phi_\beta^{-1} : \phi_\beta(U_\alpha \cap U_\beta) \longrightarrow \phi_\alpha(U_\alpha \cap U_\beta)$$

is analytic. For these details the reader is referred to Clark (1976).

The *Hankel matrix* of $[k(z),\ell(z)]$ is

$$\mathcal{H} = \begin{bmatrix} K(1) & L(1) & K(2) & L(2) & K(3) & L(3) & \cdots \\ K(2) & L(2) & K(3) & L(3) & & & \\ K(3) & L(3) & & & & & \\ \vdots & \vdots & & & & & \end{bmatrix} .$$

We shall refer to $[K(i),L(i), K(i+1),L(i+1),\ldots]$ as the i^{th} row of blocks, while the $s+m$ columns commencing with $[K(i),L(i)]$ will be the i^{th} column of blocks. Denote by $r(u,j)$ the j^{th} row of the u^{th} row of blocks of $\mathcal{H}$ (i.e. the $(u-1)s+j$ row of $\mathcal{H}$). Let $\mathcal{H}_q^p$ be the submatrix of $\mathcal{H}$ in the first p rows and q columns of blocks of $\mathcal{H}$.

LEMMA 1.3.2. (Rissanen, 1974) *If the* i^{th} *row of the Hankel matrix is a linear combination of the rows in positions* $i_1,\ldots,i_k$ *then the* $(i+s)$ *row is the* same *linear combination of the rows* $i_1+s,\ldots,i_k+s$.

The controllability and observability matrices Co, Ob from §1.2 are related to the Hankel matrix by $\mathcal{H}_n^n = Ob \cdot Co$ using (1.1.9). Then the rank of $\mathcal{H}_n^n$ is the minimum of the ranks of Ob and Co , and we obtain the following result:

LEMMA 1.3.3. (Rissanen, 1974)

(i) Rank $\mathcal{H}$ = n , *the McMillan degree of the transfer function* $[k(z),\ell(z)]$.

(ii) Rank $\mathcal{H}_n^n$ = Rank $\mathcal{H}$.

Select a basis for the space spanned by the rows of $\mathcal{H}$ having the following properties:

(1.3.1) *(i)* $r(1,j)$, $j = 1,\ldots,s$ are basis rows,

(ii) If $r(u,j)$, $u > 1$ is a basis row then $r(u-1,j)$ is also a basis row.

The first part of (1.3.1) is not really necessary. It is something of a restriction but it serves to exclude components

of the output that are uncorrelated noise. If (1.3.1*(i)*) cannot be satisfied for any choice of basis then $r(1,j_0)$ for some j_0 must be a linear combination of $r(1,j)$, $j \neq j_0$. Without loss of generality assume $j_0 = s$ and $r(1,s) + \sum_1^{s-1} \alpha_j\, r(1,j) = 0$. Then if we denote the vector $y^{(1)}(t) = (y_1(t),\ldots,y_{s-1}(t))'$ we define

$$y^{(2)}(t) = \varepsilon_s(t) + \sum_1^{s-1} \alpha_j\, \varepsilon_j(t)$$

then we have

$$\begin{bmatrix} y^{(1)}(t) \\ y^{(2)}(t) \end{bmatrix} = \begin{bmatrix} I_{s-1} & 0 \\ \alpha_1 \ldots \alpha_{s-1} \end{bmatrix} y(t) \ .$$

Then $y^{(2)}(t)$ is a trivial output in the sense that it is uncorrelated noise. Using Theorem 1.2.1 $y^{(1)}(t)$ can be written in the form

$$y^{(1)}(t) = \sum_0^\infty K^{(1)}(j)\varepsilon^{(1)}(t-j) + \sum_1^\infty L^{(1)}(j)z(t-j) \ , \ K^{(1)}(0) = I_{s-1}$$

where $\varepsilon^{(1)}(t)$ is now a $s-1$ dimensional sequence of vectors. We shall omit noise components such as $y^{(2)}(t)$ from the outputs and assume condition (1.3.1*(i)*) can always be met. A selection of a basis set satisfying (1.3.1) can be described by a partition $\alpha = \{n_1,\ldots,n_s\}$ of n (i.e. $\Sigma\, n_i = n$, $n_i > 0$) when the basis given by α is

(1.3.2) $\qquad r(i,j),\ i = 1,\ldots,n_j,\ j = 1,\ldots,s \ .$

If we assume that there are no uncorrelated noise components in the outputs as discussed above, Lemma 1.3.2 shows that if rank $\mathcal{H} = n$ it is always possible to partition n so that, for some such $\{n_i\}$ the rows given by (1.3.2) are basis rows. Usually there will be many such partitions. *Denote by* U_α *the set of elements from* $M(n)$ *having Hankel matrices where*

$\alpha = \{n_1,\ldots,n_s\}$ *selects a basis* (1.3.2). The indices $\alpha = \{n_1,\ldots,n_s\}$ are known as *structural indices*. Let $|\alpha| = \Sigma\, n_i$. The properties of the $\mathcal{U}_\alpha$ given in the next theorem are implicit in Clark (1976).

THEOREM 1.3.4.

(i) *Each* $\mathcal{U}_\alpha$, $|\alpha| = n$, *is open and dense in* $M(n)$

(ii) $\bigcup_{|\alpha| = n} \mathcal{U}_\alpha = M(n)$.

As $M(n)$ is a manifold, an obvious consequence of (i) of this theorem is that if $|\alpha| = |\beta| = n$ then $\mathcal{U}_\alpha \cap \mathcal{U}_\beta$ is dense in $M(n)$. We proceed to construct coordinate systems on the $\mathcal{U}_\alpha$. Lemma 1.3.2 shows that it is enough to construct rows $r(n_j+1,j)$, $j = 1,\ldots,s$ from the basis rows given by $\alpha = \{n_1,\ldots,n_s\}$ to recover the whole of $\mathcal{H}$ (and hence the entire transfer function). There exist unique constants $\alpha_{ij}(r)$, $r = 0,\ldots,n_j-1$, $i,j = 1,\ldots,s$ defined by

$$(1.3.3)\quad r(n_i+1,i) + \sum_{j=1}^{s} \sum_{v=1}^{n_j} \alpha_{ij}(v-1)\, r(v,j) = 0 \qquad \alpha_{ij}(r_i) = \delta_{ij}.$$

The basis rows (1.3.2) are determined from the first $s+m$ elements of each row (together with the coefficients defined in (1.3.3)). Let $\kappa_{ij}(v)$, $\lambda_{ij}(v)$ denote the (i,j) elements of $K(v), L(v)$ respectively. Then $\mathcal{U}_\alpha$ can be mapped into Euclidean space of dimension $n(2s + m)$ by the mapping

$$(1.3.4)\quad \phi_\alpha : [k(z),\ell(z)] \mapsto \left\{ \begin{array}{l} \alpha_{ij}(v),\ v = 0,1,\ldots,n_j-1,\ i,j = 1,\ldots,s \\ \kappa_{ij}(v),\ v = 1,\ldots,n_i,\ i,j = 1,\ldots,s \\ \lambda_{ij}(v),\ v = 1,\ldots,n_i,\ i = 1,\ldots,s,\ j = 1,\ldots,m \end{array} \right\}$$

The $\{\mathcal{U}_\alpha,\phi_\alpha\}$ constructed above form an atlas for $M(n)$. Thus there is a covering of $M(n)$ consisting of the $\binom{n+s-1}{s-1}$ open sets $\mathcal{U}_\alpha$, $|\alpha| = n$. It is not known whether there exists an atlas for $M(n)$ consisting of fewer charts, however it is known that unless $s = 1$, $M(n)$ is not homeomorphic to

an open subset of Euclidean space.

Akaike (1974a, 1974b) has provided an elegant construction for canonical forms of state space descriptions (1.1.8) based upon the selection of bases for certain spaces of predictors of $y(t)$. We use these ideas to relate the coordinate system defined through (1.3.4) to the state space form. Define linear predictors

$$(1.3.5) \qquad y(t+a|t) = \sum_{a}^{\infty} K(j)\varepsilon(t+a-j) + \sum_{a}^{\infty} L(j)z(t+a-j)$$

of $y(t+a)$, $a > 0$ from $y(s)$, $z(s)$, $s \leqslant t$. When (1.1.2) holds the $\varepsilon(t)$ may be constructed from $y(s)$, $z(s)$, $s \leqslant t$. When $z(t)$ is absent the predictors $y(t+a|t)$ are the *best* linear predictors of $y(t+a)$ from the past. When $z(t)$ is present there is no such obvious interpretation. Let

$$y_t' = (y(t+1|t)', y(t+2|t)', \ldots)$$

$$e_t' = (\varepsilon(t)', z(t)', \varepsilon(t-1)', z(t-1)', \ldots)$$

then $y_t = \mathcal{H} e_t$. When $[k(z), \ell(z)] \in U_\alpha$, $\alpha = \{n_1, \ldots, n_s\}$, select as basis for the predictor space y_t the components $y_j(t+i|t)$, $i = 1, \ldots, n_j$, $j = 1, \ldots, s$. This corresponds to the basis (1.3.2) for the Hankel matrix. Let $x(t)$ be the vector of these predictors, ordered first according to j and then i ,

$$x(t)' = [y_1(t+1|t), \ldots, y_1(t+n_1|t), y_2(t+1|t), \ldots, y_s(t+n_s|t)] .$$

Then $x(t)$ may be taken as the state vector in the minimum realisation of (1.1.8) where

$$F = [F_{ij}]_{i,j=1,\ldots,s}, \; F_{ij} \text{ are } n_i \times n_j \text{ matrices}$$

$$F_{ii} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & & \vdots \\ \vdots & & & \ddots & \vdots \\ \vdots & & & & 0 \\ 0 & & & & 1 \\ -\alpha_{ii}(0), & \cdots & \cdots & -\alpha_{ii}(n_i-1) & \end{bmatrix} \quad F_{ij} = \begin{bmatrix} 0 & \cdots & \cdots & 0 \\ \vdots & & & \vdots \\ \vdots & & & \vdots \\ \vdots & & & \vdots \\ 0 & \cdots & \cdots & 0 \\ -\alpha_{ij}(0) & \cdots & \cdots & -\alpha_{ij}(n_j-1) \end{bmatrix}$$

$$K = \begin{bmatrix} K_1 \\ \vdots \\ K_s \end{bmatrix} \quad K_i = \begin{bmatrix} \kappa_{i1}(1) & \cdots & \kappa_{is}(1) \\ \vdots & & \\ \kappa_{i1}(n_i) & \cdots & \kappa_{is}(n_i) \end{bmatrix} ;$$

(1.3.6)

$$L = \begin{bmatrix} L_1 \\ \vdots \\ L_s \end{bmatrix} \quad L_i = \begin{bmatrix} \lambda_{i1}(1) & \cdots & \lambda_{im}(1) \\ & & \\ \lambda_{i1}(n_i) & \cdots & \lambda_{im}(n_i) \end{bmatrix}$$

$H = [H_1,\ldots,H_s]$, H_i are $s \times n_i$ matrices ,

$$H_i = [h_{j\ell}] \ , \ h_{j\ell} = \begin{cases} 1 & j = i, \ \ell = 1 \\ 0 & \text{otherwise} \end{cases}$$

where $\alpha_{ij}(v)$, $\kappa_{ij}(v)$, $\lambda_{ij}(v)$ are the parameters appearing in (1.3.4).

A number of canonical forms that have been described (e.g. Luenberger (1967), Mayne (1972), Rissanen and Ljung (1975)) are shown to arise in this way by selecting an appropriate basis for the space of predictors y_t (Akaike, 1974a). The particular canonical form described hitherto is the 'overlapping parameterisation' considered by Rissanen and Ljung (1975). Let $\tilde{a}(z)^{-1}[\tilde{b}(z),\tilde{d}(z)]$ be ~~the~~ *a* matrix fraction description for $[k(z),\ell(z)]$. As $[k(z) - I_s,\ell(z)]$ is strictly proper

$$\deg \tilde{a}(z) = \deg \tilde{b}(z) > \deg \tilde{d}(z) \ .$$

Let

$$\tilde{a}(z) = \sum_{0}^{p} \tilde{A}(j)z^j, \quad \tilde{b}(z) = \sum_{0}^{p} \tilde{B}(j)z^j, \quad \tilde{d}(z) = \sum_{0}^{p-1} \tilde{D}(j)z^j .$$

Then the coefficient matrices in the mfd satisfy

$$(1.3.7) \qquad [\tilde{A}(0),\tilde{A}(1),\ldots,\tilde{A}(p)]\mathcal{H}_{\infty}^{p+1} = 0$$

$$(1.3.8) \quad [\tilde{A}(0),\tilde{A}(1),\ldots,\tilde{A}(p)]\begin{bmatrix} 0 & & & & & & I_s & 0 \\ \cdot & & & & & & K(1) & L(1) \\ \cdot & & & & & & & \\ \cdot & & & & & & & \\ 0 & 0 & I_s & 0 & & & & \\ I_s & 0 & K(1) & L(1) & \cdots & K(p) & & L(p) \end{bmatrix}$$

$$= [\tilde{B}(p),0,\tilde{B}(p-1),\tilde{D}(p-1),\ldots,\tilde{B}(0),\tilde{D}(0)] .$$

If $[k(z),\ell(z)] \in U_\alpha$, $\alpha = \{n_1,\ldots,n_s\}$, let $p = \max\{n_j\}$. From the linear dependence amongst the rows of $\mathcal{H}$ given in (1.3.3) we have, using (1.3.7) that

$$\tilde{A}(v) = [\alpha_{ij}(v)]_{i,j=1,\ldots,s} , \quad v = 1,\ldots,\max\{n_j\} .$$

The coefficients of $\tilde{b}(z)$, $\tilde{d}(z)$ can be recovered from (1.3.8) giving rise to a canonical mfd for $[k(z),\ell(z)] \in U_\alpha$. Let the (i,j) element of the matrix $\tilde{a}(z)$ be $\tilde{a}_{ij}(z)$. From (1.3.3) the elements of $\tilde{a}(z)$ satisfy

$$n_j = \deg\{\tilde{a}_{jj}(z)\} > \deg\{\tilde{a}_{ij}(z)\} \quad i \neq j$$

so that

$$n = \Sigma \deg\{\tilde{a}_{ii}(z)\} \geqslant \deg\{\det \tilde{a}(z)\} \geqslant n .$$

Thus this mfd is left prime from theorem 1.2.6. Finally we recover the coefficients for the ARMAX form. Let

$$[a(z),b(z),d(z)] = \operatorname{diag}\{z^{-\nu_i}\}[\tilde{a}(z),\tilde{b}(z),\tilde{d}(z)]$$

where ν_i is the highest degree in the i^{th} row of $\tilde{a}(z)$. Let the coefficient matrices of $a(z)$, $b(z)$, $d(z)$ be

$A(j)$, $B(j)$, $D(j)$ as in §1.1. As $[k(z),\ell(z)] = a(z)^{-1}[b(z), d(z)]$ we still have $A(0) = B(0)$. However, there is no guarantee that $A(0)$ is non-singular making the ARMAX form difficult to use. The other disadvantage of this form is that aside from elements of $[a(z),b(z),d(z)]$ being constrained to zero or one, and that $A(0) = B(0)$, there are other non-linear constraints on the coefficients of $b(z)$, $d(z)$. Our main interest in ARMAX forms is in their application to parameter estimation using methods based on regression procedures. Even though the non-linear constraints are easily written down, the linear nature of the estimation procedure would be lost. Of course the only constraints in state space form are zero-one constraints but then any parameter estimation procedure is based on optimising a complicated objective function. Let

$$\tilde{B}(u) = [\beta_{ij}(u)]_{i,j=1,\ldots,s} \; ; \; \tilde{D}(u) = [\delta_{ij}(u)]_{\substack{i=1,\ldots,s \\ j=1,\ldots,m}}$$

and $\alpha_{ij}(u)$ be the elements of the coefficient matrices in $\tilde{A}(z)$ as before. The following theorem due to Deistler and Hannan (1981) identifies the freely varying coefficients in the matrix fraction description.

THEOREM 1.3.5. *Let* $\alpha = \{n_1,\ldots,n_s\}$. *A set of coordinates mapping* U_α *homeomorphically into an open subset of* $\mathbb{R}^{n(2s+m)}$ *is*

$$\alpha_{ij}(v) \;,\; v = 0,1,\ldots,n_j-1 \;;\; i,j = 1,\ldots,s$$

$$\beta_{ij}(v) \;,\; v = 0,1,\ldots,n_i-1 \;;\; i,j = 1,\ldots,s$$

$$\delta_{ij}(v) \;,\; v = 0,1,\ldots,n_i-1 \;;\; j = 1,\ldots,m \;;\; i = 1,\ldots,s \;.$$

An alternative parameterisation is constructed where the

parameters occurring in theorem 1.3.5 are the *only* parameters that occur in $\tilde{b}(z)$, $\tilde{d}(z)$. *Let* V_α , $\alpha = \{n_1,\ldots,n_s\}$ *be the set of elements in* $M(n)$ *for which*

$$r(i,j) \ , \ i = 1,\ldots,n_j \ , \ j = 1,\ldots,s$$

is the first *linearly independent set of rows of* $\mathcal{H}$ *when these are ordered as*

(1.3.9) $$r(1,1),r(1,2),\ldots,r(1,s),r(2,1),r(2,2),\ldots \ .$$

This rule for selecting a basis gives rise to a basis satisfying (1.3.1). The important feature of this basis is that the row $r(n_i+1,i)$ is a linear combination of the basis rows preceding it in the ordering (1.3.9). This means that the $\alpha_{ij}(v)$ defined in (1.3.3) have the additional restriction

$$\alpha_{ij}(v) = 0 \quad \begin{cases} v > n_i-1 & j > i \\ v > n_i & j < i \end{cases}$$

thus the highest degree in the i^{th} row of $\tilde{a}(z)$ is n_i . Consequently the highest degree of the i^{th} row of $[\tilde{b}(z)-\tilde{a}(z), \tilde{d}(z)]$ is n_i-1 because $[k(z) - I_s,\ell(z)]$ is strictly proper. Thus the elements of $\tilde{a}(z)$, $\tilde{b}(z)$, $\tilde{d}(z)$ that are not constrained to be zero or one or constrained by $\tilde{a}_{ij}(n_i) = \tilde{b}_{ij}(n_i)$ are a subset of the freely varying elements described in theorem 1.3.5. As $V_\alpha \subset U_\alpha$ this theorem can be applied to show that the only constraints on the elements of $\tilde{a}(z)$, $\tilde{b}(z)$, $\tilde{d}(z)$ are the above linear constraints. The degree relationships in this mfd may be summarised in the following theorem.

THEOREM 1.3.6. *If* $[\tilde{a}(z),\tilde{b}(z),\tilde{d}(z)]$ *is the matrix fraction description of* $[k(z),\ell(z)] \in V_\alpha$, *constructed using* (1.3.7), (1.3.8), *where* $\alpha = \{n_1,\ldots,n_s\}$, *then*

$\tilde{a}_{jj}(z)$ *are monic polynomials*

$\deg\{\tilde{a}_{jj}(z)\} = n_j$, $\deg\{\tilde{a}_{ij}(z)\} < n_j$, $i \neq j$

$\deg\{\tilde{a}_{ji}(z)\} < n_j$, $i > j$; $\deg\{\tilde{a}_{ji}(z)\} \leqslant n_j$, $i \leqslant j$

$\deg\{\tilde{b}_{ji}(z)\} \leqslant n_j$; $\deg\{\tilde{d}_{ji}(z)\} < n_j$.

Coefficients of z^{n_j} *coincide in* $\tilde{a}_{ji}(z)$, $\tilde{b}_{ji}(z)$.

This canonical form is sometimes known as the *echelon form*. The spaces V_α might be called *echelon submanifolds*, since they are submanifolds (albeit not in general open) in $M(n)$. Again V_α is homeomorphic to an open subset of Euclidean space, here of dimension $d(\alpha)$ where

$$(1.3.10)\quad d(\alpha) = n(s+m+1) + \sum_{j<k} [\min\{n_j,n_k\} + \min\{n_j,n_k+1\}] .$$

The homeomorphism can be defined in an analogous way to (1.3.4) and will be denoted by the same symbol ϕ_α .

THEOREM 1.3.7.

(i) $\quad V_\alpha \subseteq U_\alpha$

(ii) $\quad \bigcup_{|\alpha|=n} V_\alpha = M(n)$

(iii) $\quad V_\alpha \cap V_\beta = \emptyset$.

The inclusion in (i) is strict except in one case. This is when the first n rows of the Hankel matrix are independent, that is if $n = ps+q$, $0 \leqslant q < s$ then $n_1 = \ldots = n_q = p+1$, $n_{q+1} = \ldots = n_s = p$. This particular echelon submanifold is open and dense in $M(n)$ and is sometimes known as "*the generic*" neighbourhood. We denote it by $V(n)$ and its structural indices by $\alpha(n)$. From theorem 1.3.5 the ARMAX representation for an echelon form in V_α is recovered by

$$[a(z),b(z),d(z)] = \mathrm{diag}\{z^{-n_i}\}[\tilde{a}(z),\tilde{b}(z),\tilde{d}(z)] .$$

In general $A(0)$ is lower triangular with units along the main diagonal. For the generic neighbourhood $V(n)$ the ARMAX form is as follows. All partitions are after the q^{th} row or column.

$$A(0) = B(0) = \begin{bmatrix} I_q & \vdots & 0 \\ \cdots & \vdots & \cdots \\ A_{21}(0) & \vdots & I_{s-q} \end{bmatrix} \quad A(1) = \begin{bmatrix} A_{11}(1) & \vdots & 0 \\ \cdots & \vdots & \cdots \\ A_{21}(1) & \vdots & A_{22}(1) \end{bmatrix}$$

(1.3.11)

$$A(p+1) = \begin{bmatrix} A_1(p+1) \\ \cdots \\ 0 \end{bmatrix} \quad B(p+1) = \begin{bmatrix} B_1(p+1) \\ \cdots \\ 0 \end{bmatrix} \quad D(p+1) = \begin{bmatrix} D_1(p+1) \\ \cdots \\ 0 \end{bmatrix}$$

$$D(0) = 0 \ , \ A(j), \ B(j), \ D(j) = 0 \quad j > p+1$$

All other $A(j)$, $B(j)$, $D(j)$ are unrestricted.

Though the echelon submanifolds give a disjoint partition of $M(n)$ so that there is a unique set of structural indices for each $[k(z),\ell(z)] \in M(n)$, permuting the components of the output vector $y(t)$ will, in general, change the structural indices. However, though the set of structural indices corresponding to a particular ordering of the outputs is not invariant, the set of structural indices remains invariant under permutation of the outputs.

THEOREM 1.3.8. *The set of structural indices for an echelon submanifold is invariant under permutation of the components of the output.*

Proof A direct proof based on the properties of the Hankel matrix is given here. The set of structural indices is completely determined by the ranks ℓ_i of the Hankel matrices $\mathcal{H}_\infty^{is}$, $i = 1,2,\ldots$. A permutation of the output corresponds to transforming the coefficients of the transfer functions

$$K(j) \rightarrow P\ K(j)P^{-1} \qquad L(j) \rightarrow P\ L(j)$$

where P is the matrix corresponding to that permutation. (Here the $\varepsilon(t)$ are permuted as well so that $K(0) = I_s$.) The ranks of the Hankel matrices $\mathcal{H}_\infty^{is}$, $i = 1,2,\ldots$ remain unchanged when the transfer functions are transformed, hence the set of structural indices remain unchanged as well. □

This proof is not restricted to invariance under permutation of outputs, indeed if $y(t) \to T\, y(t)$, $z(t) \to U\, z(t)$, T,U are invertible, then

$$K(j) \to T\, K(j) T^{-1} \qquad L(j) \to T\, L(j) U$$

and the same invariance result holds.

We shall refer to the structural indices associated with echelon submanifolds as *Kronecker indices*. This form has been variously used to describe these indices or restricted to sets of indices ordered in magnitude. In this last case, when the structural indices are ordered, the Kronecker indices are invariants. Selecting the basis rows of the Hankel matrix as the first independent set under an ordering other than (1.3.9) gives rise to other canonical forms that share the properties of V_α. One ordering that has been considered is

$$r(1,1), r(2,1), \ldots, r(n,1), r(1,2), r(2,2), \ldots \quad .$$

This has the tendency to give rise to structural indices that vary greatly in magnitude. Other coordinate systems that have been considered are the manifolds $M(p_i, q_i, r_i)$ defined in §1.2. These manifolds have been used to obtain theoretical results for maximum likelihood estimation (see references in §2.2). For parameter estimation they suffer from the difficulty that $2s + m$ indices $\{p_i, q_i, r_i\}$ need to be known or estimated compared to s indices when V_α are used.

Much of the literature concerning canonical forms

relates to the state equation of a controllable state space realisation. This approach (e.g. Luenberger (1967), Rosenbrock (1970)) parallels the scalar input-scalar output model and is based on selecting a basis for the columns of the controllability matrix according to some rule including (1.3.1). In this way we can define structural indices similar to our previous ones. When the basis columns are selected as the first linearly independent set in the natural ordering (this parallels the construction of the canonical form for echelon submanifolds), Rosenbrock (1970) shows that the structural indices obtained are the classical minimal indices (or Kronecker indices) for the singular pencil of matrices $(z\, I_n - F, [K,L])$. The invariant properties of these structural indices and the canonical forms constructed using them are discussed in Kalman (1971), Popov (1972). Another construction for these canonical forms and structural indices is based on the concept of a minimal polynomial basis for the null space of the transfer function $[k(z), \ell(z)]$. This approach gives rise to structural indices that coincide with the minimal indices of the above singular pencil of matrices (Forney (1975), see also Kailath (1980) for a further discussion).

§1.4 Structure of $\overline{M(n)}$

The topological properties in the previous section relate to an isolated $M(n)$. For the estimation problems of interest here the McMillan degree n_0 of the true model will not be known in general. Thus we consider the nature of the boundary points of $M(n)$ since we are not assuming that the

true structure is an interior point of a given $M(n)$. Let $\bar{A}$ denote the closure, with respect to T_{pt} , of a set $A \subset \bigcup_n M(n)$. Introduce a partial ordering "$\leqslant$" on the structural indices

$\alpha^{(1)} = \{n_1^{(1)},\ldots,n_s^{(1)}\} \leqslant \alpha^{(2)} = \{n_1^{(2)},\ldots,n_s^{(2)}\}$ if and only if $n_i^{(1)} \leqslant n_i^{(2)}$

for each i . Let $\Pi_\alpha : \mathbb{R}^{n(2s+m)} \to \bigcup_n M(n)$ be the mapping that constructs a transfer function by reversing the procedure used to construct ϕ_α . Then $\Pi_\alpha \circ \phi_\alpha : U_\alpha \to U_\alpha$ is the identity.

THEOREM 1.4.1.

(i) $\overline{M(n)} = \bigcup_{i \leqslant n} M(i)$

(ii) $\overline{\phi_\alpha(U_\alpha)} = \mathbb{R}^{n(2s+m)}$

(iii) $\Pi_\alpha \overline{\phi_\alpha(U_\alpha)} = \bigcup_{\beta \leqslant \alpha} U_\beta$

(iv) *T_{pt} is weaker than the manifold topology. U_α is the largest subset of $\Pi_\alpha \overline{\phi_\alpha(U_\alpha)}$ where the corresponding relative topologies coincide.*

With $d(\alpha)$ defined as in (1.3.10), we use the same symbol Π_α for the mapping $\mathbb{R}^{d(\alpha)} \to \bigcup_n M(n)$ that reconstructs the transfer function from its echelon form in V_α .

THEOREM 1.4.2.

(i) $\overline{\phi_\alpha V_\alpha} = \mathbb{R}^{d(\alpha)}$

(ii) $\Pi_\alpha \overline{\phi_\alpha V_\alpha} = \bigcup_{\beta \leqslant \alpha} V_\beta$

(iii) *V_α is the largest subset of $\Pi_\alpha \overline{\phi_\alpha V_\alpha}$ where the relative T_{pt} and manifold topologies coincide.*

Theorems 1.4.1, 1.4.2 are from Deistler and Hannan (1981).

Results for the spaces $M(p_i,q_i,r_i)$ that are analogous to theorem 1.4.2 appear in Deistler, Dunsmuir and Hannan (1978). The boundary of $M(n)$ is made up of the manifolds $M(i)$, $i < n$. We relate these theorems to parameter estimation. Let us assume we are seeking to optimise an objective function $L_T(\theta)$ defined for $\theta \in M(i)$, $i \leqslant n$. Assume the true model $\theta_0 \in U_{\alpha_0} \subset M(n_0)$, $n_0 \leqslant n$ and that when $L_T(\theta)$ is optimised over $\overline{M(n)}$ our estimation procedure leads to an estimate θ_T which is consistent. What we have in mind is that $L_T(\theta)$ is the likelihood based on T observations as defined in (2.2.1) when, as in theorem 2.2.1, the maximum likelihood estimate is consistent. To perform the optimisation we work in a coordinate system for some U_α and maximise $L_T(\Pi_\alpha(\tau))$, $\tau \in \mathbb{R}^{n(2s+m)}$. At this stage we do not consider constraints on the parameter space. Assume that $|\alpha| \geqslant n_0$ and $\theta_T \to \theta_0$. If $\liminf \| \phi_\alpha(\theta_T) \|$ remains finite then there exists a subsequence T_i along which $\phi_\alpha(\theta_{T_i})$ converges to say τ_∞. Because $\theta_T \to \theta_0$, $\Pi_\alpha(\tau_\infty) = \theta_0$ and from theorem 1.4.1 (iii), $\theta_0 \in U_\beta$, $\beta \leqslant \alpha$. Hence, even if the true McMillan degree n_0 is known but we have selected a coordinate neighbourhood U_α, $|\alpha| = n_0$ such that $\theta_0 \notin U_\alpha$, the sequence of parameter estimates diverges to infinity. The problem of selecting a suitable coordinate system in this situation is discussed in §2.2. Similarly unless $\theta_0 \in U_\beta$, $\beta \leqslant \alpha$ the sequence of parameter estimates diverges to infinity just as when the McMillan degree is known. A third case is when $|\alpha| \geqslant n_0$ and $\alpha_0 < \alpha$. Here $\liminf \| \phi_\alpha(\theta_T) \|$ may remain finite but there is no reason to expect the sequence of parameter estimates $\phi_\alpha(\theta_T)$ to converge even though $\Pi_\alpha(\phi_\alpha(\theta_T)) \to \theta_0$. This is because the Euclidean topology in

$\mathbb{R}^{n(2s+m)}$ is stronger than τ_{pt} on the space of transfer functions (Theorem 1.4.1(iv)). The discussion of these cases is from Deistler and Hannan (1981). The nature of the problem in the third case above is most easily discussed for univariate ARMA models. In the univariate case, the spaces ARMA(p,p) and $M(p)$ are equivalent and we may parameterise $M(p)$ by the coefficients $A(j)$, $B(j)$, $j = 1,\ldots,p$ of the ARMA model

$$\sum_0^p A(j)y(t-j) = \sum_0^p B(j)\varepsilon(t-j), \quad A(0) = B(0) = 1 .$$

Let $k_T(z) \in M(p_1)$, $p_1 > p$ have a prime mfd $a_T(z)^{-1}b_T(z)$. Then $k_T(z) \to k_0(z) \in M(p)$ if $[a_T(a), b_T(z)]$ converges to the set $\{[a_0(z)v(z), b_0(z)v(z)], \deg\{v(z)\} = p_1 - p\}$. This set corresponds to an affine subspace of dimension $(p_1 - p)$ in $\mathbb{R}^{2p_1}$, the parameter space for $M(p_1)$. (The models discussed here are univariate, hence $M(p)$ is isomorphic to a subset $\mathbb{R}^{2p}$.) A similar result holds for the multivariate case.

THEOREM 1.4.3. (Deistler and Hannan, 1981) *Let* $[k_0(z), \ell_0(z)] \in U_{\alpha_0}$, $\alpha_0 \leq \alpha$ *and* $|\alpha_0| = n_0$, $|\alpha| = n$. *Then the set*

$$\{\tau \in \overline{\phi_\alpha(U_\alpha)} : \Pi_\alpha \tau = [k_0(z), \ell_0(z)]\} \tag{1.4.1}$$

is an affine subspace of $\mathbb{R}^{n(2s+m)}$ *of dimension* $(n - n_0)s$.

To see how this result arises note that the construction used to obtain the embedding $\phi_\alpha : U_\alpha \to \mathbb{R}^{n(2s+m)}$ can be used to obtain the equivalence class in $\mathbb{R}^{n(2s+m)}$ corresponding to $[k_0(z), \ell_0(z)] \in U_{\alpha_0}$, $\alpha_0 < \alpha$. The coefficients $\kappa_{ij}(v)$, $\lambda_{ij}(v)$ are uniquely defined by ϕ_α just as in (1.3.4). The coefficients $\alpha_{ij}(v)$, corresponding to $[k_0(z), \ell_0(z)]$, are no longer uniquely determined by (1.3.3) since the rows $r(u,i)$, $u = 1,\ldots,n_i$, $i = 1,\ldots,s$, where $\alpha = \{n_1,\ldots,n_s\}$

are no longer independent. The independent rows are $r(u,i)$, $u = 1,\ldots,n_{0i}$, $i = 1.,\ldots,s$, $\alpha_0 = \{n_{01},\ldots,n_{0s}\}$ hence the solutions $\alpha_{ij}(v)$ of (1.3.3) give rise to a subspace of dimension $(|\alpha| - |\alpha_0|)s$ as asserted in the theorem. A similar construction applies to the equivalence classes in $\mathbb{R}^{d(\alpha)}$ corresponding to $[k_0(z),\ell_0(z)] \in V_{\alpha_0}$. In this case the dimension of the affine subspace (1.4.1) becomes

$$d(\alpha) - d(\alpha_0) - (|\alpha| - |\alpha_0|)(s+m) .$$

This aspect complicates a parameter estimation procedure when the true McMillan degree is not known, but in practice, when it is unlikely that the McMillan degree is finite anyway, it may not be such a problem. For theoretical work in later chapters we embed the parameter space $\mathbb{R}^{n(2s + m)}$ (or $\mathbb{R}^{d(\alpha)}$) into a direct sum $S_1 + S_2$ of Euclidean spaces so that the equivalence class (1.4.1) is parameterised by S_1. Then we will show that the parameters in S_2 converge. The embedding is dependent on the parameters of the true model. Here, in the univariate ARMA case, we provide an explicit construction of this parameterisation. Assume the true model $a_0y(t) = b_0\varepsilon(t)$ is in $ARMA(p_0,q_0)$. For $p \geqslant p_0$, $q \geqslant q_0$ let $\tau_0 = \{A_0(1),\ldots,A_0(p_0),0,\ldots,B_0(1),\ldots,B_0(q_0),0,\ldots,0\} \in \mathbb{R}^{p+q}$ where the entry $B_0(1)$ is in the $p+1$ position. The elements $A_0(j)$, $B_0(j)$ are the coefficients of $a_0(z)$, $b_0(z)$ as before. Let $\tau = \{A(1),\ldots,A(p),B(1),\ldots,B(q)\}$ where the $A(j)$, $B(j)$ are the coefficients of $a(z)$, $b(z)$. Define a mapping

$$L : \mathbb{R}^{p+q} \to \mathbb{R}^n , \quad n = \max\{p+q_0,q+p_0\}$$

$$: \tau-\tau_0 \mapsto \chi , \quad \chi = (\chi_1,\ldots,\chi_n)'$$

where $\Sigma \chi_j z^{-j} = a(z)b_0(z) - b(z)a_0(z)$. The null space of L consists of the vectors of coordinates for $a_0(z)\nu(z)$, $b_0(z)\nu(z)$ where $\deg \nu(z^{-1}) = \min\{p-p_0, q-q_0\} = v$. Furthermore L is a full rank mapping so it may be extended to an invertible transformation

$$
\begin{aligned}
& L_1 : \mathbb{R}^{p+q} \to \mathbb{R}^{p+q} \\
& : \tau-\tau_0 \to \begin{bmatrix} \chi \\ \psi \end{bmatrix}, \quad \psi = (\nu_1,\ldots,\nu_v)'
\end{aligned} \tag{1.4.2}
$$

where $\nu(z) = 1 + \sum_1^v \nu_j z^{-j}$. From the entries in L_1^{-1}, polynomials in z^{-1} $\xi_j^a(z), \xi_j^b(z)$ of degrees p,q respectively, can be constructed for $j = 1,\ldots,n$, so that

$$
\begin{aligned}
a(z)-a_0(z) &= \sum_{j=1}^n \xi_j^a(z)\chi_j + \sum_1^v \nu_j (z^j a_0(z)) \\
b(z)-b_0(z) &= \sum_{j=1}^n \xi_j^b(z)\chi_j + \sum_1^v \nu_j (z^j b_0(z)) \ .
\end{aligned} \tag{1.4.3}
$$

This follows from writing

$$
\begin{bmatrix} a(z)-a_0(z) \\ b(z)-b_0(z) \end{bmatrix} = \begin{bmatrix} z^{-1},\ldots,z^{-p}, & 0 \ \ldots \ 0 \\ 0 \ \ldots \ 0, & z^{-1},\ldots,z^{-q} \end{bmatrix} (\tau-\tau_0) = L_1^{-1} \begin{bmatrix} \chi \\ \psi \end{bmatrix}
$$

when the terms on the RHS of (1.4.3) are obtained by putting in turn $\chi = 0$ and $\psi = 0$. We mimic this construction in the multivariate case.

Let $\begin{bmatrix} k_0(z)' \\ \ell_0(z)' \end{bmatrix} \in U_{\beta_0}$ and $\left[\tilde{a}^\circ(z)', \begin{bmatrix} \tilde{b}^\circ(z)' \\ \tilde{d}^\circ(z)' \end{bmatrix}\right]$ be the canonical *left* mfd for $\begin{bmatrix} k_0(z)' \\ \ell_0(z)' \end{bmatrix}$ in U_{β_0} . Consequently we have the *right* mfd $[k_0(z),\ell_0(z)] = [\tilde{b}^\circ(z),\tilde{d}^\circ(z)]\tilde{a}^\circ(z)^{-1}$. Let $[a^\circ(z),b^\circ(z),d^\circ(z)]$ be the corresponding ARMAX form for U_{β_0} .

Assume $[k_0(z),\ell_0(z)] \in U_{\alpha_0}$, $|\alpha_0| = n_0$. There are $s+m$ structural indices $\{n_1^\circ,\ldots,n_{s+m}^\circ\}$ for U_{β_0}, and as $|\beta_0|$ is the column rank of the Hankel matrix for $[k_0(z),\ell_0(z)]$, $|\alpha_0| = |\beta_0|$. Define

$$\chi(z) = \Sigma\chi_j z^{-j} = a(z)[b^\circ(z),d^\circ(z)] - [b(z),d(z)]a^\circ(z)$$

where $[a(z),b(z),d(z)]$ is the canonical left mfd in U_α of $[k(z),\ell(z)]$ and $\alpha > \alpha_0$. Then $\chi(z) \equiv 0$ if and only if $a(z)^{-1}[b(z),d(z)] = [k_0(z),\ell_0(z)]$ as in the univariate case. In general, as well as elements of $\chi(z)$ being constrained to zero there will also be other linear constraints on these elements. In several cases of importance the non-zero coefficients will be freely varying. Indeed if $[a(z),b(z),d(z)]$ is in echelon form for V_α, $\alpha = \{n_1,\ldots,n_s\}$ and $\left[a^\circ(z)', \begin{bmatrix} b^\circ(z)' \\ d^\circ(z)' \end{bmatrix}\right]$ is in echelon form for V_{β_0}, $\beta_0 = \{n_1^\circ,\ldots,n_{s+m}^\circ\}$ then the degree of the polynomial in the i^{th} row and j^{th} column of $\chi(z)$ is $n_i + n_j^\circ$. Consequently $\chi(z)$ has

$$(1.4.4) \qquad \sum_{i=1}^{s}\sum_{j=1}^{s+m}(n_i+n_j^\circ) = (n+n_0)s + nm \ , \ n_0 = |\beta_0| \ , \ n = |\alpha|$$

parameters that are not restricted to zero or one. If in addition $V_\alpha = V(n)$ then the null space of the mapping

$$(1.4.5) \qquad L : [a(z),b(z),d(z)] \mapsto \chi(z)$$

has dimension $(n-n_0)s$ from theorem 1.4.3. As the parameter space for $V(n)$ has dimension $n(2s + m)$, all the parameters of $\chi(z)$ that are not constrained to zero or one must be freely varying. If V_α is not the generic neighbourhood then from the comments below theorem 1.4.3 the null space of (1.4.5) has dimension $d(\alpha)-d(\alpha_0)-(n-n_0)(s+m)$ hence combining

this with (1.4.4) we see there must be $n_0(2s+m) - d(\alpha_0)$ constraints on the parameters in $\chi(z)$ (as well as the zero-one constraints). These parameterisations were introduced in Hannan (1980,1981) where the constraints were not important. In the univariate case they were used in Hannan and Rissanen (1982).

Up till this point conditions (1.1.2) or (1.1.5) have not been imposed on the spaces of transfer functions. When these conditions are imposed denote the subspaces of $M(n)$, U_α, V_α by $M(n)$, U_α, V_α respectively. Thus $M(n)$ is a manifold having an atlas $\{\phi_\alpha, U_\alpha\}$ where ϕ_α is defined just as before. Of course the canonical forms for the U_α and echelon forms for V_α are unchanged and the theorems in §1.3 apply to U_α, V_α. In constructing the closure of $M(n)$, transfer functions with $\det k(z)$ having zeros on $|z| = 1$ need to be included in $\overline{M(n)}$ together with the lower dimensional manifolds $M(i)$, $i \leqslant n$. Thus we define $M(n)^*$ as the space consisting of $M(n)$ together with those limit points where $\det k(z) = 0$, $|z| = 1$. The spaces U_α^*, V_α^* are similarly extended from U_α, V_α. Then as an analogue of theorems 1.4.1, 1.4.2 we have

$$\overline{M(n)} = \bigcup_{i \leqslant n} M(i)^*$$

$$\Pi_\alpha \overline{\phi_\alpha U_\alpha} = \bigcup_{\beta \leqslant \alpha} U_\beta^* \qquad \Pi_\alpha \overline{\phi_\alpha V_\alpha} = \bigcup_{\beta \leqslant \alpha} V_\beta^* .$$

Of course $\overline{\phi_\alpha U_\alpha}$ is no longer the whole of $\mathbb{R}^{n(2s+m)}$ but a rather complicated unbounded subset. Then the equivalence class corresponding to (1.4.1) is a subset of the affine subspace mentioned in theorem 1.4.3. The precise nature of this subset is of little consequence for the estimation procedures that follow.

Consider the minimisation of the trace of

$$\frac{1}{2\pi}\int_{-\pi}^{\pi}(ak_0 - b - I_s)\Phi_0(ak_0 - b - I_s)^* d\omega + \int_{-\pi}^{\pi}(a\ell_0 - d)dF_z(a\ell_0 - d)^*$$

(1.4.6)

$$= \Phi_0 + \frac{1}{2\pi}\int_{-\pi}^{\pi}(ak_0 - b)\Phi_0(ak_0 - b)^* d\omega + \int_{-\pi}^{\pi}(a\ell_0 - d)dF_z(a\ell_0 - d)^* .$$

Here expressions such as $ak_0 - b$ are abbreviations for $a(e^{i\omega})k_0(e^{i\omega}) - b(e^{i\omega})$, a notation with which we will persist in expressions of this kind. Here $[k_0(z),\ell_0(z)] \in V_{\alpha_0}$ while the minimisation is carried out for $[a(z),b(z),d(z)]$ being a mfd or echelon form in V_α. If $\alpha \geqslant \alpha_0$ (1.4.6) is minimised when $a(z)^{-1}[b(z),d(z)] = [k_0(z),\ell_0(z)]$, thus the minimum is Φ_0. When $\alpha \not\geqslant \alpha_0$ the second and third terms on the RHS of (1.4.6) cannot be zero since this would mean that $[k_0(z),\ell_0(z)]$ has a mfd in V_α. Denote the minimum over V_α of (1.4.6) by Φ_α. Consequently for $\alpha \not\geqslant \alpha_0$, the minimum of (1.4.6) bounded away from Φ_0.

CHAPTER 2

STRUCTURE AND PARAMETER ESTIMATION BASED ON MAXIMUM LIKELIHOOD METHODS

§2.1 Introduction

In this chapter a number of results concerning parameter estimation and the determination of model structure are collated. Computational aspects are deferred until Chapter 3. Though maximum likelihood estimates of the transfer function $[k_0(z), \ell_0(z)]$ are consistent when only the McMillan degree is known, any optimization procedure to compute the estimates calls for a finite dimensional coordinate system for an appropriate neighbourhood of $[k_0(z), \ell_0(z)]$. The ideal situation where the McMillan degree is known a priori seems realistic only when the system obeys known physical laws. Then structure determination is no longer a question and the problems in specifying a coordinate system reduce to questions of structural identifiability (see Theorem 1.2.6 and the comments at the end of §1.2).

In the modelling situation considered here the laws governing the system would, at best, be only partly known and the problem is one of constructing a model that is adequate for the purpose the experimenter has in mind, for example, prediction, or spectral estimation. Here it is unlikely that the true transfer functions are rational. It becomes necessary to select a McMillan degree that is large enough to allow a sufficiently complex model commensurate to the quality of the data available. Thus as the amount of data increases a more complicated model would be called for. While the McMillan degree seems to be a natural characteristic, the structural indices introduced in defining U_α only play a role in determining a convenient parameter space in

which to do the required numerical computations. The Kronecker indices do have a deeper meaning in describing the structure of the null space of $[k(z),\ell(z)]$ when the latter is considered as a mapping from the spaces of input sequences to output sequences as indicated in §1.3. These aspects are of interest in control theory but will not be considered here. In the context of this work our interest in Kronecker indices arises not only because they lead to model structures that lend themselves to linear estimation techniques that will be developed in Chapter 3, but also that the V_α are often parameterised using fewer parameters than in U_α .

We discuss consistency theorems for maximum likelihood estimation of parameters as well as structural determination under very weak conditions on the inputs $\varepsilon(t)$ and $z(t)$. The input sequence $z(t)$ will be taken as a known and fixed sequence. Let F_t denote the σ-algebra of events determined by $y(s)$, $s \leqslant t$. Equivalently, as $z(t)$ is fixed and we are assuming (1.1.2), F_t is determined by $\varepsilon(s)$, $s \leqslant t$. A natural condition on the innovations is that they are *martingale differences*

$$E\{\varepsilon(t) \mid F_{t-1}\} = 0 \text{ a.s.} \tag{2.1.1}$$

This means that the best predictor of $y(t)$ from the past $y(t-j)$, $z(t-j)$, $j = 1,\ldots,\infty$ is the *best* linear predictor of $y(t)$, given the past. The former is

$$\sum_1^\infty L(j)z(t-j) + E\{\sum_0^\infty K(j)\varepsilon(t-j) \mid F_{t-1}\} .$$

If the best predictor is the best linear predictor then

$$E\{\sum_0^\infty K(j)\varepsilon(t-j) \mid F_{t-1}\} = \sum_1^\infty K(j)\varepsilon(t-j)$$

which is equivalent to (2.1.1). Thus (2.1.1) is a natural

justification for using a linear model in the first place (Hannan and Heyde, 1972). For some purposes we need

$$E\{\varepsilon(t)\varepsilon(t)' \mid \mathcal{F}_{t-1}\} = \Sigma_0, \tag{2.1.2}$$

a condition that does hold when the $\varepsilon(t)$ are Gaussian. However, except on the basis of Gaussianity, (2.1.2) seems difficult to justify from properties of the observed output. For most purposes it is replaced by the weaker condition

$$E\{\varepsilon(t)\varepsilon(t)' \mid \mathcal{F}_{-\infty}\} = \Sigma_0. \tag{2.1.3}$$

In the ARMAX case we also need

$$E\{\varepsilon_i(t)\varepsilon_j(t)\varepsilon_k(t-r) \mid \mathcal{F}_{-\infty}\} = \sigma_{ijk}(r) . \tag{2.1.4}$$

Here $\mathcal{F}_{-\infty}$ is the σ-algebra of events in the infinite past, $\mathcal{F}_{-\infty} = \bigcap_j \mathcal{F}_j$. These conditions amount to saying that $\varepsilon(t)\varepsilon(t)' - \Sigma_0$ and $\varepsilon_i(t)\varepsilon_j(t)\varepsilon_k(t-r) - \sigma_{ijk}(r)$ are purely non-deterministic in the sense that they are not influenced by the infinitely distant past. Both (2.1.3), (2.1.4) are implied by (2.1.2). Even though $z(t)$ is taken to be a fixed sequence we shall assume that the following limits exist, namely

$$\begin{aligned} &\lim_{T\to\infty} \frac{1}{T} \sum_1^T z(t)z(t+j)' = \gamma_z(j) \ , \ j = 0,1,\ldots . \\ &\lim_{T\to\infty} \frac{1}{T} \sum_1^T z(t) = 0 . \end{aligned} \tag{2.1.5}$$

The existence of these limits means that

$$\gamma_z(j) = \int_{-\pi}^{\pi} e^{ij\omega} \, dF_z(\omega)$$

where $F_z(\omega)$ is Hermitian non-negative and is non-decreasing in ω.

§2.2 Consistency of maximum likelihood estimates

In this section we state consistency results and a central limit theorem for maximum likelihood estimation for ARMAX models under the very general conditions considered in Dunsmuir and Hannan (1976), Deistler, Dunsmuir and Hannan (1978) (for ARMA models) and Hannan, Dunsmuir and Deistler (1980) (the ARMAX case). The earliest limit theorems of this nature are due to Whittle (1951), where Gaussian univariate ARMA models are considered, and subsequently, for linear models $y(t) = \Sigma K(j)\varepsilon(t-j)$, where the $\varepsilon(t)$ are independent and identically distributed, (Walker (1964)). The first theorems under the general conditions given here are for univariate ARMA models in Hannan (1973). The consistency theorem stated here relates to the estimates of the transfer functions and their convergence in terms of the natural pointwise topology $\mathcal{T}_{pt}$. Consequently a coordinate representation is not needed.

Let U be a set of transfer functions $[k(z),\ell(z)]$ having a matrix fraction description $a(z)^{-1}[b(z),d(z)]$ where the degrees p, q, r of $a(z)$, $b(z)$, $d(z)$ are a priori bounded and

$$\det a(z) \neq 0 \qquad |z| \geqslant 1$$

$$\det b(z) \neq 0 \qquad |z| \geqslant 1$$

$$k(0) = I_s \ .$$

The variance matrix Σ of the innovations process may be parameterised by $\sigma : \Sigma \to \mathbb{R}^{s(s+1)/2}$ which selects the on and above diagonal elements of Σ. Define

$$\Theta = U \times \{\sigma(\Sigma) : \Sigma > 0\} \ .$$

Let

$$\tilde{\Theta} = \bar{U} \times \{\sigma(\natural) : \natural > 0\}$$

where the closure of U is taken with respect to $\mathcal{T}_{pt}$. Though U need not have any further structure, it will usually be thought of as one of the subspaces U_α or V_α of $M(n)$ introduced in §1.4. Though Gaussian assumptions are not made, a likelihood is constructed as if the innovations were Gaussian. Put $y(t) = v(t) + u(t)$, where v and u are defined as in (1.1.3), then $v(t)$ is generated by

$$a\ v(t) = d\ z(t) .$$

Thus $v(t)$ depends on $a(z)$ and $d(z)$ together with $z(t)$, $t = 1,\ldots,T$ and a finite number of initial values $v(0)$, $v(-1),\ldots,v(-p)$, $z(0),\ldots,z(-r)$. Though ideally the likelihood could be optimised with respect to the initial values, we shall set them to zero. Let $u_T = (u(1)',\ldots,u(T)')'$ and $\Gamma_T(\theta) = E(u_T\, u_T')$, then for $\theta \in \tilde{\Theta}$, neglecting constant terms,

$$L_T(\theta) = -2T^{-1}(\log \text{ Gaussian likelihood})$$

$$= T^{-1}\log\det \Gamma_T(\theta) + T^{-1}u_T'\Gamma_T(\theta)^{-1}u_T . \tag{2.2.1}$$

The estimates $\tilde{\Theta}_T$ (or $\hat{k}_T(z)$, $\hat{\ell}_T(z)$, $\hat{\natural}_T$) obtained by minimizing $L_T(\theta)$ over $\tilde{\Theta}$ will be referred to as the maximum likelihood estimates, while, for convenience $L_T(\theta)$ will be called the likelihood.

THEOREM 2.2.1. *Assume that* $y(t)$ *is generated by an ARMAX process and* $\hat{\theta}_T$ *is the maximum likelihood estimate of* $\theta_0 \in \tilde{\Theta}$. *If* $\varepsilon(t)$ *is ergodic and* (1.2.1), (2.1.5) *hold, then*

$$\hat{\theta}_T \to \theta_0 \quad \text{a.s.}$$

If the true model has McMillan degree $j \leqslant n$ and, for example, U is taken as U_α, $|\alpha| = n$, then $\bar{U} = \overline{M(n)} = \bigcup_{j \leqslant n} M(j)^*$ so that the maximum likelihood estimate is consistent even when the McMillan degree is exceeded. Of course, from the discussion following Theorem 1.4.2, the estimates of the parameters in the canonical form for U_α will not be consistent unless $\theta_0 \in U_\alpha$. Though the discussion here is limited to ARMAX models, U can also represent a space of transfer functions parameterised by $\tau \in \Xi$. Then if models

$$(2.2.2) \qquad y(t) = \sum_0^\infty K(\tau,j)\varepsilon(t-j), \quad \sum_0^\infty \|K(\tau,j)\|^2 < \infty, \quad K(\tau,0) = I_s$$

are considered and Ξ is a compact space containing the true model τ_0 then $\hat{\tau}_T \to \tau_0$ (see Dunsmuir and Hannan, 1976). For the univariate ARMA case (Hannan, 1973) and for (2.2.2) the proof of consistency exploits the compactness of the parameter space for U. In the proof of Theorem 2.2.1 in Hannan, Dunsmuir and Deistler (1980) for ARMAX models, such compactness assumptions are not necessary. Similar results using simpler techniques but requiring Gaussian assumptions and a compact parameter space have been obtained by Rissanen and Caines (1979) for multivariate ARMA estimation.

When the true model is not in $\tilde{\Theta}$ it may happen that the limiting form of the likelihood (2.2.1) as $T \to \infty$ has several absolute minima. Call the set of absolute minima Θ_0. Indeed Astrom and Soderstrom (1974), Kabaila (1983) give examples of univariate processes where the likelihood has more than one local minimum in a particular model set $\tilde{\Theta}$. It is implicit in Dunsmuir and Hannan (1976) that if $\theta_0 \in M(n_0)$ and $\Theta = M(n) \times \{\sigma(\natural), \natural > 0\}$ where $n < n_0$, then $\hat{\theta}_T$ converges to the limit set Θ_0. However, if $[k_0, \ell_0]$

is the true transfer function then

$$(2.2.3)\quad \hat{\Sigma}_{n,T} \to \Sigma_n = \inf_{[k,\ell]\in M(n)} \left\{ \int_{-\pi}^{\pi} k^{-1} k_0 \Sigma_0 k_0^* (k^{-1})^* d\omega + \int_{-\pi}^{\pi} \ell^{-1} \ell_0 dF_z(\omega) \ell_0^* (\ell^{-1})^* \right\}$$

If $\theta_0 \in \text{int}\ \tilde{\Theta}$ then for $T > T_0$, T_0 large enough, $\hat{\theta}_T$ will remain within a neighbourhood of θ_0 which can be parameterised using the one coordinate system. Let $\hat{\tau}_T$, τ_0 be the coordinate of the transfer functions corresponding to $\hat{\theta}_T$, θ_0 respectively. Then we have the following theorem.

THEOREM 2.2.2. *Let the hypotheses of theorem 2.2.1 hold and* $\theta_0 \in \text{int}\ \tilde{\Theta}$. *If* (2.1.1), (2.1.2) *hold, then*

$$T^{\frac{1}{2}}(\hat{\tau}_T - \tau_0)$$

is asymptotically normally distributed with zero mean and the covariance matrix V_0^{-1} *that is discussed below.*

We shall construct the matrix V_0 in the simplest case where $\theta_0 \in V_\alpha$ and $\tilde{\Theta}$ is constructed with $U = V_\alpha$. Then the coordinate vectors τ will have as entries the free parameters in the echelon form for V_α. Order the elements in the matrix fraction description $[a(z),b(z),d(z)]$ first according to lag, then matrix type (whether $A(j)$, $B(j)$, $D(j)$) then column number and finally row number. Thus τ is the vector of those elements of

$$(\text{vec}[A(0)]', \text{vec}[A(1),B(1),D(1)]', \ldots, \text{vec}[A(p),B(p),D(p)]')'$$

that are not restricted to zero or one. Introduce

$$A(u,v) = \int_{-\pi}^{\pi} e^{i(u-v)\omega} \left\{ \begin{bmatrix} -k + \delta_{0,u} I_s \\ I_s \\ 0_{m,s} \end{bmatrix} \sharp_0 \left[-(k_0 - \delta_{0,v} I_s)^*, I_s, \right.\right.$$

$$\left.\left. 0'_{m,s} \right] \frac{d\omega}{2\pi} + \begin{bmatrix} -\ell_0 \\ 0'_{m,s} \\ I_m \end{bmatrix} dF_z(\omega) [-\ell_0^*, 0_{m,s}, I_m] \right\} \otimes (b_0 \sharp_0 b_0^*)'^{-1}$$

where $0_{m,s}$ is the $m \times s$ matrix of zeros. The rows of $A(u,v)$ correspond to the elements of $\mathrm{vec}[A(u),B(u),D(u)]$, similarly the columns correspond to elements of $\mathrm{vec}[A(v),B(v),D(v)]$. Form the matrix $V_0(u,v)$ by selecting those elements of $A(u,v)$ corresponding to the elements of $\mathrm{vec}[A(u),B(u),D(u)]$, $\mathrm{vec}[A(v),B(v),D(v)]$ included in τ. Then $V_0 = [V_0(u,v)]_{u,v=1,\ldots,p}$. In this case there are no constraints on the parameters in the matrix fraction description other than $A(0) = B(0)$ and the zero-one constraints imposed by the structure. When there are other constraints, for example a parameterisation using the canonical form for U_α in §1.3 is used, the covariance matrix is further complicated. Hannan, Dunsmuir and Deistler (1980) give the form of the covariance matrix when there are constraints that satisfy mild conditions and cover the case mentioned above. The condition (2.1.2) is not necessary for a central limit Theorem to hold. If (2.1.2) is replaced by (2.1.3), (2.1.4) and $\varepsilon(t)$ is assumed to have a finite 4^{th} moment, theorem 2.2.2 still holds but the covariance matrix V_0^{-1} is more complicated. See Hannan, Dunsmuir and Deistler (1980) for a discussion.

§2.3 Parameter estimation when the McMillan degree is known

Even when the McMillan degree n_0 of the true model is known, whether from prior information about the system or having estimated it from data, it is necessary to have a coordinate system capable of parameterising a neighbourhood of the true system θ_0 before any optimisation scheme can be used to obtain parameter estimates. Natural coordinate systems to consider are the overlapping neighbourhoods $\{U_\alpha : |\alpha| = n_0\}$. As each U_α is dense in $M(n)$ it is extremely unlikely that θ_0 (and also $\hat{\theta}_T$) will fall outside a particular U_α. What can happen is that θ_0 may be close to the boundary of U_α where $P(\theta_0)$, the covariance matrix of the state vector, becomes poorly conditioned. A covering of $M(n_0)$ by disjoint submanifolds such as $\{V_\alpha : |\alpha| = n_0\}$ could be used instead. In this situation, choosing a V_α other than the generic neighbourhood, $V(n)$ can have more serious consequences for $V_\alpha \neq V(n)$ has empty interior in the relative topology T_{pt} on $M(n)$. Consequently it is more important that the structural indices be correctly estimated so that the the estimation procedure is not restricted to a submanifold whose closure does not contain θ_0. Estimation procedures based on the echelon submanifolds, albeit in a somewhat different context, are examined in later chapters.

Let $V(\theta)$ be the matrix of second derivatives of the likelihood, $W(\theta)$ its expectation and $\lambda_{min}(A)$, $\lambda_{max}(A)$ denote the smallest and largest eigenvalues of A, respectively. Van Overbeek and Ljung (1982) have shown that $W(\theta)$ is singular at θ_1 if and only if $P(\theta_1)$ is singular. Moreover in a neighbourhood of θ_1 there exist constants

$\underline{c}$, $\bar{c}$ such that

$$0 < \underline{c} \leqslant \frac{\lambda_{min}(W(\theta))}{\lambda_{min}(P(\theta))} \leqslant \bar{c} < \infty \quad .$$

It is known that $\det W(\theta)$ remains invariant under changes of parameterisation (Wertz, Gevers and Hannan, 1982). Thus if $W(\theta)$ becomes poorly conditioned (i.e. the condition number of $W(\theta)$, defined as $\lambda_{max}(W(\theta))/\lambda_{min}(W(\theta))$, becomes large) in one coordinate system then $\lambda_{min}(W(\theta))$ is small and also $\lambda_{min}(P(\theta))$ is small. From a computational viewpoint it is preferable that $V(\theta)$ is well conditioned. Indeed, in a Newton-Raphson type of optimisation procedure the inverse of $V(\theta)$ occurs at each step, as well as the inverse of $W(\theta_0)$ giving the covariance matrix in the limiting distribution of the parameter estimates.

A number of criteria have been proposed to discriminate between coordinate systems all based in one way or another on the conditioning of $P(\theta)$. In essence, if we refer to the construction of the state vector in the representation (1.3.6), we are searching for a set of n_0 predictors

(2.3.1) $$y_i(t + a \mid t),\ a = 1,\ldots,n_i,\ i = 1,\ldots,s$$

as defined in (1.3.4) so that their covariance matrix is "best conditioned" in some appropriate sense.

Ljung and Rissanen (1976) consider a criterion based on the notion of the "complexity" of a random vector. Complexity is a measure of the degree of interaction between the components of a random n-vector $\underset{\sim}{x}$ (see van Emden, 1971)). If $P = E(\underset{\sim}{x}\ \underset{\sim}{x}')$ then complexity can be defined as

$$-\tfrac{1}{2} \sum_{1}^{n} \log(n\ \lambda_i)$$

where λ_i are the eigenvalues of $P/tr(P)$. If there are strong interactions between the components of $\underset{\sim}{x}$ (i.e. there are high correlations) or P is ill-conditioned, then complexity is high. Thus Ljung and Rissanen (1976) compute predictors $y_i(t + a \mid t,t - h)$ of $y_i(t + a)$ from the finite past $y(t - j),z(t - j)$, $j = 0,\ldots,h$, where h is sufficiently large so that the finite past predictors approximate the theoretical quantities (2.3.1) sufficiently well. Now vectors of predictors

$$y_i(t + a \mid t,t - h),\ a = 1,\ldots,n_i\ ,\ i = 1,\ldots,s$$

such that $\Sigma n_i = n_0$ are considered and the vector having least complexity defines the structural indices and hence the corresponding coordinate neighbourhood U_α . Such an approach requires the solution of a large number of eigenvalue problems though the amount of calculation at this stage would still be small compared to that needed in the parameter estimation stage. Gevers and Wertz (1981) consider an alternative characterisation of a "best conditioned" $P(\theta)$. In their approach the state vector is built up sequentially from the predictors $y_i(t+a \mid t,t-h)$. Assume that the predictors

(2.3.2) $$y_i(t + a \mid t,t - h),\ a = 1,\ldots,m_i\ ,\ i = 1,\ldots,s$$

have already been selected. Consider the predictors $y_i(t+m_j+1 \mid t,t-h)$, $i = 1,\ldots,s$ and select that one which has the largest projection onto the space orthogonal to the span of the previously selected rows (2.3.2). Continue in this way until the required n_0 elements have been selected. van Overbeek and Ljung (1982) have investigated a further adaption of these ideas with on-line applications. In an

off-line situation where the estimation procedure gives a sequence of estimates $\hat{\theta}_T(j)$, $j = 1,2,\ldots$ converging to $\hat{\theta}_T$, their approach may be thought of as follows. Begin in any coordinate neighbourhood U_α and monitor the conditioning of $P(\hat{\theta}_T(j))$ at each iteration. When $P(\hat{\theta}_T(j))$ becomes badly conditioned (as judged by the experimenter) select a new coordinate neighbourhood U_β and transform the parameter estimate from U_α to U_β by a known transformation $\phi_\beta \circ \phi_\alpha^{-1}$ defined in terms of (1.3.4) so that in U_β, $P(\hat{\theta}_T(j))$ is better conditioned. Van Overbeek and Ljung (1981,1982) give a model selection algorithm that selects a better (but not necessarily the best) conditioned coordinate system.

§2.4 Order estimation based on maximum likelihood

Assume there is a true model

$$y(t) = k_0 \varepsilon(t) + \ell_0 z(t) \ , \ E\{\varepsilon(t)\varepsilon(s)'\} = \delta_{st}\sharp_0$$

belonging to a model set $\Theta(n_0)$. Here $d(n)$ usually denotes the number of freely varying parameters describing an element of $\Theta(n)$. We shall usually assume $\Theta(n) \subset \overline{\Theta(m)}$ and $d(n) < d(m)$ when $n < m$. In this section $\Theta(n)$ will be often thought of as $M(n)$. Let $\hat{\sharp}_{n,T}$ be the estimate of $\sharp_0$ when the likelihood (2.2.1) is minimised over $\Theta(n)$. We seek to estimate n_0 as that integer $\hat{n}_{C(T)}$ minimising a criterion of the form

$$\log \det \hat{\sharp}_{n,T} + d(n)C(T)/T \tag{2.4.1}$$

over a prescribed set $0 \leqslant n \leqslant N$. For convenience we refer to the quantity n to be estimated using (2.4.1) as the

"*order*" of the model, regardless as to whether it is the McMillan degree, or the set of Kronecker indices $n = \{n_1,\ldots,n_s\}$ as is the case if $\theta(n) = V_n$, or the pair $n = (p,q)$ if $\theta(n) = \text{ARMA}(p,q)$. The upper bound N on the model orders considered will sometimes be a function increasing with the sample size when it will be denoted by $N(T)$. The choice of $C(T)$ is not specified at this stage; its selection may be guided by the results of this section. However two usual choices are $C(T) \equiv 2$, when (2.4.1) is known as AIC, and $C(T) = \log T$, which is known as BIC.

AIC was introduced in the univariate case by Akaike (1972,1974d). BIC was introduced by Akaike (1976) and Schwarz (1978) from an argument based on a Bayesian analysis. Rissanen (1976) derives a similar criterion, where the dominant terms give BIC, but contain an additional term that is essentially

$$-\frac{1}{N}\log\det W(\theta).$$

Here $W(\theta)$ is the inverse of the information matrix defined in §2.3. This term is constant for the various coordinate systems that could describe $M(n)$, but $W(\theta)$ will become singular when the true order is exceeded. The effect of this term on the order estimation procedure has not been investigated any further.

Several general comments can be made at this stage. If $n < n_0$ then with Σ_n as defined in (2.2.3)

$$\hat{\Sigma}_{n,T} \xrightarrow{a.s.} \Sigma_n \tag{2.4.2}$$

and Σ_n is bounded away from Σ_0 by a non-negative matrix (see §1.4). Hence for large enough T, $\hat{\Sigma}_{n,T}$ is also bounded away from $\hat{\Sigma}_{n_0,T}$, so that as long as $C(T)/T$ is

decreasing to zero the minimum of (2.4.1) cannot occur for $n < n_0$. This argument applies when the maximum likelihood estimator $\hat{\sigma}^2_{n,T}$ is replaced by any other consistent estimator of σ^2_n . If $C_1(T) > C_2(T)$ then $\hat{n}_{C_1(T)} < \hat{n}_{C_2(T)}$ so by selecting $C(T)$ in (2.4.1) to increase rapidly enough with $C(T)/T \to 0$, the frequency that $\hat{n}_{C(T)} > n_0$ will be decreased. However such an approach is of little value if the criterion is to be applied to data sets of a practical size, for (2.4.1) will frequently select a model having too simple a structure. It is of interest to find a $C(T)$ increasing as slowly as possible yet still giving consistent estimates of order.

Let the true model be a univariate Gaussian autoregression of finite order h_0, $\hat{\sigma}^2_{h,T}$ is the mean square error upon fitting an autoregression of order h , and $\hat{h}$ is the estimate of the order obtained when

$$AIC(h) = \log \hat{\sigma}^2_{h,T} + 2h/T \ , \ 0 \leqslant h \leqslant H$$

is minimised. Shibata (1976) has shown that this procedure will overestimate the true order with non-zero probability. Indeed

$$\Pr(\hat{h} = h) = \begin{cases} p_{h-h_0} q_{H-h} & H \geqslant h \geqslant h_0 \\ 0 & \text{otherwise} \end{cases}$$

where p_i, q_i are sequences of positive constants independent of the true model. Shibata (1980) considers a situation where a stationary Gaussian time series is generated by an autoregression of infinite order. He introduces a criterion $S_T(h)$ which is derived from the mean square error of prediction using an autoregression of order h estimated from one realisation and used to give a predictor for an

independent realisation of the same process. Thus, in calculating the mean square error the autoregressive constants can be held fixed. Shibata then defines a sequence h_T^* which optimises $S_T(h)$, $0 \leqslant h \leqslant H_T$, where H_T is a sequence that is increasing (but not increasing too fast) with T. He then shows that $\hat{h}_T$, the order selected by minimising $AIC(h)$, $0 \leqslant h \leqslant H_T$, satisfies $S_T(h_T)/S_T(h_T^*) \to 1$ in probability. There is slight doubt surrounding the meaning of this result because of the double sampling procedure used in deriving this criterion. Nevertheless, the result is important. Define, as an analogue to (2.2.3) for autoregression models,

$$\sharp_h = \inf \frac{1}{2\pi} \int_{-\pi}^{\pi} \phi_h k_0 \sharp_0 k_0^* \phi_h^*$$

where the infinium is taken over $\phi_h(z) = 1 + \sum_1^h \phi_h(j) z^{-j}$. Shibata (1980) shows that if $\sharp_h - \sharp_0 = c\rho_0^{2h}$ then

$$h_T \left(\frac{-2 \log \rho_0}{\log T} \right) \to 1 \text{ , in probability.} \tag{2.4.3}$$

This result is applicable to ARMA models $a_0 y(t) = b_0 \varepsilon(t)$ where ρ_0 is the modulus of a zero of $b_0(z)$ nearest $|z| = 1$. Then $\sharp_h - \sharp_0 = 0(\rho_0^{2h})$. However the exact form of $\sharp_h - \sharp_0$ is such that (2.4.3) applies. This and related results are discussed in §3.3. Shibata (1980), Hannan and Kavalieris (1984) obtain results similar to (2.4.3) for quite general linear models. It has already been mentioned that as $C(T)$ is increased the degree of overestimation (if any) is reduced. Hannan and Quinn (1979) show that if $C(T) \leqslant c \log \log T$, $c < 2$ then the estimates cannot be strongly consistent.

This gives a reasonably complete picture of the asymptotic behaviour of criteria of the form (2.4.1) when fitting autoregressive models to data that is truly obtained from an autoregression. Since it can hardly be expected that the true model is a finite order autoregression the fact that AIC selects a higher order model does not seem a serious problem when fitting a model for prediction purposes. Selecting a model having too low an order may be a more serious problem and the best approach may be to use AIC, especially when the sample size is small. For the purposes of spectral estimation, results using AIC or BIC are less clear. Shibata (1981) shows that if $\hat{f}_h(\omega)$ is the autoregressive spectral estimate

$$\hat{f}_h(\omega) = \hat{\sigma}^2_{h,T} \Big/ \Big|\sum_0^h \phi_h(j) e^{ij\omega}\Big|^2$$

where $\hat{\phi}_h(j)$ are the estimates of the coefficients of the fitted autoregression of order h, and $f_0(\omega)$ is the true spectral density, then the lower bound for the loss function

$$\int_{-\pi}^{\pi} \left(\frac{\hat{f}_h(\omega) - f_0(\omega)}{f_0(\omega)} \right)^2 d\omega$$

is attained when h is selected using AIC. Then the order selected by AIC could be determined largely by spectral structure in regions of low power (which may also be the regions of little interest). Then a high order model may be required to provide a good fit in the regions of interest. An interesting empirical study of a number of otherwise similar sets of EEG data by Steinberg, Gasser and Franke (1983) points to this kind of difficulty. When estimating the McMillan degree (or Kronecker indices) of ARMA (or ARMAX) processes the problem becomes more complex. This is because when the McMillan

degree, n, exceeds the true degree, n_0, the transfer function $k_0(z)$ corresponds to a subset of an affine subspace in $\mathbb{R}^{2ns}$ (Theorem 1.4.3). Then the parameter estimates need not converge as, in the limit, the likelihood is constant on that subspace. This difficulty in the theorem is overcome by introducing the alternative parameters χ_j as in §1.4 so that $\chi_j \to 0$. To prove the consistency theorems it is necessary to bound the zeros of $\det k_0(z)$ away from the unit circle;

$$\det k_0(z) \neq 0 \quad |z| > 1 - \delta, \ \delta > 0 . \tag{2.4.4}$$

As well as this it is necessary to introduce a prior bound on the parameters $\phi_\alpha(k_0(z))$ in an appropriate coordinate neighbourhood

$$\|\phi_\alpha(k_0(z))\| < K < \infty \tag{2.4.5}$$

where $\|.\|$ is the Euclidean norm. This last condition will mean that the coordinates used in the iterative solution of the likelihood equations will have to be bounded, a condition which would be imposed by machine limitations in any case. The conditions (2.4.4), (2.4.5) are needed when the maximum likelihood estimate $\hat{\ddagger}_{n,T}$ is used in (2.4.1). When the maximum likelihood estimate is replaced by an estimate based on regression procedures, as in Chapter 3, these conditions are no longer necessary.

THEOREM 2.4.1. *Assume* (2.4.4), (2.4.5) *hold for the model* $y(t) = k_0 \varepsilon(t)$ *where* $k_0(z)$ *has McMillan degree* n_0. *Let* $\hat{n}$ *be the estimate of the McMillan degree obtained by minimising* (2.4.1) *with* $\Theta(n) = M(n)$ *for* $0 \leqslant n \leqslant N$.

(i) If $\varepsilon(t)$ *is ergodic with a finite* 4^{th} *moment and* (2.1.1), (2.1.2) *hold,* $n_0 \leqslant N$ *and*

$$\liminf_{T\to\infty} C(T)/\{2 \log\log T\} > 1 \ , \ C(T)/T \to 0$$

then $\hat{n} \to n_0$ a.s.

If

$$\limsup_{T\to\infty} C(T)/\{2 \log\log T\} < 1 \ , \ n_0 < N$$

then a.s. *convergence does not hold.*

(ii) If $\varepsilon(t)$ *is ergodic,* $E\{\| \varepsilon(t) \|^{\gamma}\} < \infty$, $\gamma > 4$ *and* (2.1.1), (2.1.2) *hold,* $n_0 \leqslant N$ *and*

$$C(T)/T \to 0 \ , \ C(T) \to \infty$$

then $\hat{n} \to n_0$ *in probability.*

If $\varepsilon(t)$ *is ergodic with a finite* 4^{th} *moment and* (2.1.1), (2.1.2) *hold,* $n_0 < N$ *and*

$$\limsup_{T\to\infty} C(T) < \infty$$

then

$$\lim_{\delta\to 0} \lim_{T\to\infty} P(\hat{n} > n_0) = 1 \ .$$

These kinds of results are proved in the univariate case in Hannan (1980) and for multivariate systems in Hannan (1981). These proofs require independence of the $\varepsilon(t)$ sequence. Using the results of An, Chen and Hannan (1982) the independence requirement may be dropped. The condition (2.1.2) may be further relaxed to (2.1.3), (2.1.4) but then the expression 2 log log T in (i) of the theorem will need to be replaced by c_0 log log T , where c_0 is some unknown constant.

Again, when the true model is not ARMA, overestimation of order may not matter very much, but if the true model is ARMA then overestimation of order will lead to the intro-

duction of spurious poles and zeros. Then there is no expectation that the estimates of the ARMA parameters will converge, but as the transfer function still converges the poles and zeros will be close to cancelling. If these additional poles and zeros are away from the unit circle, the effect may be small if it is only the transfer function that is required. However, results of Hannan (1982) suggest that the additional zeros and poles due to overestimation of order may converge to the unit circle. Indeed in the case considered in the last reference, where a univariate ARMA model $y(t) + \alpha y(t-1) = \varepsilon(t) + \beta\varepsilon(t-1)$ is fitted to data that is truely white noise, the estimates $(\hat{\alpha},\hat{\beta})$ converge to the set $\{(1,1), (-1,-1)\}$. When this happens the estimates of the frequency response functions cannot be expected to be particularly good near 0 or π.

A theorem similar to Theorem 2.4.1 may be proved when estimating p_0, q_0 in the univariate model

$$\sum_0^{p_0} a_0(j)y(t-j) = \sum_0^{q_0} b_0(j)\varepsilon(t-j) \ , \ a_0(0) = b_0(0) = 1$$

and $\hat{p}$, $\hat{q}$ are the estimates obtained by minimising

$$\log \hat{\Sigma}_{p,q;T} + (p+q)C(T)/T, \ p \leqslant P \ , \ q \leqslant Q \ .$$

Here $\hat{\Sigma}_{p,q;T}$ is obtained by maximising the likelihood over the model set ARMA(p,q). This is the form of the theorem proved in Hannan (1980). These results can be extended to include the estimation of Kronecker indices. Now if $\hat{\Sigma}_{\alpha;T}$ is the maximum likelihood estimate of Σ_0 from the model set V_α, $d(\alpha)$ is the number of freely varying parameters in V_α given by (1.3.10), then the Kronecker indices may be estimated by minimising

(2.4.6) $\log \det \hat{\Sigma}_{\alpha;T} + d(\alpha)C(T)/T \ , \ |\alpha| \leqslant N \ .$

Though such theoretical results are available, it remains a formidable task to compute the large number of maximum likelihood estimates required. Of course, we do not need to compute $\hat{\Sigma}_{\alpha;T}$ for every α, $|\alpha| \leqslant N$. Indeed if $\hat{p}$ minimises (2.4.6) over $V(ps)$, $p = 1,2,\ldots,N$ then $\hat{p}$ is an estimate of the largest Kronecker index. Then a small number of the V_α need be examined to determine the remaining Kronecker indices. Such search strategies will be developed in §3.4. Nevertheless all of these methods require the time consuming solution of the likelihood equations. Consequently we seek an estimate of the prediction variance that is cheaply obtained and whose use in the place of the maximum likelihood estimate $\hat{\Sigma}_{\alpha;T}$ in (2.4.1) still leads to consistent estimates of the model order under reasonable conditions. This is the subject of the remainder of this thesis.

CHAPTER 3

ESTIMATION OF ARMAX MODELS

§3.1 Introduction

Because of the non-linear nature of the likelihood equations the computational burden in obtaining the maximum likelihood estimate $\hat{\xi}_{n;T}$ makes the use of the criterion (2.4.1) to identify model structure quite unattractive as $\hat{\xi}_{n;T}$ must be computed for each model set $\Theta(n)$ considered. For this reason we replace the maximum likelihood estimate $\hat{\xi}_{n;T}$ by an estimate (which we denote by the same symbol as the maximum likelihood estimate) that is obtained using linear regression methods. The procedure considered below was first introduced by Durbin (1961) to estimate the parameters of an ARMA model when the order is known. There are three stages broadly described here and investigated in detail in §§3.3, 3.4, 3.5.

Stage I Obtain an estimate of the innovations sequence by fitting a regression-autoregression model of sufficiently high order h to the data $y(t)$, $z(t)$, $t = 1,\ldots,T$. Thus we obtain estimates $\hat{\phi}_h(j)$, $\hat{\psi}_h(j)$, $j = 1,\ldots,h$ from the expression

$$\sum_0^h \phi_h(j)y(t-j) + \sum_1^h \psi_h(j)z(t-j) = e(t) \ , \ \phi_h(0) = I_s$$

so as to minimise $\mathrm{tr}\{\frac{1}{T}\Sigma e(t)e(t)'\}$. The order h is selected using a criterion such as BIC. Then the estimate of the innovations sequence is

$$\hat{\varepsilon}_h(t) = \sum_0^h \hat{\phi}_h(j)y(t-j) + \sum_1^h \hat{\psi}_h(t)z(t-j) \ .$$

State II Using the estimates $\hat{\varepsilon}_h(t)$ from Stage I, fit an ARMAX model using regression methods. For example, suppose

we fit an ARMAX model in echelon form for $\gamma(ps+q)$ (described in (1.3.11)). Then we perform the following regressions:

(i) for $i = 1,\ldots,q$ regress $y_i(t)$ on $y_j(t-1)$, $j = 1,\ldots,q$ and all components of $z(t-1)$, $\hat{\varepsilon}_h(t-1)$, $y(t-j)$, $z(t-j)$, $\hat{\varepsilon}_h(t-j)$, $j = 2,\ldots,p+1$.

(ii) for $i = q+1,\ldots,s$ regress $y_i(t)$ on $y_j(t) - \hat{\varepsilon}_{h,j}(t)$, $j = 1,\ldots,q$ and all components of $y(t-j)$, $z(t-j)$, $\hat{\varepsilon}_h(t-j)$, $j = 1,\ldots,p$.

The regressors $y_j(t) - \hat{\varepsilon}_{h,j}(t)$ occur in (ii) because $A(0) = B(0)$. For other echelon submanifolds the regressions are, in principle, similar. The residual variances from these regressions are used in (2.4.1) to estimate the model structure. Stage II also provides us with estimates, albeit not efficient, of the system and covariance parameters.

Stage III Now that the model structure has been estimated the inefficient parameter estimates from Stage II are used to initialise a maximum likelihood procedure to obtain asymptotically efficient estimates.

The essential part of the structure selection algorithm in Stage II relies only on linear optimisation techniques, so there is a great saving in computation time. Any advantage over using maximum likelihood estimates in the structure selection procedure would be lost if we were to allow non-linear constraints amongst the parameters of the ARMAX form at Stage II, for we would be forced to return to an iterative optimisation procedure. Thus this procedure is not applied to parameterisations for the overlapping neighbourhoods U_α. In some situations the calculations for Stage II can be arranged in a recursive way. Thus the parameter estimates for $\theta(n)$ are updated to give estimates for $\theta(m)$, $m > n$,

giving some savings in computation time.

For univariate models similar three stage procedures have been considered by a number of authors. Hannan and Rissanen (1982) have considered a recursive implementation of this algorithm. Here the order of an ARMA model is selected using BIC at Stage II, while at Stage I the order h is taken to be an increasing function of the sample size. The rate of increase is required to satisfy

$$(3.1.1) \qquad d \log T \leqslant h \leqslant (\log T)^a \ , \ a < \infty \ ,$$

where d is a constant. It is shown in §3.3 that the order estimate at Stage I obtained using AIC or BIC is, asymptotically, $d \log T$ where d is a constant determined by the true system parameters. Mayne and Firoozan (1982) consider such a procedure for the estimation of ARMA parameters when the ARMA order is known and h is fixed. They show the asymptotic bias of the estimates of the ARMA parameters decreases to zero as $h \to \infty$. Hannan and Kavalieris (1983b) have proved the consistency of the parameter estimates when h is chosen to satisfy (3.1.1). The algorithm has also appeared in an on-line form in Astrom and Mayne (1982) where the orders at Stages I and II are fixed and the parameter estimates are updated as new data comes to hand using a recursive least squares algorithm. For a truly on-line procedure the Stage I order must remain bounded so that the calculations can be done in a fixed amount of time. If the order of the regression-autoregression model at Stage I is bounded, there is an asymptotic bias in the parameter estimates at Stage II, so that to obtain consistent estimates of the parameters at Stage II h must be increasing to infinity.

The theorems proved in the following sections are

based on the assumption that $y(t)$ is generated by an ARMAX model, an assumption that could rarely be expected to hold. It still makes sense to fit an ARMAX model to data that is not generated by such a model since a sufficiently high order ARMAX model will approximate the true model to any desired accuracy. Our later results give only a guide to what may happen if the true structure is closely approximated by an ARMAX model. Precise results are not yet available. The results of §3.3 relating to the fitting of regression-autoregression models to data that is generated by an ARMAX process can be extended to regressive-autoregressive approximation of more general linear processes. Results of this kind are due to Shibata (1980), but also see Hannan and Kavalieris (to appear).

In proving the later theorems concerning the above three stage procedure we impose the conditions needed for the consistency of order estimates in Theorem 2.4.1 except that we no longer require (2.4.3), (2.4.4). Thus the zeros of $k(z)$ need not be bounded away from $|z| = 1$. Since we are basing our results on regression techniques, where the number of regressors is increasing as H_T (to be defined later), we shall need conditions on the input processes to ensure that the sample covariances of the $y(t)$ and $z(t)$ converge to their expected values at a rate that is uniform for H_T lags. We impose

$$(3.1.2) \qquad \sup_{0 \leqslant v \leqslant H_T} \left\| \frac{1}{T} \sum_{1}^{T} z(t) z(t+v)' - \gamma_z(v) \right\| = c_T H_T^{-1}, \quad H_T \uparrow \infty, \quad c_T \downarrow 0.$$

Here $\gamma_z(v) = E\{z(t)z(t+v)'\}$ as in (2.1.5). If $z(t)$ is itself generated by a linear process, then with quite mild conditions we may take

$$c_T = O(H_T(\log T/T)^{\frac{1}{2}})$$

(An, Chen and Hannan, 1982) but rates not nearly as strict are needed in the proofs. In §3.3 we only need $H_T \leqslant (\log T)^a$, $a < \infty$. Then if $z(t)$ is a stationary process with a fourth order spectral density that is absolutely continuous we may take $c_T = O(T^{-b}H_T^2)$ for any $b < \frac{1}{2}$ (Hannan, 1978). This rate of decrease is again much faster than is required. Thus (3.1.2) is not a strong condition. We shall always assume that

$$E\{\| z(t) \|^4\} < \infty \tag{3.1.3}$$

We also assume that $z(t)$ has an absolutely continuous spectral density $f_z(\omega)$ and

$$c_1 I_m \leqslant f_z(\omega) \leqslant c_2 I_m \ , \ 0 < c_1 \leqslant c_2 < \infty \ . \tag{3.1.4}$$

The only meaningful departure from the absolute continuity of the spectral density is that the spectrum should jump at a countable number of frequencies ω_j. Then $z(t)$ has components of the form

$$\zeta(t) = \sum_{-\infty}^{\infty} r(j)e^{it\omega_j}, \ \omega_{-j} = -\omega_j, \ r(-j) = \overline{r(j)}, \ \sum_{-\infty}^{\infty}\| r(j) \|^2 < \infty$$

and, if $ay(t) = b\varepsilon(t) + dz(t)$, then

$$a(y(t) - a^{-1}d\zeta(t)) = b\varepsilon(t) + d(z(t) - \zeta(t))$$

and $a^{-1}d\zeta(t) = \sum_{-\infty}^{\infty} s(j)e^{it\omega_j}$. If there are finitely many frequencies (ω_j), we may remove their effect from $y(t)$ and $z(t)$ by regression. The regression procedure is asymptotically as efficient as weighted least squares (see Hannan, 1970, Chapter VII). The use of the residuals will not affect the rates of convergence for the covariances of $y(t)$, $z(t)$ below. The consequence of (3.1.4) is to bound the smallest

eigenvalue of $\Gamma_{z,h} = [\gamma_z(i-j)]_{i,j=1,\ldots,h}$ away from zero and to bound its largest eigenvalue. Both bounds are uniform in h. When $z(t)$ is stationary with zero mean and finite mean square, we can write

$$z(t) = \sum_0^\infty K_z(j)\varepsilon_z(t-j);\ \sum_0^\infty \|K_z(j)\|^2 < \infty,\quad E\{\varepsilon_z(t)\varepsilon_z(s)'\} = \delta_{st}\Sigma_z.$$

When $f_z(\omega)$ is bounded away from zero, we also have

$$\{\sum_0^\infty K_z(j)z^{-j}\}^{-1} = \sum_0^\infty \Lambda(j)z^{-j},\ \sum_0^\infty \|\Lambda(j)\|^2 < \infty .$$

What will sometimes be needed is a more rapid rate of decrease of the coefficients, namely

$$\lim_{t\to\infty} t^{\frac{1}{2}} \sum_t^\infty \|\Lambda(j)\| = 0 . \tag{3.1.5}$$

The theorems in the remaining sections will be proved under one of the following sets of conditions on the innovations and inputs.

Conditions A *Assume* $\varepsilon(t)$ *is a stationary and ergodic sequence that has a finite* 4^{th} *moment and satisfies* (2.1.1), (2.1.3). *Though* $z(t)$ *is observed and considered to be a fixed sequence, it is assumed to be generated by a stationary and ergodic process satisfying* (1.2.1), (3.1.2) *for* $H_T \leq (\log T)^a$, $a < \infty$, (3.1.3), (3.1.4), (3.1.5).

Conditions B *Include all the conditions A and strengthen the conditions on* $\varepsilon(t)$ *to include* (2.1.2) *(which implies* (2.1.3)*),* $E\{\|\varepsilon(t)\|^\gamma\} < \infty$, $\gamma > 6$.

We define notation which will be adhered to in the remaining chapters. We shall need the following kinds of covariance matrices:

$$\hat{\gamma}_y(k) = \frac{1}{T}\sum_1^{T-k} y(t)y(t+k)';\ \hat{\gamma}_y(-k) = \hat{\gamma}_y(k)',\ k \geq 0$$

$$\gamma_y(k) = \lim_{T\to\infty} \hat{\gamma}_y(k),$$

$$\overset{\circ}{\gamma}_y(k) = \sum_0^\infty K(j) \overset{\circ}{\sharp} K(j+k)' + \sum_{i,j=1}^{\infty} L(i)\gamma_z(k+i-j)L(j)'$$

where $K(\ell) = 0$ if $\ell < 0$, and $\overset{\circ}{\sharp} = \frac{1}{T}\Sigma\varepsilon(t)\varepsilon(t)'$,

$$\tilde{\gamma}_y(i,j) = \frac{1}{T}\sum_1^T y(t-i)y(t-j)' .$$

Similarly we define covariances $\hat{\gamma}_z(k)$, $\hat{\gamma}_{zy}(k)$, etc with the exception that $\overset{\circ}{\gamma}_z(k) = \gamma_z(k)$, $\overset{\circ}{\gamma}_{yz}(k) = \gamma_{yz}(k)$, $\overset{\circ}{\gamma}_{zy}(k) = \gamma_{zy}(k)$. Let $\hat{\Gamma}_{y,h}(a)$ be a square matrix of $h \times h$ blocks, the (i,j) block being $\hat{\gamma}_y(i-j+a)$. Similarly we define $h \times h$ matrices of blocks for the other covariances defined above, $\overset{\circ}{\Gamma}_{y,h}(a)$ having entries $\overset{\circ}{\gamma}_y(k)$ etc.

Define matrices

$$\hat{\Gamma}_h(a) = \begin{bmatrix} \hat{\Gamma}_{y,h}(a) & \hat{\Gamma}_{yz,h}(a) \\ \hat{\Gamma}_{zy,h}(a) & \hat{\Gamma}_{z,h}(a) \end{bmatrix}$$

$$\overset{\circ}{\Gamma}_h(a) = \begin{bmatrix} \overset{\circ}{\Gamma}_{y,h}(a) & \Gamma_{yz,h}(a) \\ \Gamma_{zy,h}(a) & \Gamma_{z,h}(a) \end{bmatrix}$$

and so on for $\tilde{\Gamma}_h(a)$ and $\Gamma_h(a)$. It is convenient to put $\tilde{\Gamma}_{y,h} = \tilde{\Gamma}_{y,h}(0)$, $\Gamma_h = \Gamma_h(0)$, etc.

Let $y(t,h) = (y(t-1)', y(t-2)', \ldots, y(t-h)')'$, with corresponding meanings for $z(t,h)$, $\varepsilon(t,h)$.

The simulation results reported in this chapter were obtained on a UNIVAC 1100/82 machine. The simulated data was generated using pseudo normal innovations from the subroutine G05DDF in the NAG library. All outputs and innovations were initialised to zero and the first 100 outputs discarded from each data set to avoid the effects of this initialisation.

§3.2 Recursive parameter estimation for autoregressive processes

Suppose the true model is

$$\sum_1^\infty a(j)w(t-j) = \varepsilon(t), \quad a(0) = I_s \ , \ E\{\varepsilon(t)\varepsilon(t)'\} = \Sigma_0$$

and the $\varepsilon(t)$ satisfy conditions A or B. Let

$$a_p(z) = \sum_0^p a_p(j)z^{-j}, \ k_0(z) = \sum_0^\infty a(j)z^{-j} \ ,$$

$a_p(0) = a(0) = I_s$ and assume $\det\{k_0(z)\} \neq 0 \quad |z| \geqslant 1$.
Then to minimise

$$(3.2.1) \qquad \frac{1}{2\pi}\int_{-\pi}^{\pi} a_p k_0^{-1} \Sigma k_0^{-1*} a_p^* d\omega$$

we are led to the system of equations

$$(3.2.2) \qquad \sum_0^p a_p(j)\gamma_w(j-k) = \delta_{0k}\Sigma_p \qquad k \geqslant 0$$

where Σ_p is the minimum of (3.2.1) and $\gamma_w(j) = E\{w(t)w(t+j)'\}$.
Using estimates $G(j) = \frac{1}{T}\sum_1^{T-j} w(t)w(t+j)$, $j \geqslant 0$, $G(-j) = G(j)'$ for the $\gamma_w(j)$ these equations give estimates $\hat{a}_p(j)$, $j = 1,\ldots,p$ and S_p , the Yule-Walker estimates of $a_p(j)$, Σ_p . When $w(t)$ is scalar the Yule-Walker estimates can be obtained in a recursive way using the following algorithm, which we refer to as the Levinson-Durbin recursion.

Algorithm 3.2.1 We have available as inputs $G(j)$, $j = 0,1,\ldots$
Initialise recursion with $\hat{a}_p(0) = 1$, $S_0 = G(0)$.
At the p^{th} recursion compute

$$\hat{a}_p(p) = -\sum_0^{p-1} \hat{a}_{p-1}(j)G(p-j)/S_{p-1}$$

$$\hat{a}_p(j) = \hat{a}_{p-1}(j) + \hat{a}_p(p)\hat{a}_{p-1}(p-j) \ , \ j = 1,\ldots,p-1$$

$$S_p = (1 - \hat{a}_p(p)^2)S_{p-1} .$$

Discarding the uninteresting case when $\hat{a}_p(p) = \pm 1$ (hence $S_p = 0$ so that $w(t)$ is generated by a finite trigonometric polynomial) we have $|\hat{a}_p(p)| < 1$, $p = 1,2,\ldots$. From this we can show that $\hat{k}_p(z) = \sum_0^p \hat{a}_p(j)z^{-j}$ has all of its zeros within $|z| = 1$. This algorithm is computationally cheap using $O(p^2)$ arithmetic operations and $O(p)$ storage compared to $O(p^3)$ arithmetic operations and $O(p^2)$ storage when Choleski factorisation is used to solve (3.2.2). The numerical stability of this recursion is shown to be comparable to the stability of the Choleski factorisation algorithm (Cybenko, 1980).

When the true model is a finite order autoregression the estimates $G(j)$ in (3.2.2) give rise to asymptotically efficient parameter estimates, however for moderate sample sizes they may lead to a substantial bias. Tjostheim and Paulson (1983) obtain explicit formulae for the large sample bias of the Yule-Walker estimates in some simple cases and show, both by simulations and theory, that the bias is indeed large and the Yule-Walker estimates perform poorly compared to least squares. Alternative procedures for solving (3.2.2) when G_p is not Toeplitz (for example $G_p = [G(j,k)]_{j,k=1,\ldots,p}$, $G(j,k) = \frac{1}{T}\sum_p^T \omega(t-j)\omega(t-k)'$ which gives rise to the least squares estimates) can be constructed using the lattice filter (see the survey by Friedlander (1982) for some details). The lattice filter retains the parameter updating equation

$$\hat{a}_{p+1}(j) = \hat{a}_p(j) + K_p\, \hat{a}_p(p-j)$$

but now the K_p are computed in a different way. It has the

advantage that it can be adapted to on-line computation; where the parameter estimates are updated as new data comes to hand. There is no guarantee that the generating function $\hat{k}_p(z)$ obtained using the lattice filter will be stable except in the Toeplitz case when the K_p coincide with $\hat{a}_{p+1}(p+1)$ obtained in the Levinson-Durbin recursion. If it is considered that a stable generating function is desirable, a number of suboptimal solutions may be obtained having this property for $s = 1$ (Makhoul, 1981), among them being the maximum entropy estimate of Burg (1975).

The analogue of the Levinson-Durbin recursion in the multivariate case is due to Whittle (1963). In the algorithm that follows, the regression coefficients $F_p(j)$ correspond to the model

$$\sum_0^p F_p(j)w(t-j) = e_p(t),\ F_p(0) = I\ ,$$

and S_p is the residual mean square for this regression. The coefficients $\tilde{F}_p(j)$ correspond to the 'backwards' regression of $w(t-p-1)$ on $w(t-1),\ldots,w(t-p)$ in the model

$$\sum_0^p \tilde{F}_p(j)w(t-p-1+j) = \tilde{e}_p(t),\ \tilde{F}_p(0) = I$$

where $\tilde{S}_p$ is the corresponding residual mean square.

Algorithm 3.2.2 We have available as input $G(j)$, $j = 0,1,\ldots$

Initialise the recursion with $F_p(0) = \tilde{F}_p(0) = I$,

$S_0 = \tilde{S}_0 = G(0)$.

At the p^{th} recursion compute

$$\Delta = \sum_0^p F_p(j)G(p+1-j)'$$

$$F_{p+1}(p+1) = -\Delta\ \tilde{S}_p^{-1}\ ,\ \tilde{F}_{p+1}(p+1) = -\Delta'\ S_p^{-1}$$

$$F_{p+1}(j) = F_p(j) + F_{p+1}(p+1)\tilde{F}(p+1-j)$$

$$\tilde{F}_{p+1}(j) = \tilde{F}_p(j) + \tilde{F}_{p+1}(p+1)F(p+1-j)$$

$$S_{p+1} = \{I - F_{p+1}(p+1)\tilde{F}_{p+1}(p+1)\}S_p$$

$$\tilde{S}_{p+1} = \{I - \tilde{F}_{p+1}(p+1)F_{p+1}(p+1)\}\tilde{S}_p$$

As an analogue of the scalar case, we have

$$\det\{\sum_0^p F_p(j)z^{-j}\} \neq 0 \quad |z| \leq 1 .$$

This kind of recursive algorithm can be derived for more general Toeplitz estimation problems. For example Monden, Arimoto and Yamada (1982) consider fitting a regression-auto-regression model

$$(3.2.3) \quad \sum_0^p a_{pq}(j)y(t-j) + \sum_1^q b_{pq}(j)z(t-j) = e_{pq}(t) \ , \ a_{pq}(0) = 1$$

recursively, by an algorithm to add an extra variable (either $y(t-p-1)$ or $z(t-q-1)$) into the Toeplitz regression. The implementation of this algorithm is only slightly more complicated than the Levinson-Durbin recursion. To update (3.2.3) from order (p,q) to $(p+1,q)$ requires $O((p+q)^2)$ arithmetic operations and $O(p+q)$ storage. We leave the details to the last reference where a Pascal program is also given for the update from (p,q) to $(p+1,q)$. The order update from (p,q) to $(p,q+1)$ is accomplished using the same algorithm and only interchanging the terms $y(t)$ and $z(t)$ in (3.2.3).

§3.3 Stage I

The first stage of the procedure outlined in §3.1 is to obtain estimates of the innovations by fitting a regression-autoregression model of sufficiently high order to the data. This is conveniently done using the Levinson-Durbin recursion. As input to Algorithm 3.2.2 define

$$G(k) = \begin{bmatrix} \hat{\gamma}_y(k) & \hat{\gamma}_{yz}(k) \\ \hat{\gamma}_{zy}(k) & \hat{\gamma}_z(k) \end{bmatrix} .$$

From the p^{th} stage of Algorithm 3.2.2 we obtain outputs $F_p(j)$, $j = 1,\ldots,p$ (the $(s+m) \times (s+m)$ matrices of regression coefficients for the regression of $(y(t)',z(t)')'$ on $(y(t-j)',z(t-j)')'$, $j = 1,\ldots,p$) and the residual mean square S_p for the regression. Define the $s \times (s+m)$ matrices $[\hat{\phi}_p(j),\hat{\psi}_p(j)]$, $j = 1,\ldots,p$ as the first s rows of the $F_p(j)$. Then

$$y(t) + \sum_1^p \{\hat{\phi}_p(j)y(t-j) + \hat{\psi}_p(j)z(t-j)\} = \hat{\varepsilon}_p(t) \tag{3.3.1}$$

and the residual mean square,

$$\hat{\Sigma}_p = \frac{1}{T} \sum_1^T \hat{\varepsilon}_p(t)\hat{\varepsilon}_p(t)',$$

is the $s \times s$ submatrix consisting of the elements in the first s rows and columns of S_p. (We drop the second subscript used in §2.4, thus $\hat{\Sigma}_p = \hat{\Sigma}_{p;T}$.) To estimate the order of the model (3.3.1) we use a criterion such as AIC or BIC. Here we suggest selecting the order $\hat{h}$ of (3.3.1) as that integer minimising

$$AIC(h) = \log \det \hat{\Sigma}_h + 2hs(s+m)/T \qquad 0 \leqslant h \leqslant H_T .$$

Practical experience may indicate that a regression-autoregression model of higher or lower order than $\hat{h}$ may be better suited to particular kinds of data. For the purpose of proving theorems concerning the procedure, H_T will be taken to increase to infinity at a preassigned rate (for example as $(\log T)^a$, $a > 1$). Such a bound may be of little use in practice when the experimenter will only consider a sufficient number of values of h beyond a local minimum of $AIC(h)$ to be confident a global mimimum has indeed been found.

The use of the Levinson-Durbin recursion, as above, entails the computation of the coefficients for the regression of $z(t)$ on $y(t-j)$, $z(t-j)$, $j = 1,\ldots,p$. This suggests there may be a cheaper algorithm that yields just the required regression of $y(t)$ on $y(t-j)$, $z(t-j)$, $j = 1,\ldots,p$. However such an algorithm does not appear to be available. The recursive algorithm suggested by Monden, Arimoto and Yamada (1982) is for this kind of regression, albeit for scalar input and output, but also it requires the regression of $z(t)$ on $y(t-j)$, $z(t-j)$, $j = 1,\ldots,p$ as before and does not save computation time over the Levinson-Durbin recursion. We leave simulation results to the end of this section where they can be better discussed in the context of the theorems.

As a preparation for the asymptotic results to follow we give some properties of ARMAX systems. The functions

$$[\phi(z),\psi(z)] = \sum_0^\infty [\phi(j),\psi(j)]z^{-j}$$

$$= b_0(z)^{-1}[a_0(z),d_0(z)]$$

$$= \{\det b_0(z)\}^{-1}p(z)$$

where $a_0(z^{-1})$, $b_0(z^{-1})$, $d_0(z^{-1})$ are matrices of polynomials.

If ρ_0 is the modulus of a zero of $\det b_0(z)$ nearest $|z| = 1$. From our assumptions (1.1.2) $\rho_0 < 1$. By taking a partial fraction expansion of $\{\det b(z)\}^{-1}$ it can be seen that the coefficients $\phi(j), \psi(j)$ decrease geometrically at a rate determined by ρ_0. The property of the coefficients that is important later is

(3.3.2)
(i) $\limsup_{j\to\infty} \rho_0^{j} \|\phi(j),\psi(j)\| = c;\ 0 < c < \infty$

(ii) there exists an integer N so that for some sequence n_j such that $n_{j+1}-n_j < N$,
$\lim_{j\to\infty} \rho_0^{n_j} \|\phi(n_j),\psi(n_j)\| = c$.

As a consequence,

$$\limsup_{h\to\infty} \rho_0^{h} \sum_{h+1}^{\infty} \|\phi(j),\psi(j)\| \leqslant c_1,$$

$$\liminf_{j\to\infty} \rho_0^{n_j+1} \sum_{n_j}^{\infty} \|\phi(j),\psi(j)\| \geqslant c_2 > 0.$$

The last inequality in conjunction with (3.3.2 ii) gives

$$\liminf_{h\to\infty} \rho_0^{h} \sum_{h+1}^{\infty} \|\phi(j),\psi(j)\| \geqslant c_2\rho_0^{N} = c_3.$$

We define the quantities $\phi_h(j)$, $\psi_h(j)$ as the coefficients of $p(z)$, $q(z)$ that minimise

$$(3.3.3)\qquad \operatorname{tr}\left\{\frac{1}{2\pi}\int_{-\pi}^{\pi} pk_0 \sharp_0 k_0^{*} p_0^{*} d\omega + \int_{-\pi}^{\pi} (q+p\ell_0)\,dF_z(\omega)\,(q+p\ell)^{*}\right\}$$

where the minimum is taken over

$$p(z) = I_s + \sum_{1}^{h} p(j)z^{-j},\quad q(z) = \sum_{1}^{h} q(j)z^{-j}.$$

Similarly we define $\mathring{\phi}_h(j), \mathring{\psi}_h(j)$ by replacing $\sharp_0$ by $\mathring{\sharp} = \frac{1}{T}\sum_{1}^{T} \varepsilon(t)\varepsilon(t)'$ in (3.3.3). These quantities arise because we do not assume (2.1.2)in Theorem 3.3.1 (i.e. we do not require

$E\{\varepsilon(t)\varepsilon(t)' \mid \mathcal{F}_{t-1}\} = \Sigma_0$). As a result the most we can say about the $\hat{\gamma}_y(\ell)$ is that $\hat{\gamma}_y(\ell) - \gamma_y(\ell) = o(1)$ uniformly in ℓ. Denote $[\hat{\phi}_h(1),\ldots,\hat{\phi}_h(h),\hat{\psi}_h(1),\ldots,\hat{\psi}_h(h)]$ by $[\hat{\phi}_h,\hat{\psi}_h]$, with similar meanings for $[\overset{\circ}{\phi}_h,\overset{\circ}{\psi}_h]$, $[\phi_h,\psi_h]$. The infinite matrix $[\phi(1),\phi(2),\ldots,\psi(1),\psi(2),\ldots]$ is similarly denoted by $[\phi,\psi]$. We shall abuse this notation when we consider expressions such as $[\phi,\psi]\Gamma_h$ by which we mean $[\phi(1),\ldots,\phi(h),\psi(1),\ldots,\psi(h)]\Gamma_h$; the dimension of the matrix Γ_h will indicate which part of $[\phi,\psi]$ is intended. Similarly $([\hat{\phi}_h,\hat{\psi}_h] - [\phi,\psi])\Gamma_\infty$ will mean

$$[\hat{\phi}_h(1)-\phi(1),\ldots,\hat{\phi}_h(h)-\phi(h),-\phi(h+1),\ldots,\hat{\psi}_h(1)-\psi(1),\ldots,-\psi(h+1),\ldots]\Gamma_\infty ,$$

the matrix $[\hat{\phi}_h,\hat{\psi}_h]$ being filled out with zeros to the required size.

Our first result is of theoretical interest since the quantities $\overset{\circ}{\phi}_h(j),\overset{\circ}{\psi}_h(j)$ cannot be computed. Note too that if (2.1.2) is assumed, or we assume $\overset{\circ}{\Sigma} - \Sigma_0 = o(H_T^{-1})$, (the latter being a very weak assumption) then the $\overset{\circ}{\phi}_h(j),\overset{\circ}{\psi}_h(j)$ can be replaced by $\phi_h(j),\psi_h(j)$ in this theorem.

THEOREM 3.3.1. *Assume* $y(t)$ *is generated by an* ARMAX *process and* $z(t)$, $\varepsilon(t)$ *satisfy conditions A. If* $h \leqslant H_T$ *then, uniformly for* $i \leqslant h$

$$\sum_1^h [\hat{\phi}_h(j)-\overset{\circ}{\phi}_h(j)]\hat{\gamma}_y(j-i) + \sum_1^h [\hat{\psi}_h(j)-\overset{\circ}{\psi}_h(j)]\,\hat{\gamma}_{zy}(j-i)$$

$$= -\frac{1}{T}\sum_1^T \varepsilon(t)y(t-i)' + o\Big(h^{-1}\sum_{h+1}^{\infty} \|\phi(j),\psi(j)\|\Big) + o(T^{-\frac{1}{2}}), \text{ a.s.}$$

$$\sum_1^h [\hat{\phi}_h(j)-\overset{\circ}{\phi}_h(j)]\hat{\gamma}_{yz}(j-i) + \sum_1^h [\hat{\psi}_h(j)-\overset{\circ}{\psi}_h(j)]\hat{\gamma}_z(j-i)$$

$$= -\frac{1}{T}\sum_1^T \varepsilon(t)z(t-i)' + o\Big(h^{-1}\sum_{h+1}^{\infty} \|\phi(j),\psi(j)\|\Big) + o(T^{-\frac{1}{2}}), \text{ a.s.}$$

As the coefficients $\phi(j),\psi(j)$ are decreasing at a geometric rate,

$\sum_{h+1}^{\infty} \| \phi(j),\psi(j)\| = o(T^{-\frac{1}{2}})$ if $h \geqslant d \log T$, $d = 1/(-2 \log \rho_0)$ as defined in (3.3.2). In this case we have a rate of convergence for $\hat{\phi}_h(j),\hat{\psi}_h(j)$ equal to that in the law of the iterated logarithm. This result may be obtained by the more direct argument used in An, Chen and Hannan (1982), p. 934. However here we are concerned more with the explicit form for $\hat{\phi}_h(j)-\mathring{\phi}_h(j)$, $\hat{\psi}_h(j)-\mathring{\psi}_h(j)$ that is obtained in Theorem 3.3.1.

Corollary 3.3.2. *Assume* $y(t)$ *is generated by an ARMAX process and* $z(t)$, $\varepsilon(t)$ *satisfy conditions A. If*

$$d \log T \leqslant h \leqslant H_T;\ d > 1/(-2 \log \rho_0)$$

then

$$\sup_{j \leqslant h} \{ \| \hat{\phi}_h(j)-\phi(j) \| , \| \hat{\psi}_h(j)-\psi(j)\| \} = O(Q_T), \text{a.s.}$$

We turn to the estimation of the order h of the regression-autoregression model based on either of the criteria

$$AIC(h) = \log \det \hat{\Sigma}_h + 2hs(s+m)/T$$

$$BIC(h) = \log \det \hat{\Sigma}_h + 2hs(s+m) \log T/T .$$

For this purpose we evaluate $\hat{\Sigma}_h = \frac{1}{T} \sum_1^T \hat{\varepsilon}_h(t)\hat{\varepsilon}_h(t)'$.

Lemma 3.3.3. *Let* $y(t)$ *be generated by an ARMAX process and* $\varepsilon(t)$, $z(t)$ *satisfy conditions A. If* $h \leqslant H_T$ *then, almost surely,*

$$\hat{\Sigma}_h = \frac{1}{T} \sum_1^T \varepsilon(t)\varepsilon(t)' +$$

$$\{[\mathring{\phi}_h,\mathring{\psi}_h]-[\phi,\psi]\}\mathring{\Gamma}_\infty\{[\mathring{\phi}_h,\mathring{\psi}_h]-[\phi,\psi]\}' \ (1+o(1))$$

$$- \{\tfrac{1}{T}\Sigma\varepsilon(t)[y(t,h)', z(t,h)']\}\hat{\Gamma}_h^{-1}\left\{\tfrac{1}{T}\Sigma \begin{bmatrix} y(t,h) \\ z(t,h) \end{bmatrix} \varepsilon(t)'\right\} + o(\tfrac{h}{T}) .$$

The expression $\{[\mathring{\phi}_h,\mathring{\psi}_h]-[\phi,\psi]\}\mathring{\Gamma}_\infty\{[\mathring{\phi}_h,\mathring{\psi}_h]-[\phi,\psi]\}'$ can be thought of as the bias in $\hat{\Sigma}_h$ due to using a regression-autoregression model of order h. It is the limiting form of

$$\frac{1}{T}\sum_1^T \{\varepsilon_h(t)-\varepsilon(t)\}\{\varepsilon_h(t)-\varepsilon(t)\}' ,$$

where $\varepsilon_h(t) = y(t) + \sum_1^h \phi_h(j)y(t-j) + \sum_1^h \psi_h(j)z(t-j)$. There is a simple bound for the third term of $\hat{\Sigma}_h$. As each element of $\frac{1}{T}\Sigma\varepsilon(t)[y(t,h)'z(t,h)']$ is $O(Q_T)$ a.s. and $|\hat{\Gamma}_h^{-1}|_1<\infty$ (lemma 4.2.2) we have that

$$(3.3.4)\quad \{\frac{1}{T}\Sigma\varepsilon(t)[y(t,h)',z(t,h)']\}\hat{\Gamma}_h^{-1}\left\{\frac{1}{T}\Sigma\begin{bmatrix} y(t,h) \\ z(t,h)\end{bmatrix}\varepsilon(t)'\right\} =$$

$$o(hQ_T^2) \text{ a.s.}$$

This bound is adequate to anlayse order estimates obtained using BIC. Define $\ell_B(T)$ as the integer that minimises

$$(3.3.5)\quad \mathrm{tr}\{\Sigma_0^{-1}([\phi_h,\psi_h]-[\phi,\psi])\Gamma_\infty([\phi_h,\psi_h]-[\phi,\psi])'\} + hs(s+m)\log T/T.$$

We have $0 < c_{min}(\sum_{h+1}^\infty \|\phi(j),\psi(j)\|^2)$

$$\leqslant \|([\phi_h,\psi_h]-[\phi,\psi])\Gamma_\infty([\phi_h,\psi_h]-[\phi,\psi])'\|$$

$$\leqslant c_{max}(\sum_1^\infty \|\phi_h(j)-\phi(j),\psi_h(j)-\psi(j)\|^2)$$

$$\leqslant c(\sum_{h+1}^\infty \|\phi(j),\psi(j)\|^2)$$

where c_{min}, c_{max} are the smallest and largest eigenvalues of Γ_∞. The last inequality is a consequence of (4.2.13). When the true model is ARMAX, the discussion below (3.3.2), together with the above bounds, shows that we can evaluate $\ell_B(T)$ by minimising

$$c_h\,\rho_0^{2h} + hs(s+m)\log T/T$$

where c_h is a bounded sequence. We obtain

$$\ell_B(T)\{-2 \log \rho_0/\log T\} \to 1 .$$

Using lemma 3.3.3 and (3.3.4) we have

$$\begin{aligned} BIC(h) &= \log\{\det \tfrac{1}{T}\Sigma\varepsilon(t)\varepsilon(t)'\} \\ &+ \operatorname{tr}\{\Gamma_0^{-1}([\mathring{\phi}_h,\mathring{\psi}_h]-[\phi,\psi])\Gamma_\infty([\mathring{\phi}_h,\mathring{\psi}_h]-[\phi,\psi])'\}(1+o(1)) \\ &+ hs(s+m)\{\log T/T\}(1+o(1)) . \end{aligned}$$

THEOREM 3.3.4. *Assume* $y(t)$ *is generated by an ARMAX process and* $\varepsilon(t)$, $z(t)$ *satisfy conditions A. If* $\hat{\ell}_B(T)$ *is the integer minimising* $BIC(h)$, $0 \leqslant h \leqslant (\log T)^a$ *then*

$$\hat{\ell}_B(T)/\ell_B(T) = 1 + o(1) \text{ a.s.}$$

We can replace $BIC(h)$ by the criterion

$$\log \det \hat{\Gamma}_h + hs(s+m)c_0 \log \log T/T \tag{3.3.6}$$

and, when c_0 is chosen sufficiently large, the order estimates will behave as those obtained using $BIC(h)$. Of course c_0 is unknown (it depends on the true system parameters) so it is difficult to justify an order selection procedure based on (3.3.6).

Let $\hat{\ell}_A(T)$ be the order estimate obtained by minimising $AIC(h)$, $0 \leqslant h \leqslant (\log T)^a$. If, for $h_1 < h_2$, $BIC(h_2) < BIC(h_1)$ then $AIC(h_2) < AIC(h_1)$. Consequently, if $\hat{\ell}_B(T)$ gives the minimum for $BIC(h)$ then $AIC(\hat{\ell}_B(T)) < AIC(h)$ for $h < \hat{\ell}_B(T)$. Thus $\hat{\ell}_A(T) \geqslant \hat{\ell}_B(T)$ and $\liminf \hat{\ell}_A(T)\{-2 \log \rho_0/\log T\} \geqslant 1$. With stronger conditions on the innovations sequence we can evaluate the LHS of (3.3.4) more precisely (but only in probability) and obtain a result

analogous to Theorem 3.3.4 for $\hat{\ell}_A(T)$. We show in lemmas 5.2.1, 5.2.2 that when conditions B hold and $h \geqslant d_0 \log T$

$$(3.3.7)\quad \frac{1}{hT}\{\Sigma\varepsilon(t)[y(t,h)',z(t,h)']\}\hat{\Gamma}_h^{-1}\left\{\Sigma\begin{bmatrix} y(s,h) \\ z(s,h)\end{bmatrix}\varepsilon(s)'\right\}$$

$$= \Phi_0(s+m) + o_p(1) .$$

Using lemma 3.3.3 we have in probability

$$\begin{aligned}
AIC(h) &= \log\det\{\tfrac{1}{T}\Sigma\varepsilon(t)\varepsilon(t)'\} \\
&\quad + \operatorname{tr}\Big\{\{\Phi_0^{-1}([\mathring{\phi}_h,\mathring{\psi}_h]-[\phi,\psi])\mathring{\Gamma}_\infty([\mathring{\phi}_h,\mathring{\psi}_h]-[\phi,\psi])'\}(1+o(1)) \\
&\quad - \{h(s+m)/T\}I_s(1+o_p(1))\Big\} + 2hs(s+m)/T + o(\tfrac{h}{T}) \\
&= \log\det\{\tfrac{1}{T}\Sigma\varepsilon(t)\varepsilon(t)'\} \\
&\quad + \operatorname{tr}\{\Phi_0^{-1}([\mathring{\phi}_h,\mathring{\psi}_h]-[\phi,\psi])\mathring{\Gamma}_\infty([\mathring{\phi}_h,\mathring{\psi}_h]-[\phi,\psi])'\}(1+o(1)) \\
&\quad + \{hs(s+m)/T\}(1+o_p(1)) .
\end{aligned}$$

As before, define $\ell_A(T)$ as the integer minimising

$$\operatorname{tr}\{\Phi_0^{-1}([\phi_h,\psi_h]-[\phi,\psi])\Gamma_\infty([\phi_h,\psi_h]-[\phi,\psi])'\} + hs(s+m)/T .$$

Using the same argument as for $\ell_B(T)$, we have

$$\lim \ell_A(T) = \log T/(-2\log\rho_0) .$$

THEOREM 3.3.5. *Assume* $y(t)$ *is generated by an ARMAX process and* $\varepsilon(t)$, $z(t)$ *satisfy conditions B. If* $\hat{\ell}_B(T)$ *is the integer minimising* $AIC(h)$, $0 \leqslant h \leqslant (\log T)^a$, $a < \infty$, *then*

$$\hat{\ell}_A(T)/\ell_A(T) = 1+o(1)$$

in probability.

The only time the 'in probability' argument enters this result is in evaluating (3.3.7), and strengthening this to

an almost sure result appears to be very difficult. Though the results have been given for ARMAX models the geometric rate of decrease of the coefficients $\phi(j),\psi(j)$ is not really necessary. These results continue to hold when we only require $\Sigma j^{\frac{1}{2}} \| \phi(j),\psi(j) \| < \infty$. If these methods are to be applied to more general linear models

$$(3.3.8) \qquad \sum_{0}^{\infty}\phi(j)y(t-j) + \sum_{1}^{\infty}\psi(j)z(t-j) = \varepsilon(t)$$

with only slow rates of convergence for the $\phi(j),\psi(j)$ a higher bound on the order of the regression-autoregression models should be considered. For example instead of $H_T = (\log T)^a$ as in the ARMAX case, we might take $H_T = T^a$, $a \leqslant \frac{1}{2}$. Shibata (1980) has obtained results analogous to Theorem 3.3.5 for (3.3.8) with $H_T = T^{\frac{1}{2}}$. His results are under Gaussian assumptions and are obtained using an approach different to ours. Details of the generalisation of our approach to (3.3.8) are given in Hannan and Kavalieris (to appear).

A number of simulation results are given in table 3.3.1 for several simple models. Evidently $\log T/\{-2 \log \rho_0\}$ grossly overestimates the order estimates especially when ρ_0 is near unity. The observed median values are close to $\tilde{h}_A(T)$, $\tilde{h}_B(T)$ obtained by minimising, respectively

$$\log \sharp_h + 2h/T \ , \ \log \sharp_h + h \log T/T \ .$$

TABLE 3.3.1

Orders of regression models selected using AIC(h) and BIC(H) from 50 repetitions.

Model	$y(t) = \varepsilon(t) + .6\varepsilon(t-1)$			$y(t) = \varepsilon(t) + .8\varepsilon(t-1)$			$y(t) = \varepsilon(t) + .95\varepsilon(t-1)$		
T	100	500	1000	100	500	1000	100	500	1000
AIC min	1	2	3	2	4	6	4	8	13
med	2	4	5	5	8	10	6	14	19
max	12	20	15	12	20	15	13	30	32
$\tilde{h}_A(T)$	2	4	5	4	7	9	5	13	18
BIC min	0	2	3	1	4	4	1	4	8
med	2	3	3	2	5	6	3	7	10
max	3	5	6	6	8	10	6	12	13
$\tilde{h}_B(T)$	2	3	4	3	5	6	3	7	10
$\frac{\log T}{-2 \log \rho_0}$	5	6	7	10	14	16	45	61	67

The asymptotic formula $\log T/\{-2 \log \rho_0\}$ is obtained by treating the first term in (3.3.5) as if it is ρ_0^{2h+2} and neglecting all terms of smaller order than $\log T$. Upon differentiating (3.3.5) and taking logs the minimising value is the solution of

$$(3.3.9) \quad 2h \log \rho_0 = \log \log T - \log T - \log(-2 \log \rho_0) - 2 \log \rho_0 .$$

Neglecting smaller order terms, $h = \log T/\{-2 \log \rho_0\}$. For moderate T and ρ_0 close to unity the small order terms need not be negligible. Furthermore, treating the first term of (3.3.5) as if it were ρ_0^{2h+2} may be considerably distorting its true value. We calculate this term for the univariate moving average

$$y(t) = \varepsilon(t) + \rho_0 \varepsilon(t-1) \qquad E\{\varepsilon(t)^2\} = 1 .$$

We have $\sharp_h = \det \Gamma_{h+1}/\det \Gamma_h$ where $\gamma(0) = 1+\rho_0^2$, $\gamma(1) = \rho_0$. Then $\det \Gamma_{h+1} = (1+\rho_0^2)\det \Gamma_h - \rho_0^2 \det \Gamma_{h-1}$. Solving the difference equation with initial values $\det \Gamma_0 = 1$, $\det \Gamma_1 = \rho_0^2 + 1$ we obtain

$$\det \Gamma_h = (1 - \rho_0^{2(h+1)})/(1 - \rho_0^2) .$$

Thus

$$\sharp_h = (1 - \rho_0^{2(h+2)})/(1 - \rho_0^{2(h+1)}) .$$

Now the first term of (3.3.5) is $E\{(\varepsilon_h(t)-\varepsilon(t))^2\} = \sharp_h - 1$ which, to first order is ρ_0^{2h+2}. A better approximation is $\rho_0^{2h+2}(1 - \rho_0^2)$. This leads to the estimate from (3.3.9) being decreased by an amount $\log(1 - \rho_0^2)/2 \log \rho_0$. This improves the theoretical bound but the bound is still fairly poor when ρ_0 is near unity.

§3.4 Stage II

Using the estimates $\hat{\varepsilon}_h(t)$ of the innovations sequence obtained at Stage I we fit an ARMAX model to the data by regressing each component of $y(t)$ on certain components of $y(t) - \hat{\varepsilon}_h(t)$, $y(t-j)$, $z(t-j)$, $\hat{\varepsilon}_h(t-j)$. The selection of the components involved in these regressions is determined by the Kronecker indices of the model being fitted. Let $\hat{\Sigma}_\alpha$ be the residual mean square from the above regression when the echelon form for V_α is estimated. The Kronecker indices α are estimated as the argument minimising

$$\log \det \hat{\Sigma}_\alpha + d(\alpha)C(T)/T, \tag{3.4.1}$$

but we shall find it profitable to consider modifications of this criterion later in this section. We could consider minimising (3.4.1) over *all* α such that $|\alpha| < N$, where N is fixed a priori, but, because of the large number of echelon submanifolds to be considered even when s is quite small, we seek a more efficient approach.

The estimates $\hat{\varepsilon}_h(t)$ are not explicitly needed in Stage II; it is only their covariances that enter the calculations. Let

$$\hat{\gamma}_{\hat{\varepsilon}y}(k) = \frac{1}{T}\sum_{1}^{T-k} \hat{\varepsilon}_h(t)y(t+k) \quad k \geqslant 0$$

with corresponding meanings for $\hat{\gamma}_{\hat{\varepsilon}z}(k)$, $\hat{\gamma}_{\hat{\varepsilon}}(k)$ as in §3.1. Then

$$\hat{\gamma}_{\hat{\varepsilon}y}(k) = \begin{cases} \sum_{0}^{h} \hat{\phi}_h(j)\hat{\gamma}_y(j+k) + \sum_{1}^{h} \hat{\psi}_h(j)\hat{\gamma}_{zy}(j+k) & k > 0 \\ \hat{\Sigma}_h & k = 0 \\ 0 & k < 0 \end{cases} \tag{3.4.2}$$

$$\hat{\gamma}_{\hat{\varepsilon}z}(k) = \begin{cases} \sum_{0}^{h} \hat{\phi}_h(j)\hat{\gamma}_{yz}(j+k) + \sum_{1}^{h} \hat{\psi}_h(j)\hat{\gamma}_z(j+k) & k \geqslant 0 \\ 0 & k < 0 \end{cases}$$

(3.4.2) cont'd

$$\hat{\gamma}_{\hat{\varepsilon}}(k) = \begin{cases} \sum_{u=0}^{h} \sum_{v=0}^{k} [\hat{\phi}_h(u), \hat{\psi}(u)]\, \hat{\gamma}(u+k-v) \begin{bmatrix} \hat{\phi}_h(v)' \\ \hat{\psi}_h(v)' \end{bmatrix} & k > 0 \\ \hat{\sharp}_h & k = 0 \\ \hat{\gamma}_{\hat{\varepsilon}}(-k)' & k < 0 \end{cases}$$

$$\hat{\gamma}_{y\hat{\varepsilon}}(k) = \hat{\gamma}_{\hat{\varepsilon}y}(-k)' \ , \quad \hat{\gamma}_{z\hat{\varepsilon}}(k) = \hat{\gamma}_{\hat{\varepsilon}z}(-k)' \ .$$

We remind the reader that $V(n)$ is the generic neighbourhood having McMillan degree, and $\alpha(n)$ is the set of Kronecker indices for $V(n)$. If $n = ps+q$ then $\alpha(n) = \{n_1,\ldots,n_s\}$ where $n_1 = \ldots = n_q = p+1$, $n_{q+1} = \ldots = n_s = p$. First consider fitting ARMAX models for $V(ps)$, $p = 0,1,\ldots,P$. Construct the matrices

$$(3.4.3) \quad G(k) = \begin{bmatrix} \hat{\gamma}_y(k) & \hat{\gamma}_{yz}(k) & \hat{\gamma}_{y\hat{\varepsilon}}(k) \\ \hat{\gamma}_{zy}(k) & \hat{\gamma}_z(k) & \hat{\gamma}_{z\hat{\varepsilon}}(k) \\ \hat{\gamma}_{\hat{\varepsilon}y}(k) & \hat{\gamma}_{\hat{\varepsilon}z}(k) & \hat{\gamma}_{\hat{\varepsilon}}(k) \end{bmatrix} \quad k = 0,1,\ldots,P \ .$$

These are used as input to Algorithm 3.3.2. At the p^{th} stage, the algorithm returns coefficient matrices $F_p(j)$, $j = 1,\ldots,p$ of $2s+m$ rows and columns and the residual mean square S_p . The matrix in the first s rows and columns of S_p is $\hat{\sharp}_{\alpha(ps)}$ where $\alpha(ps) = \{p,\ldots,p\}$. Use $\hat{\sharp}_{\alpha(ps)}$ in (3.4.1) to estimate the optimal order from the class of models $V(ps)$, $p = 0,1,\ldots,P$. Denote the optimal structural indices by $\alpha(\hat{ps})$. Denote the matrix in the first s rows of

$F_{\hat{p}}(j)$ by $[\hat{A}(j),-\hat{D}(j),-\hat{B}(j)]$, $j = 1,\ldots,\hat{p}$ where the partitioning is after s, s+m columns. These are the estimated coefficient matrices for the ARMAX model in $\nu(\hat{p}s)$. Theorem 3.4.1 shows that $\hat{p}$ is a strongly consistent estimate of the largest Kronecker index. To estimate the McMillan degree calculate the residual mean squares $\hat{\sigma}_{\alpha(\hat{p}s-j)}$, $j = 1,\ldots,s-1$ from the regressions

(3.4.4)

(i) if $i = 1,\ldots,s-j$ regress $y_i(t)$ on $y_\ell(t-1)$, $\ell = 1,\ldots,s-j$ and all components of $z(t-1)$, $\hat{\varepsilon}_h(t-1)$, $y(t-\ell)$, $z(t-\ell)$, $\hat{\varepsilon}_h(t-\ell)$, $\ell = 2,\ldots,\hat{p}$.

(ii) if $i = s-j+1,\ldots,s$ regress $y_i(t)$ on $y_\ell(t)-\hat{\varepsilon}_\ell(t)$, $\ell = 1,\ldots,s-j$ and all components of $y(t-\ell)$, $z(t-\ell)$, $\hat{\varepsilon}_h(t-\ell)$, $\ell = 1,\ldots,\hat{p}-1$.

Algorithm 3.4.1 provides a convenient way to organise these calculations. Select $\hat{j}$ as that integer minimising (3.4.1) for $\alpha = \alpha(\hat{p}s-j)$, $j = 0,1,\ldots,s-1$. The function C(T) in (3.4.1) can be specified from the following theorem.

THEOREM 3.4.1. *Assume that the true transfer functions* $[k_0(z),\ell_0(z)] \in \nu(n_0)$, $n_0 = p_0 s+q_0$, $0 \leqslant q_0 < s$ *and that* $\varepsilon(t)$ *and* $z(t)$ *satisfy conditions A. The order* h *for the stage I model is selected using BIC. At Stage II* $\hat{p}$ *is selected by minimising (3.4.1) for* $\alpha = \alpha(ps)$, $p = 0,1,\ldots,P$, $P < \infty$ *and* $\hat{j}$ *is selected by minimising (3.4.1) for* $\alpha = \alpha(\hat{p}s-j)$, $j = 0,\ldots,s$. *If* $C(T) = c_0 \log T \log\log T$ *and* c_0 *is sufficiently large, then, almost surely*

(i) $$\hat{p} \to \begin{cases} p_0 , & q_0 = 0 \\ p_0+1 , & q_0 > 0 \end{cases}$$

(ii) $\hat{p}s-\hat{j} \to p_0 s+q_0$.

The constant c_0 in this theorem depends on the true system parameters, so it is not known. However, if $C(T)$ is taken to increase at a rate that is faster than $c \log T \log\log T$, $c > 0$ the conclusion of this theorem remains true. We could, for example take $C(T) = (\log T)^{1+\delta}$, $\delta > 0$. Thus we have a procedure to consistently estimate the McMillan degree of a system in $V(n_0)$ which requires $\hat{\Phi}_{\alpha(n)}$ to be calculated for $P+s$ values of n . The requirement that the true model is in generic neighbourhood may not be a particularly strong condition in practice, for the true transfer functions can hardly be expected to be rational functions. Then what we can hope for is that $\hat{\Phi}_{\alpha(n)}$ is a sufficiently good estimate of $\Phi_{\alpha(n)}$ so that the above procedure gives consistent results for different data sets from the same model. Analytic results for fitting ARMAX models to data that is not generated by rational transfer function models are not yet available. Theorem 3.4.1 is difficult to generalise to include $[k_0(z), \ell_0(z)] \in V_{\alpha_0}$, $\alpha_0 = \{n_{01},\ldots,n_{0s}\}$. The difficulty arises in evaluating $\hat{\Phi}_{\alpha(ps)}$ when $\min\{n_{0j}\} < p < \max\{n_{0j}\}$. Under stronger conditions on the innovations we give the generalisation in Theorem 3.4.4 but with convergence in probability rather than almost surely.

Algorithm 3.4.1. Assume that the ARMAX model from $V(ps)$ has been estimated using Algorithm 3.2.2 and that the outputs

$$F_{\hat{p}}(j),\ j = 1,\ldots,\hat{p}\ ;\ S_{\hat{p}}$$

$$F_{\hat{p}-1}(j),\ j = 1,\ldots,\hat{p}-1\ ;\ S_{\hat{p}-1}$$

are available. Further assume that $\hat{p}$ is a consistent estimate of $\max\{n_{0j}\}$. We introduce some complicated notation to describe the algorithm. Let $w^0(t) = (y(t)', z(t)', \hat{\varepsilon}_h(t)')'$, and the vectors

$$w^1(t) = (y_j(t),\ j = 1,\ldots,q),\quad w^2(t) = (y_j(t),\ j = q+1,\ldots,s),$$

$$w^3(t) = (\hat{\varepsilon}_{h,j}(t),\ j = 1,\ldots,q),$$

$$w^4(t) = (w^1(t)', z(t)', \hat{\varepsilon}_h(t)')' .$$

Then define the matrices

$$G^{ij}(k) = \frac{1}{T}\sum_{1}^{T-k} w^i(t)w^j(t+k)' \quad k \geqslant 0,\ i,j = 0,\ldots,4 .$$

These are submatrices of $G(k) = G^{00}(k)$ that is defined in (3.4.3). In an analogous way we define submatrices $S^{ij}_{\hat{p}}$, $S^{ij}_{\hat{p}-1}$ of $S_{\hat{p}}$, $S_{\hat{p}-1}$. For example $S^{13}_{\hat{p}}$ is the submatrix in rows $1,\ldots,q$ and columns $s+m+1,\ldots,s+m+q$ of $S_{\hat{p}}$. Define submatrices of $F_p(j)$

$F^1_p(j)$, the matrix in rows $1,\ldots,q$,

$F^2_p(j)$, the matrix in rows $q+1,\ldots,s$,

$F^3_p(j)$, the matrix in rows $s+m+1,\ldots,s+m+q$,

$F^4_p(j)$, the matrix in rows $1,\ldots,q$ and $s+1,\ldots,2s+m$.

In this notation the coefficient matrices $[\hat{A}(j), -\hat{D}(j), -\hat{B}(j)] = \begin{bmatrix} F^1_{\hat{p}}(j) \\ F^2_{\hat{p}}(j) \end{bmatrix}$. We use the terminology of linear regression to describe the algorithm, without meaning to imply that the qualities discussed are derived from ordinary regressions.

The terminology is convenient because the computations that are done in the algorithm are exactly those needed to add in or remove variables from a regression.

Step (i) Begin with the regressions

$$\begin{aligned} &(a) \quad y_j(t) \quad \text{on} \quad y(t-\ell), z(t-\ell), \hat{\varepsilon}_h(t-\ell), \\ &\ell = 1,\ldots,\hat{p} \ , \ j = 1,\ldots,q \end{aligned}$$

(3.4.5)

$$\begin{aligned} &(b) \quad y_j(t) \quad \text{on} \quad y(t-\ell), z(t-\ell), \hat{\varepsilon}_h(t-\ell), \\ &\ell = 1,\ldots,\hat{p}-1, \ j = q+1,\ldots,s \ . \end{aligned}$$

For these regressions, the residual is

$$e(t) = y(t) + \sum_{j=1}^{\hat{p}} \begin{bmatrix} F^1_{\hat{p}}(j) \\ F^2_{\hat{p}-1}(j) \end{bmatrix} w^0(t-j) \ , \ F_{\hat{p}-1}(\hat{p}) = 0 \ ,$$

and the residual mean square is

$$\hat{\Sigma} = \begin{bmatrix} S^{11}_{\hat{p}} & S^{12}_{\hat{p}} \\ S^{21}_{\hat{p}} & S^{22}_{\hat{p}-1} \end{bmatrix}$$

Step (ii) *Remove the regressors* $y_j(t-1)$, $j = q+1,\ldots,s$ *from the regressions in step (i) (a).* This gives the regressions required in (3.4.4(i)). The expression

$$\sum_0^{\hat{p}-1} F^4_{\hat{p}-1}(j) w(t-j-1), \ F_{\hat{p}-1}(0) = I_{2s+m} \tag{3.4.6}$$

is the residual from regressing $w^4(t-1)$ on $w(t-j)$, $j = 2,\ldots,\hat{p}$. The covariance of $w^2(t-1)$ and (3.4.6) is

$$\begin{aligned} \frac{1}{T} \Sigma w^2(t-1) \left\{ \sum_0^{\hat{p}-1} F^4_{\hat{p}-1}(j) w(t-j-1) \right\}' &= \sum_0^{\hat{p}-1} G^{02}(j)' F^4_{\hat{p}-1}(j)' \\ &= \sum_0^{\hat{p}-1} F^2_{\hat{p}-1}(j) G^{04}(j) = S^{24}_{\hat{p}-1} \ , \end{aligned}$$

and regressing $w^2(t-1)$ on (3.4.6) gives

$$[E_1, E_2] = -S^{24}_{\hat{p}-1} \{S^{44}_{\hat{p}-1}\}^{-1},$$

where the partition in $[E_1, E_2]$ is after q columns. The residual mean square for this last regression is

$$D_q = G^{22}(0) - S^{24}_{\hat{p}-1}\{S^{44}_{\hat{p}-1}\}^{-1}S^{42}_{\hat{p}-1},$$

and

$$[E_1, I_{s-q}, E_2]w(t-1) + [E_1, E_2]\sum_1^{\hat{p}-1} F^4_{\hat{p}-1}(j)w(t-j-1)$$

is the residual from the regression of $w^2(t-1)$ on the components of $w^1(t-1), z(t-1), \hat{\varepsilon}_h(t-1), w^0(t-j)$, $j = 2,\ldots,\hat{p}$. Finally the coefficients for the regression (3.4.4 (i)) are

$$\begin{aligned} &F^1_{\hat{p}}(1) - \hat{A}_{12}(1)[E_1, I_{s-q}, E_2] \\ &F^1_{\hat{p}}(j) - \hat{A}_{12}(1)[E_1, E_2]F^4_{\hat{p}-1}(j-1), \quad j = 2,\ldots,\hat{p}. \end{aligned} \tag{3.4.7}$$

The residual mean square when $y_j(t)$, $j = 1,\ldots,q$ are regressed on the regressors in (3.4.4 (i)) while $y_j(t)$, $j = q+1,\ldots,s$ are regressed on those in (3.4.5 *(b)*) is

$$\hat{\Sigma} + \begin{bmatrix} \hat{A}_{12}(1)\, D_q \hat{A}_{12}(1)' & \hat{A}_{12}(1) D_q \hat{A}_{22}(1) \\ \hat{A}_{22}(1)\, D_q \hat{A}_{12}(1)' & 0 \end{bmatrix} \tag{3.4.8}$$

In (3.4.7), (3.4.8) $\hat{A}_{12}(1)$ is the matrix in rows $1,\ldots,q$ and columns $q+1,\ldots,s$ of $F_{\hat{p}}(1)$ and $\hat{A}_{22}(1)$ is the matrix in rows $q+1,\ldots,s$ and columns $q+1,\ldots,s$ of $F_{\hat{p}-1}(1)$.

<u>Step (iii)</u> It remains to add the variables $y_j(t) - \hat{\varepsilon}_{h,j}(t)$, $j = 1,\ldots,q$ into the regressions in (3.4.5 *(b)*). The residual from regressing the vector of these variables on the components of $w(t-j)$, $j = 1,\ldots,\hat{p}-1$ is

$$w^1(t) - w^3(t) + \sum_1^{\hat{p}-1}\{F^1_{\hat{p}-1}(j) - F^3_{\hat{p}-1}(j)\}w(t-j) \tag{3.4.9}$$

having residual mean square that is $S = S^{11}_{\hat{p}-1} - S^{13}_{\hat{p}-1} - S^{31}_{\hat{p}-1} + S^{33}_{\hat{p}-1}$. As in step (ii) the covariance of $w^2(t)$ and (3.4.9) is

$S^{21}_{\hat{p}-1} - S^{23}_{\hat{p}-1}$. Hence the regression of $w^2(t)$ on (3.4.9) gives

$$\hat{A}_{21}(0) = -\{S^{21}_{\hat{p}-1} - S^{23}_{\hat{p}-1}\}S^{-1}$$

Thus the vector of residuals for the regression (3.4.4 (ii)) is

$$\hat{A}_{21}(0)\{w^1(t)-w^3(t)\} +$$
$$\{\hat{A}_{21}(0)[F^1_{\hat{p}-1}(j)-F^3_{\hat{p}-1}(j)] + F^2_{\hat{p}-1}(j)\}w(t-j),$$

giving the coefficients

$$\hat{A}_{21}(0),\ \hat{A}_{21}(0)\{F^1_{\hat{p}-1}(j)-F^3_{\hat{p}-1}(j)\} + F^2_{\hat{p}-1}(j) \qquad j = 1,\ldots,\hat{p}-1 \tag{3.4.10}$$

for $w^1(t)-w^3(t)$, $w(t-j)$ respectively. Finally the residual mean square for the regressions (3.4.4 (i), (ii)) is given by subtracting

$$\begin{bmatrix} 0 & S\hat{A}_{21}(0)' \\ \hat{A}_{21}(0)S & \hat{A}_{21}(0)S\hat{A}_{21}(0)' \end{bmatrix}$$

from (3.4.8). Referring to the canonical form (1.3.11) for the generic neighbourhood $\nu(n)$, $n = (\hat{p}-1)s+q$, and using (3.4.7), (3.4.8) the estimates of the coefficients are

$$\hat{A}(0) = \hat{B}(0) = \begin{bmatrix} I_q & 0 \\ \hat{A}_{21}(0) & I_{s-q} \end{bmatrix}$$

$$[\hat{A}(1),-\hat{D}(1),-\hat{B}(1)] = \begin{bmatrix} F^1_{\hat{p}}(1)-\hat{A}_{12}(1)[E_1,I_{s-q},E_2] \\ F^2_{\hat{p}-1}(1)+\hat{A}_{21}(0)\{F^1_{\hat{p}-1}(1)-F^3_{\hat{p}-1}(1)\} \end{bmatrix}$$

$$[\hat{A}(j),-\hat{D}(j),-\hat{B}(j)] = \begin{bmatrix} F^1_{\hat{p}}(j)-\hat{A}_{12}(1)[E_1,E_2]F^4_{\hat{p}-1}(j-1) \\ F^2_{\hat{p}-1}(j)+\hat{A}_{21}(0)\{F^1_{\hat{p}-1}(j)-F^3_{\hat{p}-1}(j)\} \end{bmatrix},$$

for $j = 1,\ldots,\hat{p}$, where $F_{\hat{p}-1}(\hat{p}) = 0$.

The matrix of residual mean squares and cross products is

$$\hat{\Sigma}_{\alpha(n)} = \hat{\Sigma} + \begin{bmatrix} \hat{A}_{12}(1)D_q\hat{A}_{12}(1)' & \hat{A}_{12}(1)D_q\hat{A}_{22}(1)'-S\hat{A}_{21}(0)' \\ \hat{A}_{22}(1)D_q\hat{A}_{12}(1)'-\hat{A}_{21}(0)S & -\hat{A}_{21}(0)S\hat{A}_{21}(0)' \end{bmatrix}$$

We give a number of detailed results for the univariate ARMA model which do reflect the multivariate case. Assume the true model is

$$\sum_0^{p_0} A_0(j)y(t-j) = \sum_0^{q_0} B_0(j)\varepsilon(t-j), \quad A_0(0) = B_0(0) = I$$

Put $x(t) = b_0 e(t) = \sum_0^{q_0} B_0(j)e(t-j)$ where

$$E\{e(t)\} = 0, \quad E\{e(s)e(t)\} = \delta_{st} .$$

Let ω_v^2 be the expected variance of the error of linear prediction of $x(t)$ from $x(t-j)$, $j = 1,\ldots,v$, and $\hat{\Sigma}_{p,q}$ be the residual mean square when an ARMA(p,q) model is estimated at Stage II. Define

$$BIC(p,q) = \log \hat{\Sigma}_{p,q} + (p+q)\log T/T .$$

THEOREM 3.4.2. *Assume* $\varepsilon(t)$ *satisfies conditions B and the order at Stage I is chosen to optimise* AIC(h) *or* BIC(h) , $0 \leqslant h \leqslant (\log T)^a$, $a < \infty$. *If* $\hat{p},\hat{q}$ *are chosen by optimising* BIC(p,q), $p \leqslant P$, $q \leqslant Q$, $(P,Q < \infty)$ *then*

$$\hat{p} \to p_0+v_0 \ , \ \hat{q} \to q_0+v_0$$

in probability, where v_0 *is the first integer* v *for which*

$$\omega_v^2 - \omega_{v+1}^2 < -4 \log \rho_0$$

and ρ_0 *is the modulus of a zero of* $b_0(z)$ *nearest* $|z| = 1$.

The essential part of this proof is to show that

$$(3.4.11)\quad \hat{\sigma}^2_{p,q} = \frac{1}{T}\sum_1^T \varepsilon(t)^2 + \sigma^2_0(\omega_v^2-1)h/T + o_p(\log T/T)$$

when $p \geqslant p_0$, $q \geqslant q_0$ and $v = \min\{p-p_0, q-q_0\}$. We have shown in theorem 3.3.4 that h will be $\{\log T/(-2 \log \rho_0)\} \times (1+o(1))$, in probability, so that the second term on the RHS of (3.4.1) is of the same order of magnitude as the second term of $BIC(p,q)$. Furthermore the effect of this term is to cause BIC to overestimate the order of the model, but as ω_v^2 decreases to 1 in a way that is determined only by $b_0(z)$, the degree of overestimation is determined by the true system parameters. To be more precise, if either $p < p_0$ or $q < q_0$, Hannan and Rissanen (1982) show that

$$\hat{\sigma}^2_{p,q} > \hat{\sigma}^2_{p_0,q_0} + c \ , \ c > 0$$

so that $BIC(p,q)$ has its minimum in the region $p \geqslant p_0$, $q \geqslant q_0$. When $p \geqslant p_0$, $q \geqslant q_0$, (3.4.11) shows that

$$BIC(p,q) > BIC(p_0+v, q_0+v)$$

hence we seek a minimum from $\{(p,q) = (p_0+v, q_0+v),$ $v = 0,1,\ldots,V\}$. For $v > v'$

$$BIC(p_0+v,q_0+v) - BIC(p_0+v',q_0+v')$$

$$= \{(\omega_v^2-\omega_{v'}^2)/\{-2 \log \rho_0\}+2(v-v')\}\log T/T+o_p(\log T/T)$$

giving the result.

The effect of this theorem does not become apparent unless ρ_0 is quite close to unity, for example if the true model is $y(t) = \varepsilon(t)+B_0\varepsilon(t-1)$, then v_0 is zero if $|B_0| < 0.89$. In this calculation the Stage I order is given its asymptotic value which greatly overestimates the observed

orders (see table 3.3.1). Thus for moderate sample sizes the magnitude of the second term in the RHS of (3.4.11) may be overstated in these calculations and the degree of overestimation of order may be considerably less than that stated in Theorem 3.4.2. Some simulation results showing this effect are given in Table 3.4.1. In all of the models the optimal model was selected from ARMA(p,p), $p = 0,1,\ldots,P$.

In proving Theorem 3.4.2 we have also established the following corollary concerning the behaviour of the estimates $\hat{a}(z)$, $\hat{b}(z)$.

Corollary 3.4.3 *Assume all the conditions of Theorem 3.4.3. The parameter estimates of an* ARMA(p,q) *model fitted at Stage II satisfy*

$$\hat{a}(z) = \nu(z)a_0(z) + O(\log\log T/T)^{\frac{1}{2}}, \text{ a.s.}$$

$$\hat{b}(z) = \nu(z)b_0(z) + O(\log\log T/T)^{\frac{1}{2}}, \text{ a.s.}$$

where $v = \min\{p-p_0, q-q_0\}$ and

$$\nu(z) = \begin{cases} 0 & \text{if } v < 0 \\ \nu^{\circ}(z) + o_p(1) & \text{otherwise} \end{cases}$$

and $(\nu^{\circ}(L)-1)x(t)$ *is the best linear predictor of* $x(t)$ *from* $x(t-1),\ldots,x(t-v)$.

Note that the zeros of $\nu^{\circ}(z) = \Sigma\nu^{\circ}(j)z^{-j}$ are within $|z| = 1$.

Turning to the multivariate ARMA or ARMAX case we can prove partial analogues to Theorem 3.4.2 and corollary 3.4.3. We are unable to obtain an explicit form for the degree of overestimation of the Kronecker indices when BIC is used to determine the structure, though we can show that the degree of overestimation is dependent only on the true system

Table 3.4.1

Frequencies of choice of order of ARMA(p,p) models estimated at stages II, IIa

Order	T = 200				T = 500				T = 1000			
	$\hat{p}$	$\hat{p}_c$	$\hat{p}^{(1)}$	$\hat{p}_c^{(1)}$	$\hat{p}$	$\hat{p}_c$	$\hat{p}^{(1)}$	$\hat{p}_c^{(1)}$	$\hat{p}$	$\hat{p}_c$	$\hat{p}^{(1)}$	$\hat{p}_c^{(1)}$
	$y(t) = \varepsilon(t) + 0.8\ \varepsilon(t-1)$; $v_0 = 0$											
1	89	92	94	94	83	90	92	97	87	92	96	96
2	9	6	5	5	17	10	7	3	12	7	3	3
3	2	2	1	1	-	-	1	-	1	1	1	1
	$y(t) = \varepsilon(t) + 0.9\ \varepsilon(t-1)$; $v_0 = 1$											
1	76	92	94	95	70	90	94	97	68	86	96	96
2	19	7	5	4	29	10	5	3	27	13	4	4
3	4	1	1	1	1	-	-	-	5	1	-	-
	$y(t) = \varepsilon(t) + 0.95\ \varepsilon(t-1)$; $v_0 = 1$											
1	69	90	90	92	64	81	89	91	62	86	95	96
2	25	8	7	6	32	16	7	7	32	13	5	4
3	4	1	1	1	3	2	-	-	6	1	-	-
4	2	1	2	1	1	1	1	1	-	-	-	-
≥ 5	-	-	-	-	-	-	3	1	-	-	-	-
	$y(t) = \varepsilon(t) + \varepsilon(t-1) + 0.89\ \varepsilon(t-2)$; $v_0 = 1$											
1	-	-	-	-	-	-	-	-	-	-	-	-
2	39	73	76	81	22	66	72	78	18	68	85	88
3	51	22	24	19	69	28	26	20	71	26	13	10
4	10	5	-	-	9	6	2	2	10	6	2	2
5	-	-	-	-	-	-	-	-	1	-	-	-

Criteria for estimating order

$\hat{p}$: BIC

$\hat{p}^{(1)}$: BIC at stage II, BIC at stage IIa

$\hat{p}_c$: $BIC_{p,q}(i,j)$ at stage II

$\hat{p}_c^{(1)}$: $BIC_{p,q}(i,j)$ at stage II, BIC at stage IIa

parameters.

THEOREM 3.4.4. *Assume* $[k_0(z), \ell_0(z)] \in V_{\alpha_0}, \alpha_0 = \{n_{01}, \ldots, n_{0s}\}$, $|\alpha_0| < P$, *and conditions B are satisfied. Let the order at Stage I be determined using either AIC or BIC with* $H_T \leqslant (\log T)^a$; $a < \infty$. *If* $\{\hat{n}_1, \ldots, \hat{n}_s\}$ *are the estimates of the Kronecker indices obtained by optimising*

$$\log \det \hat{\Sigma}_\alpha + d(\alpha) c_0 \log T/T, \quad |\alpha| < P < \infty \tag{3.4.12}$$

then for c_0 *sufficiently large*

$$p \lim \hat{n}_j = n_{0j} \ , \ j = 1, \ldots, s \ .$$

In proving this theorem we establish the following results:

(a) $\alpha \not\geqslant \alpha_0 \Rightarrow \hat{\Sigma}_\alpha \geqslant \Sigma_\alpha + o_p(1)$, $\Sigma_\alpha > \Sigma_{\alpha_0}$

(b) $\alpha = \alpha(ps) > \alpha_0 \Rightarrow \hat{\Sigma}_\alpha \geqslant \frac{1}{T} \sum_1^T \varepsilon(t)\varepsilon(t)' + o_p(\log T/T)$

(c) $\alpha > \alpha_0 \Rightarrow \hat{\Sigma}_\alpha \geqslant \frac{1}{T} \sum_1^T \varepsilon(t)\varepsilon(t)' + K_1(\log T/T) + o_p(\log T/T)$

(d) $\alpha = \alpha_0 \Rightarrow \Sigma_\alpha = \frac{1}{T} \sum_1^T \varepsilon(t)\varepsilon(t)' + K_0(\log T/T) + o_p(\log T/T)$.

Here K_0, K_1 are dependent only on the true system parameters. If BIC is used to estimate the structural indices (i.e. we put $c_0 = 1$ in Theorem 3.4.4) then it can happen that some of the Kronecker indices will be overestimated. Indeed if $\alpha > \alpha_0$ then

$$\begin{aligned} &BIC(\alpha_0) - BIC(\alpha) \\ &= \{tr\{\Sigma_0^{-1}(K_0 - K_1)\} + d(\alpha_0) - d(\alpha)\} \log T/T + o_p(\log T/T) \end{aligned}$$

which is positive, for sufficiently large T , when $tr\{\Sigma_0^{-1}(K_0 - K_1)\} + d(\alpha_0) - d(\alpha) > 0$. This gives a bound on the degree of overestimation that is determined by K_0 and K_1 .

To estimate all of the Kronecker indices we could proceed as follows. Assume that c_0 in (3.4.12) has been selected to be "sufficiently large" so that the estimates of the Kronecker indices are consistent, or replace $c_0 \log T$ by a function increasing more rapidly than $c_0 \log T$ (for example we could use $\log T \log\log T$.) As in Theorem 3.4.1 first consider the echelon submanifolds $\nu(ps)$, $p = 0,1,\ldots,P$. If $\alpha = \alpha(\hat{p}s)$ optimises (3.4.12) then from (a), (b), (d), $\hat{p} \to \max\{n_{0j}\}$ in probability. Now we may proceed to estimate the remaining Kronecker indices using the following scheme. Suppose the estimates $\hat{n}_{q+1},\ldots,\hat{n}_s$ of $n_{0q+1},\ldots,n_{0s}$ have already been obtained using this method. To estimate n_{0q}, optimise (3.4.12) for $\alpha = \alpha_j = \{\hat{p},\ldots,\hat{p},j,\hat{n}_{q+1},\ldots,\hat{n}_s\}$, $j = \hat{p},\hat{p}-1,\ldots,1$. If $\alpha_{\hat{j}}$ optimises (3.4.12) then $\hat{j} \to n_{0q}$ in probability. To see that this estimate is consistent, assume $\hat{n}_{q+1},\ldots,\hat{n}_s$ are consistent estimates of $n_{0q+1},\ldots,n_{0s}$. If $j > n_{0q}$ then $\alpha_j \geqslant \alpha_0$ and $\hat{\Sigma}_{\alpha_j}$ is bounded below by (c). Comparing this with $\hat{\Sigma}_{\alpha_{n_{0q}}}$ (which is evaluated by (d) if $\alpha_{n_{0q}} = \alpha_0$ or otherwise by (c)) shows that if c_0 is sufficiently large (3.4.12) increases as j increases beyond n_{0q}. When $j < n_{0q}$ then $\alpha_j \not\geqslant \alpha_0$ hence from (a) $\hat{\Sigma}_{\alpha_j} > \Sigma_0 + c$, $c > 0$. Hence as $\hat{\Sigma}_{\alpha_{n_{0q}}}$ converges to Σ_0, (3.4.12) is seen to decrease, for large T, as j increases to n_{0q}. Thus $\hat{j} \to n_{0q}$, in probability. In principle we need not consider more than $P+s\max\{n_{0j}\}$ echelon submanifolds to estimate all of the Kronecker indices. The estimation for the P submanifolds $\nu(ps)$, $p = 1,\ldots,P$ is conveniently done using the Levinson-Durbin recursion (Algorithm 3.2.2). The remaining calculations are probably best done using standard methods for adding or removing variables from a regression (see, for

example, Seber, 1977, pp. 338-342). When we are seeking to approximate the true structure by an ARMAX model in a generic neighbourhood, Algorithm 3.4.1 will probably be the most economical approach to use.

The algorithms discussed here are based on having a consistent estimate of the largest Kronecker index. In practice it should be recognised that $\hat{p}$ may be less than some n_{0j} for moderate sample sizes and a wider range of echelon submanifolds may need to be considered than those given in the estimation procedure.

The last point is brought out in the analysis of the following data consisting of sales of mink and muskrat furs by the Hudson's Bay Company for the period 1850-1911 tabulated in Jones (1914), pp. 209-214. Though it is generally considered that a linear model is not appropriate for this kind of data our main aim is to compare estimates of the Kronecker indices obtained using these methods with the estimates obtained by Cooper and Wood (1982) using the canonical correlations approach of Akaike (1974c). The data consists of two series $\tilde{y}_1(t)$, the muskrat series, and $\tilde{y}_2(t)$, the mink series, each with 62 observations. We follow Cooper and Wood and fit the models to the transformed data $y_i(t) = \log(\tilde{y}_i(t))$, $i = 1,2$.

Table 3.4.2

Estimation of Kronecker indices for Mink-Muskrat data.

Kronecker index α	No. of parameters $d(\alpha)$	$\log\det \hat{\Sigma}_\alpha + d(\alpha)C(T)/T$ $C(T) \equiv 2$	$C(T) = \log T$
{1,1}	8	-4.9495	-4.6739
{1,2}	11	-5.1429	-4.7655
{2,1}	12	-4.9506	-4.5389
{2,2}	16	-5.1230	-4.5740

The order of the regression model selected using AIC at Stage I was 5. An ARMA model from $\gamma(ps)$, $p = 1,2,\ldots$ was selected using (3.4.1). With $C(T) \equiv 2$ this criterion selects $p = 2$, or with $C(T) = \log T$ we obtain $p = 1$. Since the sample size is so small it is reasonable to doubt that these estimates have actually found the largest of the Kronecker indices so it is reasonable to examine pairs of Kronecker indices $\{j,k\}$ for $\{j,k\}$ near to $\{1,1\}$. When this was done both choices of $C(T)$ lead to $\alpha = \{1,2\}$, agreeing with the model selected by Cooper and Wood. For the $\{1,2\}$ model the parameter estimates are

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} y(t) + \begin{bmatrix} .74 & .46 \\ 0 & -1.57 \end{bmatrix} y(t-1) + \begin{bmatrix} 0 & 0 \\ -.01 & .71 \end{bmatrix} y(t-2)$$

$$= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \varepsilon(t) + \begin{bmatrix} .38 & -.22 \\ .52 & -1.10 \end{bmatrix} \varepsilon(t-1) + \begin{bmatrix} 0 & 0 \\ -.45 & .02 \end{bmatrix} \varepsilon(t-2)$$

These estimates are quite close to the maximum likelihood estimates obtained by Cooper and Wood. However, $\det \hat{b}(z)$ has a zero of modulus 1.07. Thus $\hat{b}(z)$ is unstable and as

such it cannot be used to initialise a maximum likelihood procedure. One approach we have used, which will be discussed again later, is to replace $\hat{b}(z)$ by $\hat{b}_\lambda(z) = I + \lambda(\hat{b}(z)-I)$, $0 < \lambda < 1$ where λ is selected as close to unity consistent with the stability of $\hat{b}_\lambda(z)$. This is a simple approach which appears to give quite satisfactory results.

There is a close analogue to corollary 3.4.3. In proving parts (b), (c) when $\alpha = \alpha(n)$ the coefficients of $a[b^\circ,d^\circ]-[b,d]a^\circ$ are explicitly evaluated to be $O(\log\log T/T)^{\frac{1}{2}}$ almost surely. When $a[b^\circ,d^\circ] = [b,d]a^\circ$, we have that $[a(z),b(z),d(z)] = \nu(z)[a_0(z),b_0(z),d_0(z)]$ where now $[a_0(z),b_0(z),d_0(z)]$ is the (left) ARMAX representation of $[k_0(z),\ell_0(z)]$.

COROLLARY 3.4.5. *Assume the model from* V_α , $\alpha > \alpha_0$ *that is fitted at Stage II has ARMAX representation* $[\hat{a},\hat{b},\hat{d}]$. *If* $[a_0,b_0,d_0]$ *is the ARMAX representation in* V_{α_0}, *then*

$$[\hat{a},\hat{b},\hat{d}] = \nu(z)[a_0,b_0,d_0] + O(\log\log T/T)^{\frac{1}{2}}, \text{ a.s.}$$

In general we have not been able to obtain results concerning the stability of $\nu(z)$ as we did in corollary 3.4.3. If $\alpha = \alpha(ps)$, $\alpha_0 = \alpha(p_0 s)$ the conclusions of corollary 3.4.3 carry over exactly, but otherwise there are constraints amongst the parameters of $\nu(z)$ making it difficult to evaluate the asymptotic behaviour.

As we have already mentioned the constant c_0 occuring in (3.4.12) is determined by the (unknown) system parameters. There is no difficulty in obtaining consistent estimates if $c_0 \log T$ is replaced by a function $C(T)$ that is increasing sufficiently rapidly. As we have already pointed out, if $C(T)$ is taken to be too large it will lead to underestimation of the Kronecker indices. In the univariate ARMA case our

Table 3.4.3

Frequencies of choice of order of ARMA(p,p) models at stage II

Order	T = 200		T = 500		T = 1000	
	$\hat{p}$	p^0	$\hat{p}$	p^0	$\hat{p}$	p^0
	$y(t) = \varepsilon(t) + \varepsilon(t-1) + 0.6\ \varepsilon(t-2)$					
1	26	61	-	12	-	-
2	60	38	68	79	67	88
3	14	1	30	9	32	12
4	-	-	2	-	1	-
	$y(t) = \varepsilon(t) + \varepsilon(t-1) + 0.89\ \varepsilon(t-2)$					
1	-	2	-	-	-	-
2	39	76	22	56	18	48
3	51	21	69	44	71	51
4	10	1	9	-	10	1
5	-	-	-	-	1	-

Criteria for determining order

$\hat{p}$: BIC

p^0 : $\log \hat{\sigma}^2_{p,p} + 2p \log T \log \log T/T$

simulations show that if $C(T) = \log T \log\log T$ there is less tendency to overestimate model order but there is a considerably higher frequency of underestimation when compared with $C(T) = \log T$. These simulations are tabulated as p^0 in Table 3.4.3. We now study methods for obtaining consistent estimates of the model structure which do not involve using $C(T) = \log T \log\log T$ or an arbitrary large value of c_0 but rather use $C(T) = \log T$. We only have complete results in the univariate case and we restrict our theorems to scalar ARMA models as the ARMAX case is almost the same. The gap in the proofs for the corresponding results for the estimation of Kronecker indices is that we lack a full analogue to corollary 3.4.3 for multivariate models. This problem is still being investigated.

Assume $\hat{A}(j)$, $j = 1,\ldots,\hat{p}$, $\hat{B}(j)$, $j = 1,\ldots,\hat{q}$ are the parameter estimates at order $(\hat{p},\hat{q})$ when the latter is obtained by optimising $BIC(p,q)$, $p,q \leqslant P$. Compute new estimates of the innovations as

$$(3.4.13)\quad \tilde{\varepsilon}(t) = \sum_{0}^{\hat{p}} \hat{A}(j)y(t-j) - \sum_{1}^{\hat{q}} \hat{B}(j)\tilde{\varepsilon}(t-j), \quad \hat{A}(0) = \hat{B}(0) = 1$$

which is initialised by taking $y(s)$, $\tilde{\varepsilon}(s)$ as zero for $s \leqslant 0$. Using $\tilde{\varepsilon}(t)$ in place of $\hat{\varepsilon}_h(t)$, Stage II may be repeated. We refer to the repetition of Stage II in this way as Stage IIa. Let $\tilde{\sharp}_{pq}$ be the residual mean square obtained when an $ARMA(p,q)$ model is estimated at Stage IIa and select order by optimising

$$(3.4.14)\qquad \log \tilde{\sharp}_{pq} + (p+q)\log T/T \quad p,q \leqslant P \ .$$

THEOREM 3.4.6. *Let* $s = 1$ *and the true model be in* $ARMA(p_0,q_0)$. *Assume* $\varepsilon(t)$ *satisfies conditions B and the*

order at Stage I is chosen to optimise AIC(h) *or* BIC(h) , $0 \leqslant h \leqslant H_T$. *If the optimal model at Stage II is chosen by minimising* BIC(p,q), $p,q \leqslant P < \infty$ *and a model of order* $(\tilde{p},\tilde{q})$ *is estimated at Stage IIa by optimising (3.4.14), then*

$$\tilde{p} \to p_0 \ , \ \tilde{q} \to q_0$$

in probability.

To compute estimates of the innovations using (3.4.13) it is necessary that $\hat{b}(z) = \sum_0^{\hat{q}} \hat{B}(j)z^{-j}$ be stable. As pointed out with respect to maximum likelihood estimation for the mink-muskrat data, $\hat{b}(z)$ might be replaced by $b_\lambda(z) = 1 + \lambda(b(z)-1)$ if $\hat{b}(z)$ is unstable. Here λ is selected near to unity and consistent with the stability of $b_\lambda(z)$. This simple procedure appears to perform well in the small number of simulations where $\hat{b}(z)$ was unstable. In all of these cases the value of the parameter λ required for stability of $b_\lambda(z)$ was quite close to unity. In the simulations reported in Table 3.4.1 such a procedure to compute a stable $b_\lambda(z)$ was not used. The several cases when Stage IIa led to gross overestimation of the order arose because $\hat{b}(z)$ had zeros just outside $|z| = 1$. In all of these cases the zeros were less than 1.01 in modulus.

An alternative to repeating Stage II is to attempt to estimate the second term on the RHS of (3.4.11). Of course we cannot do this directly, but we can modify BIC to take this term into account. For any p,q let $\hat{B}_{pq}(j)$ be the moving average coefficients when an ARMA(p,q) model is fitted at Stage II. Compute

$$\hat{\rho}_{p,q}(j) = \sum_{k=0}^{q-j} \hat{B}_{pq}(j)\hat{B}_{pq}(k+j) \ , \ j = 0,1,\ldots,q$$

These are the covariances of $x_{pq}(t) = \sum_0^q \hat{B}_{pq}(j)e(t-j)$ where $e(t)$ are defined above Theorem 3.4.2. Compute $\omega^2_{pq}(v)$ the variance of the error of predicting $x_{pq}(t)$ from $x_{pq}(t-j)$, $j = 1,\ldots,v$, for $v = 0,1,\ldots$. Of course an upper bound for v needs to be imposed so that both $p+v$ and $q+v$ are bounded by P. The $\omega^2_{pq}(v)$ are conveniently computed using the Levinson recursion. Thus put M = maximum of v; $G(j) = \hat{\rho}_{p,q}(j)$, $j = 0,\ldots,q$ in Algorithm 3.2.1. Identify $\omega^2_{pq}(v)$ with the output S_v. Define $BIC_{pq}(i,j)$ as

$$\begin{cases} BIC(i,j) & i \leqslant p \text{ or } j \leqslant q \\ BIC(i,j) + \hat{h}\{\omega^2_{pq}(0)-\omega^2_{pq}(v)\}/T, & v = \min\{i-p,j-q\}, \\ & i > p, j > q, \end{cases}$$

where $\hat{h}$ is the order estimated at Stage I. Define a partial ordering $\{a,b\} \leqslant \{a^1,b^1\}$ when $a \leqslant a^1$ and $b \leqslant b^1$. Select the order $\{\hat{p}_c,\hat{q}_c\}$ as the least $\{p,q\}$ for which $BIC_{pq}(i,j)$ is minimised (for $i \leqslant P$, $j \leqslant Q$) at $i = p$, $j = q$.

THEOREM 3.4.7. *Assume* $s = 1$ *and the true model is in* ARMA(p_0,q_0). *Let* $\varepsilon(t)$ *satisfy conditions B and assume the order at Stage I is chosen to optimise* AIC(h) *or* BIC(h), $0 \leqslant h \leqslant H_T$. *If* $\{\hat{p}_c,\hat{q}_c\}$ *is chosen at Stage II using* $BIC_{pq}(i,j)$ *as described above then*

$$\hat{p}_c \to p_0, \quad \hat{q}_c \to q_0$$

in probability.

This procedure has the advantage over Stage IIa in that it is not necessary to compute new estimates of the innovations. Thus the moving average polynomial $\hat{b}(z)$ need not be stable. This is a procedure that is computationally cheap and appears to give good results, thus we feel it is always worth using.

Even if $BIC_{pq}(i,j)$ is used to choose the model order at Stage II, it might be worthwhile to compute new estimates of the innovations as in (3.4.13) and to estimate order in Stage IIa. Theorem 3.4.6 still applies and the new order estimates are consistent, as before. Simulation results are denoted by $\hat{p}_c^{(1)}$ in Table 3.4.1 for this procedure.

§3.5 Stage III

Here we are primarily concerned with the problem of obtaining asymptotically efficient estimates of the system parameters once the model structure is known. The main method we consider is one based on the Gauss-Newton procedure that is widely used in numerical optimisation procedures. We commence with a discussion of some methods based on the likelihood (2.2.1). Schweppe (1965) shows that this can be conveniently calculated using the Kalman filter. Indeed if we assume a state space model

$$x(t+1) = Fx(t) + Lz(t) + K\varepsilon(t)$$

$$y(t) = Hx(t) + \varepsilon(t)$$

where the matrices F, L, K, H are parameterised by τ, the Kalman filter yields estimates $\hat{x}(t)$ of $x(t)$ (based on the observations $y(s)$, $z(s)$, $s = 1,\ldots,t-1$) and V_t, the covariance matrix of the estimated innovations $e(t) = H(x(t) - \hat{x}(t))$. Then 2(log likelihood) is given by

$$L_T(\tau) = -\sum_t \{\log \det V_t + e(t)'V_t^{-1}e(t)\} . \qquad (3.5.1)$$

We shall refer to $L_T(\tau)$ as the 'exact' likelihood for it is the one that would be obtained if the process were Gaussian.

A natural way of locating extrema is by a Newton - Raphson procedure

$$\tau^{(k+1)} = \tau^{(k)} + \lambda_k H_k^{-1} \left. \frac{\partial L_T(\tau)}{\partial \tau} \right|_{\tau = \tau^{(k)}}$$

where $\tau^{(k)}$ is the parameter estimate at the k^{th} iteration, H_k is the Hessian of $L_T(\tau)$, and λ_k a scalar step length, the choice of which will be discussed later. Engel and Watson (1981) discuss the use of the Information matrix $J(\tau^{(k)})$ evaluated at $\tau^{(k)}$ as an estimate of H_k. They show that the (i,j) element of $J(\tau^{(k)})$ is

$$J_{ij}(\tau^{(k)}) = \sum_t \mathrm{tr}\left\{ V_t^{-1} \frac{\partial V_t}{\partial \tau_i} V_t^{-1} \frac{\partial V_t}{\partial \tau_j} \right\} + \sum_t \left(\frac{\partial e(t)}{\partial \tau_i} \right)' V_t^{-1} \left(\frac{\partial e(t)}{\partial \tau_j} \right) .$$

This has the advantage that it only contains first derivatives of V_t and e_t, i.e. quantities that can be calculated using the Kalman filter. Engel and Watson indicate numerical computation of the derivatives by n passes through the Kalman filter, where n is the dimension of τ. Other approaches to computing the derivatives have been considered (Wilson and Kumar (1982) see also Astrom (1980) for further references), where the computational burden is reduced using adjoint equations. The Kalman filter is conveniently parameterised by the coefficients of $\hat{a}(z)$, $\hat{b}(z)$, $\hat{d}(z)$, the estimates of $a_0(z)$, $b_0(z)$, $d_0(z)$ (Harvey and Phillips, 1979) since we have these available. Then the question of the stability of the filter is resolved by requiring $\hat{b}(z)$ to be stable. Solo (1983) considers parameterising the Kalman filter by the parameters of $\hat{a}(z)$ ~~together with some~~ and the covariances of $y(t)$ and $z(t)$. This is the output statistics Kalman filter of (Son and Anderson, 1971). This avoids the problem of estimating anything more than the

$a_0(z)$ coefficients. It seems difficult to determine whether the filter is stable directly from this parameterisation.

The choice $L_T(\tau)$ is only one choice of an objective function to be maximised. Other choices have been considered (Dunsmuir and Hannan, 1976) which lead to the same asymptotic properties of the estimators as when $L_T(\tau)$ is optimised. The effect of using a particular objective function when the sample size is small is not clear. We consider the objective function

$$\tilde{L}(\tau) = \log \det \sharp + \frac{1}{T}\Sigma (b^{-1}ay(t)-b^{-1}dz(t))'\sharp^{-1}(b^{-1}ay(t)-b^{-1}dz(t))$$

which leads to a simpler optimisation problem. We assume that the parameters of $\sharp$ are independent of the system parameters τ. When $\tilde{L}(\tau)$ is maximised with respect to $\sharp$ by putting

$$\tilde{\sharp} = \frac{1}{T}\Sigma (b^{-1}ay(t)-b^{-1}dz(t))(b^{-1}ay(t)-b^{-1}dz(t))' .$$

Now we are left with optimising $\det \tilde{\sharp}$ with respect to the system parameters. Let $\tau^{(k)}$ be the vector of parameters in $[a,b,d]$ at the k^{th} stage of the iterative procedure optimising $\det \tilde{\sharp}$. Denote $b^{-1}ay(t)-b^{-1}dz(t)$ by $\varepsilon(t,\tau)$ and define

$$\varepsilon^{(i)}(t,\tau^1) = \left.\frac{\partial \varepsilon(t,\tau)}{\partial \tau_i}\right|_{\tau = \tau^1}$$

Approximate $\varepsilon(t,\tau)$ by

$$\varepsilon(t,\tau) = \varepsilon(t,\tau^1) + \sum_{i}^{n} \varepsilon^{(i)}(t,\tau^1)\{\tau_i-\tau_i^1\} \qquad (3.5.2)$$

where n is the dimension of τ. Let

$$\hat{F}_{\ell k}(\tau^1) = \frac{1}{T}\Sigma \varepsilon^{(\ell)}(t,\tau^1)\varepsilon^{(k)}(t,\tau^1)'$$

$$\hat{W}_\ell(\tau^1) = \frac{1}{T}\Sigma\varepsilon^{(\ell)}(t,\tau^1)\varepsilon(t,\tau^1)'$$

$$\tilde{\Psi}(\tau^1) = \frac{1}{T}\Sigma\varepsilon(t,\tau^1)\varepsilon(t,\tau^1)' \ .$$

On the basis of (3.5.2) consider the minimisation with respect to τ of

$$(3.5.3)\quad \det\left\{\frac{1}{T}\Sigma\{\varepsilon(t,\tau^1)+\sum_i\varepsilon^{(i)}(t,\tau^1)(\tau_i-\tau_i^1)\}\{\varepsilon(t,\tau^1)+\sum_i\varepsilon^{(i)}(t,\tau^1)(\tau_i-\tau_i^1)\}'\right\}$$

in the place of $\det\tilde{\Psi}$. To minimise (3.5.3) we are led to the equations

$$(3.5.4)\quad \mathrm{tr}\left\{\tilde{\Psi}(\tau)^{-1}\{\hat{W}_\ell(\tau^1)+\sum_{k=1}^{n}\hat{F}_{\ell k}(\tau^1)(\tau_k-\tau_k^1)\}\right\} = 0 \ , \qquad \ell = 1,\ldots,n$$

while in the univariate case this leads to the solution of

$$\hat{W}_\ell(\tau^1)+\sum_{k=1}^{n}\hat{F}_{\ell k}(\tau^1)(\tau_k-\tau_k^1) = 0 \qquad \ell = 1,\ldots,n \ .$$

To solve (3.5.4) approximately in the multivariate case, we replace $\tilde{\Psi}(\tau)$ by $\tilde{\Psi}(\tau^1)$ in (3.5.4), an approximation that is likely to be of little consequence provided our estimates, τ, remain in a small enough neighbourhood of τ_0. Let

$$W_\ell(\tau^1) = \mathrm{tr}\{\tilde{\Psi}(\tau^1)^{-1}\hat{W}_\ell(\tau^1)\} = \frac{1}{T}\Sigma\varepsilon(t,\tau^1)'\,\tilde{\Psi}(\tau^1)^{-1}\varepsilon^{(\ell)}(t,\tau^1)$$

$$F_{\ell k}(\tau^1) = \mathrm{tr}\{\tilde{\Psi}(\tau^1)^{-1}\hat{F}_{\ell k}(\tau^1)\} \ .$$

Define $F(\tau^1)$ to be the $n\times n$ matrix with entries $F_{\ell k}(\tau^1)$ and $W(\tau^1)$ as the vector with elements $W_\ell(\tau^1)$. Then we obtain as an approximate solution to (3.5.4)

$$(3.5.5)\qquad \tilde{\tau} = \tau^1 - F(\tau^1)^{-1}W(\tau^1) \ .$$

This leads to an iterative procedure to optimise $\tilde{L}(\tau)$

$$\tau^{(k+1)} = \tau^{(k)} - \lambda_k F(\tau^{(k)})^{-1}W(\tau^{(k)})$$

where λ_k is a step length to be discussed later. In the univariate case (3.5.5) furnishes the solution as well but the elements of $F(\tau^1)$ and $W(\tau^1)$ are $\hat{F}_{\ell k}(\tau^1)$, $\hat{W}_\ell(\tau^1)$. The parameter vector $\tilde{\tau}$ obtained from (3.5.5) will be referred to as the Gauss-Newton estimate. Under quite mild conditions it can be shown (at least in the scalar case) that if $(\tau^1-\tau_0) = O_P(a_T)$ where $a_T \downarrow 0$ then the Gauss-Newton estimate $\tilde{\tau}$ satisfies

$$\tilde{\tau}-\tau_0 = -F(\tau_0)^{-1}W(\tau_0) + O_P\{\max\{a_T^2, a_T T^{-\frac{1}{2}}\}\} .$$

(See Fuller (1976) Chapter V for a proof.) In the particular situation where $y(t)$ is generated by an ARMAX process we can arrive at this kind of result directly. We do not give a proof here, but refer to the proof of Theorem 3.6.2 where the analogous result is proved for transfer function models in the scalar case. In this situation the proof uses no different ideas. For a proof of the central limit theorem for $F(\tau_0)^{-1}W(\tau_0)$ we refer to Hannan, Dunsmuir and Deistler (1980). The iterative procedure may be initialised using the parameter estimates τ^1 obtained at Stage II. The discussion has assumed that the true model structure (e.g. the Kronecker indices) is known, thus we also assume that the Stage II estimate of model order is consistent. Let the elements of τ be ordered as in §2.1, namely τ is the vector of those elements of

$$(\text{vec}[A(0)]', \text{vec}[A(1), B(1), D(1)]', \ldots)$$

that are not restricted to zero or one when $[a(z), b(z), d(z)]$ is in echelon form. Let V_0 be constructed from $A(u,v)$ as described below Theorem 2.1.2.

THEOREM 3.5.1. *Assume* $y(t)$ *is generated by an ARMAX process and* $\varepsilon(t)$, $z(t)$ *satisfy conditions B. If* $\tilde{\tau}$ *is the Gauss-Newton estimate obtained using (3.5.5) where* τ^1 *is the Stage II parameter estimate then*

$$T^{\frac{1}{2}}(\tilde{\tau}-\tau_0) \xrightarrow{\mathcal{L}} N(0,V_0^{-1}) \ .$$

If the procedure is iterated the same proof shows that all successive estimates are also asymptotically efficient.

We proceed to obtain an explicit form for the quantities used (3.5.5). Calculate $\varepsilon(t,\tau^1)$ recursively using

$$\sum_0^p \hat{B}(j)\varepsilon(t-j,\tau^1) = \sum_0^p \hat{A}(j)y(t-j) - \sum_1^p \hat{D}(j)z(t-j) \qquad (3.5.6)$$

where τ^1 is the system parameter estimate obtained at Stage II. This recursion is initialised by putting $\varepsilon(s,\tau^1) = 0$, $y(s) = 0$, $z(s) = 0$, $s \leqslant 0$. Since $\hat{a}(z)$, $\hat{b}(z)$, $\hat{d}(z)$ is in echelon form, $\hat{A}(0)-I_s$ only has non-zero elements below the main diagonal. As $\hat{A}(0) = \hat{B}(0)$ this means that the i^{th} component of $\varepsilon(t,\tau^1)$ is

$$\varepsilon_i(t,\tau^1) = \sum_{j=1}^{i-1} \hat{A}_{ij}(0)(y_j(t)-\varepsilon_j(t,\tau^1)) + \sum_1^p \hat{A}(j)y(t-j) - \sum_1^p \hat{B}(j)\varepsilon(t-j,\tau^1) - \sum_1^p \hat{D}(j)z(t-j)$$

so that to compute $\varepsilon_i(t,\tau^1)$ we only need $\varepsilon_j(t,\tau^1)$, $j < i$ together with $y(t)$ and 'past' $y(s)$, $z(s)$, $\varepsilon(s,\tau^1)$, $s < t$. Denote by e_{ij} the matrix with all of its entries zero except for "1" in the (i,j) place. The context where e_{ij} is used will specify its dimension. Then with, $\varepsilon(t,\tau^1) = \hat{b}^{-1}\hat{a}y(t) - \hat{b}^{-1}\hat{d}z(t)$, we have

$$\left.\frac{\partial\varepsilon(t,\tau)}{\partial A_{ij}(k)}\right|_{\tau=\tau^1} = \hat{b}^{-1}e_{ij}y(t-k) \ , \qquad k > 0$$

$$\left.\frac{\partial\varepsilon(t,\tau)}{B_{ij}(k)}\right|_{\tau=\tau^1} = -\hat{b}^{-1}e_{ij}(\hat{b}^{-1}\hat{a}y(t-k)-\hat{b}^{-1}\hat{d}z(t-k))$$

$$= -\hat{b}^{-1} e_{ij} \varepsilon(t,\tau^1) \ , \qquad k > 0$$

$$\left.\frac{\partial \varepsilon(t,\tau)}{\partial A_{ij}(0)}\right|_{\tau = \tau^1} = \hat{b}^{-1} e_{ij} \{y(t) - \varepsilon(t,\tau^1)\}$$

$$\left.\frac{\partial \varepsilon(t,\tau)}{\partial D_{ij}(k)}\right|_{\tau = \tau^1} = -\hat{b}^{-1} e_{ij} z(t-k) \ .$$

In these expressions $A_{ij}(k)$, $B_{ij}(k)$, $D_{ij}(k)$ represent the (i,j) elements of the matrices $A(k)$, $B(k)$, $D(k)$ that occur in the parameter vector τ. The calculation of these quantities may be streamlined by computing recursively

$$(3.5.7) \quad \sum_0^p \hat{B}(j)[\eta(t-j), \xi(t-j), \zeta(t-j)] = [y(t)', -\varepsilon(t,\tau^1)', -z(t)'] \otimes I_s$$

where $\eta(t)$, $\xi(t)$, $\zeta(t)$ have respectively s^2, s^2, sm columns and are set to zero when $t \leqslant 0$. For example, the $i+s(j-1)$ column of $\eta(t)$ is

$$\hat{b}^{-1}(0,\ldots,0,y_j(t),0,\ldots,0)' = \hat{b}^{-1} e_{ij} y(t)$$

where the vector $(0,\ldots,0,y_j(t),0,\ldots,0)'$ has its non-zero element in the i^{th} place. Let $X(t)'$ be the matrix consisting of those columns of $\eta(t-j)$, $\xi(t-j)$, $\zeta(t-j)$ (and $\eta(t) + \xi(t)$ if $\hat{A}(0) - I_s$ is not zero) corresponding to the freely varying elements of $\hat{A}(j)$, $\hat{B}(j)$, $\hat{D}(j)$ in τ^1. Then the Gauss-Newton estimate $\tilde{\tau}$ given by (3.5.5) is

$$(3.5.8) \quad \tilde{\tau} = \tau^1 - \{\frac{1}{T} \Sigma X(t) \{\tilde{\Sigma}(\tau^1)^{-1} \otimes I_s\} X(t)'\}^{-1}$$
$$\{\frac{1}{T} \Sigma X(t) \{\tilde{\Sigma}(\tau^1)^{-1} \otimes I_s\} \varepsilon(t,\tau^1)'\}$$

To implement this procedure the essential requirement is that $\hat{b}(z)$ be stable so that the $\varepsilon(t,\tau^1)$ and $\eta(t)$, $\xi(t)$, $\zeta(t)$ can be formed. As we have indicated in §3.4 that

one way to achieve this is to replace $\hat{b}(z)$ by $b_\lambda(z) = I_s + \lambda(\hat{b}(z)-I_s)$ where λ is selected close to unity and consistent with stability. It should be noted that if the zeros of $\det \hat{b}(z)$ (or $\det b_\lambda(z)$) are very close to $|z| = 1$ then the effects due to setting $\tilde{\varepsilon}(t)$, $y(t)$, $z(t)$ to zero, $t \leqslant 0$ in computing $\varepsilon(t,\tau^1)$ in (3.5.6) and initiating $\eta(t)$, $\xi(t)$, $\zeta(t)$ to zero $t \leqslant 0$ in (3.5.7) will die out very slowly. This effect can lead to poor estimates of $\varepsilon(t,\tau^1)$ and $\eta(t)$, $\xi(t)$, $\zeta(t)$ which in turn can lead to poor performance of the Gauss-Newton estimate, particularly when the sample size is small. For this reason it may be worthwhile to select λ above so that the zeros of $\det b_\lambda(z)$ are bounded some small distance away from $|z| = 1$.

The iterative procedure includes a 'step length' parameter λ_k the choice of which we discuss here. The Gauss-Newton procedure locates a direction along which the objective function decreases, but this direction and unit step length is only optimal if the objective function is truely quadratic. As $\det \tilde{\Phi}(\tau)$ is not quadratic, we stand to reduce the number of iterations of the Gauss-Newton procedure by selecting the step length in the given direction so that the objective function is decreased as much as possible. That is, if

$$\tau^{(k+1)}(\lambda) = \tau^{(k)} - \lambda F(\tau^{(k)})^{-1} W(\tau^{(k)})$$

we seek to find a λ to minimise the quantity $\det \tilde{\Phi}(\tau^{(k+1)}(\lambda))$. How accurately we should evaluate λ depends on the relative cost of this evaluation and the cost of one iteration of the Gauss-Newton procedure. The choice of step length has one other important consequence, namely that if in the echelon form $[a^{(k)}, b^{(k)}, d^{(k)}]$ corresponding to $\tau^{(k)}$,

$b^{(k)}(z)$ is stable, then for small enough step length λ, $b^{(k+1)}(z)$ will also be stable. For our problem a single evaluation of the objective function is considerably faster than one iteration of the Gauss-Newton procedure, so it is generally worthwhile expending some effort in selecting the step length. One economical approach is as follows. At the k^{th} iteration we have already evaluated $\det \tilde{\Phi}(\tau^{(k+1)}(0)) = \det \tilde{\Phi}(\tau^{(k)})$. At the $(k+1)^{st}$ iteration evaluate $\det \tilde{\Phi}(\tau^{(k+1)}(1))$ and

$$G = \{(\tau^{(k+1)}(1)-\tau^{(k)})W(\tau^{(k)}\}/\| \tau^{(k+1}(1)-\tau^{(k)} \|$$

This is the directional derivative of $\det \tilde{\Phi}(\tau)$ at $\tau^{(k)}$ in the direction $\tau^{(k+1)}(1)-\tau^{(k)}$. Now we fit a quadratic curve through the points $\{0, \det \tilde{\Phi}(\tau^{(k)})\}$, $\{1, \det \tilde{\Phi}(\tau^{(k+1)}(1))\}$ having gradient G at 0. The stationary point of this curve is at

$$\lambda_1 = G/2\{\det \tilde{\Phi}(\tau^{(k+1)}(1))-\det \tilde{\Phi}(\tau^{(k)})-G\} .$$

Several cases arise.

(i) $\det \tilde{\Phi}(\tau^{(k+1)}(1)) \geqslant \det \tilde{\Phi}(\tau^{(k)}))$. In this case $\lambda_1 \in [0,1]$. Evaluate successively $\det \tilde{\Phi}(\tau^{(k+1)}(\lambda))$ for $\lambda = \lambda_1, \lambda_1/2, \lambda_1/4, \ldots$ selecting the step length as the first such λ for which

$$\det \tilde{\Phi}(\tau^{(k+1)}(\lambda)) < \det \tilde{\Phi}(\tau^{(k)}) .$$

(ii) $\det \tilde{\Phi}(\tau^{(k+1)}(1)) < \det \tilde{\Phi}(\tau^{(k)})$. If $\lambda_1 < 0$ then it is most likely that the optimal step length is >1. In this case proceed to evaluate $\det \tilde{\Phi}(\tau^{(k+1)}(\lambda))$ for at least one step length $\lambda_2 > 1$ (take for example $\lambda_2 = 2$) and select the step length as 1 or λ_2 for which

$\det \tilde{\Sigma}(\tau^{(k+1)}(\lambda))$ is smallest. The density of the line search will depend on the cost of evaluating the objective function.

If $\lambda_1 > 0$ select the step length as λ_1 when

$\det \tilde{\Sigma}(\tau^{(k+1)}(\lambda_1)) < \det \tilde{\Sigma}(\tau^{(k+1)}(1))$ and 1 otherwise.

Of course many other methods for selecting step lengths are available. For a discussion of some of them, and more details and refinements on this one, see Bard (1974).

The greatest computational burden in the maximum likelihood calculation is in computing the quantities $\eta(t)$, $\xi(t)$, $\zeta(t)$, $\varepsilon(t,\tau)$ and their covariances. In comparison to this the additional work in solving (3.5.8) (where a rather large matrix is to be inverted) is quite small and it hardly seems worthwhile in attempting to obtain a recursive solution of (3.5.8) particularly in the multivariate case when any such recursion would be exceedingly complicated. When fitting an ARMA(p,p,p) model (or one from $\mathcal{V}(ps)$) the Levinson-Whittle recursion can be adpated to give the recursive solution of (3.5.6). The recursive solution can be justified in these cases when we want to estimate model order at Stage III as we shall now describe. First we give the recursive algorithm.

Denote by $C_p(j)$, $j = 1,\ldots,p$ the coefficients obtained by regressing $\varepsilon(t,\tau^1)$ on $w(t-1),\ldots,w(t-p)$ in the model

$$\varepsilon(t,\tau^1) + \sum_1^p C_p(j)w(t-j) = e_p(t) \tag{3.5.9}$$

and let $\tilde{\Sigma}_p$ be the residual variance. Here

$$w(t) = \begin{pmatrix} \eta(t) \\ \xi(t) \\ \zeta(t) \end{pmatrix} .$$

Calculate

$$G(k) = \frac{1}{T}\Sigma w(t)\{\tilde{\Phi}(\tau^1)^{-1}\}w(t+k)'$$

$$G_{\varepsilon w}(k) = \frac{1}{T}\Sigma \varepsilon(t,\tau^1)\{\tilde{\Phi}(\tau^1)^{-1}\}w(t-k)'$$

where the $G(k)$ are used as inputs to the Levinson-Durbin recursion as in algorithm 3.2.2 . At the p^{th} stage of the recursive solution of (3.5.8) assume we have calculated $C_p(j)$, $j = 1,\ldots,p$ and the residual mean square $\tilde{\Phi}_p$. As well as this, the coefficients $\tilde{F}_p(j)$, $j = 1,\ldots,p$ and the residual mean square $\tilde{S}_p$ from the regression of $w(t-p-1)$ on $w(t-1),\ldots,w(t-p)$ have been obtained as output from the Levinson-Durbin recursion. At the $(p+1)^{st}$ stage we add $w(t-p-1)$ into the regression (3.5.9). The residual

$$\sum_0^p \tilde{F}_p(p-j)w(t-j-1), \quad \tilde{F}_p(0) = I$$

is Toeplitz orthogonal to the space spanned by the regressors $w(t-1),\ldots,w(t-p)$ hence

$$C_{p+1}(p+1) = -\{\sum_0^p G_{\varepsilon w}(j+1)'\tilde{F}_p(p-j)'\}\tilde{S}_p^{-1} .$$

Then

$$e_{p+1}(t) = \varepsilon(t,\tau^1) + \sum_1^p C_p(j)w(t-j) + C_{p+1}(p+1)\{\sum_0^p \tilde{F}_p(p-j)w(t-j-1)\}$$

and

$$C_{p+1}(j) = C_p(j) + C_{p+1}(p+1)\tilde{F}_p(p-j+1) .$$

The residual variance can be updated as

$$\tilde{\Phi}_{p+1} = \tilde{\Phi}_p - C_{p+1}(p+1)\tilde{S}_p C_{p+1}(p+1)' .$$

Now another pass through the Levinson-Durbin recursion is needed to compute $\tilde{F}_{p+1}(j)$, $j = 1,\ldots,p+1$ and $\tilde{S}_{p+1}$ in

preparation for adding $w(t-p-2)$ into the regression (3.5.9) at the next stage. Instead of using the Levinson-Durbin recursion to fit ARMA(p,p) models, we could use the procedure of Monden, Yamada and Arimoto (1982) to introduce, separately, $\eta(t-p-1)$ or $\xi(t-p-1)$ and then we could fit ARMA(p,q) models in a recursive way.

The recursions described here are of particular importance if we are to obtain order estimates at Stage III as described below. We have shown in Theorem 3.4.2 that if BIC is used to select model order the estimated orders may be too high. Similar results can be obtained in the multivariate case (see the comments below theorem 3.4.4). Here we consider another method for obtaining consistent order estimates that is based on the Gauss-Newton estimates obtained at Stage III. We shall restrict the discussion to the univariate ARMA case. Assume $\hat{p},\hat{q}$ are the order estimates obtained at Stage II and τ^1 is the vector of parameter estimates $\hat{a}(j)$, $\hat{b}(j)$. Using our previous notation (defined above 3.5.2)

$$\begin{aligned}\sum_{i=1}^{\hat{p}+\hat{q}} \tau_i^1 \varepsilon^{(i)}(t,\tau^1) &= \sum_1^{\hat{p}} \hat{a}(j)\eta(t-j) + \sum_1^{\hat{q}} \hat{b}(j)\xi(t-j) \\ &= (\hat{a}-1)\hat{b}^{-1}y(t) - (\hat{b}-1)\hat{b}^{-1}\varepsilon(t,\tau^1) \\ &= -\eta(t) - \xi(t)\end{aligned} \tag{3.5.10}$$

In the univariate case using the above identity we can reduce (3.5.3) to

$$\frac{1}{T}\Sigma\{\varepsilon(t,\tau^1)+\eta(t)+\xi(t) + \sum_1^{\hat{p}+\hat{q}} \tau_i \varepsilon^{(i)}(t,\tau^1)\}^2$$

This minimisation lends itself to recursive solution using the algorithm described above with

$$G(k) = \frac{1}{T} \Sigma \begin{bmatrix} \eta(t) \\ \xi(t) \end{bmatrix} [\eta(t+k), \xi(t+k)]$$

$$G_{\varepsilon w}(k) = \frac{1}{T} \Sigma (\varepsilon(t,\tau^1)+\eta(t)+\xi(t)) [\eta(t-k), \xi(t-k)].$$

Let $\tilde{\Sigma}_{pq}$ be the residual mean square obtained by regressing $\varepsilon(t,\tau^1)+\eta(t)+\xi(t)$ on $\eta(t-j)$, $j = 1,\ldots,p$, $\xi(t-j)$, $j = 1,\ldots,q$. Select the order $\{\tilde{p},\tilde{q}\}$ as that pair of integers minimising

$$(3.5.11) \qquad \log \tilde{\Sigma}_{pq} + (p+q)\log T/T \ , \ p,q < P \ .$$

THEOREM 3.5.2. *Assume* $y(t)$ *is generated by a univariate ARMA process and* $\varepsilon(t)$ *satisfies conditions B. If the orders* $\{\tilde{p},\tilde{q}\}$ *at Stage III are obtained by optimising (3.5.11) then*

$$\tilde{p} \to p_0 \ , \quad \tilde{q} \to q_0$$

in probability.

This order estimation procedure could be applied at any stage of the iterative maximum likelihood procedure above. It has certain appeal because it is based on estimates of the parameters that are closer to the maximum likelihood estimates. Some results of simulations comparing the order estimates obtained at Stages II, IIa and III are given in Table 3.5.1. In these simulations ARMA(p,p) models were estimated and the orders at each stage estimated using BIC. From these limited results it appears that there is some merit in this approach.

Table 3.5.1

Frequencies of choice of order of ARMA(p,p) models.

	Stage II				Stage IIa			Stage III		
Order	1	2	3	⩾4	1	2	3	1	2	3
	$y(t) = \varepsilon(t) + .8\ \varepsilon(t-1)$									
Sample Size										
200	44	4	2	-	48	1	1	49	1	-
500	47	2	1	-	50	-	-	50	-	-
1000	44	6	-	-	47	3	-	48	2	-
	$y(t) = \varepsilon(t) + .9\ \varepsilon(t-1)$									
200	42	7	1	-	49	1	-	50	-	-
500	43	7	-	-	49	1	-	50	-	-
1000	37	12	-	1	49	1	-	49	1	-
	$y(t) = \varepsilon(t) + .95\ \varepsilon(t-1)$									
200	41	8	1	-	49	1	-	50	-	-
500	38	11	1	-	47	3	-	50	-	-
1000	31	17	1	1	49	1	-	49	1	-
	$y(t) = \varepsilon(t) + \varepsilon(t-1) + .89\ \varepsilon(t-2)$									
200	-	25	34	1	-	43	7	-	46	4
500	-	19	28	3	-	46	4	-	49	1
1000	-	8	38	4	-	45	5	-	48	2

§3.6 Estimation of transfer function models

In this section we consider the estimation of the transfer function model

$$y(t) = \ell_0 z(t) + \eta(t) \; ; \; \ell_0(z) = \sum_1^\infty L_0(j) z^{-j} \tag{3.6.1}$$

where $\ell_0(z)$ is a rational transfer function satisfying (1.1.4). Assume the noise model is an ARMA process

$$a_0 \eta(t) = b_0 \dot{\varepsilon}(t) \tag{3.6.2}$$

where $a_0(z)$, $b_0(z)$ satisfy (1.1.2). Here we consider only models with scalar input and scalar output. Estimation of this kind of model has been considered in Young (1976), Ljung and Soderstrom (1983). These authors are concerned with algorithms that are recursive in time. They also assume that the model structure is known. Here we consider the problem of structure estimation under the quite general conditions used in the previous sections. To describe the model structure, orders need to be specified for the output and noise models separately. The structure estimation algorithms proposed here, like our algorithms for ARMAX estimation are based on obtaining an estimate of the innovations sequence.

Stage (i) *Obtain estimates* $\hat{\phi}_h(j)$, $\hat{\psi}_h(j)$, $j = 1,\ldots,h$ *to define*

$$\hat{\varepsilon}_h(t) = y(t) + \sum_1^h \hat{\phi}_h(j) y(t-j) - \sum_1^h \hat{\psi}_h(j) z(t-j)$$

exactly as in Stage I. The order h *is selected by optimising* AIC(h) *or* BIC(h) *as before.*

The rational function $\ell_0(z)$ is decomposed as $c_0(z)^{-1}d_0(z)$, where $c_0(z) = 1 + \sum_1^{n_{03}} c_0(j)z^{-j}$, $d_0(z) = \sum_1^{n_{04}} d_0(j)z^{-j}$, $c_0(z)$, $d_0(z)$ have no common factors and satisfy (1.1.2). Similarly let $a_0(z) = 1 + \sum_1^{n_{01}} a_0(j)z^{-j}$, $b_0(z) = 1 + \sum_1^{n_{02}} b_0(j)z^{-j}$. Let $\hat{\phi}_h(z) = 1 + \sum_1^h \hat{\phi}_h(j)z^{-j}$, $\hat{\psi}_h(z) = \sum_1^h \hat{\psi}_h(j)z^{-j}$ where h is the optimal order used to calculate $\hat{\varepsilon}_h(t)$ in Stage (i). The functions $\hat{\phi}_h$, $\hat{\psi}_h$ are estimates of $\phi_0 = b_0^{-1}a_0$, $\psi_0 = (b_0c_0)^{-1}a_0d_0$ respectively. Cancelling common factors in b_0c_0, a_0d_0, $\psi_0 = (\tilde{b_0c_0})^{-1}(\tilde{a_0d_0})$. Let ρ_0 be the modulus of a zero of b_0 or $(\tilde{b_0c_0})$ nearest $|z| = 1$. When conditions A hold and BIC(h) is used to determine the order h at Stage (i),

$$(3.6.3) \qquad h = \{\log T/(-2 \log \rho_0)\}(1+o(1)) \text{ a.s.}$$

as in Theorem 3.3.4. Similarly, when conditions B hold and AIC(h) is used to estimate order, (3.6.3) holds in probability (theorem 3.3.6).

We have the relationship

$$\phi_0 y(t) + (c_0 - 1)\{\phi_0 y(t) - \varepsilon(t)\} - d_0\phi_0 z(t) = \varepsilon(t) .$$

This suggests that the coefficients $c_0(j)$, $d_0(j)$ can be estimated by regressing $\hat{\phi}_h y(t)$ on $\hat{\phi}_h y(t-j) - \hat{\varepsilon}_h(t-j)$, $j = 1,\ldots,n_{03}$, $\hat{\phi}_h z(t-j)$, $j = 1,\ldots,n_{04}$. Here $\hat{\psi}_h z(t) = \hat{\phi}_h y(t) - \hat{\varepsilon}_h(t)$. This leads to the following procedure to estimate the orders n_{03}, n_{04} as well as the parameters of $c_0(z)$, $d_0(z)$.

Stage (iia) *For each pair* $\{n_3, n_4\}$ *where* $n_3, n_4 \leqslant N$, N *fixed, minimise*

(3.6.4) $$\frac{1}{T}\sum_{1}^{T}\{\hat{\phi}_h y(t)+(c-1)\hat{\psi}_h z(t)-d\hat{\phi}_h z(t)\}^2$$

for $c(z) = 1 + \sum_{1}^{n_3} c(j)z^{-j}$, $d(z) = \sum_{1}^{n_4} d(j)z^{-j}$. *Let* $\hat{\mathfrak{k}}_{n_3,n_4}$ *be the minimum of (3.6.4). The estimates* $\{\hat{n}_3,\hat{n}_4\}$ *are obtained by optimising the criterion*

$$\log \hat{\mathfrak{k}}_{n_3,n_4} + (n_3+n_4)\log T/T \ , \ n_3,n_4 \leqslant N \ .$$

Let

$$\hat{c}(z) = 1+\sum_{1}^{\hat{n}_3}\hat{c}(j)z^{-j} \ , \ \hat{d}(z) = \sum_{1}^{\hat{n}_4}\hat{d}(j)z^{-j}$$

be the estimates of $c_0(z)$, $d_0(z)$ *at the optimal order* $\{\hat{n}_3,\hat{n}_4\}$.

The minimisation of (3.6.4) is effected by regression. Thus, as in Stage II, estimates of the required covariances are computed using formulae analogous to (3.4.2) without the necessity of forming the filtered quantities $\hat{\phi}_h z(t)$, $\hat{\psi}_h z(t)$, $\hat{\phi}_h y(t)$. If we choose to restrict $n_3 = n_4$, then the computations may be effectively done using the Levinson-Durbin recursion as discussed in §3.5. Estimate the noise model as follows:

<u>Stage (iib)</u> *Let* $\hat{\varepsilon}_h(t)$ *be defined as in Stage (i) and form* $\hat{\eta}(t) = y(t) - \hat{c}^{-1}\hat{d}z(t)$. *For* $n_1,n_2 \leqslant N$ *minimise*

(3.6.5) $$\frac{1}{T}\Sigma\{a\hat{\eta}(t) - (b-1)\hat{\varepsilon}(t)\}^2$$

for $a(z) = 1+\sum_{1}^{n_1}a(j)z^{-j}$, $b(z) = 1+\sum_{1}^{n_2}b(j)z^{-j}$. *Let* $\hat{\mathfrak{k}}_{n_1,n_2}$ *be the minimum of (3.6.5). Now the orders* $\{\hat{n}_1,\hat{n}_2\}$ *are estimated using the techniques described in Stage II,* §3.4. *Let*

$$\hat{a}(z) = 1 + \sum_1^{n_1} \hat{a}(j) z^{-j}, \hat{b}(z) = 1 + \sum_1^{n_2} \hat{b}(j) z^{-j}$$

be the estimates of $a_0(z)$, $b_0(z)$ *at the optimal orders.*

It is not possible to use formulas such as (3.4.2) to compute the covariances required in minimising (3.6.5), hence we need to compute $\hat{\varepsilon}_h(t)$, $\hat{\eta}(t)$ explicitly. The estimates $\hat{\eta}(t)$ are computed recursively from

$$\hat{\eta}(t) = y(t) + \sum_1^{\hat{n}_3} \hat{c}(j)\{y(t-j) - \hat{\eta}(t-j)\} - \sum_1^{\hat{n}_4} \hat{d}(j) z(t-j),$$

after initialising $y(s)$, $z(s)$, $\hat{\eta}(s)$ to zero for $s \leqslant 0$.

THEOREM 3.6.1. *Assume the true model is given by* (3.6.1), (3.6.2) *where* a_0, b_0, c_0, d_0 *have orders* n_{01}, n_{02}, n_{03}, n_{04} *respectively. Assume that* $z(t)$, $\varepsilon(t)$ *satisfy conditions A and the order* h *from Stage (i) is selected by optimising* $\mathrm{BIC}(h)$, $h \leqslant H_T$. *If the order* $\{\hat{n}_3, \hat{n}_4\}$ *at Stage (iia) is selected by optimising*

$$\log \hat{\sigma}^2_{n_3,n_4} + (n_3+n_4) c_0 \log\log T/T \ , \ n_3, n_4 \leqslant N, \ N < \infty$$

and c_0 *is sufficiently large, then, almost surely*

$$\hat{n}_3 \to n_{03} \ , \ \hat{n}_4 \to n_{04} \ .$$

Turning to Stage (iib) we shall show that all the results pertaining to Stage II apply in this case as well. In particular we are interested in Theorem 3.4.2 and corollary 3.4.3 as well as the results for the alternative procedures from Theorems 3.4.5, 3.4.6. These theorems are all proved by evaluating particular terms occurring in regressions. Define $\tilde{\varepsilon}(t)$ as the residual when a high order autoregression is fitted to the true noise sequence $\eta(t)$ as in Stage I.

The proofs of the theorems in §3.4 carry over to this case if we show that replacing $\hat{\eta}(t)$ by $\eta(t)$ and $\hat{\varepsilon}_h(t)$ by $\tilde{\varepsilon}(t)$ in Stage (iib) produces sufficiently small changes in the covariances used in constructing the regressions required in the proofs. We leave these details to §5.4.

The parameter estimates obtained in Stages (iia), (iib) are not efficient. We consider a third stage closely related to Stage III to obtain asymptotically efficient parameter estimates. The estimates $\{\hat{n}_1,\hat{n}_2\}$ from Stage (iia) are consistent; here we also assume that the method chosen to give estimates $\{\hat{n}_3,\hat{n}_4\}$ in Stage (iib) is also consistent. As in §3.5 we consider optimising an objective function

$$\tilde{L}(\tau) = \log \sharp + \frac{1}{T} \Sigma\{b^{-1}a(y(t)-c^{-1}dz(t))\}^2\sharp^{-1}$$

where $\tau^1 = [a(1),\ldots,a(n_1),b(1),\ldots,b(n_2),c(1),\ldots,d(n_4)]$. Assuming $\sharp$ is independent of the system parameters we are led to minimise

$$\frac{1}{T} \Sigma\{b^{-1}a(y(t)-c^{-1}dz(t))\}^2 .$$

The motivation for the following algorithm is similar to that leading to 3.5.5.

Stage (iii) *Using the orders* $\hat{n}_i$, $i = 1,\ldots,4$ *and the estimates* $\hat{a}$, $\hat{b}$, $\hat{c}$, $\hat{d}$ *from Stages* (iia), (iib) *construct the quantities*

$$\xi_a(t) = -\hat{b}^{-1}\hat{\eta}(t) \qquad \xi_b(t) = \hat{b}^{-1}\tilde{\varepsilon}(t)$$

(3.6.6)
$$\xi_c(t) = -\hat{c}^{-1}\,\tilde{\psi}z(t) \qquad \xi_d(t) = \hat{c}^{-1}\,\tilde{\phi}z(t)$$

$$\hat{\eta}(t) = y(t)-\hat{c}^{-1}\hat{d}z(t) \ , \quad \tilde{\varepsilon}(t) = \tilde{\phi}\,\hat{\eta}(t)$$

where $\tilde{\phi} = \hat{b}^{-1}\hat{a}$ *and* $\tilde{\psi} = \hat{c}^{-1}\hat{d}\,\tilde{\phi}$. *Let*

$$\xi_{ab}(t) = [\xi_a(t-1),\ldots,\xi_a(t-\hat{n}_1),\xi_b(t-1),\ldots,\xi_b(t-\hat{n}_2)]'$$

$$\xi_{cd}(t) = [\xi_c(t-1),\ldots,\xi_c(t-\hat{n}_3),\xi_d(t-1),\ldots,\xi_d(t-\hat{n}_4)]'$$

and $\hat{\tau} = [\hat{a}(1),\ldots,\hat{a}(\hat{n}_1),\hat{b}(1),\ldots,\hat{d}(\hat{n}_4)]'$. *Then*

$$\Delta\tau = \begin{bmatrix} \frac{1}{T}\Sigma\xi_{ab}(t)\xi_{ab}(t)' & 0 \\ 0 & \frac{1}{T}\Sigma\xi_{cd}(t)\xi_{cd}(t)' \end{bmatrix}^{-1} \begin{bmatrix} \frac{1}{T}\Sigma\xi_{ab}(t)\tilde{\varepsilon}(t) \\ \frac{1}{T}\Sigma\xi_{cd}(t)\tilde{\varepsilon}(t) \end{bmatrix}$$

and

$$\tilde{\tau} = \hat{\tau}+\Delta\tau$$

are the required asymptotically efficient estimates.

The limiting form for the covariance matrix of the maximum likelihood estimator τ_{ML} of $\tau_0 = [a_0(1),\ldots,a_0(n_{01}), b_0(1),\ldots,d_0(n_{04})]'$, for the related model where $\ell_0(z)$ has finite order, is obtained in Pierce (1971) under stronger conditions on the innovations sequence. Using a similar procedure (see also Hannan, Dunsmuir and Deistler, 1980) we proceed to construct the covariance matrix for the maximum likelihood estimate. Define

$$A_{11}(\ell,k) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{e^{i(\ell-k)\omega}}{|a_0|^2}\,d\omega$$

$$A_{12}(\ell,k) = -\frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{e^{i(\ell-k)\omega}}{a_0 b_0^*}\,d\omega\ ;\ A_{21}(\ell,k) = A_{12}(k,\ell)$$

$$A_{22}(\ell,k) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{e^{i(\ell-k)\omega}}{|b_0|^2}\,d\omega$$

$$A_{33}(\ell,k) = \tfrac{1}{4}\pi^{-1}\int_{-\pi}^{\pi} \frac{|\psi_0|^2}{|c_0|^2}\, e^{i(\ell-k)\omega}\, dF_z(\omega)$$

$$A_{34}(\ell,k) = -\tfrac{1}{4}\pi^{-1}\int_{-\pi}^{\pi} \frac{\psi_0\phi_0}{|c_0|^2}\, e^{i(\ell-k)\omega}\, dF_z(\omega)\ ;\ A_{43}(\ell,k) = A_{34}(k,\ell)$$

$$A_{44}(\ell,k) = \tfrac{1}{4}\pi^{-1}\int_{-\pi}^{\pi} \frac{|\phi_0|^2}{|c_0|^2}\, e^{i(\ell-k)\omega}\, dF_z(\omega)$$

and $A_{ij}(\ell,k) = 0$ for other $i,j = 1,\ldots,4$. Let A_{ij} be the $n_i \times n_j$ matrix with the (k,ℓ) entry being $A_{ij}(k,\ell)$, $k = 1,\ldots,n_i$, $\ell = 1,\ldots,n_j$. Finally A is defined to be the matrix of blocks $[A_{ij}]$, $i,j = 1,\ldots,4$. Then from Hannan, Dunsmuir and Deistler (1980)

$$T^{\frac{1}{2}}(\tau_{ML} - \tau_0) \xrightarrow{\mathcal{L}} N(0,A^{-1}) .$$

We prove the following theorem when $E\{\varepsilon(t)^2 \mid F_{t-1}\} = \tfrac{1}{4}\pi$ so as to have the simpler form for the covariance matrix (see Theorem 2.4.1).

THEOREM 3.6.2. *Assume conditions B hold. Let the order at Stage (i) be selected using* AIC *or* BIC. *Assume the order estimates from Stages (iia), (iib) are consistent. If* $\tilde{\tau}$ *is the vector of parameters obtained at Stage (iii) then*

$$T^{\frac{1}{2}}(\tilde{\tau} - \tau_0) \xrightarrow{\mathcal{L}} N(0,A^{-1}) .$$

As we have already discussed in §3.5 it would be worthwhile iterating the Stage (iii) procedure. Let the output after the i^{th} iteration of Stage (iii) be $\tilde{a}_{(i)}$, $\tilde{b}_{(i)}$, $\tilde{c}_{(i)}$, $\tilde{d}_{(i)}$ which give $\tilde{\tau}_{(i)}$. Define $\tilde{\phi}_{(i)} = \tilde{b}_{(i)}^{-1}\tilde{a}_{(i)}$ and $\tilde{\psi}_{(i)} = \tilde{c}_{(i)}^{-1}\tilde{d}_{(i)}\tilde{\phi}_{(i)}$. These quantities are used in (3.6.6) in

place of $\hat{a},\hat{b},\ldots$. Then the coordinates of the i+1 iteration are given by $\tilde{\tau}_{(i+1)} = \tilde{\tau}_{(i)} + \lambda\Delta\tau$. Here λ is a step length which should be adjusted to ensure stability of $\tilde{a}_{(i)}$, $\tilde{c}_{(i)}$ and the decrease of the objective function. These points have been amplified in §3.5.

CHAPTER 4

LIMIT THEOREMS AND TECHNICAL LEMMAS

§4.1 Uniform convergence of covariances

The key requirement necessary to establish the theorems in chapter 3 is the uniform rate of convergence of the covariances and cross covariances of $y(t)$ and $z(t)$. For many purposes the strongest possible rate of convergence will be needed for these covariances, namely the rate in the law of the iterated logarithm. This rate will need to be uniform for lags up to $H_T = (\log T)^a$, $a < \infty$. Results of this kind, under conditions similar to those in chapter 3, first appeared in An, Chen and Hannan (1982). In proving these results (2.1.2) (i.e. $E\{\varepsilon(t)\varepsilon(t)' \mid \mathcal{F}_{t-1}\} = \Sigma_0$) was imposed. Define $\mathring{\gamma}_u(k) = \sum_{j=0}^{\infty} K(j)\mathring{\Sigma}K(j+k)'$, where $\mathring{\Sigma} = \frac{1}{T}\sum_1^T \varepsilon(t)\varepsilon(t)'$ as in §3.1 and

$$(4.1.1) \qquad u(t) = \sum_0^{\infty} K(j)\varepsilon(t-j), \quad \sum_0^{\infty} \|K(j)\|^2 < \infty, \quad K(0) = I_s .$$

When (2.1.2) is weakened to (2.1.3) (i.e. $E\{\varepsilon(t)\varepsilon(t)' \mid \mathcal{F}_{-\infty}\} = \Sigma_0$) we have the following results from Hannan and Kavalieris (1983a):

THEOREM 4.1.1. *Assume* (4.1.1), (2.1.1), (2.1.3) *hold*, $E\{\|\varepsilon(t)\|^4\} < \infty$, *and* $k(z) = \sum_{j=0}^{\infty} K(j)z^{-j}$ *is rational. Let* $H_T = (\log T)^a$, $a < \infty$. *Then, almost surely,*

$$\sup_{0 \leq t \leq H_T} \|\hat{\gamma}_u(t) - \mathring{\gamma}_u(t)\| = O\{(\log\log T/T)^{\frac{1}{2}}\} .$$

THEOREM 4.1.2. *Assume* (4.1.1), (2.1.1), (2.1.3), (3.1.5) *hold*, $E\{\|\varepsilon(t)\|^4\} < \infty$. *Then, almost surely,*

$$\sup_{0 \leq t \leq \infty} \|\hat{\gamma}_u(t) - \mathring{\gamma}_u(t)\| = O\{(\log T/T)^{\frac{1}{2}}\} .$$

(Here $\hat{\gamma}_u(t) = 0$ if $t \geqslant T$.) The effect of (2.1.2) is to make $\varepsilon(t)\varepsilon(t)' - \Sigma_0$ into martingale differences (which are stationary and have finite variance) when the law of the iterated logarithm for martingales (Stout, 1970) shows that $\| \mathring{\Sigma} - \Sigma_0 \| = O\{(\log \log T/T)^{\frac{1}{2}}\}$ a.s. Thus when (2.1.2) holds $\mathring{\gamma}_u(t)$ may be replaced by $\gamma_u(t)$ in theorems 4.1.1, 4.1.2, giving the results in An, Chen and Hannan (1982). When $u(t)$ is scalar these results may be written in terms of the autocorrelations as in Hannan and Kavalieris (1983a). Indeed put $r_u(t) = \hat{\gamma}_u(t)/\hat{\gamma}_u(0)$, $\rho_u(t) = \gamma_u(t)/\gamma_u(0) = \mathring{\gamma}_u(t)/\mathring{\gamma}_u(0)$ then if $t \leqslant H_T$

$$
\begin{aligned}
r_u(t) - \rho_u(t) &= \{\mathring{\gamma}_u(0)\hat{\gamma}_u(t) - \hat{\gamma}_u(0)\mathring{\gamma}_u(t)\}/\hat{\gamma}_u(0)\mathring{\gamma}_u(0) \\
&= \{\mathring{\gamma}_u(0)(\hat{\gamma}_u(t) - \mathring{\gamma}_u(t)) + \mathring{\gamma}_u(t)(\mathring{\gamma}_u(0) - \hat{\gamma}_u(0)\}/\hat{\gamma}_u(0)\mathring{\gamma}_u(0) \\
&= O\{(\log \log T/T)^{\frac{1}{2}}\} \quad \text{a.s.}
\end{aligned}
$$

using theorem 4.1.1 and noting that $\hat{\gamma}_u(0)\mathring{\gamma}_u(0)$ is bounded away from zero. Thus under the conditions of theorems 4.1.1, 4.1.2 we have, respectively

$$\sup_{1 \leqslant t \leqslant H_T} |r_u(t) - \rho_u(t)| = O\{(\log \log T/T)^{\frac{1}{2}}\} \text{ a.s.}$$

$$\sup_{1 \leqslant t < \infty} |r_u(t) - \rho_u(t)| = O\{(\log T/T)^{\frac{1}{2}}\} \quad \text{a.s.}$$

We also need similar rates of convergence of $\frac{1}{T}\Sigma\varepsilon(t)z(t+k)'$. The following theorem is proved in Hannan and Kavalieris (1984a).

THEOREM 4.1.3. *Let* $\varepsilon(t)$ *satisfy* (2.1.1), (2.1.3), $E\{\|\varepsilon(t)\|^4\} < \infty$ *and* $z(t)$ *be stationary and ergodic and satisfy* (2.1.5) *and* $E\{\|z(t)\|^4\} < \infty$. *If* $H_T = (\log T)^a$, $a < \infty$ *then*

$$\sup_{|k| \leq H_T} \| \frac{1}{T} \sum_{m_1}^{T+m_2} \varepsilon(t) z(t-k)' \| = O\{(\log \log T/T)^{\frac{1}{2}}\} \text{ , a.s.}$$

$m_1 = k+1$, $m_2 = 0$ if $k > 0$, or $m_1 = 1$, $m_2 = k$ if $k \leq 0$.
We have the following lemma, the proof of which will be used repeatedly in chapter 5.

LEMMA 4.1.4. *Let* $u(t) = \sum_0^\infty K(j)\varepsilon(t-j)$, $v(t) = \sum_1^\infty L(j)z(t-j)$ *where the transfer functions* $k(z) = \sum_0^\infty K(j)z^{-j}$, $\ell(z) = \sum_1^\infty L(j)z^{-j}$ *are rational. Assume both* $\varepsilon(t)$, $z(t)$ *are stationary and ergodic with a finite* 4^{th} *moment. If* $K_T = c_0 \log T$, c_0 *sufficiently large, then almost surely and uniformly for* $n \leq (\log T)^a$, $a < \infty$,

$$(4.1.2) \quad \frac{1}{T} \sum_{n+1}^{T} u(t)v(t-n)' = \sum_{\ell,k=1}^{K_T} \sum K(\ell) \{\frac{1}{T} \sum_{m_1}^{T-m_2} \varepsilon(t)z(t-n+\ell-k)'\}L(k)' + o(T^{-\frac{1}{2}})$$

where $m_1 = n-\ell+k+1$, $m_2 = 0$ if $n+k > \ell$, or $m_1 = 1$, $m_2 = \ell-n-k$ *if* $n+k \leq \ell$.

Proof: The (i,j) element of $\frac{1}{T} \sum_{n+1}^{T} \varepsilon(t-\ell)z(t-n-k)'$ is

$$\frac{1}{T} \sum_{n+1}^{T} \varepsilon_i(t-\ell)z_j(t-n-k) \leq \frac{1}{T} \left\{ \sum_{n+1}^{T} \varepsilon_i(t-\ell)^2 \right\}^{\frac{1}{2}} \left\{ \sum_{n+1}^{T} z_j(t-n-k)^2 \right\}^{\frac{1}{2}}$$

$$\leq \frac{1}{T} \{\sum_1^T \varepsilon_i(t)^2 + \sum_{-\ell}^{0} \varepsilon_i(t)^2\}^{\frac{1}{2}} \{\sum_1^T z_j(t)^2 + \sum_{-n-k}^{0} z_j(t)^2\}^{\frac{1}{2}}$$

$$= o\{(1 + \ell/T)^{\frac{1}{2}}(1 + (n+k)/T)^{\frac{1}{2}}\} .$$

The last bound is because both $\varepsilon(t)$ and $z(t)$ are ergodic. Now if c_0 is sufficiently large, $\sum_{K_T}^{\infty} \ell^{\frac{1}{2}} \| K(\ell) \| = o(T^{-\frac{1}{2}})$, $\sum_{K_T}^{\infty} \ell^{\frac{1}{2}} \| L(\ell) \| = o(T^{-\frac{1}{2}})$ hence

$$\| \sum_{\ell,k=K_T}^{\infty} \sum K(\ell)(1 + \ell/T)^{\frac{1}{2}}(1 + (n+k)/T)^{\frac{1}{2}} L(k)' \| = o(T^{-1})$$

$$\Big\| \sum_{\ell=K_T}^{\infty} \sum_{k=1}^{\infty} K(\ell)(1+\ell/T)^{\frac{1}{2}}(1+(n+k)/T)^{\frac{1}{2}} L(k)' \Big\| = o(T^{-\frac{1}{2}}) \ .$$

To establish (4.1.2) it remains to evaluate

$$(4.1.3) \qquad \frac{1}{T} \sum_{n+1}^{T} \varepsilon(t-\ell) z(t-n-k)' - \frac{1}{T} \sum_{m_1}^{T-m_2} \varepsilon(t) z(t-n+\ell-k)'$$

for $k,\ell \leqslant K_T$, $n \leqslant H_T$. Take, for example, $n-\ell+k > 0$, when (4.1.3) is

$$\frac{1}{T}\left\{ \sum_{n+1}^{n+k+1} \varepsilon(t-\ell) z(t-n-k)' - \sum_{T+1}^{T+\ell} \varepsilon(t-\ell) z(t-n-k)' \right\} \ .$$

As $\varepsilon(t)$ and $z(t)$ have finite 4^{th} moments, $\varepsilon(t) = o(T^{\frac{1}{4}})$, $z(t) = o(T^{\frac{1}{4}})$ and $\sum_{T+1}^{T+\ell} \varepsilon(t-\ell) z(t-n-k)' = \ell o(T^{\frac{1}{2}})$. The bound we have established above for $\frac{1}{T} \sum_{n+1}^{T} \varepsilon(t-\ell) z(t-n-k)'$ shows that

$$\frac{1}{T} \sum_{n+1}^{n+k+1} \varepsilon(t-\ell) z(t-n-k)' = o(T^{-\frac{1}{2}}) \ .$$

In a similar way, when $\ell \geqslant n+k$, (4.1.3) has the bound $o(T^{-\frac{1}{2}}) + (n+k)o(T^{-\frac{1}{2}})$. Since $\sum_{0}^{\infty} \ell \| K(\ell) \| < \infty$ we have

$$\sum_{k,\ell=0}^{K_T} \sum K(\ell) \left\{ \frac{1}{T} \sum_{n+1}^{T} \varepsilon(t-\ell) z(t-n-k)' - \frac{1}{T} \sum_{m_1}^{T-m_2} \varepsilon(t) z(t-n+\ell-k)' \right\} L(k) = o(T^{-\frac{1}{2}})$$

a.s.

This completes the proof of the lemma. □

Directly from (4.1.2) and Theorem 4.1.3 we have

THEOREM 4.1.5. *Under conditions of theorem 4.1.3, if* $H_T = (\log T)^a$, $a < \infty$ *then*

$$\sup_{|k| \leqslant H_T} \Big\| \frac{1}{T} \sum_{m_1}^{T+m_2} u(t) v(t-k)' \Big\| = O\{(\log \log T/T)^{\frac{1}{2}}\} \quad \text{a.s.}$$

where $m_1 = k+1$, $m_2 = 0$ *if* $k > 0$, *or* $m_1 = 1$, $m_2 = k$ *if* $k \leqslant 0$.

§4.2 Properties of $\hat{\Gamma}_h$

For a matrix A denote the smallest and largest eigenvalues by $\lambda_{min}\{A\}$, $\lambda_{max}\{A\}$ respectively. As in chapter 3 put $Q_T = (\log\log T/T)^{\frac{1}{2}}$, and put $H_T = (\log T)^a$, $a < \infty$. The eigenvalues of $\mathring{\Gamma}_{u,h}$ are bounded away from zero and infinity, the bounds being uniform in h. For example, the smallest eigenvalue is

$$\text{(4.2.1)} \qquad \inf \frac{1}{2\pi} \int_{-\pi}^{\pi} p^*(e^{i\omega})k(e^{i\omega}) \mathring{\sharp} k^*(e^{i\omega})p(e^{i\omega})d\omega$$

where the infimum is taken over all s-vectors of polynomials having degree h that satisfy

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} p^*(e^{i\omega})p(e^{i\omega}) = I .$$

Here (4.2.1) is bounded below by $\inf_\omega \lambda_{min}\{k(e^{i\omega}) \mathring{\sharp} k^*(e^{i\omega})\}$ which, because of (1.1.4), is positive. Similarly the upper bound on (4.2.1) is $\sup_\omega \lambda_{max}\{k(e^{i\omega}) \mathring{\sharp} k^*(e^{i\omega})\}$. The Euclidean norm $\|\hat{\Gamma}_{u,h} - \mathring{\Gamma}_{u,h}\| = hO(Q_T)$ under the conditions of theorem 4.1.1. Hence, if $h \leq (\log T)^a$, $a < \infty$, then

$$\text{(4.2.2)} \qquad 0 < \underline{c} \leq \lambda_{min}\{\hat{\Gamma}_{u,h}\} \leq \lambda_{max}\{\hat{\Gamma}_{u,h}\} \leq \bar{c} < \infty \text{ , a.s.}$$

Of course for this kind of result to hold, the rate of convergence of $\hat{\gamma}_u(k) \to \mathring{\gamma}_u(k)$ need only be $o(H_T^{-1})$ so that $\|\hat{\Gamma}_{u,h} - \mathring{\Gamma}_{u,h}\| = o(1)$. Indeed such a rate is all we need in the next lemma.

LEMMA 4.2.1. *Assume conditions A hold. If* $h \leq H_T$ *then*

$$0 < \underline{c} \leq \lambda_{min}\{\hat{\Gamma}_h\} \leq \lambda_{max}\{\hat{\Gamma}_h\} \leq \bar{c} < \infty \text{ , a.s.}$$

The constants $\underline{c}$, $\bar{c}$ in this lemma are different from those in (4.2.2).

Proof: Here

$$\hat{\Gamma}_h = \begin{bmatrix} \hat{\Gamma}_{u,h} & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} \hat{\Gamma}_{v,h} & \hat{\Gamma}_{vz,h} \\ \hat{\Gamma}_{zv,h} & \hat{\Gamma}_{z,h} \end{bmatrix} + \begin{bmatrix} \hat{\Gamma}_{uv,h}+\hat{\Gamma}_{vu,h} & \hat{\Gamma}_{uz,h} \\ \hat{\Gamma}_{zu,h} & 0 \end{bmatrix}.$$

The last matrix has elements that are $O(Q_T)$, by Theorem 4.1.5. A similar argument to that used to deduce Theorem 4.1.5 from Theorem 4.1.3, shows that the sample covariances $\hat{\gamma}_v(k)$, $\hat{\gamma}_{vz}(k)$ differ from their limiting values by a quantity that is $o(H_T^{-1})$ because $\|\hat{\gamma}_z(k)-\gamma_z(k)\| = o(H_T^{-1})$ by (3.1.2). Consequently $\|\hat{\Gamma}_h-\mathring{\Gamma}_h\| = o(1)$ so we proceed to bound the eigenvalues of

$$\mathring{\Gamma}_h = \begin{bmatrix} \mathring{\Gamma}_{u,h} & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} \Gamma_{v,h} & \Gamma_{vz,h} \\ \Gamma_{zv,h} & \Gamma_{z,h} \end{bmatrix}.$$

As $f_z(\omega)$ is bounded away from zero and infinity (3.1.4) the eigenvalues of $\Gamma_{z,h}$ are also bounded away from zero and infinity, hence $\lambda_{min}\{\Gamma_h\} \geqslant \underline{c} > 0$. To bound the largest eigenvalue of $\mathring{\Gamma}_h$, we need only bound the largest eigenvalue of the second matrix in the decomposition of $\mathring{\Gamma}_h$ above (this is because of (4.2.2)). This bound follows from the same kind of argument used to bound (4.2.1) because the eigenvalues of $\ell(e^{i\omega})f_z(\omega)\ell^*(e^{i\omega})$, $f_z(\omega)$ are bounded for each ω. □

Define the norm $|.|_1$ *for a matrix* $A = [a_{ij}]_{i,j=1,\ldots,h}$ *as* $|A|_1 = \sup_i \sum_j |a_{ij}|$. If A is partitioned into $m \times m$ blocks, $A = [A_{ij}]_{i,j=1,\ldots,h}$ then define $|A|_m = \sup_i \sum_j \|A_{ij}\|$, where $\|.\|$ is the Euclidean norm. As $|A|_1 \leqslant \sqrt{m}\|A\|$, then $|A|_1 \leqslant \sqrt{m}|A|_m$. These norms satisfy $|AB|_m \leqslant |A|_m|B|_m$. We have $\sum_0^\infty \|\gamma_z(k)\| < \infty$,

$\sum_0^\infty ||\gamma_y(k)|| < \infty$ hence $|\Gamma_h|_1$ is bounded uniformly in h. Similarly, $|\mathring{\Gamma}_h|_1$ is uniformly bounded because $\sum_0^\infty ||\mathring{\gamma}_y(k)|| < \infty$. Under conditions A the sample covariances $\hat{\gamma}_y(k)$, $\hat{\gamma}_z(k)$, $\hat{\gamma}_{yz}(k)$ converge to $\mathring{\gamma}_y(k)$, $\gamma_z(k)$, $\gamma_{yz}(k)$ at a rate that is no slower than $o(H_T^{-1})$ hence

$$(4.2.3) \qquad |\hat{\Gamma}_h|_1 \leqslant c < \infty \ , \quad h \leqslant H_T \ .$$

To obtain a corresponding bound for $|\hat{\Gamma}_h^{-1}|_1$ is more difficult.

LEMMA 4.2.2. *If* $z(t)$ *is stationary with zero mean and finite variance and* (3.1.5) *holds then*

$$|\Gamma_{z,h}^{-1}|_1 \leqslant c < \infty$$

uniformly in $h = 1,2,\ldots$.

Proof: Recall that we have put

$$z(t) = \sum_0^\infty K_z(j)\varepsilon_z(t-j) \ , \ E\{\varepsilon_z(t)\varepsilon_z(s)'\} = \delta_{st}\sharp_z$$

and $m(z) = \sum_0^\infty K_z(j)z^{-j}$, $\Lambda(z) = m(z)^{-1} = \sum_0^\infty \Lambda(j)z^{-j}$. Let $z_{t,h} = (z(t+1)',\ldots,z(t+h)')'$, $\varepsilon_{z;t,h} = (\varepsilon_z(t+1)',\ldots,\varepsilon_z(t+h)')'$ and $z_t = (z(t)',z(t-1)',\ldots)'$. If

$$\Lambda_h = \begin{bmatrix} I & 0 & \cdots & \cdots & 0 \\ \Lambda(1) & I & 0 & & \vdots \\ \vdots & & & & \vdots \\ \vdots & & & I & 0 \\ \Lambda(h-1) & \cdots & \cdots & \Lambda(1) & I \end{bmatrix}$$

then $\Lambda_h z_{t,h} = \varepsilon_{z;t,h} + H\, z_t$ where $H = \mathcal{H}_\infty^h$ and $\mathcal{H}$ is the Hankel matrix of $m(z)$. As $E\{z_t\varepsilon_{z;t,h}'\} = 0$ we have

$$(4.2.4) \qquad \Lambda_h\Gamma_{z,h}\Lambda_h' = I_h \otimes \sharp_z + H\,\Gamma_{z,\infty}\,H' \ .$$

We proceed to invert the RHS of (4.2.4). The (i,j) block of $H\,\Gamma_{z,\infty}\,H'$ is

$$M_{ij} = \sum_{u,v=0}^{\infty}\sum \Lambda(i+u)\gamma_z(u-v)\Lambda(j+v)' \ .$$

As $|\Gamma_{z,\infty}|_1 < \infty$, $\Sigma\| \Lambda(j)\| < \infty$, we have

$$\sum_{u,v=0}^{\infty}\sum \|\gamma_z(u-v)\Lambda(j+v)'\| \leqslant c < \infty \tag{4.2.5}$$

uniformly in $j = 1,2,\ldots$.

Then

$$t^{\frac{1}{2}} \sum_{i=t}^{\infty} \|M_{ij}\| \leqslant \sum_{u=0}^{\infty} \{(t^{\frac{1}{2}} \sum_{i=t}^{\infty} \|\Lambda(i+u)\|)(\sum_{v=0}^{\infty} \|\gamma_z(u-v)\Lambda(j+v)'\|)\}$$

As the bound in (4.2.5) is uniform in j we have

$$t^{\frac{1}{2}} \max_j \sum_{i=t}^{\infty} \|M_{ij}\| \to 0 \tag{4.2.6}$$

because of (3.1.5). Partition $I_h \otimes \sharp_z + H\,\Gamma_{z,\infty}\,H'$ as $\begin{bmatrix} A & B \\ B' & C \end{bmatrix}$ where the partition is after t rows and columns of $(m \times m)$ blocks, and

$$\begin{bmatrix} A & B \\ B' & C \end{bmatrix}^{-1} = \begin{bmatrix} A^{-1} + A^{-1}B(C-B'A^{-1}B)^{-1}B'A^{-1} & -A^{-1}B(C-B'A^{-1}B)^{-1} \\ -(C-B'A^{-1}B)^{-1}B'A^{-1} & (C-B'A^{-1}B)^{-1} \end{bmatrix} \tag{4.2.7}$$

In this expression

$$C = I_{h-t} \otimes \sharp_z + M_t \ , \ M_t = [M_{ij}]_{i,j = t+1,\ldots,h}$$

and taking t large (but fixed) $|M_t|_m$ may be made arbitrarily small (by (4.2.6)), hence $|(I_{h-t} \otimes \sharp_z^{-1})M_t|_m$ may be taken to be arbitrarily small. This bound is independent of h . Thus expanding

(4.2.8) $$\{I_{(h-t)m} + (I_{h-t} \otimes \sharp_z^{-1})M_t\}^{-1}$$

in a Neumann series in terms of $(I_{h-t} \otimes \sharp_z^{-1})M_t$, (4.2.8) may be written as $(I_{(h-t)m} + C_1)$ where $|C_1|_m$ is bounded uniformly in h. Thus $C^{-1} = (I_{h-t} \otimes \sharp_z^{-1})(I_{(h-t)m} + C_1)$ so that $|C^{-1}|_m \leqslant c < \infty$. Denote the j^{th} row of $(m \times m)$ blocks of B' by $b_j' = [M_{j_1}, M_{j_2}, \ldots, M_{j_t}]$. Then

$$|B'A^{-1}B|_m = \max_i \{\sum_j \|b_i' A^{-1} b_j\|\}$$

$$\leqslant \lambda_{max}\{A^{-1}\} \max_i \|b_i\| \sum_j \|b_j\| .$$

Now $$\sum_j \|b_j\| = \sum_{j=t+1}^{h} \left(\sum_{i=1}^{t} \|M_{ij}\|^2 \right)^{\frac{1}{2}}$$

$$\leqslant \sum_{j=t+1}^{h} \sum_{i=1}^{t} \|M_{ij}\|$$

$$\leqslant t \max_i \left\{ \sum_{j=t+1}^{\infty} \|M_{ij}\| \right\} .$$

Hence from (4.2.6), $|B'A^{-1}B|_m$ can be made arbitrarily small by taking t sufficiently large. Consequently $|C^{-1}B'A^{-1}B|_m$ is small and again we can expand $(I - C^{-1}B'A^{-1}B)^{-1}$ in a Neumann series giving a bound, uniformly in h, for $|(C - B'A^{-1}B)^{-1}|_m$. As t is fixed, $|A^{-1}|_m < \infty$ so that using the norm bounds on B, B', $(C - B'A^{-1}B)^{-1}$ we conclude that each expression in the partitioned matrix (4.2.7) is bounded (uniformly in h) in the norm $|.|_m$, hence $|(\Lambda_h \Gamma_{z,h} \Lambda_h')^{-1}|_m \leqslant c < \infty$. As $|\Lambda_h|_m \leqslant c_1 < \infty$ we conclude that $|\Gamma_{z,h}^{-1}|_m \leqslant c < \infty$ as required. □

This is the only time that (3.1.5) is necessary in the theorems given in chapter 3. We extend this result to show that

$|\Gamma_h^{-1}|_1 \leqslant c < \infty$. Here

$$\begin{bmatrix} k(z) & \ell(z)m(z) \\ 0 & m(z) \end{bmatrix}^{-1} = \begin{bmatrix} k(z)^{-1} & -k(z)^{-1}\ell(z) \\ 0 & m(z)^{-1} \end{bmatrix} = W(z)$$

and

$$W(z)\begin{bmatrix} y(t) \\ z(t) \end{bmatrix} = \begin{bmatrix} \varepsilon(t) \\ \varepsilon_z(t) \end{bmatrix} .$$

The coefficients of $k(z)^{-1}$, $k(z)^{-1}\ell(z)$ are decreasing geometrically while the coefficients of $m(z)^{-1}$ satisfy (3.1.5), hence $t^{\frac{1}{2}} \sum_t^\infty \|W(j)\| \to 0$, where $W(z) = \sum_0^\infty W(j)z^{-j}$. Thus lemma 4.2.2 applies unchanged to show that $|\Gamma_h^{-1}|_1 \leqslant c < \infty$, and $|\mathring{\Gamma}_h^{-1}|_1 \leqslant c_1 < \infty$. To extend these results to $|\hat{\Gamma}_h^{-1}|_1$ we have

$$\begin{aligned} |\hat{\Gamma}_h^{-1}|_1 &\leqslant |\mathring{\Gamma}_h^{-1}|_1 + |\hat{\Gamma}_h^{-1} - \mathring{\Gamma}_h^{-1}|_1 \\ &\leqslant |\mathring{\Gamma}_h^{-1}|_1 + |\hat{\Gamma}_h^{-1}|_1 |\hat{\Gamma}_h - \mathring{\Gamma}_h|_1 |\mathring{\Gamma}_h^{-1}|_1 , \end{aligned}$$

which gives

$$|\hat{\Gamma}_h^{-1}|_1 \{1 - |\hat{\Gamma}_h - \mathring{\Gamma}_h|_1 |\mathring{\Gamma}_h^{-1}|_1\} \leqslant |\mathring{\Gamma}_h^{-1}|_1 \leqslant c < \infty .$$

Now if $h \leqslant H_T = (\log T)^a$, $a < \infty$ then under conditions A $|\hat{\Gamma}_h - \mathring{\Gamma}_h|_1 = o(1)$, a.s. thus we have proved

LEMMA 4.2.3. *Under conditions A, and uniformly in* $h \leqslant H_T$,

$$|\Gamma_h^{-1}|_1 \leqslant c < \infty , \quad |\hat{\Gamma}_h^{-1}|_1 \leqslant c_1 < \infty .$$

Though we have stated these results for the norm $|.|_1$ they apply, without change in the proof, to $|.|_i$ for any fixed i .

To conclude this section we investigate some properties of the functions $\phi_h(z), \psi_h(z)$ defined by minimising

(3.3.3). Given an absolutely summable sequence $g(k)$ let $a_h(\omega) = \sum_1^h a_h(j)e^{-ij\omega}$ satisfy

$$(4.2.9) \qquad \int_{-\pi}^{\pi} a_h(\omega)f(\omega)e^{ik\omega}d\omega = g(k) \ , \ k = 1,\ldots,h \ .$$

Under mild conditions on $f(\omega)$, Baxter (1963) shows that

$$(4.2.10) \qquad \sum_1^h \|a_h(j)\| \leqslant M \sum_1^h \|g(k)\|$$

and M is independent of the particular sequence $g(k)$ or h involved. Though Baxter's theorem is for the univariate case the proof readily generalizes to a_h, f, g being matrices, a_h and g not necessarily square, the essential properties of $f(\omega)$ being that it has an absolutely convergent Fourier series and $f(\omega) = f_1(\omega)f_2(\omega)$ where $f_1(\omega) = \sum_{-\infty}^{0} \hat{f}_1(j)e^{ij\omega}$, $f_2(\omega) = \sum_1^{\infty} \hat{f}_2(j)e^{ij\omega}$. Let $a(\omega) = \sum_1^{\infty} a(j)e^{-ij\omega}$ satisfy (4.2.9) when $h = \infty$. Then

$$\int_{-\pi}^{\pi} \sum_1^h (a_h(j) - a(j))e^{-ij\omega}f(\omega)e^{ik\omega}d\omega$$

$$= \int_{-\pi}^{\pi} \sum_{h+1}^{\infty} a(j)e^{-ij\omega}f(\omega)e^{ik\omega}d\omega \ , \ k = 1,\ldots,h \ ,$$

and, as

$$\left\| \int_{-\pi}^{\pi} \sum_{h+1}^{\infty} a(j)e^{-ij\omega}f(\omega)e^{ik\omega}d\omega \right\| \leqslant c \sum_{h+1}^{\infty} \|a(j)\|$$

we conclude, from (4.2.10), that

$$(4.2.11) \qquad \sum_1^h \|a_h(j) - a(j)\| \leqslant M_1 \sum_{h+1}^{\infty} \|a(j)\| \ .$$

Put

$$(4.2.12) \qquad f(\omega) = \begin{bmatrix} k_0 \Phi_0 k_0^* + \ell_0 dF_z \ell_0^* & \ell_0 dF_z \\ dF_z \ell_0^* & dF_z \end{bmatrix}$$

and let the coefficients of $a_h(z)$ be $s \times (s+m)$ matrices. In this case

$$a(z) = [k_0^{-1}(z), -k_0^{-1}(z)\ell_0(z)] = [\phi(z), \psi(z)]$$

while $a_h(z) = [\phi_h(z), \psi_h(z)]$ as in (3.3.3). Then we have,

$$(4.2.13) \qquad \sum_0^h \| [\phi_h(j), \psi_h(j)] - [\phi(j), \psi(j)] \| \leq M_1 \sum_{h+1}^{\infty} \| \phi(j), \psi(j) \| .$$

Furthermore

$$\Phi_h - \Phi_0 = \int_{-\pi}^{\pi} \{ [\phi_h(e^{i\omega}), \psi_h(e^{i\omega})] - [\phi(e^{i\omega}), \psi(e^{i\omega})] \} f(\omega) d\omega$$

hence from (4.2.13)

$$(4.2.14) \qquad \| \Phi_h - \Phi_0 \| \leq M_1 \sum_{h+1}^{\infty} \| \phi(j), \psi(j) \| , \quad h = 1,2,\ldots .$$

If Φ_0 in (4.2.12) is replaced by $\mathring{\Phi} = \frac{1}{T} \Sigma \varepsilon(t) \varepsilon(t)'$ then $a(z)$ remains unchanged and $a_h(z) = [\mathring{\phi}_h(z), \mathring{\psi}_h(z)]$ as in §3.3. Thus we also have

$$(4.2.15) \qquad \sum_1^h \| [\mathring{\phi}_h(j), \mathring{\psi}_h(j)] - [\phi(j), \psi(j)] \| \leq M_1 \sum_{h+1}^{\infty} \| \phi(j), \psi(j) \|$$

Of course the constants M_1 in (4.2.13), (4.2.14), (4.2.15) are not necessarily the same.

Define another norm $\| . \|_1$ for the matrix $A = [a_{ij}]_{i,j = 1,\ldots,h}$ as $\|A\|_1 = \Sigma\Sigma_{ij} |a_{ij}|$. This norm satisfies

$$\|AB\|_1 \leq \|A\|_1 \|B\|_1 ; \quad \|AB\|_1 \leq \|A\|_1 |B|_1$$

We proceed to obtain a representation of Γ_h^{-1} that will be needed in lemma 5.1.1.

Let

$$A_h = \begin{bmatrix} I_s & \hat{\phi}_{h-1}(1) & \cdots & \hat{\phi}_{h-1}(h-1) & 0 & \hat{\psi}_{h-1}(1) & \cdots & \hat{\psi}_{h-1}(h-1) \\ 0 & I_s & \hat{\phi}_{h-2}(1) & \cdots & \vdots & \vdots & & \vdots \\ \vdots & & & \ddots & \vdots & \vdots & & \vdots \\ & & & I_s & \hat{\psi}_1(1) & 0 & \cdots & 0 \quad \hat{\psi}_1(1) \\ 0 & \cdots & \cdots & 0 & I_s & 0 & \cdots & 0 \end{bmatrix}$$

(4.2.16)

The coefficients, $\hat{\phi}_p(j)$, $\hat{\psi}_p(j)$ are obtained as solutions of

$$\sum_{j=0}^{p} [\hat{\phi}_p(j), \hat{\psi}_p(j)] \gamma_w(j-k) = \delta_{0k} \hat{\sigma}_p \quad , \quad k = 0,1,\ldots,p$$

where $w(t) = [y(t)', z(t)']'$. When $p \geq q$

$$\sum_{j=0}^{p} \sum_{k=0}^{q} [\hat{\phi}_p(j), \hat{\psi}_p(j)] \gamma_w(j-k+q-p) [\hat{\phi}_q(k), \hat{\psi}_q(k)]' = \delta_{pq} \hat{\sigma}_p .$$

Hence

(4.2.17) $$A_h \Gamma_h A_h' = S_h$$

where $S_h = \begin{bmatrix} \hat{\sigma}_{h-1} & & & 0 \\ & \hat{\sigma}_{h-2} & & \\ & & \ddots & \\ & & & \hat{\sigma}_1 \\ 0 & & & \hat{\gamma}_y(0) \end{bmatrix}$. Define $\tilde{A}_h = \begin{bmatrix} A_h \\ \cdots \\ 0 \quad I_{hm} \end{bmatrix}$

then

$$\Gamma_h^{-1} = \tilde{A}_h' \begin{bmatrix} S_h^{-1} & 0 \\ 0 & \Gamma_{z,h}^{-1} \end{bmatrix} \tilde{A}_h$$

as $\tilde{A}_h$ is invertible. Define the matrix $\tilde{A}$ by replacing $\phi_i(j)$ by $\phi(j)$, $\psi_i(j)$ by $\psi(j)$, $i = 1,\ldots,h$ in $\tilde{A}_h$ and put $\tilde{S} = I_h \otimes \mathring{\sigma}$, $\mathring{\sigma} = \frac{1}{T} \sum_1^T \varepsilon(t)\varepsilon(t)'$. From theorem 3.3.1, and

~~lemma 4.2.4~~ using (4.2.15)

$$\sum_{j=1}^{p} \|\hat{\phi}_p(j)-\phi(j), \hat{\psi}_p(j)-\psi(j)\| = O(pQ_T) + O\Big(\sum_{p+1}^{\infty} \|\phi(j),\psi(j)\|\Big).$$

hence

$$\|\tilde{A} - \tilde{A}_h\|_1 = O(h^2 Q_T) + O\Big\{\sum_{t=0}^{h} \sum_{j=t+1}^{\infty} \|\phi(j),\psi(j)\|\Big\}$$

which converges to zero as $\|\phi(j),\psi(j)\|$ decrease geometrically. (Here we are assuming conditions A.) Similarly from lemma 3.3.3 and (3.3.4) we have that

$$\hat{\Sigma}_p - \overset{\circ}{\Sigma} = O(hQ_T) + O\Big(\sum_{p+1}^{\infty} \|\phi(j),\psi(j)\|\Big)$$

Again this shows that $\|S_h - \tilde{S}\|_1$ converges to zero. Thus

$$\hat{\Gamma}_h^{-1} = \tilde{A}_h' \begin{bmatrix} S_h^{-1} & 0 \\ 0 & \hat{\Gamma}_{z,h}^{-1} \end{bmatrix} \tilde{A}_h'$$

$$= \tilde{A}' \begin{bmatrix} I_h \otimes \overset{\circ}{\Sigma}^{-1} & 0 \\ 0 & \hat{\Gamma}_{z,h}^{-1} \end{bmatrix} \tilde{A} + D, \quad \|D\|_1 \leq c < \infty \tag{4.2.18}$$

where D is composed of terms that are products, for example

$$(\tilde{A}_h - \tilde{A})' \begin{bmatrix} S_h^{-1} & 0 \\ 0 & \hat{\Gamma}_{z,h}^{-1} \end{bmatrix} \tilde{A}_h$$

where at least one factor is bounded in the norm $\|.\|_1$ while the remaining factors are bounded in $|.|_1$. Thus $\|D\|_1$ is bounded. Without real changes this same argument can be used to obtain

$$\Gamma_h^{-1} = \tilde{A}' \begin{bmatrix} I_h \otimes \overset{\circ}{\sharp}^{-1} & 0 \\ 0 & \Gamma_{z,h}^{-1} \end{bmatrix} \tilde{A} + D_1 \ , \quad \|D_1\|_1 \leqslant c < \infty \ .$$

The fact that the coefficients $\phi(j),\psi(j)$ decrease at a geometric rate can be considerably relaxed and the decomposition (4.2.17) continues to hold. Our argument only requires $t(\log t)^{1+\delta} \sum_t^\infty \|\phi(j),\psi(j)\| \to 0$.

We also need a factorisation

$$(4.2.19) \qquad \tilde{A}\hat{\Gamma}_h(a)\tilde{A}' = \begin{bmatrix} I_h(a) \otimes \overset{\circ}{\sharp} & 0 \\ 0 & \hat{\Gamma}_{z,h}(a) \end{bmatrix} + D, \quad \|D\|_1 < \infty \ .$$

Here $\hat{\Gamma}_h(a)$, $\hat{\Gamma}_{z,h}(a)$ is defined in §3.1 while $I_h(a) = [\delta_{i,j-a}]_{i,j=1,\ldots,h}$. In (4.2.17) replace S_h by $I_h \otimes \overset{\circ}{\sharp}$, let B_h be the matrix in the *last* (h-a)s rows of A_h and C_h be the matrix in the *first* (h-a)s rows of A_h . Then as $\|S_h - I_h \otimes \overset{\circ}{\sharp}\|_1 \leqslant c < \infty$,

$$B_h\hat{\Gamma}_h C_h' = I_{h-a}(a) \otimes \overset{\circ}{\sharp} + D \ , \quad \|D\|_1 \leqslant c$$

Let $\tilde{C}_h'$ be the matrix obtained from C_h by replacing columns $(h-a)s+1,\ldots,hs$, $hs+(h-a)m+1,\ldots,h(s+m)$ by zeros. Then $\|C_h - \tilde{C}_h'\| \leqslant c < \infty$ by arguments similar to those used before. Now

$$\|B_h\hat{\Gamma}_h(C_h - \tilde{C}_h)'\|_1 \leqslant c < \infty$$

hence $B_h\hat{\Gamma}_h\tilde{C}_h' = I_{h-a}(a) \otimes \overset{\circ}{\sharp} + D_2$, $\|D_2\|_1 \leqslant c$. But deleting the columns of zeros from B_h, $\tilde{C}_h$ and the corresponding columns from $\hat{\Gamma}_h$ we have $A_{h-a}\hat{\Gamma}_{h-a}(a)A_{h-a}' = B_h\hat{\Gamma}_h\tilde{C}_h'$, and

$$\tilde{A}_{h-a}\hat{\Gamma}_{h-a}(a)\tilde{A}'_{h-a} = \begin{bmatrix} I_{h-a}(a) \otimes \mathring{\Sigma} & 0 \\ 0 & \Gamma_{z,h-a} \end{bmatrix} + D_2 \quad , \quad \|D_2\| \leq c < \infty$$

Replacing h-a by h , the construction is completed as before by putting $\tilde{A}$ in place of $\tilde{A}_h$ giving (4.2.19).

CHAPTER 5

PROOFS OF THEOREMS FROM CHAPTER 3

§5.1 Stage I

For brevity we shall denote $(\log\log T/T)^{\frac{1}{2}}$ by Q_T throughout the proofs in this chapter.

(i) Theorem 3.3.1

We need to evaluate

$$\{[\hat{\phi}_h,\hat{\psi}_h]-[\mathring{\phi}_h,\mathring{\psi}_h]\}\hat{\Gamma}_h \quad , \tag{5.1.1}$$

the notation being defined above the statement of theorem 3.3.1. As

$$[\hat{\phi}_h,\hat{\psi}_h]\hat{\Gamma}_h = -[\hat{\gamma}_y(-1),\ldots,\hat{\gamma}_y(-h),\hat{\gamma}_{yz}(-1),\ldots,\hat{\gamma}_{yz}(-h)]$$

(5.1.1) can be written as the sum of the following terms

$$-[\phi,\psi]\hat{\Gamma}_h \tag{5.1.2}$$

$$-[\hat{\gamma}_y(-1),\ldots,\hat{\gamma}_{yz}(-h)]+\{[\phi,\psi]-[\mathring{\phi}_h,\mathring{\psi}_h]\}\mathring{\Gamma}_h \tag{5.1.3}$$

$$\{[\phi,\psi]-[\mathring{\phi}_h,\mathring{\psi}_h]\}\ (\hat{\Gamma}_h-\mathring{\Gamma}_h) \tag{5.1.4}$$

First consider (5.1.4). The elements of $\hat{\Gamma}_h-\mathring{\Gamma}_h$ converge uniformly to zero at a rate at least as fast as $o(H_T^{-1})$. Using (4.2.15) to bound $\sum_1^h \|\phi(j)-\mathring{\phi}_h(j),\ \psi(j)-\mathring{\psi}_h(j)\|$, every element of (5.1.4) converges to zero at a rate

$$o(H_T^{-1}\sum_{h+1}^{\infty}\|\phi(j),\psi(j)\|) \quad , \text{ a.s.}$$

Consider (5.1.3). The equations defining $\mathring{\phi}_h(j),\mathring{\psi}_h(j)$ are

$$[\mathring{\phi}_h,\mathring{\psi}_h]\mathring{\Gamma}_h = -[\mathring{\gamma}_y(-1),\ldots,\mathring{\gamma}_y(-h),\gamma_{yz}(-1),\ldots,\gamma_{yz}(-h)]$$

and, as above (4.2.15), $[\phi,\psi]\mathring{\Gamma}_\infty = -[\mathring{\gamma}_y(-1),\ldots,\gamma_{yz}(-h)]$,

hence the i, i+h blocks of (5.1.3) are, respectively,

$$(5.1.5)\qquad -\sum_{h+1}^{\infty}\phi(j)\mathring{\gamma}_y(j-i) - \sum_{h+1}^{\infty}\psi(j)\gamma_{zy}(j-i)-\hat{\gamma}_y(-i)$$

$$(5.1.6)\qquad -\sum_{h+1}^{\infty}\phi(j)\gamma_{yz}(j-i) - \sum_{h+1}^{\infty}\psi(j)\gamma_{z}(j-i)-\hat{\gamma}_{yz}(-i)\ .$$

In these expressions replace $\mathring{\gamma}_y(k),\gamma_{zy}(k),\gamma_{yz}(k),\gamma_z(k)$ by the sample covariances $\hat{\gamma}_y(k),\hat{\gamma}_{zy}(k),\hat{\gamma}_{yz}(k),\hat{\gamma}_z(k)$ and upon adding these terms to (5.1.2) we obtain a matrix of blocks where the i , i+h blocks are, respectively,

$$-\sum_0^{\infty}\phi(j)\hat{\gamma}_y(j-i)-\sum_1^{\infty}\psi(j)\hat{\gamma}_{zy}(j-i)+o(H_T^{-1}\sum_{h+1}^{\infty}\|\phi(j),\psi(j)\|)$$

$$-\sum_0^{\infty}\phi(j)\hat{\gamma}_{yz}(j-i)-\sum_1^{\infty}\psi(j)\hat{\gamma}_{z}(j-i)+o(H_T^{-1}\sum_{h+1}^{\infty}\|\phi(j),\psi(j)\|)\ .$$

It remains to replace the covariances $\hat{\gamma}_y(i-j),\hat{\gamma}_{zy}(i-j),\ldots$ by $\tilde{\gamma}_y(i,j),\tilde{\gamma}_{zy}(i,j),\ldots$ thus

$$\sum_0^{\infty}\phi(j)\hat{\gamma}_y(j-i)+\sum_1^{\infty}\psi(j)\hat{\gamma}_{zy}(j-i)$$

$$= \frac{1}{T}\sum_1^T\varepsilon(t)y(t-i)'+\sum_0^{\infty}\phi(j)\{\hat{\gamma}_y(j-i)-\frac{1}{T}\sum_1^T y(t-j)y(t-i)'\}$$

$$(5.1.7)\qquad +\sum_1^{\infty}\psi(j)\{\hat{\gamma}_{zy}(j-i)-\frac{1}{T}\sum_1^T z(t-j)y(t-i)'\}$$

Exactly as in lemma 4.1.4,

$$\tilde{\gamma}_{zy}(j,i) = \frac{1}{T}\sum_1^T z(t-j)y(t-i)' = O\{(1+\tfrac{j}{T})^{\frac{1}{2}}(1+\tfrac{i}{T})^{\frac{1}{2}}\}$$

hence, as $\hat{\gamma}_{zy}(k)$ is bounded we may truncate the sum in the last term of (5.1.7) at $K_T = c\log T$, where c is sufficiently large so that $\sum_{K_T+1}^{\infty}\psi(j)\{\hat{\gamma}_{zy}(j-i)-\tilde{\gamma}_{zy}(j,i)\} = o(T^{-\frac{1}{2}})$. If $j \geqslant i$

$$(5.1.8)\quad \hat{\gamma}_{zy}(j-i)-\tilde{\gamma}_{zy}(j,i) = \sum_{T+1}^{T+i}\{z(t-j)y(t-i)'-\sum_{1}^{j}z(t-j)y(t-i)'\}.$$

Exactly as in the evaluation of (4.1.3) the first term of the RHS of (5.1.8) is $i\ o(T^{-\frac{1}{2}})$ while the second term is $o(T^{-\frac{1}{2}})$. Similarly, when $j < i$, $\hat{\gamma}_{zy}(j-i)-\tilde{\gamma}_{zy}(j,i) = j\ o(T^{-\frac{1}{2}})+o(T^{-\frac{1}{2}})$. Consequently

$$\sum_{1}^{K_T}\psi(j)\{\hat{\gamma}_{zy}(j-i)-\tilde{\gamma}_{zy}(j,i)\} = o(T^{-\frac{1}{2}}) .$$

The second term of the RHS of (5.1.7) is similarly bounded hence

$$\sum_{0}^{\infty}\phi(j)\hat{\gamma}_{y}(j-i)+\sum_{1}^{\infty}\psi(j)\hat{\gamma}_{zy}(j-i) = \frac{1}{T}\sum_{1}^{T}\varepsilon(t)y(t-i)'+o(T^{-\frac{1}{2}}) .$$

In a similar way

$$\sum_{0}^{\infty}\phi(j)\hat{\gamma}_{yz}(j-i) + \sum_{1}^{\infty}\psi(j)\hat{\gamma}_{z}(j-i) = \frac{1}{T}\sum_{1}^{T}\varepsilon(t)z(t-i)' + o(T^{-\frac{1}{2}}) .$$

Hence (5.1.1) becomes

$$-[\tfrac{1}{T}\Sigma\varepsilon(t)y(t-1)',\ldots,\tfrac{1}{T}\Sigma\varepsilon(t)y(t-h)',$$

$$\tfrac{1}{T}\Sigma\varepsilon(t)z(t-1)',\ldots,\tfrac{1}{T}\Sigma\varepsilon(t)z(t-h)']$$

$$+ o(H_T^{-1}\sum_{h+1}^{\infty}\|\phi(j),\psi(j)\|) + o(T^{-\frac{1}{2}}) .$$

All of the above bounds hold almost surely. □

(ii) Corollary 3.3.2

When $d \geqslant 1/(-2\log\rho_0)$, $h \geqslant d\log T$ then $\sum_{h+1}^{\infty}\|\phi(j),\psi(j)\| = O(T^{-\frac{1}{2}})$. This, together with theorems 4.1.3 and 4.1.5 shows that the elements of

$$\{[\hat{\phi}_h,\hat{\psi}_h]-[\overset{\circ}{\phi}_h,\overset{\circ}{\psi}_h]\}\hat{\Gamma}_h$$

are, uniformly $O(Q_T)$, a.s. Since $|\Gamma_h^{-1}|_1$ is bounded (lemma 4.2.2) then

$$\sup_j \|\hat{\phi}_h(j)-\mathring{\phi}_h(j),\hat{\psi}_h(j)-\mathring{\psi}_h(j)\| = O(Q_T) \text{ a.s.}$$

From ~~(4.1.5)~~ (4.2.18)

$$\sum_{j=1}^{h} \|\mathring{\phi}_h(j)-\phi(j),\mathring{\psi}_h(j)-\psi(j)\| = O(T^{-\frac{1}{2}})$$

hence

$$\sup_j \|\hat{\phi}_h(j)-\phi(j),\hat{\psi}_h(j)-\psi(j)\| = O(Q_T) \text{ a.s.}$$

as required. □

From the discussion below (3.3.2) we have that

$$0 < \underline{c} \leqslant \liminf_{h\to\infty} \rho_0^h \sum_{h+1}^{\infty} \|\phi(j),\psi(j)\|$$

$$\leqslant \limsup_{h\to\infty} \rho_0^h \sum_{h+1}^{\infty} \|\phi(j),\psi(j)\| \leqslant \bar{c} < \infty$$

and a similar result for the superior and inferior limits of

$$\rho_0^{2h} \sum_{h+1}^{\infty} \|\phi(j),\psi(j)\|^2$$

As a result the inferior and superior limits, as $h\uparrow\infty$, of

$$\{\sum_{h+1}^{\infty} \|\phi(j),\psi(j)\|^2\}/\{\sum_{h+1}^{\infty} \|\phi(j),\psi(j)\|\}^2 \tag{5.1.9}$$

are bounded away from zero and infinity respectively. This will be used repeatedly in what follows.

(iii) Lemma 3.3.3

Denote the following expressions by $S_1,\ldots,S_5$ respectively

$$\frac{1}{T}\Sigma\{\overset{\circ}{\varepsilon}_h(t)-\varepsilon(t)\}\{\overset{\circ}{\varepsilon}_h(t)-\varepsilon(t)\}';\ \frac{1}{T}\Sigma\varepsilon(t)\{\overset{\circ}{\varepsilon}_h(t)-\varepsilon(t)\}'$$

$$\frac{1}{T}\Sigma\{\hat{\varepsilon}_h(t)-\overset{\circ}{\varepsilon}_h(t)\}\{\overset{\circ}{\varepsilon}_h(t)-\varepsilon(t)\}';\ \frac{1}{T}\Sigma\varepsilon(t)\{\hat{\varepsilon}_h(t)-\overset{\circ}{\varepsilon}_h(t)\}'$$

$$\frac{1}{T}\Sigma\{\hat{\varepsilon}_h(t)-\overset{\circ}{\varepsilon}_h(t)\}\{\hat{\varepsilon}_h(t)-\overset{\circ}{\varepsilon}_h(t)\}'$$

where $\overset{\circ}{\varepsilon}_h(t) = \sum_0^h \overset{\circ}{\phi}_h(j)y(t-j) + \sum_1^h \overset{\circ}{\psi}_h(j)z(t-j)$. Then

$$\hat{\Phi}_h = \frac{1}{T}\sum_1^T \varepsilon(t)\varepsilon(t)' + S_1 + (S_2+S_2') + (S_3+S_3') + (S_4+S_4'+S_5) .$$

Write S_1 as

$$\{[\overset{\circ}{\phi}_h,\overset{\circ}{\psi}_h]-[\phi,\psi]\}\tilde{\Gamma}_\infty\{[\overset{\circ}{\phi}_h,\overset{\circ}{\psi}_h]-[\phi,\psi]\}' \tag{5.1.10}$$

For example, we can show that when $K_T = c \log T$, c sufficiently large,

$$\sum_{i,j=K_T}^{\infty}\sum \phi(i)\frac{1}{T}\sum_1^T y(t-i)z(t-j)'\psi(j)' = o(T^{-1})$$

using the arguments in lemma 4.1.4. Hence we truncate the infinite series in (5.1.10) to obtain

$$S_1 = \{[\overset{\circ}{\phi}_h,\overset{\circ}{\psi}_h]-[\phi,\psi]\}\tilde{\Gamma}_{K_T}\{[\overset{\circ}{\phi}_h,\overset{\circ}{\psi}_h]-[\phi,\psi]\}' + o(T^{-1})$$

The elements in $\tilde{\Gamma}_{K_T} - \hat{\Gamma}_{K_T}$ are $o(H_T T^{-\frac{1}{2}})$ as in the evaluation of (5.1.8) while those in $\hat{\Gamma}_{K_T} - \overset{\circ}{\Gamma}_{K_T}$ are uniformly $o(H_T^{-1})$. Hence the largest eigenvalue of $\tilde{\Gamma}_{K_T} - \overset{\circ}{\Gamma}_{K_T}$ is $o(1)$ and

$$S_1 = \{[\overset{\circ}{\phi}_h,\overset{\circ}{\psi}_h]-[\phi,\psi]\}\overset{\circ}{\Gamma}_{K_T}\{[\overset{\circ}{\phi}_h,\overset{\circ}{\psi}_h]-[\phi,\psi]\}'(1+o(1)) + o(T^{-1}) \text{ a.s.}$$

The same argument used to truncate the infinite sums allows us to restore them, thus $\overset{\circ}{\Gamma}_{K_T}$ may be replaced by $\overset{\circ}{\Gamma}_\infty$ in the last expression.

Consider S_2. The arguments used above allow us to truncate the infinite sums so that S_2' is

$$\frac{1}{T}\sum_1^T \left\{\sum_1^{K_T} \{[\overset{\circ}{\phi}_h(j)-\phi(j)]y(t-j) + [\overset{\circ}{\psi}_h(j)-\psi(j)]z(t-j)\}\right\}\varepsilon(t)'+o(T^{-1})$$

Now $\frac{1}{T}\sum_1^T y(t-j)\varepsilon(t)'$, $\frac{1}{T}\sum_1^T z(t-j)\varepsilon(t)'$ are almost surely $O(Q_T)$ uniformly in $j \leqslant H_T$ by theorem 4.1.3, (4.1.4) and noting that

$$\frac{1}{T}\sum_1^j y(t-j)\varepsilon(t)' \ , \ \frac{1}{T}\sum_1^j z(t-j)\varepsilon(t)' = o(T^{-\frac{1}{2}}) \ , \ \text{a.s.}$$

Hence, using (4.2.15) we obtain

$$S_2 = O(Q_T \sum_{h+1}^{\infty} \|\phi(j),\psi(j)\|) + o(T^{-1}) \ \text{a.s.}$$

We can write

$$S_3 = \{[\hat{\phi}_h,\hat{\psi}_h]-[\overset{\circ}{\phi}_h,\overset{\circ}{\psi}_h]\}\tilde{\Gamma}_\infty\{[\overset{\circ}{\phi}_h,\overset{\circ}{\psi}_h]-[\phi,\psi]\}'.$$

As before, the infinite sums may be truncated at K_T , i.e.

$$S_3 = \{[\hat{\phi}_h,\hat{\psi}_h]-[\overset{\circ}{\phi}_h,\overset{\circ}{\psi}_h]\}\tilde{\Gamma}_{K_T}\{[\overset{\circ}{\phi}_h,\overset{\circ}{\psi}_h]-[\phi,\psi]\}'+o(T^{-1}).$$

Just as $|\hat{\Gamma}_{K_T}|_1$ is bounded, $|\tilde{\Gamma}_{K_T}|_1$ is bounded as the elements of $\hat{\Gamma}_{K_T} - \tilde{\Gamma}_{K_T}$ are, uniformly, $o(H_T T^{-\frac{1}{2}})$ as in (5.1.8). The elements of $[\hat{\phi}_h-\overset{\circ}{\phi}_h],[\hat{\psi}_h-\overset{\circ}{\psi}_h]$ are evaluated in theorem 3.3.1 to be, uniformly, $O(Q_T) + o(h^{-1}\sum_{h+1}^{\infty} \|\phi(j),\psi(j)\|)$ hence $\{[\hat{\phi}_h,\hat{\psi}_h]-[\overset{\circ}{\phi}_h,\overset{\circ}{\psi}_h]\}\tilde{\Gamma}_{K_T}$ is also of this order. Bounding $\sum_1^{\infty} \|\overset{\circ}{\phi}_h(j)-\phi(j),\overset{\circ}{\psi}_h(j)-\psi(j)\|$ as in (4.2.15) gives

$$S_3 = O(Q_T \sum_{h+1}^{\infty} \|\phi(j),\psi(j)\|)+o(h^{-1}\{\sum_{h+1}^{\infty} \|\phi(j),\psi(j)\|\}^2)+o(T^{-1}), \ \text{a.s.}$$

Finally we turn to the term $S_4+S_4'+S_5$. Recall $y(t,h) = [y(t-1)',\ldots,y(t-h)']'$ as in §3.1.

We denote

$$([\hat{\phi}_h,\hat{\psi}_h]-[\mathring{\phi}_h,\mathring{\psi}_h])+\{\frac{1}{T}\sum_1^T\varepsilon(t)[y(t,h)',z(t,h)']\}\hat{\Gamma}_h^{-1}$$

by F. Thus

$$\hat{\varepsilon}_h(t)-\mathring{\varepsilon}_h(t)$$

$$=\sum_1^h\{[\hat{\phi}_h(j)-\mathring{\phi}_h(j)]y(t-j)+[\hat{\psi}_h(j)-\mathring{\psi}_h(j)]z(t-j)\}$$

$$=-\frac{1}{T}\sum_1^T\varepsilon(t)[y(t,h)',z(t,h)']\hat{\Gamma}_h^{-1}\begin{bmatrix}y(t,h)\\z(t,h)\end{bmatrix}+F\begin{bmatrix}y(t,h)\\z(t,h)\end{bmatrix}.$$

Now

$$S_5=\{\frac{1}{T}\sum_1^T\varepsilon(t)[y(t,h)',z(t,h)']\}\hat{\Gamma}_h^{-1}\left\{\frac{1}{T}\sum_1^T\begin{bmatrix}y(t,h)\\z(t,h)\end{bmatrix}\varepsilon'(t)\right\}$$

$$-F\left\{\frac{1}{T}\sum_1^T\begin{bmatrix}y(t,h)\\z(t,h)\end{bmatrix}\varepsilon'(t)\right\}-\left\{\frac{1}{T}\sum_1^T\varepsilon(t)[y(t,h)',z(t,h)']\right\}F'$$

$$+F\hat{\Gamma}_hF'+o(T^{-1})$$

and

$$S_4=-\{\frac{1}{T}\sum_1^T\varepsilon(t)[y(t,h)',z(t,h)']\}\hat{\Gamma}_h^{-1}\left\{\frac{1}{T}\sum_1^T\begin{bmatrix}y(t,h)\\z(t,h)\end{bmatrix}\varepsilon'(t)\right\}$$

$$+\{\frac{1}{T}\sum_1^T\varepsilon(t)[y(t,h)',z(t,h)']\}F'+o(T^{-1}).$$

The terms that are $o(T^{-1})$ in S_4 and S_5 arise from replacing the covariances $\tilde{\gamma}_y(i,j),\tilde{\gamma}_{yz}(i,j),\ldots,\frac{1}{T}\sum_1^T\varepsilon(t)y(t-j)'$ by $\hat{\gamma}_y(i-j),\hat{\gamma}_{yz}(i-j),\ldots,\frac{1}{T}\sum_{j+1}^T\varepsilon(t)y(t-j)'$. Thus $S_4+S_4'+S_5$ is

$$-\{\frac{1}{T}\sum_1^T\varepsilon(t)[y(t,h)',z(t,h)']\}\hat{\Gamma}_h^{-1}\left\{\frac{1}{T}\sum_1^T\begin{bmatrix}y(t,h)\\z(t,h)\end{bmatrix}\varepsilon'(t)\right\}+F\hat{\Gamma}_hF'$$

$$+o(T^{-1}).$$

As the elements of F are $o(h^{-1}\sum_{h+1}^{\infty}\|\phi(j),\psi(j)\|)+o(T^{-\frac{1}{2}})$, and $|\hat{\Gamma}_h|_1$ is finite

$$(5.1.11)\qquad F\,\hat{\Gamma}_h\,F' = o(h/T)+o(h^{-1}\{\sum_{h+1}^{\infty}\|\phi(j),\psi(j)\|\}^2)$$

From the discussion below (3.3.5) $c\sum_{h+1}^{\infty}\|\phi(j),\psi(j)\|^2 \leqslant \|S_1\|$, $c>0$ hence the discussion around (5.1.9) shows that

$$h^{-1}\{\sum_{h+1}^{\infty}\|\phi(j),\psi(j)\|\}^2 = o(\|\{[\mathring{\phi}_h,\mathring{\psi}_h]-[\phi,\psi]\}\mathring{\Gamma}_\infty\{[\mathring{\phi}_h,\mathring{\psi}_h]-[\phi,\psi]\}'\|).$$

It remains to show that term $O(Q_T\{\sum_{h+1}^{\infty}\|\phi(j),\psi(j)\|\})$ in S_2, S_3 is dominated by the larger of h/T, $\|\{[\mathring{\phi}_h,\mathring{\psi}_h]-[\phi,\psi]\}\mathring{\Gamma}_\infty\{[\mathring{\phi}_h,\mathring{\psi}_h]-[\phi,\psi]\}'\|$. As the latter term is bounded below by $c\{\sum_{h+1}^{\infty}\|\phi(j),\psi(j)\|^2\}$, $c>0$ we show

$$P(T) = Q_T\{\sum_{h(T)+1}^{\infty}\|\phi(j),\psi(j)\|\}/\max\{h(T)/T,\ \sum_{h(T)+1}^{\infty}\|\phi(j),\psi(j)\|^2\}$$

converges to zero. Here we have put $h(T)=h$ to emphasize the dependence of h on T. As usual, c represents a constant, not necessarily the same one. If $\limsup P(T) = c>0$, then along a subsequence $\{T_i\}$, $P(T_i)\to c$. If $\limsup h(T_i)/\log T \geqslant d = 1/(-2\log\rho_0)$, then there is a subsequence $\{\tilde{T}_i\}$ of $\{T_i\}$ where $\tilde{T}_i^{\frac{1}{2}}\{\sum_{h(\tilde{T}_i)+1}^{\infty}\|\phi(j),\psi(j)\|\}\to c_1<\infty$.

Thus $P(\tilde{T}_i) \leqslant Q_{\tilde{T}_i}\{\sum_{h(\tilde{T}_i)+1}^{\infty}\|\phi(j),\psi(j)\|\}/\{h(\tilde{T}_i)\tilde{T}_i^{-1}\}$

$$= O(Q_{\tilde{T}_i}\tilde{T}_i^{\frac{1}{2}}/h(\tilde{T}_i)).$$

This contradicts $P(T_i)\to c$. Assume $\liminf h(T_i)/\log T_i = c_1 < d$. Then along some subsequence $\{\tilde{T}_i\}$, $h(\tilde{T}_i)/\log T_i \to c_1$ and $\tilde{T}_i^b\{\sum_{h(\tilde{T}_i)+1}^{\infty}\|\phi(j),\psi(j)\|\}\to c>0$ for some $b<\frac{1}{2}$. From (5.1.9) $\liminf \tilde{T}_i^{2b}\sum_{h(\tilde{T}_i)+1}^{\infty}\|\phi(j),\psi(j)\|^2$ is bounded away from zero.

Consequently

$$P(\tilde{T}_i) \leqslant Q_{\tilde{T}_i} \sum_{h(\tilde{T}_i)+1}^{\infty} \|\phi(j),\psi(j)\| / \sum_{h(\tilde{T}_i)+1}^{\infty} \|\phi(j),\psi(j)\|^2$$

$$= O(Q_{\tilde{T}_i} \tilde{T}_i^{b})$$

again contradicting $P(T_i) \to c > 0$. Thus lim sup $P(T) = 0$ as required. Thus the terms S_2, S_3 are

$$o(\max\{h/T, \|\{[\mathring{\phi}_h,\mathring{\psi}_h]-[\phi,\psi]\}\mathring{\Gamma}_\infty\{[\mathring{\phi}_h,\mathring{\psi}_h]-[\phi,\psi]\}'\}\|)$$

concluding the lemma. □

(iv) Theorems 3.3.4, 3.3.5

Let $\mathring{\ell}_B(T)$ minimise

$$\mathrm{tr}\{\mathring{\Sigma}_0^{-1}([\mathring{\phi}_h,\mathring{\psi}_h]-[\phi,\psi])\mathring{\Gamma}_\infty([\mathring{\phi}_h,\mathring{\psi}_h]-[\phi,\psi])'\} + hs(s+m)\log T/T$$

Again we have

$$(5.1.12) \quad \underline{c}\ \rho_0^{2h} \leqslant \mathrm{tr}\{\mathring{\Sigma}_0^{-1}([\mathring{\phi}_h,\mathring{\psi}_h]-[\phi,\psi])\mathring{\Gamma}_\infty([\mathring{\phi}_h,\mathring{\psi}_h]-[\phi,\psi])'\} \leqslant \bar{c}\ \rho_0^{2h},$$

$0 < \underline{c} \leqslant \bar{c} < 0$, hence $\mathring{\ell}_B(T)/\ell_B(T) = 1 + o(1)$ by evaluating $\mathring{\ell}_B(T)$ as $\log T\{1/(-2\log \rho_0)\}(1+o(1))$ in the same way as we obtained $\ell_B(T)$. If lim sup $\hat{\ell}_B(T)/\mathring{\ell}_B(T) > 1$ then along some subsequence $\{T_i\}$ $\hat{\ell}_B(T_i)/\mathring{\ell}_B(T_i) \to a > 1$, and the upper and lower bounds in (5.1.12) are $\bar{c}\ T_i^{-a}$, $\underline{c}\ T_i^{-a}$ respectively.

$$(5.1.13) \quad \mathrm{BIC}(\hat{\ell}_B(T_i)) - \mathrm{BIC}(\mathring{\ell}_B(T_i))$$

$$\geqslant \underline{c}\ T_i^{-a}(1+o(1)) - \bar{c}\ T_i^{-1}(1+o(1)) + s(s+m)(\hat{\ell}_B(T_i)-\mathring{\ell}_B(T_i))\{\log T_i/T_i\}(1+o(1))$$

The dominant term in the RHS of (5.1.13) is $(\hat{\ell}_B(T_i)-\mathring{\ell}_B(T_i))\{\log T_i/T_i\}$ hence (5.1.13) is positive and BIC(h) could

not achieve a minimum at infinitely many values, $\hat{\ell}_B(T_i)$. Similarly when $\liminf \hat{\ell}_B(T)/\mathring{\ell}_B(T) < 1$ then $a < 1$ in (5.1.13) and then the dominant term is $\underline{c}\, T_i^{-a} > 0$ hence $\liminf \hat{\ell}_B(T)/\mathring{\ell}_B(T)$ cannot be less than unity. Thus $\lim \hat{\ell}_B(T)/\mathring{\ell}_B(T) = 1$ and Theorem 3.3.4 is proved. □

The proof of theorem 3.3.5 is simplified because we already have $\liminf \hat{\ell}_A(T)/\ell_A(T) \geqslant 1$ a.s. When $\limsup \hat{\ell}_A(T)/\mathring{\ell}_A(T) > 1$ (define $\mathring{\ell}_A(T)$ in an analogous way to $\mathring{\ell}_B(T)$) we have, in the place of (5.1.13),

$$\mathrm{AIC}(\hat{\ell}_A(T_i)) - \mathrm{AIC}(\mathring{\ell}_A(T_i))$$

$$\geqslant \underline{c}\, T_i^{-a}(1+o(1)) - \bar{c}\, T_i^{-1}(1+o(1) + s(s+m)\{(\hat{\ell}_A(T_i)-\mathring{\ell}_A(T_i))/T_i\}(1+o_p(1))$$

which is again dominated by the last term as when $a > 1$, $\{\hat{\ell}_A(T_i)-\mathring{\ell}_A(T_i)\}/\log T_i \to c > 0$. Thus Theorem 3.3.5 is established. □

(v) The remaining lemmas in this section are needed to evaluate the expressions

$$\frac{1}{h}\sum_1^T \{\hat{\varepsilon}_h(t)-\varepsilon(t)\}\{\hat{\varepsilon}_h(t-a)-\varepsilon(t-a)\}', \quad \frac{1}{h}\sum_1^T \varepsilon(t)\{\hat{\varepsilon}_h(t-a)-\varepsilon(t-a)\}$$

that are fundamental for the theorems in Stages II and III. These lemmas are included here because when $a = 0$ they are needed to evaluate (3.3.7) in the expressions for $\mathrm{AIC}(h)$.

Lemma 5.1.1. Assume $y(t)$ is generated by an ARMAX process and $\varepsilon(t)$, $z(t)$ satisfy conditions A. Then if $h \geqslant c \log T$, $c > 0$, and $\mathring{\Gamma} = \frac{1}{T}\sum_1^T \varepsilon(t)\varepsilon(t)'$ we have, almost surely,

$$\frac{1}{hT}\sum_{s,t=1}^{T\,\Sigma\Sigma} \varepsilon(s)[y(s,h)',z(s,h)']\tilde{\Gamma}_h^{-1}\tilde{\Gamma}_h(a)\tilde{\Gamma}_h^{-1}\begin{bmatrix} y(t,h) \\ z(t,h) \end{bmatrix}\varepsilon(t)'$$

$$= \frac{1}{hT} \sum_{j=1}^{h-a} \Big(\sum_{s=1}^{T} \varepsilon(s)\varepsilon(s-j)' \Big) \mathring{\ddagger}^{-1} \Big(\sum_{t=1}^{T} \varepsilon(t-j-a)\varepsilon(t)' \Big)$$

$$+ \frac{1}{hT} \sum_{s,t=1}^{T} \varepsilon(s) z(s,h)' \tilde{\Gamma}_{z,h}^{-1} \tilde{\Gamma}_{z,h}(a) \tilde{\Gamma}_{z,h}^{-1} z(t,h)\varepsilon(t)' + O\left(\frac{\log \log T}{\log T} \right)$$

Proof: Define $\tilde{A}$, $\tilde{A}_h$ as in §4.2, let

$$H = \begin{bmatrix} \phi(h+1) & \psi(h+1) & \phi(h+2) & \psi(h+2) & \dots \\ \phi(h) & \psi(h) & \phi(h+1) & \dots & \dots \\ \vdots & & & & \\ \phi(1) & \psi(1) & \psi(2) & \dots & \dots \end{bmatrix}$$

and

$$w_t = [y(t-h-1)', z(t-h-1)', y(t-h-2)', \dots]' .$$

Then

$$\tilde{A}_h \begin{bmatrix} y(t,h) \\ z(t,h) \end{bmatrix} = \varepsilon(t,h) + Hw_t . \tag{5.1.14}$$

Let $\tilde{A} = \begin{bmatrix} A_h & \\ 0 & I_{hm} \end{bmatrix}$ as in §4.2. As in the evaluation of (5.1.8) we have, for example $\hat{\gamma}_{zy}(i-j) - \hat{\gamma}_{zy}(i,j) = o(hT^{-\frac{1}{2}})$, hence $\| \hat{\Gamma}_h - \tilde{\Gamma}_h \|_1 = o(1)$. Thus in (4.2.18), (4.2.19) we can replace $\hat{\Gamma}_h, \hat{\Gamma}_h^{-1}, \dots$ by $\tilde{\Gamma}_h, \tilde{\Gamma}_h^{-1}, \dots$ and obtain

$$\tilde{\Gamma}_h^{-1} = \tilde{A}' \begin{bmatrix} I_h \otimes \mathring{\ddagger}^{-1} & 0 \\ 0 & \tilde{\Gamma}_{z,h}^{-1} \end{bmatrix} \tilde{A} + D , \quad \|D\|_1 \leq c < \infty$$

$$\tilde{A}\tilde{\Gamma}_h(a)\tilde{A}' = \begin{bmatrix} I_h(a) \otimes \mathring{\ddagger} & 0 \\ 0 & \tilde{\Gamma}_{z,h}(a) \end{bmatrix} + D_1 , \quad \|D_1\|_1 \leq c < \infty .$$

Thus

$$\sum_s \varepsilon(s)[y(s,h)', z(s,h)']\tilde{\Gamma}_h^{-1}$$

$$= \sum_s \varepsilon(s)[y(s,h)', z(s,h)']\left\{\tilde{A}'\begin{bmatrix} I_h \otimes \overset{\circ}{\Sigma}{}^{-1} & 0 \\ 0 & \tilde{\Gamma}_{z,h}^{-1} \end{bmatrix}\tilde{A} + D\right\}$$

$$= \sum_s \varepsilon(s)[\varepsilon(s,h)' + w_t'H', z(s,h)']\left\{\begin{bmatrix} I_h \otimes \overset{\circ}{\Sigma}{}^{-1} & 0 \\ 0 & \tilde{\Gamma}_{z,h}^{-1} \end{bmatrix}\tilde{A} + D_2\right\} ,$$

hence

$$\sum_{s,t}\sum \varepsilon(s)[y(s,h)', z(s,h)']\tilde{\Gamma}_h^{-1}\tilde{\Gamma}_h(a)\tilde{\Gamma}_h^{-1}\begin{bmatrix} y(t,h) \\ z(t,h) \end{bmatrix}\varepsilon(t)'$$

$$= \sum_{s,t}\sum \varepsilon(s)[\varepsilon(s,h)', z(s,h)']\begin{bmatrix} I_h \otimes \overset{\circ}{\Sigma}{}^{-1} & 0 \\ 0 & \tilde{\Gamma}_{z,h}^{-1} \end{bmatrix}\begin{bmatrix} I_h(a) \otimes \overset{\circ}{\Sigma} & 0 \\ 0 & \hat{\Gamma}_{z,h}(a) \end{bmatrix}$$

(5.1.15)

$$\cdot \begin{bmatrix} I_h \otimes \overset{\circ}{\Sigma}{}^{-1} & 0 \\ 0 & \tilde{\Gamma}_{z,h}^{-1} \end{bmatrix}\begin{bmatrix} \varepsilon(t,h) \\ z(t,h) \end{bmatrix}\varepsilon(t)'$$

$$+ \text{ other terms.}$$

It remains to evaluate the "other terms" in (5.1.15). These terms are of the form $x_1'X\,x_2$. Here x_1, x_2 are either $\sum_s \varepsilon(s)[y(s,h)', z(s,h)']$ or $\sum_s \varepsilon(s)w_s'H'$ and consequently are matrices whose entries are, uniformly, $O(TQ_T)$, a.s. The matrix X is a product of three terms selected from

$$\begin{bmatrix} I_h \otimes \overset{\circ}{\Sigma}^{-1} & 0 \\ 0 & \tilde{\Gamma}^{-1}_{z,h} \end{bmatrix}, \begin{bmatrix} I_h(a) \otimes \overset{\circ}{\Sigma} & 0 \\ 0 & \tilde{\Gamma}_{z,h}(a) \end{bmatrix}, D, D_1 .$$

If the product contains either D or D_1 then $\|X\|_1 \leq c < \infty$ as $\|D\|$, $\|D_1\|_1$ are bounded and the other matrices are bounded in the norm $|.|_1$. In this case, $x_1'X\,x_2 = O(T^2Q_T^2)$ $= O(T \log\log T)$, a.s. thus $x_1'X\,x_2 = o(hT)$ as $h > c \log T$. The only other possibility is that

$$x_1'X\,x_2 = \sum_{s,t}\{\varepsilon(s)w_sH'\}(I_h \otimes \overset{\circ}{\Sigma}^{-1})(I_h(a) \otimes \overset{\circ}{\Sigma})(I_h \otimes \overset{\circ}{\Sigma}^{-1})\{Hw_t\varepsilon(t)'\} .$$

Here

$$\|H\|_1 \leq \sum_{i=1}^{h}\sum_{j=i}^{\infty}\|\phi(j),\psi(j)\| \leq c < \infty ,$$

consequently

$$\Big\|\sum_s\{\varepsilon(s)w_sH'\}\Big\|_1 = O(TQ_T), \text{ a.s.}$$

and then $x_1'X\,x_2 = O(TQ_T)$. Thus we have established that the small order terms in (5.1.15) are $O(T \log\log T)$, almost surely. □

We are only using this lemma in conjunction with lemma 5.1.2 where conditions B *are* assumed, in which case Σ_0 will replace $\overset{\circ}{\Sigma}$ throughout lemma 5.1.1. The expression $\overset{\circ}{\Sigma}$ only arises through our argument to obtain (4.2.18), (4.2.19).

LEMMA 5.1.2. *Assume* $\varepsilon(t)$ *and* $z(t)$ *satisfy conditions B. If* $0 \leq a < A$, *then, uniformly for* $c \log T \leq h \leq H_T$

(i) $$\frac{1}{hT}\sum_{j=1}^{h-a}\Big(\sum_{s=1}^{T}\varepsilon(s)\varepsilon(s-j)'\Big)\Sigma_0^{-1}\Big(\sum_{t=1}^{T}\varepsilon(t-j-a)\varepsilon(t)'\Big) = \delta_{0a}\Sigma_0 s + o_p(1)$$

(ii) $$\frac{1}{hT}\sum_{s,t=1}^{T}\varepsilon(s)z(s,h)'\tilde{\Gamma}^{-1}_{z,h}\tilde{\Gamma}_{z,h}(a)\tilde{\Gamma}^{-1}_{z,h}z(t,h)\varepsilon(t)' = \delta_{0a}\Sigma_0 m + o_p(1)$$

Proof: (i) We partition the expression

$$\frac{1}{T}\sum_{u,v=1}^{T}\varepsilon(u)\varepsilon(u-j)'\sharp_0^{-1}\varepsilon(v-j-a)\varepsilon(v)'$$

into three parts S_A, S_B, $S_C(j)$ where A, B, C represent the ranges of the summations

$$A = \{(u,v) : u = v\}$$

$$B = \{(u,v) : u \neq v \text{ but } u = v-j-a \text{ or } v = u-j\}$$

$$C = \{(u,v) : u \neq v,\ u \neq v-j-a,\ v \neq u-j\}$$

We proceed by evaluating moments of

$$S_A - \delta_{0a}\sharp_0 s \ , \ S_B \ , \ S_C(j)$$

Take the case when $a = 0$ and consider $S_A - \sharp_0 s$:

$$\frac{1}{T}\sum_u \varepsilon(u)\varepsilon(u-j)'\sharp_0^{-1}\varepsilon(u-j)\varepsilon(u)' - \sharp_0 s$$

$$= \frac{1}{T}\sum_u \{\varepsilon(u)\varepsilon(u)' - \sharp_0\}\mathrm{tr}\{\sharp_0^{-1}\varepsilon(u-j)\varepsilon(u-j)'\} \tag{5.1.16}$$

$$+ \sharp_0\ \mathrm{tr}\{\sharp_0^{-1}\ \frac{1}{T}\sum\varepsilon(u-j)\varepsilon(u-j)' - I_s\}$$

The last term converges (almost surely and uniformly in j) to zero at a rate $O(Q_T)$ so we can ignore it. For $2 > p > 1$ we evaluate the p^{th} moment of the (ℓ,k) element of the remaining part of (5.1.16). Let $\sharp_{\ell k}$ denote the (ℓ,k) element of $\sharp_0$.

$$E\{[\frac{1}{T}\sum_u(\varepsilon_\ell(u)\varepsilon_k(u) - \sharp_{\ell k})\mathrm{tr}(\sharp_0^{-1}\varepsilon(u-j)\varepsilon(u-j)')]^p\}$$

$$\leqslant \frac{c}{T^p}E\{[\sum_u(\varepsilon_\ell(u)\varepsilon_k(u) - \sharp_{\ell k})^2(\mathrm{tr}(\sharp_0^{-1}\varepsilon(u-j)\varepsilon(u-j)'))^2]^{p/2}\}$$

(from Burkholder's inequality, Hall and Heyde (1980), p.23)

$$\leq \frac{c}{T^p} \sum_u E\{(\varepsilon_\ell(u)\varepsilon_k(u) - \sharp_{\ell k})^p (\mathrm{tr}(\sharp_0^{-1}\varepsilon(u-j)\varepsilon(u-j)')^p\}$$

Using Hölder's inequality the expectation in this term is bounded by

$$[E\{(\varepsilon_\ell(u)\varepsilon_k(u) - \sharp_{\ell k})^{2p}\}]^{\frac{1}{2}}[E\{(\mathrm{tr}(\sharp_0^{-1}\varepsilon(u-j)\varepsilon(u-j)'))^{2p}\}]^{\frac{1}{2}}.$$

The bound is independent of u and j. This argument requires the existence of the $4p$ moment of $\varepsilon(t)$. Hence the p^{th} moment of (5.1.16) is c/T^{p-1} and thus (5.1.16) converges to zero in probability uniformly in j. Hence

$$\frac{1}{h} \sum_{j=1}^{h} S_A \to 0$$

uniformly in h.

Consider the sum over B. Without loss of generality we take $v = u-j$ in S_B. Then $E\{S_B S_B'\}$ is

$$(5.1.17) \quad E\{\varepsilon(u)\varepsilon(u-j)'\sharp_0^{-1}\varepsilon(u-2j)\varepsilon(u-j)'\varepsilon(v-j)\varepsilon(v-2j)'\sharp_0^{-1}\varepsilon(v-j)\varepsilon(v)'\}.$$

Unless $u = v$ this term is zero. For example, take $u > v$. Then (5.1.17) is of the form $E\{\varepsilon(u)f(u,v,j)\}$ and $f(u,v,j)$ is measurable with respect to F_{u-1}. Thus $E\{\varepsilon(u)f(u,v,j)\}$ $= E\{E(\varepsilon(u) \mid F_{u-1})f(u,v,j)\} = 0$. Hence (5.1.17) is

$$\sharp_0 \; E\{\varepsilon(u-j)'\sharp_0^{-1}\varepsilon(u-2j)\varepsilon(u-j)'\varepsilon(u-j)\varepsilon(u-2j)'\sharp_0^{-1}\varepsilon(u-j)\}$$

The terms within this expectation are composed of linear combinations of $e(u-j)f(u-2j)$ where $e(s) = \varepsilon_{\ell_1}(s)\varepsilon_{\ell_2}(s)\varepsilon_{\ell_3}(s)\varepsilon_{\ell_4}(s)$, $f(s) = \varepsilon_{\ell_5}(s)\varepsilon_{\ell_6}(s)$. Using Burkholder and Hölder inequalities as above for $2 > p > 1$ we evaluate

$$E\{[\frac{1}{T} \sum_u e(u-j)f(u-2j)]^p\}$$

$$\leq \frac{c}{T^p} E\{[\Sigma_u e(u-j)^2 f(u-2j)^2]^{p/2}\}$$

$$\leq \frac{c}{T^p} \Sigma_u E\{e(u-j)^p f(u-2j)^p\}$$

$$\leq \frac{c}{T^p} \Sigma_u \{E[e(u-j)^{3p/2}]\}^{2/3} \{E[f(u-2j)^{3p}]\}^{1/3} .$$

The summands in the last expression are finite as $\varepsilon(t)$ has a moment of order $6p$, hence the p^{th} moment of S_B is c/T^{p-1} and S_B converges to zero in probability, uniformly in j. Consequently

$$\frac{1}{h} \sum_{j=1}^{h} S_B \to 0$$

uniformly in h.

Finally consider the sum $S_C(j)$. First we make the observation that if $p(u) = \varepsilon(u_1)\varepsilon(u_2)'\varepsilon(u_3)\varepsilon(u_4)'$ where $u_1 > u_2 > u_3 > u_4$ then $E(p(u)p(v)') = 0$ unless $u_i = v_i$, $i = 1,\ldots,4$. The argument below (5.1.17) shows this. By a similar argument we evaluate $E(p(u)p(u)') = \Sigma_0 tr\Sigma_0 tr(\Sigma_0^2)$. For a given h, the variance of $\sum_{j=1}^{h} S_C$ is

$$(5.1.18) \qquad \frac{1}{T^2} \sum_{j,k=1}^{h} \sum\sum\sum\sum_{\substack{(u_1,u_2)\in C \\ (u_3,u_4)\in C}} E\{\varepsilon(u_1)\varepsilon(u_1-j)'\Sigma_0^{-1}\varepsilon(u_2-j)\varepsilon(u_2)' \varepsilon(u_3)\varepsilon(u_3-k)'\Sigma_0^{-1}\varepsilon(u_4-k)\varepsilon(u_4)'\}$$

Without loss of generality, assume $u_1 > u_2$, $u_4 > u_3$. Then the summands are zero unless $u_1 = u_4$, $u_2 = u_3$, $k = j$. As

$$E\{\varepsilon(u_1)\varepsilon(u_1-j)'\Sigma_0^{-1}\varepsilon(u_2-j)\varepsilon(u_2)'\varepsilon(u_2)\varepsilon(u_2-j)'\Sigma_0^{-1}\varepsilon(u_1-j)\varepsilon(u_1)'\}$$

$$= s\Sigma_0 tr\Sigma_0$$

and there are hT^2 non zero terms in the summation in (5.1.18) we obtain $\|\mathrm{var}\{\sum_{j=1}^{h} S_C(j)\}\| \leq c_1 h$. It remains to show that

$\frac{1}{h}\sum_{j=1}^{h} S_C(j) \to 0$ uniformly for $c \log T \leqslant h \leqslant H_T$. Let ξ denote $(c \log T)^{\frac{1}{2}}$ and define p by $\xi^{p-1} < H_T \leqslant \xi^p$. Then

$$P\{\max_{\xi \leqslant h \leqslant H_T} \| \frac{1}{h} \sum_{1}^{h} S_C(j) \| > \varepsilon\}$$

$$\leqslant \sum_{k=2}^{p} P\{\max_{\xi^k \leqslant h < \xi^{k+1}} \| \frac{1}{h} \sum_{1}^{h} S_C(j) \| > \varepsilon\}$$

$$\leqslant \frac{1}{\varepsilon^2} \sum_{k=2}^{p} E\{[\max_{\xi^k \leqslant h \leqslant \xi^{k+1}} \| \frac{1}{h} \sum_{1}^{h} S_C(j) \|]^2\}$$

Now Theorem B of Moricz (1976) applies because $\| \mathrm{var}\{\sum_{j=1}^{h} S_C(j)\| \leqslant c_1 h$, giving us

$$E\{[\max_{\xi^k \leqslant h \leqslant \xi^{k+1}} \| \frac{1}{h} \sum_{1}^{h} S_C(j) \|]^2\}$$

$$\leqslant c_2 \, \xi^{(k+1)(1+\delta)} / \xi^{2k} \, , \quad \delta > 0$$

$$= c_2 \, \xi^{k(\delta-1)+1+\delta}$$

Consequently

$$P\{\max_{\xi \leqslant h \leqslant H_T} \| \frac{1}{h} \sum_{1}^{h} S_C(j) \| > \varepsilon\} \leqslant \frac{c_2}{\varepsilon^2} \sum_{k=2}^{p} \xi^{k(\delta-1)+1+\delta} \to 0$$

as required.

When $a > 0$ cases S_B and S_C are essentially unchanged from what has been presented above while S_A is simplified by dropping the mean correction term when the proof continues as before.

(ii) Put

$$P(u,v) = z(u,h)' \tilde{\Gamma}_{z,h}^{-1} \tilde{\Gamma}_{z,h}(a) \tilde{\Gamma}_{z,h}^{-1} z(v,h) \, .$$

Consider the second moment of

$$S_A = \frac{1}{hT} \sum_{\substack{u,v=1 \\ u>v}}^{T} \varepsilon(u) P(u,v) \varepsilon(v)' \ .$$

From the discussion below (5.1.17),

$$E\{\varepsilon(u_1) P(u_1,v_1) \varepsilon(v_1)' \varepsilon(v_2) P(u_2,v_2) \varepsilon(u_2)'\}$$

is zero unless $u_1 = u_2$, $v_1 = v_2$. Thus if

$\frac{1}{hT^2} \sum_{u,v=1}^{T} P(u,v)^2 < \infty$ then

$$E(S_A S_A') = \frac{1}{h^2 T^2} \sum_{u,v=1}^{T} P(u,v)^2 \Phi_0 \operatorname{tr} \Phi_0 = o(\frac{1}{h}).$$

Here we have evaluated

$$\frac{1}{T^2} \sum_{u,v=1}^{T} P(u,v)^2 = \operatorname{tr}\{\tilde{\Gamma}_{z,h}^{-1} \tilde{\Gamma}_{z,h}(a) \tilde{\Gamma}_{z,h}^{-1} \tilde{\Gamma}_{z,h}(a)\} \ .$$

Thus $S_A \to 0$ in probability and uniformly in h . Take

$$S_B = \frac{1}{hT} \sum_{1}^{T} \varepsilon(u) P(u,u) \varepsilon(u)' \ .$$

The variance of S_B is

$$\frac{1}{h^2 T^2} \left\{ \sum_{u,v=1}^{T} E\{\varepsilon(u) P(u,u) \varepsilon(u)' \varepsilon(v) P(v,v) \varepsilon(v)'\} - \left(\sum_{1}^{T} P(u,u)\right)^2 \Phi_0^2 \right\}$$

When $u \neq v$, $E\{\varepsilon(u)\varepsilon(u)'\varepsilon(v)\varepsilon(v)'\} = \Phi_0^2$, and let $E\{\varepsilon(u)\varepsilon(u)'\varepsilon(u)\varepsilon(u)'\} = S$. Then

$$\operatorname{var}(S_B) = \frac{1}{h^2 T^2} \sum_{1}^{T} P(u,u)^2 (S - \Phi_0^2)$$

which we show is $O(\frac{1}{hT})$, thus $S_B \to E(S_B)$ in probability. To evaluate $E(S_B)$ we have

$$\frac{1}{T} \sum_{1}^{T} P(u,u) = \operatorname{tr}\{\tilde{\Gamma}_{z,h}^{-1} \tilde{\Gamma}_{z,h}(a) \tilde{\Gamma}_{z,h}^{-1} \frac{1}{T} \Sigma z(u,h) z(u,h)'\}$$

$$= \operatorname{tr}\{\tilde{\Gamma}_{z,h}^{-1} \tilde{\Gamma}_{z,h}(a)\}$$

When $a = 0$ this term is hm while when $a \neq 0$ there are only a columns with non-zero entries on the main diagonal of $\tilde{\Gamma}_{z,h}^{-1}\tilde{\Gamma}_{z,h}(a)$. The non trivial entries in $\tilde{\Gamma}_{z,h}^{-1}\tilde{\Gamma}_{z,h}(a)$ are uniformly bounded, hence $\mathrm{tr}\{\tilde{\Gamma}_{z,h}^{-1}\tilde{\Gamma}_{z,h}(a)\} \leqslant c\, a$, $c < \infty$. The proof of (ii) is complete once we show that

$$\frac{1}{hT^2} \sum_{u,v=1}^{T} P(u,v)^2 < \infty$$

But this expression is $\mathrm{tr}\{\tilde{\Gamma}_{z,h}^{-1}\tilde{\Gamma}_{z,h}(a)\tilde{\Gamma}_{z,h}^{-1}\tilde{\Gamma}_{z,h}(a)\}/h$ which equals m when $a = 0$. When $a > 0$, $\tilde{\Gamma}_{z,h}^{-1}\tilde{\Gamma}_{z,h}(a)\tilde{\Gamma}_{z,h}^{-1}\tilde{\Gamma}_{z,h}(a)$ has at most $2a$ non zero elements along the main diagonal hence

$$\frac{1}{hT^2} \sum_{u,v=1}^{T} P(u,v)^2 \leqslant m .$$

□

LEMMA 5.1.3. *Assume* $y(t)$ *is generated by an ARMAX process and* $\varepsilon(t)$, $z(t)$ *satisfy conditions B. If* $d \log T \leqslant h \leqslant H_T$, $d = 1/(-2 \log \rho_0)$ *and* $0 \leqslant a$ *then, uniformly in* h ,

$$\frac{1}{h} \sum_{1}^{T} \{\hat{\varepsilon}_h(t) - \varepsilon(t)\}\{\hat{\varepsilon}_h(t-a) - \varepsilon(t-a)\}' = \delta_{0a} \sharp_0(s+m) + o_p(1)$$

$$\frac{1}{h} \sum_{1}^{T} \varepsilon(t)\{\hat{\varepsilon}_h(t-a) - \varepsilon(t-a)\}' = -\delta_{0a} \sharp_0(s+m) + o_p(1) .$$

Proof: Take the case $a = 0$. First truncate the series

$$\hat{\varepsilon}_h(t) - \varepsilon(t) = \sum_{1}^{\infty} \{(\hat{\phi}_h(j) - \phi(j))y(t-j) + (\hat{\psi}_h(j) - \psi(j))z(t-j)\}$$

after h terms. Then

$$\frac{1}{h} \sum_{1}^{T} \{\hat{\varepsilon}_h(t) - \varepsilon(t)\}\{\hat{\varepsilon}_h(t) - \varepsilon(t)\}'$$

(5.1.19)

$$= \frac{T}{h}\{[\hat{\phi}_h, \hat{\psi}_h] - [\phi, \psi]\}\tilde{\Gamma}_h\{[\hat{\phi}_h, \hat{\psi}_h] - [\phi, \psi]\}' + O\left(\frac{(\log \log T)^{\frac{1}{2}}}{h}\right)$$

Similarly we can replace $[\phi, \psi]$ in the above expression by $[\mathring{\phi}_h, \mathring{\psi}_h]$ as

$$\sum_{1}^{h} \| \overset{\circ}{\phi}_h(j)-\phi(j), \overset{\circ}{\psi}_h(j)-\psi(j) \| \leq M \sum_{h+1}^{\infty} \| \phi(j),\psi(j) \| = O(T^{-\frac{1}{2}})$$

and $|\tilde{\Gamma}_h|_1 < \infty$. This alters (5.1.1) by $O(h^{-1})$. From theorem 3.3.1

$$(5.1.20) \quad \{[\hat{\phi}_h,\hat{\psi}_h]-[\overset{\circ}{\phi},\overset{\circ}{\psi}]\} = -\frac{1}{T}\sum_{1}^{T}\varepsilon(t)[y(t,h)',z(t,h)']\hat{\Gamma}_h^{-1} + o(T^{-\frac{1}{2}})$$

so that (5.1.19) becomes

$$(5.1.21) \quad \frac{1}{hT}\{\sum_{1}^{T}\varepsilon(t)[y(t,h)',z(t,h)']\}\hat{\Gamma}_h^{-1}\tilde{\Gamma}_h\hat{\Gamma}_h^{-1}\left\{\sum_{1}^{T}\begin{bmatrix} y(t,h) \\ z(t,h) \end{bmatrix}\varepsilon(t)'\right\} + o_p(1)$$

The $o_p(1)$ term in the above expression is made of two dominant parts, viz

$$(5.1.22) \quad \begin{aligned} &(T/h)x'\hat{\Gamma}_h^{-1}\tilde{\Gamma}_h\hat{\Gamma}_h^{-1}x \ , \\ &(T/h)\ \frac{1}{T}\{\sum_{1}^{T}\varepsilon(t)[y(t,h)',z(t,h)']\}\hat{\Gamma}_h^{-1}\tilde{\Gamma}_h\hat{\Gamma}_h^{-1}x \end{aligned}$$

where

$$\begin{aligned} x &= \{[\hat{\phi}_h,\hat{\psi}_h]-[\overset{\circ}{\phi}_h,\overset{\circ}{\psi}_h]\}\hat{\Gamma}_h + \frac{1}{T}\sum_{1}^{T}\varepsilon(t)[y(t,h)',z(t,h)'] \\ &= o(T^{-\frac{1}{2}}) \ . \end{aligned}$$

The first of the expressions in (5.1.24) is $o(1)$ as $|\hat{\Gamma}_h^{-1}\tilde{\Gamma}_h\hat{\Gamma}_h^{-1}|_1 < \infty$. To evaluate the second expression in (5.1.24), we have

$$(5.1.23) \quad \begin{aligned} &\|\frac{1}{T}\sum_{1}^{T}\varepsilon(t)[y(t,h)',z(t,h)']\hat{\Gamma}_h^{-1}\tilde{\Gamma}_h\hat{\Gamma}_h^{-1}x\| \\ &\leq \frac{1}{T}\ \|\sum_{1}^{T}\varepsilon(t)[y(t,h)',z(t,h)']\hat{\Gamma}_h^{-1}\tilde{\Gamma}_h\hat{\Gamma}_h^{-1}\left\{\frac{1}{T}\sum_{1}^{T}\begin{bmatrix} y(t,h) \\ z(t,h) \end{bmatrix}\varepsilon(t)'\right\}\|_1^{\frac{1}{2}} \times \\ &\qquad \|\ x'\hat{\Gamma}_h^{-1}\tilde{\Gamma}_h\hat{\Gamma}_h^{-1}x\|_1^{\frac{1}{2}} \end{aligned}$$

so that as $\|x'\hat{\Gamma}_h^{-1}\tilde{\Gamma}_h\hat{\Gamma}_h^{-1}x\|_1 = o(\frac{h}{T})$, and since (5.1.21) is

bounded ((5.1.21) is evaluated below), we have that (5.1.23) is $o_p(h/T)$. Now evaluate (5.1.21). We can replace $\hat{\Gamma}_h^{-1}$ by $\tilde{\Gamma}_h^{-1}$ altering (5.1.21) by $o(1)$. Exactly the same argument follows through when $a > 0$ so that we obtain

$$\frac{1}{h}\sum_1^T \{\hat{\varepsilon}_h(t)-\varepsilon(t)\}\{\hat{\varepsilon}_h(t-a)-\varepsilon(t-a)\}$$

$$= \frac{1}{hT}\sum_{u,v=1}^{T} \varepsilon(u)[y(u,h)',z(u,h)']\tilde{\Gamma}_h^{-1}\tilde{\Gamma}_h(a)\tilde{\Gamma}_h^{-1}\begin{bmatrix} y(v,h) \\ z(v,h)\end{bmatrix}\varepsilon(v)' + o_p(1)$$

Now lemmas 5.1.1, 5.1.2 show that this is $\delta_{0a}\Gamma_0(s+m) + o_p(1)$ uniformly in h , as required.

To prove the second statement we have, when $a = 0$

$$\frac{1}{h}\sum_1^T \varepsilon(t)\{\hat{\varepsilon}_h(t)-\varepsilon(t)\}'$$

$$= \frac{1}{h}\sum_1^T \varepsilon(t)[y(t,h)',z(t,h)']\{[\hat{\phi}_h,\hat{\psi}_h]-[\mathring{\phi}_h,\mathring{\psi}_h]\}' + O\left(\frac{(\log\log T)^{\frac{1}{2}}}{h}\right)$$

$$= -\frac{1}{hT}\sum_1^T \varepsilon(t)[y(t,h)',z(t,h)']\tilde{\Gamma}_h^{-1}\left\{\sum_1^T \begin{bmatrix} y(t,h) \\ z(t,h)\end{bmatrix}\varepsilon(t)'\right\} + o_p(1)$$

using exactly the same arguments as before. Lemmas 5.1.1, 5.1.2 complete the proof. □

§5.2 Stage II

Throughout these proofs we use the terminology of linear regression when minimising particular quadratic forms. We point out that our usage is only a convenient way to describe the computations rather than to suggest that our procedures are what would usually be called "linear regressions".

(i) Theorem 3.4.1.

In the proof of this theorem let $p_1 = \begin{cases} p_0 & \text{if } q_0 = 0 \\ p_0+1 & \text{if } q_0 > 0 \end{cases}$.

The optimal model from $\gamma(n)$ is obtained by minimising the trace of

$$(5.2.1)\quad \frac{1}{T}\sum_1^T \{ay(t)-dz(t)-(b-I)\hat{\varepsilon}_h(t)\}\{ay(t)-dz(t)-(b-I)\hat{\varepsilon}_h(t)\}'$$

over $\gamma(n)$ where $[a(z),b(z),d(z)]$ is the ARMAX representation of $[k(z),\ell(z)] \in \gamma(n)$. When $n = ps$, the minimisation of (5.2.1) is achieved by regressing $y(t)$ on $y(t-j),z(t-j),\hat{\varepsilon}_h(t-j)$, $j = 1,\ldots,p$. In any other case we perform the regressions given in (3.4.4i,ii). As $\hat{\varepsilon}_h(t)-y(t)$ is a linear combination of $y(t-j),z(t-j)$, $j = 1,\ldots,h$ a lower bound for the residual mean square for the regressions in 3.4.4(ii) is obtained by regressing $y_i(t)$ on $y(t-j),z(t-j)$, $j = 1,\ldots,h+p+1$, $n = ps+q$, $q < s$. This same reasoning can be used to replace the regressor $\hat{\varepsilon}_h(t)$ by $y(t),y(t-j)$, $z(t-j)$, $j = 1,\ldots,h$. Thus for any $\gamma(n)$, $n = ps+q$, $q < s$, a lower bound for (5.2.1) is obtained by regressing $y(t)$ on $y(t-j),z(t-j)$, $j = 1,\ldots,h+p+1$. (The same construction gives a lower bound for (5.2.1) over any γ_α.) Denote this lower bound by $\hat{S}_n$ and the minimum of (5.2.1) by $\hat{\sharp}_{\alpha(n)}$. Then as $h = \{\log T/(-2\log \rho_0)\}(1+o(1))$ a.s., lemma 3.3.3 and the evaluation of (3.3.4) give

$$(5.2.2)\qquad \hat{\sharp}_{\alpha(n)} \geqslant \hat{S}_n = \frac{1}{T}\sum_1^T \varepsilon(t)\varepsilon(t)' + O(Q_T^2 \log T) \text{ a.s.}$$

When $n = n_0$, an upper bound for $\hat{\sharp}_{\alpha(n_0)}$ is given by

$$\frac{1}{T}\sum_1^T \{a_0 y(t)-d_0 z(t)-(b_0-I)\hat{\varepsilon}_h(t)\}\{a_0 y(t)-d_0 z(t)-(b_0-I)\hat{\varepsilon}_h(t)\}'$$

$$= \frac{1}{T}\sum_1^T \{\varepsilon(t)-(b_0-I)(\hat{\varepsilon}_h(t)-\varepsilon(t))\}\{\varepsilon(t)-(b_0-I)(\hat{\varepsilon}_h(t)-\varepsilon(t))\}'$$

$$= \frac{1}{T}\sum_1^T \varepsilon(t)\varepsilon(t)' + O(Q_T^2 \log T) \text{ a.s. ,}$$

$(a_0^{-1}[b_0,d_0] = [k_0,\ell_0])$. Here we have used the expressions in the proof of lemma 3.3.3 together with (3.3.4) to evaluate $\frac{1}{T}\Sigma\varepsilon(t)(\hat{\varepsilon}_h(t-j)-\varepsilon(t-j))'$, $\frac{1}{T}\Sigma(\hat{\varepsilon}_h(t)-\varepsilon(t))(\hat{\varepsilon}_h(t-j)-\varepsilon(t-j))'$ as $O(Q_T^2 \log T)$. This is also an upper bound for $\hat{\Sigma}_{\alpha(n)}$, $n \geqslant n_0$, in particular

$$\hat{\Sigma}_{\alpha(p_1 s)} \geqslant \frac{1}{T}\sum_1^T \varepsilon(t)\varepsilon(t)' + O(Q_T^2 \log T) \text{ , a.s.} \tag{5.2.3}$$

Now if $p > p_1$ then

$$\log\det \hat{\Sigma}_{\alpha(p_1 s)} - \log\det \hat{\Sigma}_{\alpha(ps)} = O(Q_T^2 \log T) \text{ a.s.}$$

Therefore $\log\det \hat{\Sigma}_{\alpha(ps)} + ps(2s+m)C(T)/T$ is increasing as p increases beyond p_1 , and, asymptotically, $\hat{p}$ cannot be greater than p_1 , almost surely. Similarly we show $\log\det \hat{\Sigma}_{\alpha(n)} + n(2s+m)C(T)/T$ increases as n increases beyond n_0 hence, for sufficiently large, T $\hat{p}s-\hat{j}$ cannot be greater than n_0 , almost surely.

When $p < p_1$ we evaluate (5.2.1) over $\gamma(ps)$ by regressing $y(t)$ on $y(t-j), z(t-j), \hat{\varepsilon}_h(t-j)$, $j = 1,\ldots,p$. Let

$$\hat{w}(t) = [y(t-1)',\ldots,y(t-p)',z(t-1)',\ldots,z(t-p)',\hat{\varepsilon}_h(t-1)',\ldots,\hat{\varepsilon}_h(t-p)']'$$

and define $w(t)$ similarly, by replacing $\hat{\varepsilon}_h(s)$ by $\varepsilon(s)$, $s = t-1,\ldots,t-p$ in $\hat{w}(t)$. Then $\frac{1}{T}\sum_1^T \hat{w}(t)\hat{w}(t)'$ converges to a non-singular matrix $E(w(t)w(t)')$ where the expectation is taken with respect to the σ-fields generated by the $\varepsilon(s)$ and $z(s)$. Indeed, if $E(w(t)w(t)')$ is singular then there exists a non-trivial vector $\underset{\sim}{a}$ such that

$$\underset{\sim}{a}'E(w(t),w(t)')\underset{\sim}{a} = E\{(\underset{\sim}{a}'w(t))^2\} = 0$$

hence $\underset{\sim}{a}'w(t) = 0$ but this means that we can add $\underset{\sim}{a}'$ to any of the last $s-q_0$ rows of

$$(5.2.4) \qquad [A_0(1),\ldots,A_0(p),-D_0(1),\ldots,-D_0(p),-B_0(1),\ldots,B_0(p)]$$

and obtain a different ARMAX representation of $[k_0(z),\ell_0(z)]$ which is the *same* echelon form as $[a_0(z),b_0(z),d_0(z)]$, which is evidently impossible. Furthermore $\frac{1}{T}\sum_1^T y(t)\hat{w}(t)'$ converges to $E\{y(t)w(t)'\}$ so that $\hat{\sharp}_{\alpha(ps)}$ is

$$\hat{\gamma}_y(0) - \frac{1}{T}\sum_1^T y(t)\hat{w}(t)'\Big[\frac{1}{T}\sum_1^T \hat{w}(t)\hat{w}(t)'\Big]^{-1}\frac{1}{T}\sum_1^T \hat{w}(t)y(t)'$$

$$= \hat{\gamma}_y(0) - E\{y(t)w(t)'\}[E\{w(t)w(t)'\}]^{-1}E\{w(t)y(t)'\} + o(1) \text{ a.s.}$$

$$= \sharp_{\alpha(ps)} + o(1) \text{ a.s.}$$

From the discussion below (1.4.6) $\sharp_{\alpha(ps)} > \sharp_{\alpha(n_0)} = \frac{1}{T}\sum_1^T \varepsilon(t)\varepsilon(t)' + o(1)$. Hence, for sufficiently large T, $\log\det\hat{\sharp}_{\alpha(ps)} - \log\det\hat{\sharp}_{\alpha(p_1 s)} \geqslant c > 0$ and $\log\det\hat{\sharp}_{\alpha(ps)} + ps(2s+m)C(T)/T$ decreases as p increases to p_1. This completes the proof of (i) of the theorem. The last part of the proof holds so long as $C(T)/T \downarrow 0$.

To prove (ii) it remains to show that $\hat{\sharp}_{\alpha(n)} = \sharp_{\alpha(n)} + o(1)$ when $n = p_1 s - j < n_0$. Then, as $\sharp_{\alpha(n)} > \sharp_{\alpha(n_0)}$ the above argument completes the proof. Let

$$\hat{w}_1(t) = [y_1(t-1),\ldots,y_{s-j}(t-1),z(t-1)',\hat{\varepsilon}_h(t-1)',y(t-2)',\ldots,\hat{\varepsilon}_h(t-p_1)']'$$

$$\hat{w}_2(t) = [y_1(t)-\hat{\varepsilon}_{h,1}(t),\ldots,y_{s-j}(t)-\hat{\varepsilon}_{h,s-j}(t),y(t-1)',\ldots,\hat{\varepsilon}_h(t-p_1-1)']'$$

correspond to the regressors in (3.4.4i), (3.4.4ii) respectively, and $w_1(t),w_2(t)$ be obtained by replacing $\hat{\varepsilon}_h(s)$ by $\varepsilon(s)$ in $\hat{w}_1(t),\hat{w}_2(t)$ as before. It remains to show that $E\{w_1(t),w_1(t)'\}$, $E\{w_2(t)w_2(t)'\}$ are non singular. If

$E\{w_1(t)w_1(t)'\}$ is singular then there is a non zero vector $\underset{\sim}{a}$ such that $\underset{\sim}{a}'w_1(t) = 0$. Then $\underset{\sim}{a}'$ can be added to the first s-j rows of (5.2.4) (after j zeros are included in $\underset{\sim}{a}$ after the $s-j^{th}$ element) to obtain a different ARMAX representation of $[k_0(z),\ell_0(z)]$ in the same echelon form as $[a_0(z),b_0(z),d_0(z)]$. If $E\{w_2(t)w_2(t)'\}$ is singular then a non-trivial row vector $\underset{\sim}{a}'$ may be added to the first s-j rows of

$$[A_0(1)-B_0(1),A_0(2),\ldots,A_0(p_1),-D_0(2),\ldots,-D_0(p_1),-B_0(2),\ldots,-B_0(p_1)]$$

(after $\underset{\sim}{a}$ has j zeros inserted after the $s-j^{th}$ element), thus arriving at the same contradiction as before. Thus we conclude that $\hat{\sharp}_{\alpha(n)} = \sharp_{\alpha(n)} + o(1)$ a.s. as before and the proof of (ii) can be completed. □

(ii) Theorem 3.4.2 and corollary 3.4.3

We present the proof in the ARMA case for notational simplicity. For ARMAX estimation there are no essential changes in the approach. We proceed to estimate $\hat{\sharp}_{p,q}$ as in (3.4.11).

$$\hat{\sharp}_{p,q} = \min \frac{1}{T} \sum_1^T \{ay(t)-(b-1)\hat{\varepsilon}_h(t)\}^2$$

where the minimisation is over the set ARMA(p,q) . Consider $p \geqslant p_0$, $q \geqslant q_0$ and define $v = \min\{p-p_0,q-q_0\}$. Let $a_0(z) = 1 + \sum_1^{p_0} A_0(j)z^{-j}$, $b_0(z) = 1 + \sum_1^{q_0} B_0(j)z^{-j}$ when

$$(5.2.5) \quad ay(t)-(b-1)\hat{\varepsilon}_h(t) = (ab_0-ba_0)b_0^{-1}y(t)$$

$$+ \varepsilon(t)-(b_0-1)\{\hat{\varepsilon}_h(t)-\varepsilon(t)\} - (b-b_0)\{\hat{\varepsilon}_h(t)-\varepsilon(t)\}$$

Let the elements χ_j, $j = 1,\ldots,u = \max\{p+q_0,q+p_0\}$ be defined by $\Sigma \chi_j z^{-j} = a(z)b_0(z)-b(z)a_0(z)$. We have shown

in (1.4.3) that

$$b(z)-b_0(z) = (\nu(z)-1)b_0(z) + \sum_j \chi_j \xi_j^b(z)$$

thus we rewrite (5.2.5) as

$$\sum_j \chi_j\{b_0^{-1}y(t-j) - \xi_j^b(\hat{\varepsilon}_h(t)-\varepsilon(t))\} + \varepsilon(t)-(b_0-1)\{\hat{\varepsilon}_h(t)-\varepsilon(t)\} - (\nu-1)b_0\{\hat{\varepsilon}_h(t)-\varepsilon(t)\}.$$

Denote $b_0^{-1}y(t-j)-\xi_j^b(\hat{\varepsilon}_h(t)-\varepsilon(t))$ by $R_j(t)$. We have evaluated

$$\frac{1}{h}\sum_1^T \varepsilon(t)\{\hat{\varepsilon}_h(t-j)-\varepsilon(t-j)\} \ , \ \frac{1}{h}\sum_1^T \{\hat{\varepsilon}_h(t)-\varepsilon(t)\}\{\hat{\varepsilon}_h(t-j)-\varepsilon(t-j)\}$$

in lemma 5.1.3, thus we obtain

$$\frac{1}{T}\sum_1^T [\varepsilon(t)-(b_0-1)\{\hat{\varepsilon}_h(t)-\varepsilon(t)\}]^2$$

$$= \frac{1}{T}\sum_1^T \varepsilon(t)^2 + \frac{h}{T}\Phi_0 \sum_1^{q_0} B_0(j)^2 + o_p(\log T/T)$$

$$= \frac{1}{T}\sum_1^T \varepsilon(t)^2 + \frac{h}{T}\Phi_0(\omega_0^2-1) + o_p(\log T/T) \ .$$

We evaluate $\hat{\Phi}_{p,q}$ by regressing $\zeta(t) = \varepsilon(t)-(b_0-1)\{\hat{\varepsilon}_h(t)-\varepsilon(t)\}$ on

$$R_j(t), \ j = 1,\ldots,u \ ; \ b_0\{\hat{\varepsilon}_h(t-j)-\varepsilon(t-j)\}, \ j = 1,\ldots,v \ .$$

First regress $\zeta(t)$ on $R_j(t)$, $j = 1,\ldots,u$. By the same argument used to evaluate S_1 in §5.1 (iii) we have

$$\frac{1}{T}\Sigma\{\hat{\varepsilon}_h(t)-\varepsilon(t)\}\{\hat{\varepsilon}_h(t-j)-\varepsilon(t-j)\}$$

(5.2.6)

$$= \{\hat{\phi}_h-\phi\}\hat{\Gamma}_{y,h}(j)\{\hat{\phi}_h-\phi\}'(1+o(1)) + o(T^{-1}) \text{ a.s.}$$

From corollary 3.3.2, $\hat{\phi}_h(j)-\phi(j) = O(Q_T)$, a.s. as $h \geqslant \{\log T/(-2\log \rho_0)\}(1+o(1))$, a.s., hence (5.2.6) is $O(hQ_T^2)$ a.s. Moreover this result is uniform for $j \leqslant H_T$. In a similar way,

(5.2.7) $$\frac{1}{T}\Sigma\varepsilon(t)\{\hat{\varepsilon}_h(t-j)-\varepsilon(t-j)\}$$

$$= \frac{1}{T}\Sigma\varepsilon(t)y(t-j,h)'\{\hat{\phi}_h-\phi\}'(1 + o(1)) + o(T^{-1})$$

which is almost surely $O(Q_T^2 h)$ when $j > 0$ and $O(Q_T)$ when $j \leqslant 0$. From (5.2.6) we have

$$\frac{1}{T}\Sigma\{\xi_j^b(\hat{\varepsilon}_h(t)-\varepsilon(t))\}\{\xi_k^b(\hat{\varepsilon}_h(t)-\varepsilon(t))\} = O(hQ_T^2) \text{ a.s.}$$

As $b_0^{-1}y(t) = a_0^{-1}\varepsilon(t)$, we can truncate the expansion of $a_0^{-1}(z)$ after K_T terms, where $K_T = c \log T$, c sufficiently large, so that we can use (5.2.7) to show

$$\frac{1}{T}\Sigma\{b_0^{-1}y(t-j)\}\{\xi_k^b(\hat{\varepsilon}_h(t)-\varepsilon(t))\} = O(Q_T) \text{ a.s.}$$

This enables us to conclude that

$$\frac{1}{T}\Sigma R_j(t)R_k(t) = \frac{1}{T}\Sigma\{b_0^{-1}y(t-j)\}\{b_0^{-1}y(t-k)\} + O(Q_T) \text{ a.s.}$$

and hence the variance-covariance matrix R_T of $\{R_j(t),$ $j = 1,\ldots,u\}$ converges to a non-singular limit. The last conclusion arises because $b_0^{-1}y(t)$ has spectral density $\frac{1}{2\pi}|a_0(e^{i\omega})|^2\sharp_0$ which is bounded away from zero by (1.1.2). For $j = 1,\ldots,u$

$$\frac{1}{T}\Sigma\zeta(t)R_j(t) = O(Q_T) \text{ a.s.}$$

from the evaluations of (5.2.6), (5.2.7). Thus

$$[\chi_1,\ldots,\chi_u] = -\frac{1}{T}\Sigma\zeta(t)[R_1(t),\ldots,R_u(t)]R_T^{-1}$$

$$= O(Q_T) \text{ , a.s.}$$

and the regression mean square is $O(Q_T^2)$. To complete the evaluation of $\hat{\sharp}_{p,q}$ we regress $\zeta(t)$ on $b_0\{\hat{\varepsilon}_h(t-j)-\varepsilon(t-j)\}$, $j = 1,\ldots,v$ after the latter have been made orthogonal to the regressors $R_j(t)$, $j = 1,\ldots,u$. From the evaluations

of (5.2.6), (5.2.7) we have

$$\frac{1}{T}\Sigma b_0\{\hat{\varepsilon}_h(t-j)-\varepsilon(t-j)\}R_k(t) = O(Q_T) .$$

Thus there exist $\alpha_j(k) = O(Q_T)$, $j = 1,\ldots,v$, $k = 1,\ldots,u$ so that

$$S_j(t) = b_0\{\hat{\varepsilon}_h(t-j)-\varepsilon(t-j)\} + \Sigma_k \alpha_j(k)R_k(t)$$

are the residuals from regressing $b_0\{\hat{\varepsilon}_h(t-j)-\varepsilon(t-j)\}$ on $R_k(t)$, $k = 1,\ldots,u$. Then, as

$$\frac{1}{h}\sum_1^T\{\Sigma_k \alpha_j(t)R_k(t)\}\{\Sigma_k \alpha_j(k)R_k(t)\} = O(\log\log T/h)$$

and

$$\frac{1}{h}\sum_1^T\{b_0(\hat{\varepsilon}_h(t)-\varepsilon(t))\}\{\Sigma_k \alpha_j(k)R_k(t)\} = O(\log\log T/h)$$

we have

$$\frac{1}{h}\sum_1^T S_j(t)S_k(t)$$

$$= \frac{1}{h}\sum_1^T\{b_0(\hat{\varepsilon}_h(t-j)-\varepsilon(t-j))\}\{b_0(\hat{\varepsilon}_h(t-k)-\varepsilon(t-k))\} + O(\log\log T/h) \quad \text{a.s.}$$

Using lemma 5.1.3 this expression becomes

$$\text{(5.2.8)} \qquad \sharp_0\Sigma_\ell B_0(\ell+k)B_0(\ell+j) + o_p(1) .$$

In a similar way, as $k \geqslant 1$,

$$\frac{1}{h}\Sigma\zeta(t)S_k(t)$$

$$\text{(5.2.9)} \qquad = -\frac{1}{h}\sum_1^T\{(b_0-1)(\hat{\varepsilon}(t)-\varepsilon(t))\}\{b_0(\hat{\varepsilon}_h(t-k)-\varepsilon(t-k))\}+o_p(1)$$

$$= -\sharp_0\Sigma_\ell B_0(\ell)B_0(\ell-k) + o_p(1)$$

With $x(t) = b_0e(t)$, as above the statement of theorem 3.4.2, we see that (5.2.8) is $E\{x(t-j)x(t-k)\}\sharp_0 + o_p(1)$ while (5.2.9) is $E\{x(t)x(t-k)\}\sharp_0 + o_p(1)$. Then the coefficients

ν_j, $j = 1,\ldots,v$ obtained by regressing $\zeta(t)$ on $S_j(t)$, $j = 1,\ldots,v$ converge in probability to the coefficients of $\nu^\circ(z)$. Here $\nu^\circ(z) = 1 + \sum_1^v \nu_j^\circ z^{-j}$ has its zeros within $|z| = 1$ and minimises $\frac{1}{2\pi}\int_{-\pi}^{\pi} |\nu^\circ|^2 |b_0|^{-2} d\omega$. The residual mean square is $\sharp_0(\omega_0^2-\omega_v^2)h/T + o_p(h/T)$. Consequently the residual mean square is

$$\hat{\sharp}_{p,q} = \frac{1}{T}\sum_1^T \varepsilon(t)^2 + \frac{h}{T}\sharp_0(\omega_v^2-1) + o_p(\log T/T)$$

as required in theorem 3.4.2. □

To prove corollary 3.4.3 we have shown that, almost surely $\chi_j = O(Q_T)$, therefore from (1.4.3) the first statements in the corollary are established. The second part is established in the comments concluding the theorem. □

(iii) Theorem 3.4.4 and corollary 3.4.5

The strategy of the proof follows theorem 3.4.2 quite closely. Here we prove (a),...,(d) .

(a) Here $\alpha = \{n_1,\ldots,n_s\} \not\ni \alpha_0$ and $\hat{\sharp}_\alpha$ is obtained by regressing each component of $y(t)$ on certain components of $y(t)-\hat{\varepsilon}_h(t), y(t-j), z(t-j), \hat{\varepsilon}_h(t-j)$. Components of $y(t)-\hat{\varepsilon}_h(t)$ occur because $A(0) = B(0)$ and $A(0)$ need not be the identity. The particular sets of regressors are determined by α , for example if $\alpha = \alpha(ps+q)$, the ARMAX representation is given by (1.3.11) and the regressions are described in §3.1. Let us suppose $y_i(t)$ is to be regressed upon a set, $\hat{R}_i(t)$, of components of $y(t)-\hat{\varepsilon}_h(t), y(t-j), z(t-j), \hat{\varepsilon}_h(t-j)$. Let $R_i(t)$ be derived from $\hat{R}_i(t)$ by replacing the $\hat{\varepsilon}_{h,i}(s)$ by $\varepsilon_i(s)$ whenever they occur in $\hat{R}_i(t)$. Let $S(t)$ be made up of all components of $\hat{\varepsilon}_h(t-j)-\varepsilon(t-j)$,

$j = 0,\ldots,p = \max\{n_j\}$. The union of $R_i(t)$ and $S(t)$ spans a space containing the set $\hat{R}_i(t)$ so we obtain a lower bound on $\hat{\Sigma}_\alpha$ by regressing on $R_i(t) \cup S(t)$. The regressors in $R_i(t)$ may not form an independent set because it may happen that $n_j > n_{0j}$ for some j . If this is the case, we remove the redundant variables and arrive at a linearly independent set of regressors, $R_i^\circ(t)$. Hence the matrix $\frac{1}{T}\sum_1^T R_i^\circ(t)R_i^\circ(t)'$ converges to a non singular limit. The regression of $y_i(t)$ on $R_i^\circ(t)$ produces residuals $r_i(t)$ such that

$$\frac{1}{T}\sum_1^T r(t)r(t)' = \Sigma_\alpha + o(1);\ r(t)' = (r_1(t),\ldots,r_s(t))$$

as required. However this residual will be further reduced by regressing $y_i(t)$ on the residuals of $S(t)$ after the latter have been regressed upon $R_i^\circ(t)$. As in (5.2.6), (5.2.7)

$$\frac{1}{T}\sum_1^T\{\hat{\varepsilon}_h(t)-\varepsilon(t)\}R_i^\circ(t)' = O(Q_T)$$

Hence the residuals of $S(t)$ are $s(t) = S(t) + \alpha R_i^\circ(t)$ where the elements of α are $O(Q_T)$. Now obtain the regression mean square due to the regression of $y_i(t)$ on $s(t)$. Again from (5.2.6), (5.2.7)

$$\frac{1}{T}\sum_1^T y_i(t)s(t)' = O(Q_T)$$

while

$$\frac{1}{T}\sum_1^T s(t)s(t)' = \frac{1}{T}\sum_1^T S(t)S(t)' + O(Q_T^2)$$

$$= (s+m)(I_p \otimes \Sigma_0)h/T + o_p(1)$$

Thus the regression mean square is $O_p(\log\log T/\log T)$ hence

$$\hat{\Sigma}_\alpha \geqslant \Sigma_\alpha + o_p(1)$$

as required. □

(b) and (c) From the definitions of $a^{\circ}(z), b^{\circ}(z), d^{\circ}(z)$ above (1.4.5)

$$ay(t)-dz(t)-(b-I_s)\hat{\varepsilon}_h(t)$$

(5.2.10)

$$= \{a[b^{\circ},d^{\circ}]-[b,d]a^{\circ}\}(a^{\circ})^{-1}\begin{bmatrix}\varepsilon(t)\\ z(t)\end{bmatrix}-(b-I_s)\{\hat{\varepsilon}_h(t)-\varepsilon(t)\}+\varepsilon(t)$$

The covariance matrix of $S_j(t) = (a^{\circ})^{-1}\begin{bmatrix}\varepsilon(t-j)\\ z(t-j)\end{bmatrix}$, $j = 1,\ldots,p$ converges to a non singular limit as in theorem 3.4.2. From (5.2.6), (5.2.7)

$$\frac{1}{T}\Sigma\{\hat{\varepsilon}_h(t-k)-\varepsilon(t-k)\}S_j(t)' = O(Q_T) .$$

Thus we can find (by regression) coefficients $A_j(k) = O(Q_T)$ such that the variables

$$R_j(t) = \{\hat{\varepsilon}_h(t-j)-\varepsilon(t-j)\} + \Sigma_k A_j(k)S_k(t)$$

are orthogonal to $S_k(t)$, $k = 1,\ldots,n$, the upper bound for k will be determined later.

When $\alpha = \alpha(n)$ there are only zero-one constraints on the coefficients of $a[b^{\circ},d^{\circ}]-[b,d]a^{\circ}$ (see the discussion around (1.4.4), (1.4.5)) and the residual mean square due to regressing $\varepsilon(t)$ on the appropriate components of the $S_j(t)$ may be computed directly. When V_α is not a generic neighbourhood (i.e. $\alpha \neq \alpha(n)$ for any n) there are other constraints (albeit linear ones) on the coefficients of $a[b^{\circ},d^{\circ}]-[b,d]a^{\circ}$. A lower bound on the residual may be obtained by regressing $\varepsilon(t)$ on $S_j(t)$ as if all the elements of $a[b^{\circ},d^{\circ}]-[b,d]a^{\circ}$ which are not assigned to zero

or one, are unconstrained. As

$$\frac{1}{T}\sum_{1}^{T}\varepsilon(t)S_j(t)' = O(Q_T) \text{ a.s. } j = 1,2,\ldots$$

and the covariance matrix of the $S_j(t)$ converges to a non-singular limit, the residual mean square from regressing $\varepsilon(t)$ on the components of $S_j(t)$ is bounded below by

$$\frac{1}{T}\sum_{1}^{T}\varepsilon(t)\varepsilon(t)' + O(Q_T^2) \text{ a.s.} \tag{5.2.11}$$

We point out that this argument shows that the estimated components of $a[b^\circ,d^\circ]-[b,d]a^\circ$ are $O(Q_T)$ almost surely when $\alpha = \alpha(n)$ as required to prove corollary 3.4.5.

Now obtain a further reduction to (5.3.11) by regressing $\varepsilon(t)$ on the components of $R_j(t)$. The coefficients in $b-I_s$ do not vary freely as they are tied to the elements of $a[b^\circ,d^\circ] - [b,d]a^\circ$ but we obtain an upper bound to the further reduction of (5.2.11) by allowing them to vary freely. First consider (b). Then $B(0) = I_s$ so we need to regress $\varepsilon(t)$ on $R_j(t)$, $j \geqslant 1$. From lemma 5.1.3

$$\frac{1}{T}\sum_{1}^{T}\varepsilon(t)R_j(t)' = o_p(\log T/T)$$

As in (5.2.8)

$$\frac{1}{T}\sum_{1}^{T}R_j(t)R_k(t)' = \Sigma_0\, h/T + o_p(\log T/T)$$

so that the regression mean square of $\varepsilon(t)$ on $R_j(t)$, $j \geqslant 1$ is $o_p(\log T/T)$ completing the proof of (b). In (c) $B(0)-I_s \neq 0$ so that $\varepsilon(t)$ is regressed on certain components of $R_0(t)$. As

$$\frac{1}{T}\sum_{1}^{T}\varepsilon(t)R_0(t)' = -(s+m)\Sigma_0 h/T + o_p(\log T/T)$$

the regression mean square is bounded by $K_1 \log T/T$ where K_1 is constructed from certain components of Σ_0 in a way

determined by α . However, an upper bound on K_1 is $(s+m)\sharp_0/(-2 \log \rho_0)$. This completes the proof of (c). □

(d) Here we consider $\alpha = \alpha_0$. Again we commence from (5.2.10). If $[a(z),b(z),d(z)]$ is in echelon form for V_α then the null space of the mapping

$$[a,b,d] \rightarrow a[b^\circ,d^\circ]-[b,d]a^\circ \tag{5.2.12}$$

is trivial. If $\{\chi_1,\ldots,\chi_n\}$ is a set of independent coefficients from $a[b^\circ,d^\circ]-[b,d]a^\circ$ that parameterise the image of (5.2.12), there exist matrices $J_j(z)$ (of polynomials in z^{-1}) such that $b(z)-b_0(z) = \Sigma\chi_j J_j(z)$. This is the vector analogue of (1.4.3) when $v = 0$. $J(z)$ depends only on the true system parameters. From (1.4.4) the coefficients in $a[b^\circ,d^\circ]-[b,d]a^\circ$ that are not constrained to zero or one are not freely varying unless $\alpha = \alpha(n)$. However the constraints are linear, hence $\{a[b^\circ,d^\circ]-[b,d]a^\circ\}(a^\circ)^{-1}\begin{bmatrix}\varepsilon(t)\\ z(t)\end{bmatrix}$ can be written as $\Sigma\chi_j\zeta_j(t)$ where $\zeta_j(t)$ are matrices whose elements are linear combinations of the components of $(a^\circ)^{-1}\begin{bmatrix}\varepsilon(t-j)\\ z(t-j)\end{bmatrix}$, $j = 1,2,\ldots$. Now (5.2.10) can be written as

$$\Sigma_j \chi_j\{\zeta_j(t) + J_j(\hat{\varepsilon}_h(t)-\varepsilon(t))\} + \varepsilon(t)-(b_0-I_s)\{\hat{\varepsilon}_h(t)-\varepsilon(t)\}.$$

From (5.2.6), (5.2.7),

$$\frac{1}{T}\Sigma\{\zeta_j(t)+J_j(\hat{\varepsilon}_h(t)-\varepsilon(t))\}\{\zeta_k(t)+J_k(\hat{\varepsilon}_h(t)-\varepsilon(t))\}'$$

$$= \frac{1}{T}\Sigma\zeta_j(t)\zeta_k(t)' + O(Q_T)$$

The covariance matrix of $\zeta_j(t)$, $j = 1,\ldots,n$ converges to a non-singular limit for otherwise $\Sigma_j a_j\zeta_j(t) \equiv 0$ for some non zero a_j , which contradicts the independence of the χ_j .

As above

$$\frac{1}{T}\Sigma\{\zeta_j(t)+J_j(\hat{\varepsilon}_h(t)-\varepsilon(t))\}\{\varepsilon(t)-(b_0-I_s)(\hat{\varepsilon}_h(t)-\varepsilon(t))\}'$$

$$= O(Q_T)$$

hence the regression mean square for the regression of $\varepsilon(t)-(b_0-I_s)\{\hat{\varepsilon}_h(t)-\varepsilon(t)\}$ on $\zeta_j(t)+J_j(\hat{\varepsilon}_h(t)-\varepsilon(t))$, $j=1,\ldots,n$ is $O(Q_T^2)$. Hence the residual mean square is

$$\hat{\Phi}_{\alpha_0} = \frac{1}{T}\Sigma\{\varepsilon(t)-(b_0-I_s)(\hat{\varepsilon}_h(t)-\varepsilon(t))\}\{\varepsilon(t)-(b_0-I_s)(\hat{\varepsilon}_h(t)-\varepsilon(t))\}' + O(Q_T^2)$$

$$= \frac{1}{T}\Sigma\varepsilon(t)\varepsilon(t)' + K_0 h/T + o_p(\log T/T)$$

where

$$K_0 = (s+m)\sum_{\ell} B_0(\ell)\Phi_0 B_0(\ell) - (B_0(0)-I_s)\Phi_0(B_0(0)-I_s)' - \Phi_0$$

as in general $B_0(0) \neq I_s$. This concludes (d). □

The argument below (5.2.12) used to construct $\zeta_j(t)$, $j = 1,\ldots,n$ and to show its covariance matrix converges to a non singular limit can be used to evaluate the χ_j, $j = 1,\ldots,n$ in (b). This completes the remaining step in proving corollary 3.4.5. □

(iv) Theorem 3.4.6

Here $\tilde{\varepsilon}(t) = \hat{b}^{-1}\hat{a}y(t)$, where $\hat{a}$, $\hat{b}$ are estimates obtained from stage II. We need to minimise

$$\frac{1}{T}\Sigma\{ay(t)-(b-1)\tilde{\varepsilon}(t)\}^2 \qquad (5.2.13)$$

over ARMA(p,q) when $p \geq p_0$, $q \geq q_0$. We have

$$ay(t)-(b-1)\tilde{\varepsilon}(t) = (a\hat{b}-b\hat{a})\hat{b}^{-1}y(t) + \tilde{\varepsilon}(t).$$

When $\hat{p} > p$, $\hat{q} > q$, all of the parameters in $\hat{a}\hat{b}-\hat{b}\hat{a}$ are not freely varying, however we obtain a lower bound for (5.2.13) by assuming they are free. Thus we regress $\tilde{\varepsilon}(t)$ on $\hat{b}^{-1}y(t-j)$, $j = 1,\ldots,n$, where $n = \max\{\hat{p}+q\ ,\ \hat{q}+p\}$. The covariance matrix of the $\hat{b}^{-1}y(t-j)$, $j = 1,\ldots,n$ converges to a non singular limit because, for sufficiently large T , the zeros of $\hat{b}(z)$ are within $|z| = 1$, in probability, (corollary 3.4.3) hence $\hat{b}^{-1}y(t)$ is a stationary process having a spectral density that is bounded away from zero. The coefficients of $(\hat{a}b_0-\hat{b}a_0)\hat{b}^{-1}$ are decreasing geometrically, indeed they are $O_p(Q_T\rho^{-j})$, $\rho > 1$, as the coefficients of $\hat{a}b_0-\hat{b}a_0$ are $O(Q_T)$. Hence

$$\frac{1}{T}\sum_1^T\{\tilde{\varepsilon}(t)-\varepsilon(t)\}\hat{b}^{-1}y(t-j)$$

$$= \frac{1}{T}\sum_1^T\{(\hat{a}b_0-\hat{b}a_0)b_0^{-1}\hat{b}^{-1}y(t)\}\{\hat{b}^{-1}y(t-j)\}$$

$$= O_p(Q_T)$$

and

$$\frac{1}{T}\sum_1^T\tilde{\varepsilon}(t)\hat{b}^{-1}y(t-j) = \frac{1}{T}\sum_1^T\varepsilon(t)\hat{b}^{-1}y(t-j) + O_p(Q_T)$$

$$= O_p(Q_T)\ .$$

Consequently the residual mean square of the regression of $\tilde{\varepsilon}(t)$ on $\hat{b}^{-1}y(t-j)$ is bounded below by

$$\frac{1}{T}\sum_1^T\tilde{\varepsilon}(t)^2 + O_p(Q_T^2)\ . \tag{5.2.14}$$

Now

$$\frac{1}{T}\sum_1^T(\tilde{\varepsilon}(t)-\varepsilon(t))^2 = \frac{1}{T}\sum_1^T\{(\hat{a}b_0-\hat{b}a_0)b_0^{-1}\hat{b}^{-1}y(t)\}^2$$

$$= O_p(Q_T^2)$$

and

$$\frac{1}{T}\sum_1^T(\tilde{\varepsilon}(t)-\varepsilon(t))\varepsilon(t) = \frac{1}{T}\sum_1^T\{(\hat{a}b_0-\hat{b}a_0)b_0^{-1}\hat{b}^{-1}y(t)\}\varepsilon(t)$$

$$= O_p(Q_T^2) ,$$

the latter equality is because $\hat{a}b_0-\hat{b}a_0$ has no constant term so that only $y(t-j)$, $j \geqslant 1$ occur in that expression. Hence (5.2.14) is

$$\frac{1}{T}\sum_1^T\varepsilon(t)^2 + O_p(Q_T^2) .$$

When either $p \leqslant p_0$ or $q \leqslant q_0$ minimise (5.2.13) by regress-on $y(t)$ on $y(t-j)$, $j = 1,\ldots,p$, $\tilde{\varepsilon}(t-j)$, $j = 1,\ldots,q$. Denote the vector of regressors by $\tilde{R}(t)$ and derive $R(t)$ from $\tilde{R}(t)$ by replacing $\tilde{\varepsilon}(t-j)$ by $\varepsilon(t-j)$ in $\tilde{R}(t)$. Using the above results

$$\frac{1}{T}\sum_1^T\tilde{R}(t)\tilde{R}(t)' = \frac{1}{T}\sum_1^T R(t)R(t)' + O_p(Q_T) .$$

As either $p \leqslant p_0$ or $q \leqslant q_0$, $\frac{1}{T}\sum_1^T R(t)R(t)'$ converges to a non singular limit. As

$$\frac{1}{T}\sum_1^T y(t)\tilde{R}(t)' = \frac{1}{T}\sum_1^T y(t)R(t)' + O_p(Q_T)$$

the residual mean square for the regression procedure is the same as

$$(5.2.15) \qquad \min \frac{1}{2\pi}\int_{-\pi}^{\pi} I(\omega)\left|(ab_0-(b-1)a_0)b_0^{-1}\right|^2 d\omega + O_p(Q_T^2)$$

where

$$I(\omega) = \frac{1}{T}\sum_1^T\left|y(t)e^{it\omega}\right|^2$$

and the minimum is over the set ARMA(p,q). When $p = p_0$, $q = q_0$, put $a = a_0$, $b = b_0$ in (5.2.15) giving an upper bound $\frac{1}{T}\sum_1^T\varepsilon(t)^2 + O_p(Q_T^2)$,

but when either $p < p_0$ or $q < q_0$ (5.2.15) is bounded above the latter quantity. Thus (3.4.14) is minimised at $p = p_0$, $q = q_0$ in probability. □

(v) Theorem 3.4.7

If either $p < p_0$ or $q < q_0$ then

$$BIC_{pq}(p,q)-BIC_{pq}(p_0,q_0) = BIC(p,q)-BIC(p_0,q_0) + O_p(\log T/T)$$

which is positive as $\hat{\Sigma}_{pq}$ is bounded above $\hat{\Sigma}_{p_0q_0}$. Thus $BIC_{pq}(i,j)$ does not attain its minimum at (p,q). Let $\min\{i-p_0, j-q_0\} = v > 0$ then

$$BIC_{p_0q_0}(p_0,q_0)-BIC_{p_0q_0}(i,j)$$

$$= \log \hat{\Sigma}_{p_0q_0} - \log \hat{\Sigma}_{ij} - \hat{h}\{\omega^2_{p_0q_0}(0)-\omega^2_{p_0q_0}(v)\}/T + (p_0-i+q_0-j)\log T/T .$$

From (3.4.11) this quantity is

$$(p_0-i+q_0-j)\log T/T + o_p(\log T/T) < 0 .$$

When $i < p_0$, or $j < p_0$ the above argument shows that

$$BIC_{p_0q_0}(p_0q_0) < BIC_{p_0q_0}(i,j) .$$

Thus the minimum of $BIC_{p_0q_0}(i,j)$ is attained at (p_0,q_0) and the theorem is proved. □

§5.3 Stage III

(i) Theorem 3.5.2

Asymptotically, the order estimates $\{\hat{p},\hat{q}\}$ obtained at stage II exceed the true orders. From theorem 3.4.2, corollary 3.4.3 we have that

$$\hat{a}(z) = \nu(z)a_0(z)+O(Q_T) \ , \ \hat{b}(z) = \nu(z)b_0(z) + O(Q_T)$$

and, for sufficiently large T , $\nu(z)$ is stable. Hence the coefficients of

$$\hat{b}^{-1}-\nu^{-1}b_0^{-1} = (\nu b_0-\hat{b})\hat{b}^{-1}b_0^{-1}\nu^{-1}$$

$$\hat{b}^{-2}\hat{a}-\nu^{-1}b_0^{-2}a_0 = (\nu b_0(\hat{a}b_0-\hat{b}a_0)+(\nu b_0-\hat{b})\hat{b}a_0)\hat{b}^{-2}b_0^{-2}\nu^{-1}$$

are decreasing as $O(Q_T\rho^{-j})$, $\rho > 1$. As in §5.2 (iv) this means that the covariances of $\eta(t),\xi(t),\tilde{\varepsilon}(t)$ are estimated, up to a term that is $O(Q_T)$, by the corresponding covariances of $\eta_0(t) = \nu^{-1}b_0^{-1}y(t)$, $\xi_0(t) = -\nu^{-1}b_0^{-1}\varepsilon(t),\varepsilon(t)$. We estimate

$$\tilde{\phi}_{pq} = \inf\{\tfrac{1}{T}\Sigma\{\tilde{\varepsilon}(t) + a\eta(t) + b\xi(t)\}^2\} \ ,$$

(the infinum is taken over ARMA(p,q)) by regressing $\tilde{\varepsilon}(t)+\eta(t)+\xi(t)$ on $\eta(t-j)$, $j = 1,\ldots,p$; $\xi(t-j)$, $j = 1,\ldots,q$. The calculation is similar to §5.2 (iv) so we omit many of the details.

When either $p < p_0$ or $q < q_0$ or when $p = p_0$ and $q = q_0$ the covariance matrix of the regressors $\eta(t-j)$, $j = 1,\ldots,p$; $\xi(t-j)$, $j = 1,\ldots,q$ converges to a non singular limit, so in this case

$$(5.3.1) \qquad \tilde{\phi}_{pq} = \inf \frac{1}{T}\sum_1^T\{\varepsilon(t) + a\nu^{-1}b_0^{-1}y(t) - b\nu^{-1}b_0^{-1}\varepsilon(t)\}^2 + O(Q_T^2)$$

where the infinum is over ARMA(p,q) . When $p = p_0$, $q = q_0$ put $a = a_0$, $b = b_0$ in the above expression, when

$$\tilde{\phi}_{p_0q_0} \leqslant \frac{1}{T}\sum_1^T\varepsilon(t)^2 + O(Q_T^2) \ .$$

When either $p < p_0$ or $q < q_0$ (5.3.1) converges to

$$\inf \frac{\sharp_0}{2\pi} \int_{-\pi}^{\pi} |1-\nu^{-1}a_0^{-1}b_0^{-1}(ab_0-ba_0)|^2 d\omega$$

$$= \sharp_0 + \inf \frac{\sharp_0}{2\pi} \int_{-\pi}^{\pi} |\nu^{-1}a_0^{-1}b_0^{-1}(ab_0-ba_0)|^2 d\omega$$

$$> \sharp_0 + c \ , \ c > 0 \ .$$

Of course the above infimum is over ARMA(p,q) . Thus (3.5.11) cannot attain a minimum if either $p < p_0$ or $q < q_0$.

Notice that both $\eta(t)$, $\xi(t)$ are linear combinations of $\hat{b}^{-2}y(t-j)$, $j = 1,\ldots,\max\{\hat{p},\hat{q}\}$. Thus the span of

$$\hat{b}^{-2}y(t-j), \ j = 1,\ldots,\max\{\hat{p}+q,\hat{q}+p\} \tag{5.3.2}$$

contains the regressors $\eta(t-j)$, $j = 1,\ldots,p$; $\xi(t-j)$, $j = 1,\ldots,q$. Furthermore the covariance matrix of the variables (5.3.2) converges to a non singular limit. When $p > p_0$ and $q > q_0$ we obtain a lower bound for $\tilde{\sharp}_{pq}$ by regressing $\tilde{\varepsilon}(t)+\eta(t)+\xi(t)$ on (5.3.2). As above, we estimate the bound up to terms that are $O(Q_T^2)$ by regressing $\varepsilon(t)+\eta_0(t)+\xi_0(t)$ on $\nu^{-2}b_0^{-2}y(t-j)$, $j = 1,\ldots,\max\{\hat{p}+q,\hat{q}+p\}$. But $\eta_0(t)+\xi_0(t) = (b_0-a_0)\nu^{-2}b_0^{-2}y(t-j)$ which belongs to the span of the regressors. Thus we are left with regressing $\varepsilon(t)$ on $\nu^{-2}b_0^{-2}y(t-j)$, $j = 1,\ldots,\max\{\hat{p}+q,\hat{q}+p\}$. Here

$$\frac{1}{T}\sum_1^T \varepsilon(t)\{\nu^{-2}b_0^{-2}y(t-j)\} = O(Q_T) \ , \ j > 1$$

hence the residual mean square is

$$\frac{1}{T}\sum_1^T \varepsilon(t)^2 + O(Q_T^2)$$

which is the lower bound for $\tilde{\sharp}_{pq}$. Thus (3.5.11) cannot attain a minimum when $p > p_0$, $q > q_0$ and the theorem is complete. □

§5.4 Transfer function models

(i) Theorem 3.6.1 (stage (iia))

For any pair $\{n_3, n_4\}$

$$\hat{\sharp}_{n_3,n_4} = \inf\{\frac{1}{T}\sum_1^T [\hat{\phi}_h y(t) + (c-1)\hat{\psi}_h z(t) - d\hat{\phi}_h z(t)]^2\}$$

where the infinum is taken over $[c(z), d(z)] \in ARMA(n_3, n_4)$. As $\hat{\phi}_h y(t) = \hat{\varepsilon}_h(t) + \hat{\psi}_h z(t)$ the expression to be minimised is

$$\frac{1}{T}\sum_1^T [\hat{\varepsilon}_h(t) + (c\hat{\psi}_h - d\hat{\phi}_h) z(t)]^2 .$$

Consider $n_3 > n_{03}$, $n_4 > n_{04}$. A lower bound for $\hat{\sharp}_{n_3,n_4}$ is obtained by regressing $\hat{\varepsilon}_h(t)$ on $z(t-j)$, $j = 1, \ldots, p = \max\{n_3, n_4\} + h$. The smallest eigenvalue of $\hat{\Gamma}_{z,p}$ is bounded away from zero as in lemma 4.2.1. When $0 \leqslant j \leqslant H_T$

$$\frac{1}{T}\Sigma \hat{\varepsilon}_h(t) z(t-j) = \frac{1}{T}\Sigma \varepsilon(t) z(t-j) + \frac{1}{T}\Sigma\{\hat{\varepsilon}_h(t) - \varepsilon(t)\} z(t-j)$$

$$= O(Q_T) \text{ , a.s.}$$

For $1 \leqslant j \leqslant h$ note that

$$\frac{1}{T}\Sigma \hat{\varepsilon}_h(t) z(t-j)$$

$$= \hat{\gamma}_{yz}(-j) + \sum_1^h [\hat{\phi}_h(k)\hat{\gamma}_{yz}(k-j) + \hat{\psi}_h(k)\hat{\gamma}_z(k-j)] = 0$$

Note that in calculating expressions such as $\frac{1}{T}\Sigma \hat{\varepsilon}_h(t) z(t-j)$ we are acting as if $z(t)$ is zero outside $0 < s \leqslant T$. Thus only a finite member of terms in $\frac{1}{T}\Sigma \hat{\varepsilon}_h(t) z(t,p)$ are non zero and

$$\hat{\sharp}_{n_3,n_4} \geqslant \frac{1}{T}\Sigma \hat{\varepsilon}_h(t)^2 - \{\frac{1}{T}\Sigma \hat{\varepsilon}_h(t) z(t,p)'\}\hat{\Gamma}_{z,p}^{-1}\{\frac{1}{T}\Sigma \hat{\varepsilon}_h(t) z(t,p)\}$$

(5.4.1)

$$= \frac{1}{T}\Sigma \hat{\varepsilon}_h(t)^2 + O(Q_T^2) \text{ a.s.}$$

Consider the case when either $n_3 < n_{03}$ or $n_4 < n_{04}$ or when $n_3 = n_{03}$, $n_4 = n_{04}$. Let $\hat{R}_{n_3,n_4}(t) = [\hat{\psi}_h z(t-1),\ldots,\hat{\psi}_h z(t-n_3),\hat{\phi}_h z(t-1),\ldots,\hat{\phi}_h z(t-n_4)]'$ and derive $R_{n_3,n_4}(t)$ from $\hat{R}_{n_3,n_4}(t)$ by replacing $\hat{\phi}_h$, $\hat{\psi}_h$ by ϕ_0,ψ_0 in $\hat{R}_{n_3,n_4}(t)$. As $h = \{\log T/(-2\log \rho_0)\}(1+o(1))$,

$$\sum_1^\infty \|\hat{\phi}_h(j)-\phi(j),\hat{\psi}_h(j)-\psi(j)\| = O(hQ_T) \text{ a.s.}$$

from corollary 3.3.2. The covariances of $\hat{\phi}_h z(t-j),\hat{\psi}_h z(t-k)$ are approximated up to $O(Q_T)$ by the corresponding covariances of $\phi_0 z(t-j)$, $\psi_0 z(t-k)$. Here

$$\frac{1}{T}\Sigma\{(\hat{\phi}_h-\phi_0)z(t)\}\{\hat{\phi}_h z(t-k)\}$$

(5.4.2)

$$= [\hat{\phi}_h-\phi_0]\hat{\Gamma}_{z,K_T}(k)\hat{\phi}_h' + o(T^{-1})$$

by truncating the sums $\sum_0^\infty \phi_0(j)z(t-j)$ at $K_T = c\log T$ as in evaluating (5.2.6), (5.2.7). As $|\hat{\Gamma}_{z,K_T}|_1 \leqslant c < \infty$, (5.4.2) is $O(Q_T)$, a.s. Similarly $[\hat{\phi}_h-\phi_0]\hat{\Gamma}_{z,K_T}[\hat{\phi}_h-\phi_0] = O(hQ_T^2)$ a.s. hence

(5.4.3) $$\frac{1}{T}\Sigma\{\hat{\phi}_h z(t)\}\{\hat{\phi}_h z(t-j)\} = \frac{1}{T}\Sigma\{\phi_0 z(t)\}\{\phi_0 z(t-j)\} + O(Q_T)$$ a.s.

The covariance matrix

(5.4.4) $$\hat{C}(T) = \frac{1}{T}\Sigma \hat{R}_{n_3,n_4}(t)\hat{R}_{n_3,n_4}(t)' = \frac{1}{T}\Sigma R_{n_3,n_4}(t)R_{n_3,n_4}(t)' + O(Q_T)$$

which converges to a non singular limit. Indeed, if the limit is singular, then for some non zero vector $\alpha = (c(1),\ldots,c(n_3),-d(1),\ldots,-d(n_4))'$, $\alpha'\hat{C}(T)\alpha \to 0$. Put $c(z) = \sum_1^{n_3} c(j)z^{-j}$, $d(z) = \sum_1^{n_4} d(j)z^{-j}$ then the limit of $\alpha'\hat{C}(T)\alpha$ is

$$\int_{-\pi}^{\pi} |c\psi_0 - d\phi_0|^2 dF_z(\omega) = \int_{-\pi}^{\pi} |\phi_0|^2 |cd_0 - dc_0|^2 |c_0|^{-2} dF_z(\omega)$$

which cannot be zero unless $n_3 > n_{03}$, $n_4 \geqslant n_{04}$. Denote $\frac{1}{T}\Sigma R_{n_3,n_4}(t) R_{n_3,n_4}(t)'$ by $C(T)$. Using an argument similar to that justifying (5.4.3) we can show

$$\frac{1}{T}\Sigma \hat{\varepsilon}_h(t)\hat{\phi}_h z(t-j) = \frac{1}{T}\Sigma \varepsilon(t)\phi_0 z(t-j) + O(Q_T) \text{ a.s.}$$

thus

$$\frac{1}{T}\Sigma\{\hat{\psi}_h y(t)\}\{\hat{\psi}_h z(t-j)\} = \frac{1}{T}\Sigma\{\hat{\phi}_h z(t)+\hat{\varepsilon}_h(t)\}\{\hat{\psi}_h z(t-j)\}$$

$$= \frac{1}{T}\Sigma\{\phi_0 z(t)\}\{\psi_0 z(t-j)\}+O(Q_T) \text{ a.s.}$$

We evaluate $\hat{\sharp}_{n_{03},n_{04}}$ by regressing $\hat{\phi}_h y(t)$ on $\hat{R}_{n_{03},n_{04}}(t)$. The vector of regression coefficients is

$$\begin{aligned}
\hat{\alpha} &= -\hat{C}(T)^{-1}\{\tfrac{1}{T}\Sigma\hat{\phi}_h y(t)\hat{R}_{n_{03},n_{04}}(t)\} \\
&= -\{C(T)+O(Q_T)\}^{-1}\{\tfrac{1}{T}\Sigma\{\psi_0 z(t)\}R_{n_{03},n_{04}}(t)+O(Q_T)\} \\
&= -C(T)^{-1}\{\tfrac{1}{T}\Sigma\{\psi_0 z(t)\}R_{n_{03},n_{04}}(t)\} + O(Q_T) \\
&= -[c_0(1),\ldots,c_0(n_{03}),-d_0(1),\ldots,-d_0(n_{04})]'+O(Q_T) \text{ a.s.}
\end{aligned} \tag{5.4.5}$$

Thus the residual mean square is

$$\begin{aligned}
&\tfrac{1}{T}\Sigma\{\hat{\varepsilon}_h(t)+\hat{\psi}_h z(t)\}^2-\hat{\alpha}'\hat{C}(T)\hat{\alpha} + O(Q_T^2) \text{ a.s.} \\
&= \tfrac{1}{T}\Sigma\{\hat{\varepsilon}_h(t)+\hat{\psi}_h z(t)\}^2-\hat{\alpha}'C(T)\hat{\alpha} + O(Q_T^2) \text{ a.s.}
\end{aligned} \tag{5.4.6}$$

Now $\hat{\alpha}' R_{n_{03},n_{04}}(t) = -(c_0-1)\psi_0 z(t)+d\phi_0 z(t)$

$$= \psi_0 z(t)$$

thus (5.4.6) becomes

$$\frac{1}{T}\Sigma\hat{\varepsilon}_h(t)^2 + \frac{2}{T}\Sigma\hat{\varepsilon}_h(t)\{\hat{\psi}_h z(t)\} + O(Q_T^2) \text{ a.s.}$$

The second term is identically zero as $\frac{1}{T}\Sigma\hat{\varepsilon}(t)z(t-j) = 0$, $j = 1,\ldots,h$ hence

$$\hat{\sharp}_{n_{03},n_{04}} = \frac{1}{T}\Sigma\hat{\varepsilon}_h(t)^2 + O(Q_T^2) \text{ a.s.}$$

This, together with (5.4.1) shows that for c_0 sufficiently large,

$$\log \hat{\sharp}_{n_3,n_4} + (n_3+n_4)c_0 \log\log T/T$$

is increasing as n_3, n_4 increase beyond n_{03}, n_{04}.

When $n_3 < n_{03}$ or $n_4 < n_{04}$ we have shown that the regression mean square for the regression of $\hat{\phi}_h y(t)$ on $\hat{R}_{n_3,n_4}(t)$ converges to

$$\{\frac{1}{T}\Sigma\psi_0(z)R_{n_3,n_4}(t)'\}C(T)^{-1}\{\frac{1}{T}\Sigma\psi_0(z)R_{n_3,n_4}(t)\}$$

$$= \inf \int_{-\pi}^{\pi} |(c-1)\psi_0 - d\phi_0|^2 dF_z(\omega) + o(1)$$

where the infinum is taken over $c(z) = 1 + \sum_1^{n_3} c(j)z^{-j}$, $d(z) = \sum_1^{n_4} d(z)z^{-j}$. Then the residual mean square is

$$\hat{\sharp}_{n_3,n_4} = \frac{1}{T}\Sigma\hat{\varepsilon}(t)^2 + \inf \int_{-\pi}^{\pi} |c\psi_0 - d\phi_0|^2 dF_z(\omega) + o(1)$$

$$> \frac{1}{T}\Sigma\hat{\varepsilon}(t)^2 + c \ , \ c > 0$$

as $n_3 < n_{03}$ or $n_4 < n_{04}$. Thus

$$\log \hat{\sharp}_{n_3,n_4} + (n_3+n_4)C(T)/T > \log \hat{\sharp}_{n_{03},n_{04}} + (n_{03}+n_{04})C(T)/T$$

if $C(T)/T \to 0$ and the proof is complete. □

(ii) Stage (iib)

We consider the minimisation of

$$\frac{1}{T}\Sigma\{a\hat{\eta}(t)-(b-1)\hat{\varepsilon}_h(t)\}^2$$

for $[a(z),b(z)] \in ARMA(n_1,n_2)$. We follow the notation and definitions of theorem 3.4.2 as closely as possible and only discuss the essential changes here.

$$a\hat{\eta}(t)-(b-1)\hat{\varepsilon}_h(t) = \Sigma_j \chi_j R_j(t) + S(t) + (\nu-1)U(t)$$

where

$$R_j(t) = b_0^{-1}\hat{\eta}(t-j) + \xi_j^b[a_0 b_0^{-1}\hat{\eta}(t)-\hat{\varepsilon}_h(t)]$$

$$S(t) = a_0 b_0^{-1}\hat{\eta}(t)-(b_0-1)[a_0 b_0^{-1}\hat{\eta}(t)-\hat{\varepsilon}_h(t)]$$

$$U(t) = b_0(a_0 b_0^{-1}\hat{\eta}(t)-\hat{\varepsilon}_h(t))$$

We have shown (5.4.5) that the coefficients of $\hat{c}-c_0$, $\hat{d}-d_0$ are $O(Q_T)$, a.s. hence for sufficiently large T , $\hat{c}(z)$ is stable. Denote $a_0 b_0^{-1}\hat{c}^{-1}c_0^{-1}(\hat{c}d_0-\hat{d}c_0)$ by $g(z)$. The coefficients of $g(z)$ decrease as $O(Q_T\rho^{-j})$, $\rho > 1$, and

$$a_0 b_0^{-1}\hat{\eta}(t) = gz(t) + \varepsilon(t) .$$

Because of this we can show that

$$\frac{1}{T}\Sigma\{a_0 b_0^{-1}\hat{\eta}(t)-\hat{\varepsilon}_h(t)\}\{a_0 b_0^{-1}\hat{\eta}(t-j)-\hat{\varepsilon}_h(t-j)\}$$

(5.4.7)

$$= \frac{1}{T}\Sigma\{\varepsilon(t)-\hat{\varepsilon}_h(t)\}\{\varepsilon(t-j)-\hat{\varepsilon}_h(t-j)\} + O(Q_T^2) .$$

To establish (5.4.7) note that $\frac{1}{T}\Sigma\{gz(t)\}\{gz(t-j)\} = O(Q_T^2)$ as in §5.2 (iv) and

$$\frac{1}{T}\Sigma\{gz(t)\}\{\varepsilon(t-j)-\hat{\varepsilon}_h(t-j)\} = O(Q_T^2)$$

because $\frac{1}{T}\Sigma z(t-k)\{\varepsilon(t-j)-\hat{\varepsilon}_h(t-j)\} = O(Q_T)$.

We check the steps in proving theorem 3.4.2.

(a) Regress $S(t)$ on $R_j(t)$, $j = 1,\ldots,n$. This requires that the covariance matrix of the $R_j(t)$ converges to a non singular limit. As

$$\frac{1}{T}\Sigma R_j(t)R_k(t) = \frac{1}{T}\Sigma\{b_0^{-1}\hat{\eta}(t-j)\}\{b_0^{-1}\hat{\eta}(t-k)\} + O(Q_T^2)$$

from (5.4.7). Similarly $b_0^{-1}\hat{\eta}(t-j) = a_0^{-1}\varepsilon(t-j) + a_0^{-1}gz(t-j)$ hence

$$\frac{1}{T}\Sigma R_j(t)R_k(t) = \frac{1}{T}\Sigma\{a_0^{-1}\varepsilon(t-j)\}\{a_0^{-1}\varepsilon(t-k)\} + O(Q_T^2)$$

which shows the required covariance matrix has a non singular limit. Similarly,

$$\frac{1}{T}\Sigma S(t)R_j(t) = O(Q_T)$$

hence $\chi_j = O(Q_T)$ a.s. as in theorem 3.4.2.

(b) Obtain the residuals $U(t-j) + \Sigma\alpha_j(k)R_k(t)$ from regressing $U(t-j)$ on the $R_k(t)$. We only require that $\alpha_j(k) = O(Q_T)$ which holds because $\frac{1}{T}\Sigma U(t)R_j(t) = O(Q_T)$.

(c) Regress $S(t)$ on $U(t-j) + \Sigma\alpha_j(k)R_k(t)$. Here

$$\frac{1}{T}\Sigma\{U(t-j)+\Sigma\alpha_j(k)R_k(t)\}\{U(t-\ell)+\Sigma\alpha_\ell(k)R_k(t)\}$$

$$= \frac{1}{T}\Sigma U(t-j)U(t-\ell) + O(Q_T^2)$$

$$= \frac{1}{T}\Sigma\{b_0(\varepsilon(t-j)-\hat{\varepsilon}_h(t-j))\}\{b_0(\varepsilon(t-\ell)-\hat{\varepsilon}_h(t-\ell))\}+O(Q_T^2)$$

from (5.4.7). Similarly

$$\frac{1}{T}\Sigma S(t)\{U(t-j) + \Sigma\alpha_j(k)R_k(t)\}$$

$$= \frac{1}{T}\Sigma\{\varepsilon(t)-(b_0-1)[\varepsilon(t)-\hat{\varepsilon}_h(t)]\}\{b_0[\varepsilon(t-j)-\hat{\varepsilon}_h(t-j)]\} + O(Q_T^2)$$

using the same argument as in (5.4.7). Now exactly the same conclusions of theorem 3.4.2 hold since the covariances need

to be evaluated to an accuracy of $o_p(\log T/T)$.

The conclusions of theorems 3.4.6, 3.4.7 are similarly justified. □

(iii) Theorem 3.6.2, stage (iii)

Partition $\tilde{\tau} = \{\tilde{\tau}_N', \tilde{\tau}_S'\}'$ where $\tilde{\tau}_N$ consists of the parameters of the noise model and $\tilde{\tau}_S$ contains the parameters of the output model. Partition τ_0, $\tilde{\tau}$ similarly. Let

$$A_N = \frac{1}{T} \Sigma \xi_{ab}(t) \xi_{ab}(t)' .$$

In stage (iii) we calculate

$$\tilde{\tau}_N - \hat{\tau}_N = A_N^{-1} \{\frac{1}{T} \Sigma \xi_{ab}(t) \tilde{\varepsilon}(t)\}$$

hence

$$\tilde{\tau}_N - \tau_{0N} = A_N^{-1} \{\frac{1}{T} \Sigma \xi_{ab}(t) \varepsilon(t)\}$$

(5.4.8)

$$+ A_N^{-1} \{\frac{1}{T} \Sigma \xi_{ab}(t) [(\tilde{\varepsilon}(t) - \varepsilon(t)) + \xi_{ab}(t)' (\hat{\tau}_N - \tau_{0N})]\} .$$

We show $A_N = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} + o(1)$ a.s. A typical element of A_N is

$$\frac{1}{T} \Sigma \{\hat{b}^{-1} \hat{\eta}(t)\} \{\hat{b}^{-1} \tilde{\varepsilon}(t)\}$$

(5.4.9)
$$= \frac{1}{T} \Sigma \{a_0^{-1} \varepsilon(t)\} \{b_0^{-1} \varepsilon(t)\} + O(Q_T)$$

$$= A_{12}(j,j) + o(1)$$

Here $\hat{b}^{-1} \hat{\eta}(t) = a_0^{-1} g z(t) + a_0^{-1}(t)$ where $g(z)$ is defined in §5.4 (ii) and hence the coefficients of $a_0^{-1} g$ decrease at a rate $O(Q_T \rho^{-j})$, $\rho > 1$. Furthermore

$$\hat{b}_0^{-1}\tilde{\varepsilon}(t) = b_0^{-1}\varepsilon(t)+b_0^{-1}\hat{b}^{-1}(a_0\hat{b}-b_0\hat{a})a_0^{-1}\varepsilon(t)+b_0^{-1}gz(t) \ .$$

Again the coefficients of $b_0^{-1}\hat{b}^{-1}(a_0\hat{b}-b_0\hat{a})a_0^{-1}$, $b_0^{-1}g$ are decreasing as $O(Q_T\rho^{-j})$, $\rho > 1$. Now the argument used to establish (5.4.7) serves to establish (5.4.9). As in Dunsmuir and Hannan (1976), Hannan, Dunsmuir and Deistler (1980) we can establish that

$$A_N^{-1}\{\frac{1}{\sqrt{T}}\Sigma\xi_{ab}(t)\varepsilon(t)\} \xrightarrow{\mathcal{L}} N(0,\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22}\end{bmatrix}) \ .$$

The theorem is proved if we show

$$(5.4.10) \quad \frac{1}{T}\Sigma\xi_{ab}[(\tilde{\varepsilon}(t)-\varepsilon(t))+\xi_{ab}(t)'(\hat{\tau}_N-\tau_{0N})] = o(T^{-\frac{1}{2}})$$

As $\tau'_{0N}\xi_{ab}(t) = (a_0-1)\xi_a(t)+(b_0-1)\xi_b(t)$,

$$\tilde{\varepsilon}(t)-\varepsilon(t)+\xi_{ab}(t)'(\hat{\tau}_N-\tau_{0N})$$

$$= \tilde{\varepsilon}(t)-\varepsilon(t)+(\hat{a}-a_0)\xi_a(t)+(\hat{b}-b_0)\xi_b(t)$$

$$= \{\hat{b}^{-2}b_0^{-1}(\hat{a}b_0-\hat{b}a_0)(\hat{b}-b_0)\}\hat{\eta}(t)+\{\hat{c}^{-1}c_0^{-1}(c_0\hat{d}-d_0\hat{c})\}z(t)$$

The coefficients of $\{\hat{b}^{-2}b_0^{-1}(\hat{a}b_0-\hat{b}a_0)(\hat{b}-b_0)\}$ are decreasing geometrically as $O(Q_T^2\rho^{-j})$, $\rho > 1$ thus we obtain

$$\frac{1}{T}\Sigma\xi_{ab}\{[\hat{b}^{-2}b_0^{-1}(\hat{a}b_0-\hat{b}a_0)(\hat{b}-b_0)]\hat{\eta}(t)\} = O(Q_T^2)$$

As

$$\frac{1}{T}\Sigma\hat{\eta}(t)z(t-j) = \frac{1}{T}\Sigma(\hat{\eta}(t)-\eta(t))z(t-j) + O(Q_T)$$

using lemma 4.1.4, and $\hat{\eta}(t)-\eta(t) = (c_0^{-1}d_0-\hat{c}^{-1}\hat{d})z(t)$, where the coefficients of $c_0^{-1}\hat{c}^{-1}(\hat{c}d_0-\hat{d}c_0)$ are decreasing as $O(Q_T\rho^{-j})$, $\rho > 1$, $\frac{1}{T}\Sigma\hat{\eta}(t)z(t-j) = O(Q_T)$. Consequently

$$\frac{1}{T}\Sigma\xi_a(t)\{\hat{c}^{-1}c_0^{-1}(c_0\hat{d}-d_0\hat{c})\}z(t)$$

$$= -\frac{1}{T}\Sigma\hat{b}^{-1}\hat{\eta}(t)\{\hat{c}^{-1}c_0^{-1}(c_0\hat{d}-d_0\hat{c})\}z(t) \quad = O(Q_T^2) \ .$$

In a similar way, as $\xi_b(t) = \hat{b}^{-2}\hat{a}\hat{\eta}(t)$,

$$\frac{1}{T}\Sigma\xi_b(t)\{\hat{c}^{-1}c_0^{-1}(c_0\hat{d}-d_0\hat{c})\}z(t) = O(Q_T^2) \ .$$

Hence (5.4.10) is $o(T^{-\frac{1}{2}})$ as required. The same argument carries over to prove the asymptotic efficiency for the parameter estimates of the output model. □

REFERENCES

Akaike, H. (1969), Fitting autoregressive models for prediction, *Ann.Inst.Statist.Math.* 21, 243

Akaike, H. (1972), Information theory and an extension of the maximum likelihood principle, in *Proc. 2nd International Symposium on Information Theory*, supplement to Problems of Control and Information Theory, 267-281

Akaike, H.(1974a), Stochastic theory of minimal realisation, *IEEE-AC 19, 667*

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*

Akaike, H. (1974c), Canonical correlation analysis of time series and the use of an information criterion, in *System Identification: Advances and Case Studies.* (R.K. Mehra and D.G. Lainiotis, eds), Academic Press, NY, 27-96

Akaike, H. (1974d), A new look at the statistical model identification, *IEEE-AC 19, 716*

Akaike, H. (1976), On entropy maximisation principle, in *Applications of Statistics*, P.R. Krishnaiah ed, North Holland, New York

An, Hong-zhi, Chen, Zhao-guo and Hannan, E.J. (1982), Autocorrelation, autoregression and autoregressive approximation, *Ann.Statist. 10, 926*

Astrom, K.J. (1980), Maximum likelihood and prediction error methods, *Automatica, 16, 551*

Astrom, K.J. and Mayne, D.Q. (1982), A new algorithm for recursive estimation of controlled ARMA process, *Proc. 6th IFAC symposium on identification and systems parameter estimation*, Arlington, Virginia, USA, 122-126

Astrom, K.J. and Soderstrom, T. (1974), Uniqueness of the maximum likelihood estimates of the parameters of an ARMA model, *IEEE-AC 19, 769*

Bard, Y. (1974), *Non linear Parameter Estimation*, Academic Press, New York

Baxter, G. (1963), A norm inequality for a "finite section" Wiener-Hopf equation, *Ill.J.Math. 7, 97*

Box, G.E.P. and Jenkins, G.M. (1970), *Time Series Analysis, Forecasting and Control*, Holden-Day, San Francisco

Burg, J.P. (1975), *Maximum Entropy Spectral Analysis*, PhD dissertation, Dept. Geophysics, Stanford Univ. California

Clark, J.M.C. (1976), The consistent selection of parameterisation in systems identification, *Paper presented at 1976 JACC*, Purdue University, Indiana

Cooper, D.M. and Wood, E.F. (1982), Identifying multivariate time series models, *J. Time Series Anal. 3, 153*

Cybenko, G. (1980), The numerical stability of the Levinson-Durbin algorithm for Toeplitz systems of equations, *SIAM J.Sci.Stat.Comput. 1, 303*

Deistler, M. (1981), The structure of ARMA systems in relation to estimation in *Geometry and Identification*, Proc. APSM workshop on system geometry, system identification and parameter estimation, Eds. Caines, P.E. and Hermann, R., Math Sci Press, Mass., 49-62

Deistler, M., Dunsmuir, W.M. and Hannan, E.J. (1978), Vector linear time series models: corrections and extensions, *Adv.Appl.Prob. 10, 360*

Deistler, M. and Hannan, E.J. (1981), Some properties of the parameterisation of ARMA systems with unknown order, *J. Multivariate Anal. 11, 474*

Dunsmuir, W.M. and Hannan, E.J. (1976), Vector linear time series models, *Adv.Appl.Prob.* 8, 339

Durbin, J. (1961), Fitting of time series models, *Rev.Inst. Statist.* 28, 233

Engel, R. and Watson, M. (1981), One factor multivariate time series models, *JASA* 76, 774

Forney, G.D. (1975), Minimal bases of rational vector spaces with applications to multivariable linear systems, *SIAM J. Control* 13, 493

Friedlander, B. (1982), Lattice methods for spectral estimation, *Proc. IEEE* 70, 990

Fuller, W.A. (1976), *Introduction to Statistical Time Series*, Wiley, New York

Gevers, M. and Wertz, V. (1981), Overlapping parameterisation for the representation of multivariate stationary time series, in *Geometry and Identification*, Proc. APSM workshop on system geometry, system identification and parameter estimation, eds Caines, P.E. and Hermann, R., Math Sci Press, Mass., 73-100

Hall, P. and Heyde, C.C. (1980), *Martingale limit theory and its application*, Academic Press, New York

Hannan, E.J. (1969), The identification of vector mixed autoregressive moving average systems, *Biometrika* 56, 223

Hannan, E.J. (1970), *Multiple Time Series*, Wiley, New York.

Hannan, E.J. (1971), The identification problem for multiple equation systems with moving average errors, *Econometrika*, 39, 751

Hannan, E.J. (1973), The asymptotic theory of linear time series models, *J.Appl.Prob.* 10, 130

Hannan, E.J. (1978), Rates of convergence for time series

regression, *Adv.Appl.Prob.* 10, 740

Hannan, E.J. (1980), The estimation of the order of an ARMA process, *Ann.Statist.* 8, 1071

Hannan, E.J. (1981), Estimating the dimension of a linear system, *J. Multivariate Anal.* 11, 458

Hannan, E.J. (1982), Testing for autocorrelation and Akaike's criterion, in *Essays in Statistical Science* (The Moran Festschrift), eds. Gani, J. and Hannan, E.J., 403-412

Hannan, E.J., Dunsmuir, W.M. and Deistler, M. (1980), Estimation of vector ARMAX models, *J. Multivariate Anal.* 10, 275

Hannan, E.J. and Heyde, C.C. (1972), On limit theorems for quadratic functions of discrete time series, *Ann.Math. Statist.* 43, 2058

Hannan, E.J. and Kavalieris, L. (1983a), The convergence of autocorrelations and autoregressions, *Austral.J.Statist.* 25, 287

Hannan, E.J. and Kavalieris, L. (1983b), Linear estimation of ARMA processes, *Automatica* 19, 447

Hannan, E.J. and Kavalieris, L. (1984a), A method for autoregressive-moving average estimation, *Biometrika* 71,

Hannan, E.J. and Kavalieris, L. (1984b), Multivariate linear time series models, *Adv.Appl.Prob.*

Hannan, E.J. and Kavalieris, L. (to appear), Regression, autoregression models, *J. Time Series Anal.*

Hannan, E.J. and Quinn, B.G. (1979), The determination of the order of an autoregression, *JRSS (B)*41, 190

Hannan, E.J. and Rissanen, J. (1982), Recursive estimation of ARMA order, *Biometrika* 69, 81

Harvey, A.C. and Phillips, G.P.A. (1979), Maximum likelihood

estimates of regression models with ARMA disturbances, *Biometrika 66, 49*

Hazewinkel, M. and Kalman, R.E. (1975), On invariants, canonical forms and moduli for linear, constant, finite dimensional, dynamical systems, in *Mathematical System Theory, Undine*, Lecture notes in economics and mathematical systems, 131, Springer, Berlin, 48-60

Jones, J.W. (1914), *Fur Farming in Canada*, Commission of conservation, Ottawa

Kabaila, P. (1983), Parameter values of ARMA models minimising the one step ahead prediction error when the true system is not in the model set, to appear

Kailath, T. (1980), *Linear Systems*, Prentice-Hall, New Jersey

Kalman, R.E. (1963), Mathematical description of linear dynamical systems. *SIAM J. Control 1, 152*

Kalman, R.E. (1971), Kronecker invariants and feedback, in *Ordinary Differential Equations*, Ed. L. Weiss, Academic Press, New York

Kalman, R.E. (1974), Algebraic geometric description of a class of linear systems of constant dimension, *8th Annual Princeton Conference on Information Science and Systems*, Princeton

Kashyap, R.L. (1977) A Bayesian comparison of different classes of dynamic models using empirical data. IEEE-AC 27, 715.

Ljung, J. and Rissanen, J. (1976), On canonical forms, parameter identifiability and the concept of complexity, *4th IFAC symposium on identification and system parameter estimation*, Tbilisi, USSR, 58-69

Ljung, J. and Soderstrom, T. (1983), *Theory and Practice of Recursive Identification*, MIT Press, Cambridge, Mass.

Luenberger, D.G. (1967), Canonical forms for linear multivariable systems *IEEE-AC 12, 290*

MacDuffee, C.C. (1956), *The Theory of Matrices*, Chelsea, New York

Makhoul, J. (1981), Lattice methods in spectral estimation, in *Applied Time Series Analysis II* ed. D.F. Findley, 301-325

Mayne, D.Q. (1972), A canonical model for identification of multivariable linear systems, *IEEE-AC 17, 728*

Mayne, D.Q. and Firoozan, F. (1982), Linear identification of ARMA processes, *Automatica 18, 461*

Monden, Y., Yamada, M. and Arimoto, S. (1982), Fast algorithms for identification of an ARX model and its order determination, *IEEE-ASSP 30, 390*

Moricz, F. (1976), Moment inequalities and the strong laws of large numbers, *Z. Wahrscheinlichkeitstheorie verw. Gebiete 35, 299*

Pierce, D.A. (1971), Least squares estimation in the regression model with autoregressive-moving average errors, *Biometrika 58, 299*

Popov, V.N. (1969), Some properties of control systems with irreducible matrix transfer functions, in *Seminar on Differential Equations and Dynamical Systems II*, Lecture notes in mathematics 144, Springer, Berlin, 169-180

Popov. V.N. (1972), Invariant description of linear time-invariant controllable systems, *SIAM J. Control 10, 252*

Rissanen, J. (1974), Basis of invariants and canonical forms for linear dynamical systems, *Automatica 10, 175*

Rissanen, J. (1976), Minimax entropy estimation of models for vector processes, in *System Identification: Advances and Case Studies* (R.K. Mehra and D.G. Lainiotis, eds) Academic Press, New York, 97-100

Rissanen, J. and Caines, P.E. (1979), Strong consistency for maximum likelihood estimators for ARMA processes, *Ann. Statist.* 7, 297

Rissanen, J. and Ljung, L. (1975), Estimates of optimal structures and parameters for linear systems, in *Mathematical Systems Theory, Undine,* Lecture notes in Economics and Mathematical Systems, 131, Springer, Berlin, 75-91

Rosenbrock, H.H. (1970), *State Space and Multivariable Theory,* Wiley, New York

Rozanov, Yu.A. (1967), *Stationary Random Processes,* Holden-Day, San Francisco

Schwarz, G. (1978), Estimating the dimension of a model, *Ann. Statist.* 6, 461

Schweppe, F.C. (1965), Evaluation of likelihood functions for Gaussian signals, *IEEE-IT* 11, 61.

Seber, G.A.F. (1977), *Linear Regression Analysis,* Wiley, New York

Shibata, R. (1976), Selection of the order of an autoregressive model by Akaike's information criterion, *Biometrika 63, 117*

Shibata, R. (1980), Asymptotically efficient selection of the order of the model for estimating parameters of a linear process, *Ann.Statist.* 8, 147

Shibata, R. (1981), An optimal autoregression spectral estimate, *Ann.Statist.* 9, 300

Solo, V. (1983), The exact likelihood for a multivariate ARMA model, *J. Multivariate Anal.* (to appear)

Son, L.H. and Anderson, B.D.H. (1971), Design of Kalman filters using signal model statistics, *Proc. IEE* 120, 312

Steinberg, H.W., Gasser, T. and Franke, J. (1983), Fitting autoregressive processes to EEG time series: an empirical comparison of estimates of the order. *Universität Heidelberg, Preprint 236*

Stout, W.F. (1970), The Hartman-Wintner law of the iterated logarithm for martingales, *Ann.Math.Statist. 41, 2158*

Tjostheim, D. and Paulson, J. (1983), Bias of some commonly used time series estimates, *Biometrika 70, 389*

van Emden, M. (1971), *Analysis of Complexity*, Math.Centr. Tracts, 35, Amsterdam

van Overbeek, A.J.M. and Ljung, L. (1981), On line structure selection for multivariable state space models, in *Geometry and Identification*, Proc. APSM workshop on system geometry, system identification and parameter estimation, eds Caines, P.E. and Hermann, R., Math Sci Press, Mass., 153-184

van Overbeek, A.J.M. and Ljung. L. (1982), On line structure selection for multivariable state space models, *Automatica 18, 529*

Walker, A.M. (1964), Asymptotic properties of least squares estimates of the parameters of the spectrum of a stationary non-deterministic time series, *J.Austral.Math.Soc. 4, 363.*

Wertz, V., Gevers, M. and Hannan, E.J. (1982), The determination of optimal structures for the state space representation of multivariate stochastic processes, *IEEE-AC 27, 1200*

Whittle, P. (1951), *Hypothesis Testing in Time Series Analysis*, Almqvist and Wiksell, Uppsala

Whittle, P. (1963), On the fitting of multivariate autoregressions and the approximate cononical factorisation of a

spectral density matrix, *Biometrika*, *50*, *129*

Wilson, D.A. and Kumar, A. (1982), Derivative computations for the log-likelihood function, *IEEE-AC*, *27*, *230*

Young, P.C. (1976), Some observations on instrumental variable methods of time series analysis, *Int.J. Control*, *23*, *593*

INDEX TO DEFINITIONS

$\alpha(n)$	27	observable	12
ARMAX model	5	observational equivalence	10
ARMA(p,q), ARMAX(p,q,r)	16	order	15
		output	4
controllable	11	output model	7
echelon form	27	prediction error form	8
echelon submanifold	27	prime matrix	14
endogenous variables	4	proper, strictly proper	17
exogenous variables	4		
generic neighbourhood	27	realisation	10
Hankel matrix	18	stable	6
		state space form	7
identifiable	15	state vector	7
input	4	structural identifiability	15
		structure indices	21
Kronecker indices	29	system similarity	11
linear innovations	6	transfer function	6
		transfer function model	6
matrix fraction description,		U_α	20
mfd	9	$\mathbf{U}_\alpha$	37
McMillan degree	15	unimodular	14
$M(n)$	17		
$\mathbf{M}(n)$	37	V_α	26
		$\mathbf{V}_\alpha$	37
Noise model	7	$V(n)$	27