The Annals of Statistics
1992, Vol. 20, No. 1, 195–215

IDENTIFICATION OF ECHELON CANONICAL FORMS FOR VECTOR LINEAR PROCESSES USING LEAST SQUARES

BY D. S. POSKITT

Australian National University

In this paper a method of identifying stationary and invertible vector autoregressive moving-average time series is proposed. The models are presumed to be represented in (reversed) echelon canonical form. Consideration is given to both parameter estimation and the determination of structural indices, the evaluations being based on the use of closed form least squares calculations. Consistency of the technique is shown and the operational characteristics of the procedure when employed as a means of approximating more general processes is discussed.

1. Modelling vector time series. When modelling vector time series it is natural to think in terms of multivariate generalisations of the class of univariate autoregressive moving-average (ARMA) models [Box and Jenkins (1976)] and vector ARMA models have been widely studied in the literature. For example, estimation of such models is discussed in Tunnicliffe-Wilson (1973), Dunsmuir and Hannan (1976), Nicholls (1977) and Kohn (1978); diagnostic checking procedures have been examined in Hosking (1980) and Poskitt and Tremayne (1982); and the problem of model construction is considered in Akaike (1976), Chan and Wallis (1978), Tiao and Box (1981) and Cooper and Wood (1982), amongst others. The structure theory relating to such processes is also known to be far more involved than a simple, naive extrapolation of univariate ideas would indicate [see Deistler (1985) for a detailed review and extensive references], and it is this structure that introduces an inherent complexity not found in the univariate situation. A basic purpose of the present paper is to indicate how, despite this complexity, the model parameters can be estimated (identified) and the internal structure of the process determined (identified) using a simple finite algorithm.

Heretofore, much of the discussion of the identification of multivariate models has addressed problems and techniques associated with the use of canonical correlations due to Akaike (1976); see Box and Tiao (1977) and Tsay and Tiao (1985), for example. Such methods often involve the solution of a large number of eigenvalue problems. Here we shall adopt the general approach taken by Hannan and Kavalieris (1984). The latter authors consider estimating the coefficients of an ARMA model using regression-type methods and determining the integer structural parameters using model selection

Received May 1988; revised June 1991.
AMS 1980 *subject classifications*. Primary 62M10; secondary 62F12, 62J05, 93B30, 93E12.
Key words and phrases. Autoregressive moving-average, echelon canonical form, Kronecker indices, identification, least squares regression, consistency, linear process, approximation.

criteria, such as AIC [Akaike (1974)] and BIC [Schwarz (1978) and Rissanen (1978)], that are evaluated in terms of the one-step-ahead prediction error variance and a penalty adjustment for the number of coefficients fitted. The technique adopted by Hannan and Kavalieris (1984) to identify the structural parameters is to first select the overall order, or observability index, and then optimise the model selection criterion over a range of models delineated by the value so chosen. An alternative method of structure determination in which the integer parameters are assessed sequentially from smallest to largest is considered here; see Poskitt (1987b) and Pötscher (1990) for discussion of a similar approach to univariate time series. Consistency of the method is established and the asymptotic behaviour of the procedure when it is used to identify an ARMA model which is fitted as an approximation to a data generating process whose transfer function is not rational is also examined. The notion that a model set does not contain the truth but serves as a vehicle for describing the salient features of the process under study, some approximation to reality, is becoming more prevalent in time series analysis. In this context we hope to indicate the first few steps along a path outlined in the survey by Hannan (1987), extending to multivariate ARMA modelling some of the ideas introduced by Shibata (1980) and further examined in Hannan and Kavalieris (1986).

The plan of the paper is as follows. The next section sets out notation, defines terms and presents the basic assumptions used in subsequent developments. Section 3 analyses the asymptotic properties of the least squares estimates when the model structure is given. This is followed in the subsequent section by an outline of the structure determination procedure. This section also incorporates a proof of consistency. In Section 5 the operational behaviour of the procedure when used as an approximation method is presented. The final section contains some closing remarks pertaining to the practical implementation of the techniques.

2. Preliminaries. Let $x(t) = (x_1(t), \ldots, x_v(t))'$, $t = 1, \ldots, T$, denote a realisation of a stationary and ergodic vector stochastic process containing v components. To avoid excessive notation $x(t)$ is employed to denote both a given process and a realised value of that process. Assume that any deterministic components have been removed so that, without loss of generality, $x(t)$ is a zero mean regular process. Wold's decomposition [Rozanov (1967)] implies that

$$x(t) = \sum_{j=0}^{\infty} k(j)\varepsilon(t-j), \qquad k(0) = I_v, \qquad t \in \mathbb{Z}, \tag{2.1}$$

where the transfer function

$$K(z) = \sum_{j=0}^{\infty} k(j)z^{-j}$$

is analytic and $\det K(z) \neq 0$, $|z| > 1$, $\varepsilon(t)$ being the innovation process. Interpreting z^{-1} as the unit lag operator, that is, $z^{-1}x(t) = x(t-1)$, (2.1) can be

reexpressed more succinctly as $x(t) = K(z)\varepsilon(t)$. The innovation satisfies $E[\varepsilon(t)] = 0$ and $E[\varepsilon(t)\varepsilon(s)'] = \delta_{t,s}\Omega$, where $\delta_{t,s}$ denotes Kronecker's delta and $\Omega > 0$. The second order properties of $x(t)$ are uniquely characterised by its impulse response, the sequence of $v \times v$ matrix coefficients $\{I_v, k(1), \ldots, k(j), \ldots\}$ and the innovation variance-covariance matrix Ω. When modelling $x(t)$ it will be necessary to capture the fundamental features of these. Using $\|\cdot\|$ for the Frobenius norm of a matrix argument, suppose that

$$\sum_{j=0}^{\infty} \|k(j)\|^2 < \infty$$

and strengthen the condition on the zeroes of $K(z)$ to $\det K(z) \neq 0$, $|z| \geq 1$; see Hannan and Poskitt (1988) for some justification of the latter assumption. We shall refer to the conditions given above as assumption (A1).

The class of multivariate ARMA models are defined by a specification of the form

$$\sum_{j=0}^{p} a(j)x(t-j) = \sum_{j=0}^{p} m(j)\eta(t-j), \qquad t \in \mathbb{Z}, \tag{2.2}$$

a particular ARMA structure being obtained by fixing a value for the integer p, the observability index and allotting numerical values to the elements of the coefficient matrices $a(j)$ and $m(j)$, $j = 0, 1, \ldots, p$. In this model the observed process is expressed as a linear transformation of an unobservable disturbance $\eta(t)$ with a proper, rational transfer function $\Phi(z) = \sum_{j\geq 0}\phi(j)z^{-j} = N(z)/d(z)$, where the numerator matrix $N(z)$ and denominator $d(z)$ are determined from the operators

$$A(z) = \sum_{j=0}^{p} a(j)z^{-j} \quad \text{and} \quad M(z) = \sum_{j=0}^{p} m(j)z^{-j} \tag{2.3}$$

via $N(z) = \text{adj}\{A(z)\}M(z)$ and $d(z) = \det A(z)$. In relation to (2.2), the following assumptions are made.

(A2). The disturbance process $\eta(t)$ is a stationary, ergodic martingale difference sequence. Thus if $\mathscr{F}_t$ denotes the σ-field generated by $\eta(s)$, $s \leq t$, then $E[\eta(t)|\mathscr{F}_{t-1}] = 0$. Moreover, $E[\eta(t)\eta(t)'|\mathscr{F}_{t-1}] = \Sigma > 0$ and $E[\eta_j(t)^4] < \infty$, $j = 1, \ldots, v$.

(A3). The matrices $A(z)$ and $M(z)$ satisfy $\det A(z) \neq 0$ and $\det M(z) \neq 0$, $|z| \geq 1$. The pair $[A(z) : M(z)]$ are (left) coprime and in (reversed) echelon canonical form.

Writing $A_{rc}(z)$ for the r, cth element of $A(z)$, $r, c = 1, \ldots, v$, and similarly $M(z) = [M_{rc}(z)]$, the (reversed) echelon form has the following properties that

define $[A(z): M(z)]$:

$$A_{rr}(z) = 1 + a_{rr}(1)z^{-1} + \cdots + a_{rr}(n_r)z^{-n_r}, \tag{2.4a}$$

$$A_{rc}(z) = a_{rc}(n_r - n_{rc}^{*} + 1)z^{n_{rc}-n_r-1} + \cdots + a_{rc}(n_r)z^{-n_r}, \tag{2.4b}$$

$$M_{rc}(z) = m_{rc}(0) + m_{rc}(1)z^{-1} + \cdots + m_{rc}(n_r)z^{-n_r}, \tag{2.4c}$$

with $m_{rc}(0) = a_{rc}(0)$, wherein

$$n_{rc} = \begin{cases} \min(n_r + 1, n_c), & r \geq c, r, c = 1, \ldots, v, \\ \min(n_r, n_c), & r < c, r, c = 1, \ldots, v. \end{cases} \tag{2.4d}$$

The integers n_r, $r = 1, \ldots, v$, are the Kronecker indices associated with $\Phi(z)$ and completely determine the structure of $[A(z): M(z)]$ from (2.4). Writing $A(z)$ and $M(z)$ as in (2.3), $p = \max_r(n_r)$, note that the row degrees of $[A(z): M(z)]$, $\delta_r[A(z): M(z)]$, equal n_r, $r = 1, \ldots, v$. The degree of $d(z)$, $\delta(d(z))$, is $\sum n_r = n$ and is an invariant of $\Phi(z)$ called the order or McMillan degree. Setting $\nu = \{n_1, \ldots, n_v\}$, the number of independent parameters in $[A(z): M(z)]$ is given by

$$\begin{aligned} d(\nu) &= n(v+1) + \sum \sum_{r<c=1}^{v} \{\min(n_r, n_c) + \min(n_c + 1, n_r)\} \\ &\leq 2nv. \end{aligned} \tag{2.5}$$

For more detailed particulars on the concepts discussed in this paragraph, we refer again to Deistler (1985); see also Hannan and Deistler [(1988), Chapter 2].

Now suppose that P is a $v \times v$ permutation matrix such that $P[A(z): M(z)]$ permutes the rows of $[A(z): M(z)]$. By appropriate choice of P it is clear that an observationally equivalent ARMA representation of $\Phi(z)$ can be found with row degrees that are ordered from smallest to largest, $n_{r(1)} \leq n_{r(2)} \leq \cdots \leq n_{r(v)}$, with $r(j) = c(j)$, $j = 1, \ldots, v$, being the permutation of $1, \ldots, v$ induced by P. With a slight abuse of notation, we shall write the permuted structural index ν as $P\nu = \{n_{r(1)}, \ldots, n_{r(v)}\}$. Observe that premultiplying $[A(z): M(z)]$ by P amounts to a simple reordering of the equations in (2.4) and does not change the magnitude of the Kronecker indices. By way of contrast, however, rearranging the elements of $x(t)$, which corresponds to a premultiplication of $\Phi(z)$ by some P, will in general alter the n_r, $r = 1, \ldots, v$, but still leave the $n_{r(j)}$, $j = 1, \ldots, v$, unchanged. It is this fact, that the $n_{r(j)}$, $j = 1, \ldots, v$, constitute a set of invariants for $\Phi(z)$, that forms the basis of the structure determination procedure presented below. In particular, when the indices are ordered according to their magnitude additional structure can be determined for $[A(z): M(z)]$ which can be exploited to advantage. Thus, consider the $r(u)$th row of $A(z)$,

$1 \le u \le v$. From (2.4d),

$$n_{r(u)c(j)} = \begin{cases} n_{c(j)}, & j = 1, \ldots, u-1, \\ n_{r(u)} + 1, & r(u) \ge c(j), n_{r(u)} < n_{c(j)}, j = u, \ldots, v, \\ n_{r(u)}, & r(u) \ge c(j), n_{r(u)} = n_{c(j)}, j = u, \ldots, v, \\ n_{r(u)}, & r(u) < c(j), j = u, \ldots, v. \end{cases}$$

Therefore

$$\text{(2.6a)} \qquad A_{r(u)c(j)}(z) = \sum_{s=s(u,j)}^{n_{r(u)}} a_{r(u)c(j)}(s) z^{-s}, \qquad j = 1, \ldots, v,$$

where

$$\text{(2.6b)} \quad s(u,j) = \begin{cases} n_{r(u)} - n_{c(j)} + 1, & j = 1, \ldots, u-1, \\ 1, & r(u) < c(j) \text{ or} \\ & r(u) \ge c(j) \text{ and } n_{r(u)} = n_{c(j)}, \\ 0, & r(u) \ge c(j) \text{ and } n_{r(u)} < n_{c(j)}, \\ & j = u, \ldots, v. \end{cases}$$

In the $r(u)$th row, the maximum lag of all variables is $n_{r(u)}$, with all lags included if the Kronecker index associated with the $c(j)$th variable, $n_{c(j)} \ge n_{r(u)}$, but the smallest lags deleted if $n_{c(j)} < n_{r(u)}$ so as to give only $n_{c(j)}$ nonzero coefficients. Also, if $r(u) > c(j)$ and $n_{r(u)} = n_{c(j)}$, the contemporaneous influence of $x_{c(j)}(t)$ on $x_{r(u)}(t)$ does not appear. Interchanging the roles of row and column we find from the $c(u)$th column that $x_{c(u)}(t)$ appears in all equations with a maximum lag determined by $n_{r(j)}$, the Kronecker index of that equation or row, but with the leading terms of the operator truncated to give only $n_{c(u)}$ nonzero coefficients if $n_{c(u)} < n_{r(j)}$:

$$\text{(2.7a)} \qquad A_{r(j)c(u)}(z) = \sum_{s=s(j,u)}^{n_{r(j)}} a_{r(j)c(u)}(s) z^{-s},$$

where

$$\text{(2.7b)} \quad s(j,u) = \begin{cases} n_{r(j)} - n_{c(u)} + 1, & j = u, \ldots, v, \\ 1, & r(j) < c(u) \text{ or} \\ & r(j) \ge c(u) \text{ and } n_{r(j)} = n_{c(u)}, \\ 0, & r(j) \ge c(u) \text{ and } n_{r(j)} < n_{c(u)}, \\ & j = 1, \ldots, u-1. \end{cases}$$

Counting the number of freely varying parameters in the $r(u)$th equation, we

obtain

$$d_{r(v)}(\nu) = \sum_{j=1}^{u-1} n_{r(j)} + (2v - u + 1)n_{r(u)},$$

assuming for simplicity that the $n_{r(j)}$, $j = 1, \ldots, v$, are distinct.

By defining constraints on the rows of $[A(z) : M(z)]$, the (reversed) echelon form provides a reduction in the number of system parameters required to represent $\Phi(z)$. More significantly, however, equations (2.6) and (2.7) imply that the specification of the $r(u)$th equation, $1 \leq u \leq v$, depends on $n_{r(j)}$, $j = 1, \ldots, u$, but not on any larger index and $\Sigma d_{r(j)}(\nu)$ when compared to (2.5) indicates that the number of freely varying parameters is minimised for the ordered Kronecker indices or Kronecker invariants. This suggests how the problem of simultaneous inference implicit in model identification might be handled and a parsimonious model constructed by building up the Kronecker indices in order of magnitude from smallest to largest. Before developing this idea in detail in Section 4, we will require the following construction in the next section.

Employing the notation and conventions of Neudecker (1969), let $\bar{\alpha} = \text{vec}(a(1) : \cdots : a(p))$ and $\bar{\mu} = \text{vec}(m(1) : \cdots : m(p))$, where $a(j)$ and $m(j)$, $j = 1, \ldots, p$, are the coefficient matrices of (2.3) with $[A(z) : M(z)]$ in (reversed) echelon form. If $\zeta_p(z) = (z^{-1}, \ldots, z^{-p})'$, then

$$\text{vec } A(z) = \left(\zeta_p(z)' \otimes I_{v^2}\right)\bar{\alpha} + \text{vec } a(0),$$

$$\text{vec } M(z) = \left(\zeta_p(z)' \otimes I_{v^2}\right)\bar{\mu} + \text{vec } m(0).$$

Now let $S_{a(\nu)}$ be a selection matrix with rows equal to v^2p element unit vectors of the form $(0, \ldots, 0, 1, 0, \ldots, 0)$, the ones appearing in those columns corresponding to the nonzero elements of $\bar{\alpha}$ found as one moves down the vector. Put $\alpha = S_{a(\nu)}\bar{\alpha}$. From the relation $S_{a(\nu)}S'_{a(\nu)} = I$, it follows that $\bar{\alpha} = S'_{a(\nu)}\alpha$. Similarly, set $\lambda = S_{f(\nu)}\bar{\lambda}$, $\bar{\lambda} = \text{vec}(a(0) - I_v)$, where $S_{f(\nu)}$ selects the nonzero elements of $a(0)$ below the leading diagonal appearing in $\bar{\lambda}$. Then

$$\text{(2.8a)} \qquad \text{vec } A(z) = \left(\zeta_p(z)' \otimes I_{v^2}\right)S'_{a(\nu)}\alpha + S'_{f(\nu)}\lambda + \text{vec } I_v.$$

Using what is now an obvious notation we also have

$$\text{(2.8b)} \qquad \text{vec } M(z) = \left(\zeta_p(z)' \otimes I_{v^2}\right)S'_{m(\nu)}\mu + S'_{f(\nu)}\lambda + \text{vec } I_v,$$

where $\mu = S_{m(\nu)}\bar{\mu}$, $S_{m(\nu)}$ being defined in a manner similar to $S_{a(\nu)}$ and $S_{f(\nu)}$. These representations prove useful in simplifying the exposition of the estimation procedure discussed in the next section. Here we note that α, λ and μ contain the freely varying parameters of $[A(z) : M(z)]$ in $\mathbb{R}^{d(\nu)}$ not restricted to be either 0 or 1. Thus if σ is a vector containing the $v(v + 1)/2$ distinct elements of Σ, then $\text{ARMA}_E(\nu) = \{(\alpha' : \lambda' : \mu')' \times \sigma \in \mathbb{R}^{d(\nu)} \times \mathbb{R}^{(1/2)v(v+1)} : \Phi(z) = A(z)^{-1}M(z), \Sigma > 0\}$ is the set of all ARMA models in (reversed) echelon form with (structural index = {Kronecker indices}) $\nu = \{n_1, \ldots, n_v\}$. The process $x(t)$ admits an ARMA representation if there exists a $\nu_0 = \{n_{10}, \ldots, n_{v0}\}$, $n_{j0} < \infty$, $j = 1, \ldots, v$, with associated parameter values α_0, λ_0, μ_0 and σ_0 such

that $\Phi_0(z) = K(z)$ a.e., $|z| = 1$ and $\Sigma_0 = \Omega$, implying that $\eta(t) = \varepsilon(t)$ a.s. This establishes the link between the process generating the data (2.1) and the model (2.2). In this case, the $\text{ARMA}_E(\nu_0)$ model will be said to obtain or to hold and the true autoregressive and moving-average operators will be denoted A_0 and M_0, respectively. As here, the indeterminant z will often be omitted from polynomials and power series, where this causes no confusion and the appendage of the subscript 0 will be used throughout to denote true values.

Finally, for any two processes $\xi(t)$ and $\nu(t)$, we shall write

$$\Gamma_{\xi\nu}(z) = \sum_{j=-\infty}^{\infty} \gamma_{\xi\nu}(j) z^{-j}$$

for the cross-covariance generating function between them. In particular,

$$\Gamma_{xx}(z) = \sum_{j=-\infty}^{\infty} \gamma_{xx}(j) z^{-j} = K(z)\Omega K(z^{-1})'$$

and

$$\Gamma_{x\varepsilon}(z) = \sum_{j=0}^{\infty} \gamma_{x\varepsilon}(j) z^{-j} = K(z)\Omega.$$

3. Estimation of $\text{ARMA}_E(\nu)$ models. Suppose that the structural index $\nu = \{n_1, \ldots, n_v\}$ is a fixed a priori. In order to estimate the parameters in $(\alpha' : \lambda' : \mu')'$ and, implicitly, σ, the following two-stage technique, which is a variant of the method proposed in Hannan and Kavalieris (1984), can be employed.

STAGE 1. Regress $x(t)$ on $x(t-j)$, $j = 1, \ldots, h$, to obtain residuals

$$\hat{e}_h(t) = x(t) - \sum_{j=1}^{h} \hat{\psi}(j) x(t-j)$$

and set

$$\text{AIC}(h) = \log\det \frac{1}{T} \sum_{t=1}^{T} \hat{e}_h(t)\hat{e}_h(t)' + \frac{2hv^2}{T}.$$

Define h_T by the requirement that $\text{AIC}(h) \geq \text{AIC}(h_T)$, $0 \leq h, h_T < H_T$, $H_T = (\log T)^a$, $1 < a < \infty$. Put $\hat{\varepsilon}_T(t) = \hat{e}_{h_T}(t)$, $t = 1, \ldots, T$.

The purpose of this stage is to provide estimates of the unobserved innovation process. In the light of the results of Shibata (1980) and Hannan and Kavalieris (1986) an autoregressive approximation is used, the order of the approximation being chosen by reference to AIC. Now set

$$\eta_T(t) = A(z)x(t) - (M(z) - I_v)\hat{\varepsilon}_T(t). \tag{3.1}$$

Using the elementary result that a column vector is its own vectorisation, the rule $\text{vec } ABC = (C' \otimes A)\text{vec } B$ and (2.8) the right-hand side of (3.1) can be

reexpressed as

$$x(t) + X_\nu(t)\alpha + \hat{F}_\nu(t)\lambda - \hat{E}_\nu(t)\mu,$$

where

$$\begin{aligned} X_\nu(t)\alpha &\equiv \operatorname{vec}[A(z) - a(0)]x(t) \\ &= (x(t)' \otimes I_v)\operatorname{vec}\left[\sum_{j=1}^{p} a(j)z^{-1}\right] \\ &= \left(\zeta_p(z)' \otimes x(t)' \otimes I_v\right)S'_{a(\nu)}\alpha \end{aligned}$$

defines $X_\nu(t)$ and similar identities define

$$\hat{F}_\nu(t) = \left([x(t) - \hat{\varepsilon}_T(t)]' \otimes I_v\right)S'_{f(\nu)}$$

and

$$\hat{E}_\nu(t) = \left(\zeta_p(z)' \otimes \hat{\varepsilon}_T(t)' \otimes I_v\right)S'_{m(\nu)}.$$

This leads to:

STAGE 2. Let the parameter vector $\theta = (\alpha' : \lambda' : \mu')'$ and set the regressor variables $\hat{R}_\nu(t)' = [-X_\nu(t) : -\hat{F}_\nu(t) : \hat{E}_\nu(t)]$, $t = 1, \ldots, T$. Minimise the residual sum of squares

$$\begin{aligned} \sum_{t=1}^{T} \|\eta_T(t)\|^2 &= \sum_{t=1}^{T} \|A(z)x(t) - (M(z) - I_v)\hat{\varepsilon}_T(t)\|^2 \\ &= \sum_{t=1}^{T} \left\|x(t) - \hat{R}_\nu(t)'\theta\right\|^2 \end{aligned}$$

with respect to the $d(\nu)$ freely varying elements of θ ($[A; M]$) to obtain the least squares estimate $\hat{\theta}_T$ ($[\hat{A}_T : \hat{M}_T]$).

In describing the procedure, the nomenclature and ideas of least squares regression are employed since these will frequently be called upon in subsequent theoretical developments. For technical reasons also, presample values will be assumed to be zero. This has an asymptotically negligible effect but could, as is well known, be of some significance in applications. We shall return to this issue which impinges on the question of practical implementation and algorithm construction below.

THEOREM 3.1. *Suppose that $x(t)$ admits an ARMA representation satisfying conditions* (A1)–(A3) *of Section 2. Then* $[\hat{A}_T : \hat{M}_T] = [A_\infty : M_\infty] + O(Q_T)$ *a.s.*, $Q_T = (\log\log T/T)^{1/2}$, *where* $[A_\infty : M_\infty]$, *which may depend on* T, *minimises* $\operatorname{tr} \Psi(\nu)$

$$\Psi(\nu) = \frac{1}{2\pi}\int_{-\pi}^{\pi} (AK - M)\Omega(AK - M)^* \, d\omega. \tag{3.2}$$

For ease of presentation, the argument $z = e^{i\omega}$ is omitted in the integrand of expression (3.2) and an asterisk is used to denote the joint operations of transposition and complex conjugation, a notation with which we will persist in expressions of this kind. It is possible to give more detail about the form and behavior of $[A_\infty : M_\infty]$ but these are not required here and are, therefore, omitted until needed. Note immediately that if $\nu = \nu_0$, meaning $n_r = n_{r0}$, $r = 1, \ldots, v$, then $[A_\infty : M_\infty] = [A_0 : M_0]$, giving rise to the following.

COROLLARY 3.1. *If $ARMA_E(\nu_0)$ obtains and $\nu = \nu_0$, then $\hat{\theta}_T$ ($[\hat{A}_T : \hat{M}_T]$) provides a strongly consistent estimator of θ_0 ($[A_0 : M_0]$).*

The proof of Theorem (3.1) depends on the following results.

LEMMA 3.2. *Under the conditions of Theorem* (3.1):

(i) *uniformly in $h \leq (\log T)^a$, $1 < a < \infty$,*

$$\frac{1}{T}\sum_{t=1}^{T}\{\hat{e}_h(t) - \varepsilon(t)\}\{\hat{e}_h(t-j) - \varepsilon(t-j)\} = O(Q_T^2 h), \qquad \forall \textit{ fixed } j,$$

$$\frac{1}{T}\sum_{t=1}^{T}\varepsilon(t)\{\hat{e}_h(t) - \varepsilon(t)\} = O(Q_T^2 h);$$

(ii) $h_T = c_0 \log T\{1 + o(1)\}$ *a.s.*, $c_0 > 0$.

PROOF. See Hannan and Kavalieris (1984, 1986). □

Now set $S_\nu = \operatorname{diag}(S_{a(\nu)}; S_{f(\nu)}; S_{m(\nu)})$, a $d(\nu) \times v^2(2p+1)$ selection matrix and let

$$G = \frac{1}{2\pi}\int_{-\pi}^{\pi} RR^*\, d\omega = \begin{bmatrix} G_{\alpha\alpha} & G_{\alpha\lambda} & G_{\alpha\mu} \\ & G_{\lambda\lambda} & G_{\lambda\mu} \\ & & G_{\mu\mu} \end{bmatrix},$$

the upper triangle only being indicated because $G' = G$,

$$R(z) = S_\nu \begin{bmatrix} -\zeta_p(z)' \otimes K(z)\Omega^{1/2} \otimes I_v \\ -1 \otimes (K(z) - I_v)\Omega^{1/2} \otimes I_v \\ \zeta_p(z) \otimes \Omega^{1/2} \otimes I_v \end{bmatrix},$$

with $\Omega^{1/2}$ the Cholesky lower triangular factor of Ω. Also put

$$g = \begin{bmatrix} g_\alpha \\ g_\lambda \\ g_\mu \end{bmatrix} = \frac{1}{2\pi}\int_{-\pi}^{\pi} S_\nu \begin{bmatrix} \zeta_p \otimes \operatorname{vec}(K\Omega K^*)' \\ 1 \otimes \operatorname{vec}(K\Omega(K - I_v)^*)' \\ \zeta_p \otimes \operatorname{vec}(\Omega K^*)' \end{bmatrix} d\omega.$$

LEMMA 3.3. *Under the conditions of Theorem* 3.2, *the matrix*

$$\hat{G}_T = \frac{1}{T}\sum_t \hat{R}_\nu(t)\hat{R}_\nu(t)' = G + O(Q_T) \quad a.s.$$

and the vector

$$\hat{g}_T = \frac{1}{T}\sum_t \hat{R}_\nu(t)x(t) = g + O(Q_T) \quad a.s.$$

PROOF. Consider a typical element of the top right-hand subblock of $\hat{G}_T$ corresponding to $G_{\alpha\mu}$. This is composed of an appropriate selection from terms of the form

$$\hat{\gamma}_{x\hat{\varepsilon}_T}(k-j) = T^{-1}\sum_{k,j}{}' x(t-j)\hat{\varepsilon}_T(t-k)', \qquad j,k = 1,\ldots,p,$$

$\Sigma'_{k,j}$ indicating summation over $t = \max(j,k)+1,\ldots,T$. Expanding $\hat{\gamma}_{x\hat{\varepsilon}_T}(k-j)$ as $\hat{\gamma}_{x\varepsilon}(k-j) + \hat{\gamma}_{x(\hat{\varepsilon}_T-\varepsilon)}(k-j)$, the first term is $\gamma_{x\varepsilon}(k-j) + O(Q_T)$ a.s. [An, Chen and Hannan (1982) and Hannan and Kavalieris (1983)] and from the Cauchy–Schwarz inequality the second has components majorised by factors involving the elements of $\hat{\gamma}_{(\varepsilon_T-\varepsilon)(\hat{\varepsilon}_T-\varepsilon)}(0)$, which by Lemma 3.2 are $O(Q_T^2 \log T)$. Applying the same argument to the remaining subblocks of $\hat{G}_T$ establishes the first statement of the lemma and the second is shown analogously. □

PROOF OF THEOREM 3.1. By definition $\hat{\theta}_T$ is obtained as a solution of the normal equations $\hat{G}_T\theta = \hat{g}_T$. We will suppose, without loss of generality, that $\hat{\theta}_T$ is taken as the minimum norm least squares solution, Rao and Mitra [(1971), Chapter 3] and satisfies $\|\hat{\theta}_T\| < \infty$. From Lemma 3.3 it is straightforward to show that $G\hat{\theta}_T = g + O(Q_T)$ a.s. Thus, if G is nonsingular, $\theta_\infty = G^{-1}g$ is unique and $\hat{\theta}_T = \theta_\infty + O(Q_T)$ a.s. If G is singular, then $\hat{\theta}_T$ differs from a member of the set $\Theta_\infty = \{\theta\colon G\theta = g\}$ by a term that is $O(Q_T)$ a.s. and converges to Θ_∞ in the sense that any subsequence of $\hat{\theta}_T$ has a sub-subsequence converging to a point in Θ_∞. Thus we may write $\hat{\theta}_T = \theta_{\infty T} + O(Q_T)$ a.s., where $\theta_{\infty T} \in \Theta_\infty$. It remains to show that a solution of the system of equations $G\theta = g$ minimises $\operatorname{tr}\Psi(\nu)$. To this end, note that $\operatorname{tr}\Psi(\nu)$ is a norm for $\{A(z)K(z) - M(z)\}\Omega^{1/2}$ which we denote by $\|\cdot\|^2_{\mathscr{L}^2}$ [Rosenberg (1963)] and after a little algebra

$$\begin{aligned}\operatorname{vec}\{A(z)K(z) - (M(z) - I)\}\Omega^{1/2} &= \operatorname{vec} E(z) \quad \text{(say)}\\ &= \left((K(z)\Omega^{1/2})' \otimes I\right)\operatorname{vec} I - R(z)'\theta.\end{aligned}$$

Some simple analysis shows that

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} EE^*\,d\omega = \Omega + \Psi(\nu).$$

Using the equality $\operatorname{tr}(BB^*) = (\operatorname{vec} B)'\operatorname{vec} \overline{B}$, where $\overline{B}$ is the complex conjugate of the matrix B, Parsevel's relation and natural Hilbert space generalisations of the least squares results presented in Rao and Mitra [(1971), Chapter 3.7], we deduce that the equation system $G\theta = g$ generates the least squares solution to the minimisation of $\|\hat{E}(z)\|^2_{\mathscr{L}^2}$ and hence minimises $\operatorname{tr} \Psi(\nu)$ as required. □

Although the estimator $[\hat{A}_T : \hat{M}_T]$ is consistent, it can be shown to be asymptotically inefficient relative to the Gaussian maximum likelihood estimator; see McDougall (1988) for the case $v = 1$. An efficient estimate can be obtained either by use of a full maximum likelihood procedure or by implementing Gauss–Newton-type recursions. Examination of the least squares calculations is motivated partly by the need to provide consistent preliminary estimates that can be used to initiate such iterative calculations. Moreover, it is in general extremely unlikely that ν_0 will be known and it seems prudent to contemplate identifying ν_0 before complicated, open-ended algorithms are employed to obtain efficient estimates. We shall therefore consider identifying ν_0 using the closed-form least squares calculations described in Stages 1 and 2 above.

4. Determination of Kronecker indices. Since the ordered Kronecker indices are invariants of Φ, the objective of the procedure proposed here is to identify $n_{r(1)} \leq n_{r(2)} \leq \cdots \leq n_{r(v)}$ in sequence, exploiting the additional structure that such an ordering imposes on the specification. Let $\hat{\sigma}_r^2(\nu)$, $r = 1, \ldots, v$, denote the estimate of the one-step ahead prediction error variance for the rth variable. For $\nu = \{0\} = \{0, \ldots, 0\}$, $\hat{\sigma}_r^2(\{0\}) = T^{-1}\sum x_r(t)^2$. We consider using a model selection criterion for the rth equation of the form

$$\Lambda_r(\nu) = \log \hat{\sigma}_r^2(\nu) + d_r(\nu)C(T)/T,$$

where $C(T)$ is a real-valued, nonnegative, possibly stochastic function of T. It is convenient to think of the procedure described immediately below as taking place sequentially. It is designed so that the smallest index will be identified in the first pass through Steps 1 and 2 and in subsequent passages the remaining indices will be determined in increasing order of magnitude.

Set $\hat{\nu}(1, n) = \{n, \ldots n\} = \nu(n)$ for all $n \geq 0$. Commencing at $u = 1$, $n = 1$, repeat Steps 1 and 2 until all $\hat{n}_{r(j)}$, $j = 1, \ldots, v$, have been evaluated.

STEP 1. For $r(j)$, $j = u, \ldots, v$, regress $x_{r(j)}(t)$ on $x_{r(m)}(t - n + \hat{n}_{r(m)} - 1)$, $\ldots, x_{r(m)}(t - n)$, $\hat{\varepsilon}_{r(m)T}(t - 1), \ldots, \hat{\varepsilon}_{r(m)T}(t - n)$, $m = 1, \ldots, u - 1$, $u \geq 2$ and $x_{r(m)}(t - l)$, $\hat{\varepsilon}_{r(m)T}(t - l)$, $l = 1, \ldots, n$, plus $x_{r(m)}(t) - \hat{\varepsilon}_{r(m)T}(t)$ if $r(m) < r(u)$, $m = u, \ldots, v$, and evaluate $\Lambda_{r(j)}\{\hat{\nu}(u, n)\}$, $j = u, \ldots, v$.

STEP 2. If

$$\Lambda_{r(j)}\{\hat{\nu}(u, n)\} < \Lambda_{r(j)}\{\hat{\nu}(u, n - 1)\} \quad \text{for all } j \in \{u, \ldots, v\},$$

increment n by 1 and return to Step 1. If $\Lambda_{r(j_i)}\{\hat{\nu}(u, n)\} \geq \Lambda_{r(j_i)}\{\hat{\nu}(u, n - 1)\}$,

$j_i \in \{u, \ldots, v\}$, $i = 1, \ldots, s$, set $\hat{n}_{r(u)} = \cdots = \hat{n}_{r(u+s-1)} = n - 1$, $r(u) = r(j_1), \ldots, r(u + s - 1) = r(j_s)$. Increment u by s and let $\hat{\nu}(u, n) = \hat{P}_u \nu = \{\hat{n}_{r(1)}, \ldots, \hat{n}_{r(u-1)}, n, \ldots, n\}$, where $\hat{P}_u$ determines the permutation $\{r(1), \ldots, r(u - 1), r(u), \ldots, r(v)\}$ of $\{1, \ldots, v\}$. Return to Step 1.

A feature of the identification process is that the estimates $\hat{n}_{r(j)}$, $j = 1, \ldots, v$, are characterised by the fact that $\Lambda_{r(j)}\{\hat{\nu}(r(j), n)\} > \Lambda_{r(j)}\{\hat{\nu}(r(j), n + 1)\}$, $n < \hat{n}_{r(j)}$, $\Lambda_{r(j)}\{\hat{\nu}(r(j), \hat{n}_{r(j)})\} \leq \Lambda_{r(j)}\{\hat{\nu}(r(j), \hat{n}_{r(j)} + 1)\}$, $j = 1, \ldots, v$. Each index is the right-hand endpoint of the interval over which its associated criterion function is strictly decreasing and is defined without reference to an upper bound. The algorithm has also been presented in terms of the estimation of individual equations by breaking down the general notation used to describe Stage 2 into its component parts. In particular, a careful examination of the rows of $\hat{R}_\nu(t)'$ reveals that the nonzero elements in the rth row correspond to regressor variables appropriate to the rth equation. Similarly, each column of $\hat{R}_\nu(t)'$ contains only one nonzero entry, a number in the rth row corresponding to the value of a regressor for the rth equation. Thus, the equation system $\hat{G}_T\hat{\theta}_T = \hat{g}_T$ can be rearranged into v subsystems, which we write using an obvious notation as $\hat{G}_{rT}\hat{\theta}_{rT} = \hat{g}_{rT}$, $r = 1, \ldots, v$, one for each row of $[A : M]$. The freely varying parameters of any particular equation in the system are determined independently of those of any other. It is this property, in conjuction with the structure implied by (2.6) and (2.7), that ultimately leads to the following result justifying the procedure.

THEOREM 4.1. *Suppose that* $ARMA_E(\nu_0)$ *obtains and assumptions* (A1)–(A3) *are satisfied. Then if* $C(T)/\log T \log\log T \to \infty$ *such that* $C(T)/T \to 0$ *a.s. as* $T \to \infty$, *then* $\hat{n}_{r(j)} = n_{r(j)0}$, $j = 1, \ldots, v$, *with probability* 1.

PROOF. First we adapt an argument of Hannan and Kavalieris [(1984), page 550] to obtain a lower bound to $\hat{\sigma}_r^2(\nu)$, $r = 1, \ldots, v$. For any $\hat{\nu}(u, n)$, $u = 1, \ldots, v$, the variables being regressed upon at either Step 1 or Step 2 are from $x(t - l)$ and $\hat{\varepsilon}_T(t - l)$, $1 \leq l \leq n$, together with certain other elements of $x(t) - \hat{\varepsilon}_T(t)$. From Stage 1, however, $x(t) - \hat{\varepsilon}_r(t)$ and $\hat{\varepsilon}_T(t - l)$, $l = 1, \ldots, n$, are linear combinations of $x(t - l)$, $1 \leq l \leq (n + h_T)$. Therefore, if $\hat{\xi}(t)$ is the residual from the unrestricted autoregression of $x(t)$ on $x(t - l)$, $l = 1, \ldots, (n + h_T)$,

$$\frac{1}{T}\sum_{t=1}^{T}\hat{\eta}_T(t)\hat{\eta}_T(t)' \geq \frac{1}{T}\sum_{t=1}^{T}\hat{\xi}(t)\hat{\xi}(t)' = \frac{1}{T}\sum_{t=1}^{T}\varepsilon(t)\varepsilon(t)' + O(Q_T^2 \log T) \quad \text{a.s.}, \tag{4.1}$$

from Lemma (3.2).

Now let $[A_\infty : M_\infty]$ be as in Theorem (3.1) and $\theta_\infty = (\alpha'_\infty : \lambda'_\infty : \mu'_\infty)$. Then for any permutation matrix P,

$$(4.2)\quad P\hat{\eta}_T(t) = P\{A_\infty(z)x(t) - (M_\infty(z) - I)\hat{\varepsilon}_T(t)\} + P\hat{R}_\nu(t)(\theta_\infty - \hat{\theta}_T).$$

To further analyse (4.2), we appeal to:

LEMMA 4.2. *Under the conditions of Theorem* 4.1:

(i) *If* $n_r < n_{r0}$, $r = 1, \ldots, \nu$, *then* $[A_\infty : M_\infty]$, *which is unique, is such that* $\|\{A_\infty(z)K(z) - M_\infty(z)\}\Omega^{1/2}\|_{\mathscr{L}_2} > 0$.

(ii) *If* $n_{r(j)} = n_{r(j)0}$, $j = 1, \ldots, u$, $n_{r(j)} = n \le n_{r(u)0}$, $j = u+1, \ldots, v$, $1 \le u \le v$, *then*

$$P[A_\infty : M_\infty] = \begin{bmatrix} A^P_{1\infty} : M^P_{1\infty} \\ A^P_{2\infty} : M^P_{2\infty} \end{bmatrix} \begin{matrix} u, \\ v-u, \end{matrix}$$

where $[A^P_{1\infty} : M^P_{1\infty}]$ *satisfies* $A^P_{1\infty}(z)K(z) - M^P_{1\infty}(z) = 0$ *a.s.*, $|z| = 1$ *and* $[A^P_{2\infty} : M^P_{2\infty}]$ *minimises* $\|\{A^P_2(z)K(z) - M^P_2(z)\}\Omega^{1/2}\|_{\mathscr{L}_2} \ge 0$, *where* $[A^P_2 : M^P_2]$ *are the last* $v - u$ *rows of the permuted polynomial operator pair* $P[A : M]$, $\{n_{r(1)}, \ldots, n_{r(v)}\} = P\nu = \nu(u+1, n)$.

From Lemma 4.2,

$$P\{A_\infty(z)x(t) - (M_\infty(z) - I)\hat{\varepsilon}_T(t)\} = \begin{bmatrix} \varepsilon^P_{1\infty}(t) \\ \varepsilon^P_{2\infty}(t) \end{bmatrix} + P(M_\infty(z) - I)\{\hat{\varepsilon}_T(t) - \varepsilon(t)\},$$

where

$$\varepsilon^P_{1\infty}(t) = [I_u : 0]P\varepsilon(t),$$

because $A^P_{1\infty}(z)K(z) = M^P_{1\infty}(z)$ and

$$\varepsilon^P_{2\infty}(t) = [0 : I_{v-u}]P\varepsilon(t) + \{A^P_{2\infty}(z)K(z) - M^P_{2\infty}(z)\}\varepsilon(t).$$

Using Lemma (3.2) together with Theorem (3.1), we obtain

$$\frac{1}{T}\sum_{t=1}^{T} P\hat{\eta}_T(t)\hat{\eta}_T(t)'P' = \frac{1}{T}\sum_{t=1}^{T}\left[\varepsilon^P_{1\infty}(t)\varepsilon^P_{2\infty}(t)\right]\left(\varepsilon^P_{1\infty}(t)' : \varepsilon^P_{2\infty}(t)'\right) + \begin{bmatrix} O(Q_T^2 \log T) & O(Q_T) \\ O(Q_T) & O(Q_T) \end{bmatrix}. \quad (4.3)$$

We also have, via ergodicity,

$$\text{(4.4a)} \qquad \frac{1}{T}\sum \varepsilon_{1\infty}^{P}(t)\varepsilon_{1\infty}^{P}(t)' = [I_u : 0]P\Omega P'[I_u : 0]' + o(1) \quad \text{a.s.},$$

$$\text{(4.4b)} \qquad \frac{1}{T}\sum \varepsilon_{1\infty}^{P}(t)\varepsilon_{2\infty}^{P}(t)' = [I_u : 0]P\Omega P'[0 : I_{v-u}]' + o(1) \quad \text{a.s.},$$

$$\text{(4.4c)} \qquad \begin{aligned} &\frac{1}{T}\sum \varepsilon_{2\infty}^{P}(t)\varepsilon_{2\infty}^{P}(t)' \\ &\quad = [0 : I_{v-u}]P\Omega P'[0 : I_{v-u}]' \\ &\qquad + \frac{1}{2\pi}\int_{-\pi}^{\pi}\left(A_{2\infty}^{P}K - M_{2\infty}^{P}\right)\Omega\left(A_{2\infty}^{P}K - M_{2\infty}^{P}\right)^{*} d\omega + o(1) \quad \text{a.s.}, \end{aligned}$$

the form of (4.4b) and (4.4c) resulting from the fact that $A(z)K(z) - M(z)$ is strictly proper.

Suppose that $\hat{n}_{r(j)} = n_{r(j)0}$, $j = 1, \ldots, s-1$, $2 \le s \le v$. Let $R_v(t)'$ be defined as for $\hat{R}_v(t)'$, except that $\hat{\varepsilon}_T(t)$ is replaced by $\varepsilon(t)$ everywhere it occurs. Regarding the columns of $R_v(t)'$ as elements of the Hilbert space $\mathscr{L}_2(P)$ of random variables and recalling the isometry with $\mathscr{L}_2(\Omega\, d\omega)$ using the conventional norms as in Rosenberg (1963) and Rozanov (1967), it is easily seen that G is the Grammian of the regressors $R_v(t)'$. Write $M_r(\nu)$ for the $\mathscr{L}_2(P)$ closure of the linear manifold spanned by the subset of regressors appearing in the rth row of $R_v(t)'$. The squared norm of the residual from the projection of $x_{r(s)}(t)$ on $M_{r(s)}\{\nu(s,n)\}$, the residual variance $\sigma^2_{r(s)}\{\nu(s,n)\}$, is given by the $r(s)$th diagonal element of (4.4) with $u = s-1$ if $n \ne n_{r(s)0}$ and $u = s$, $n = n_{r(s)0}$. Extending the methods employed to prove Lemma (4.2) and using least squares theory in conjunction with an adaptation of the arguments of Pötscher [(1990), page 175] indicates that $\sigma^2_{r(s)}\{\nu(s,n)\}$ is monotonically decreasing in n for $n < n_{r(s)0}$, see below. Thus for $T > T' < \infty$, $\log \hat{\sigma}^2_{r(s)}\{\nu(s,n)\} > \log \hat{\sigma}^2_{r(s)}\{\nu(s,n+1)\}$ a.s. in view of the convergence of $\hat{\sigma}^2_{r(s)}\{\nu(s,n)\}$ to $\sigma^2_{r(s)}\{\nu(s,n)\}$ implied by (4.3). This means that $\Lambda_{r(s)}\{\nu(s,n)\} > \Lambda_{r(s)}\{\nu(s,n+1)\}$ for $0 \le n < n_{r(s)0}$, because $C(T)/T \to 0$ and hence $\hat{n}_{r(s)} \ge n_{r(s)0}$ a.s. When $n = n_{r(s)0}$, $\hat{\sigma}^2_{r(s)}\{\nu(s,n)\} = T^{-1}\sum \varepsilon_{r(s)}(t)^2 + O(Q_T^2 \log T)$ a.s. from (4.3) and (4.1) indicates that to terms $O(Q_T^2 \log T)$ the lower bound to the residual sum of squares, which is independent of ν, is attained. It follows that

$$\begin{aligned} &\Lambda_{r(s)}\{\nu(s, n_{r(s)0}+1)\} - \Lambda_{r(s)}\{\nu(s, n_{r(s)0})\} \\ &\qquad \ge (2v - s + 1)C(T)/T + O(Q_T^2 \log T) \end{aligned}$$

is strictly positive for T sufficiently large if $C(T)/\log T \log\log T \to \infty$ and therefore we can conclude that $\hat{n}_{r(s)} = n_{r(s)0}$ a.s. A completely analogous argument applied to $\Lambda_r\{\nu(n)\}$, $r = 1, \ldots, v$, recalling that $\nu(n) = \{n, \ldots, n\}$, shows that $\hat{n}_{r(1)}$ behaves as stated in the theorem and a simple induction completes the proof.

Reverting to the proof of Lemma (4.2), suppose $n_r < n_{r0}$, $r = 1, \ldots, v$, and assume that there exists a nonnull polynomial pair $[A : M]$ such that $\|\{A(z)K(z) - M(z)\}\Omega^{1/2}\|_{\mathscr{L}_2} = 0$. This implies $A(z)K(z) = M(z)$ a.s., $|z| = 1$,

since $\Omega > 0$, contradicting the coprimeness of $[A_0 : M_0]$ since $\delta_r[A(z) : M(z)] < n_{r0}$, $r = 1, \ldots, v$. Therefore $\|\{A_\infty(z)K(z) - M_\infty(z)\}\Omega^{1/2}\| > 0$. A similar argument shows that G is nonsingular: For if there exists a nonzero vector $\beta' = (c' : e' : d')$, partitioned conformably with the rows and columns of G, such that $\beta' G = 0$, then

$$\beta' G \beta = \frac{1}{2\pi}\int_{-\pi}^{\pi} \operatorname{tr}(CK - D)\Omega(CK - D)^* \, d\omega,$$

where

$$\operatorname{vec} C(z) = \left(\zeta_p(z)' \otimes I_{v^2}\right) S'_{a(v)} c + S'_{f(v)} e,$$

$$\operatorname{vec} D(z) = \left(\zeta_p(z)' \otimes I_{v^2}\right) S'_{m(v)} d + S'_{f(v)} e,$$

leading to the same contradiction. This establishes (i).

To show (ii), observe from the structure indicated in (2.6) and (2.7) that the specification of the operators in equations $r(j)$, $j = 1, \ldots, u$, are correct and are not influenced by any possible misspecification in rows $r(j)$, $j = u + 1, \ldots, v$. Using an analogous argument to that employed immediately above we can show that $\det G_{r(j)} \neq 0$, $j = 1, \ldots, u$, so that the first u rows of $P[A_\infty : M_\infty]$ are unique. Furthermore, Theorem (3.1) tells us that $[A_\infty : M_\infty]$ minimises $\|\{A(z)K(z) - M(z)\}\Omega^{1/2}\|_{\mathscr{L}_2}$. From the uniqueness of $[A^P_{1\infty} : M^P_{1\infty}]$ and $P[A_0 : M_0]$ it is clear that $[A^P_{1\infty} : M^P_{1\infty}] = [A^P_{10} : M^P_{10}]$. Thus $A^P_{1\infty}(z)K(z) - M^P_{1\infty}(z) = 0$ a.s., $|z| = 1$ and $[A^P_{2\infty} : M^P_{2\infty}]$ minimises $\|\{A^P_2(z)K(z) - M^P_2(z)\}\Omega^{1/2}\|_{\mathscr{L}_2} \geq 0$ as required.

It remains for us to show that $\sigma^2_{r(s)}\{\nu(s, n)\}$ form a strictly decreasing sequence in n when $n < n_{r(s)0}$. Set $P_s = [I_s : 0]P$ and write $[A_\infty(z) : M_\infty(z)]$ for the operator pair that minimises $\operatorname{tr} \Psi(\nu)$, where $P\nu = \nu(s, n)$. Since $n_{r(j)} = n_{r(j)0}$, $j = 1, \ldots, s - 1$, we can deduce from Lemma (4.2) with $u = s - 1$ that $\sigma^2_{r(s)}\{\nu(s, n)\} = w_{r(s)r(s)} + \|P_s\{A_\infty(z)K(z) - M_\infty(z)\}\Omega^{1/2}\|^2_{\mathscr{L}_2}$. Let $[A(z) : M(z)]$ be any echelon canonical form with permuted structural index $\nu(s, n + 1)$. Using the generalised Bezout equality, Kailath [(1980), page 382], we can determine from $[A_\infty(z) : M_\infty(z)]$ a unimodular coordinate transformation between $[A(z) : M(z)]$ and $[X(z) : Y(z)]$, say, given by

$$[A(z) : M(z)] = [X(z) : Y(z)]\begin{bmatrix} -H_\infty(z) & G_\infty(z) \\ A_\infty(z) & M_\infty(z) \end{bmatrix}$$

$$= [Y(z)A_\infty(z) - X(z)H_\infty(z) : Y(z)M_\infty(z) + X(z)G_\infty(z)].$$

Now assume that $\sigma^2_{r(s)}\{\nu(s, n)\} = \sigma^2_{r(s)}\{\nu(s, n + 1)\}$. By definition of $\sigma^2_{r(s)}\{\nu(s, n + 1)\}$ this means that for every operator pair $[X(z) : Y(z)]$,

$$\begin{aligned} &\left\| P_s\{A_\infty(z)K(z) - M_\infty(z)\}\Omega^{1/2}\right\|^2_{\mathscr{L}_2} \\ &\leq \left\| P_s\{Y(z)[A_\infty(z)K(z) - M_\infty(z)] \right. \\ &\qquad \left. - X(z)[H_\infty(z)K(z) + G_\infty(z)]\}\Omega^{1/2}\right\|^2_{\mathscr{L}_2}. \end{aligned} \tag{4.5}$$

The minimum is obviously achieved when $[X(z) : Y(z)] = [0 : I]$. Treating the

right-hand side of (4.5) as a functional of $[X(z):Y(z)]$, it is straightforward to show that it has Frechet differential

$$dF\{X(z):Y(z):\Delta X(z):\Delta Y(z)\}$$
$$= \frac{1}{\pi}\operatorname{tr}\int_{-\pi}^{\pi} \mathscr{R}eP_s\{Y(\overset{*}{A}_\infty K - M_\infty) - X(H_\infty K + G_\infty)\}$$
$$\times\Omega\left[P_s\{\Delta Y(A_\infty K - M_\infty) - \Delta X(H_\infty K + G_\infty)\}\right]^* d\omega.$$

Because $[0, I]$ provides an extreme point of the functional we know, via Vainberg [(1964), Theorem 9.1], for example, that $dF\{0:I:\Delta X(z):\Delta Y(z)\} = 0$ for arbitrary $[\Delta X(z):\Delta Y(z)]$. Choosing $[\Delta X(z):\Delta Y(z)] = [0:I]$, we are lead to the conclusion that $\|P_s\{A_\infty(z)K(z) - M_\infty(z)\}\Omega^{1/2}\|_{\mathscr{L}_2} = 0$. This implies, however, that $K(z)$ can be expressed in terms of an echelon canonical form whose first s Kronecker invariants are $n_{r(j)} = n_{r(j)0}$, $j = 1, \ldots, s-1$, and $n_{r(s)} = n$, which is evidently not possible if $n < n_{r(s)0}$. Hence equality between $\sigma^2_{r(s)}\{\nu(s,n)\}$ and $\sigma^2_{r(s)}\{\nu(s,n+1)\}$ is excluded when $n < n_{r(s)0}$. Because $M_{r(s)}\{\nu(s,n)\} \subseteq M_{r(s)}\{\nu(s,n+1)\}$ for all n, it now follows from standard least squares (Hilbert space) theory [Rao and Mitra (1971); Seber (1977)] that the residual variances satisfy $\sigma^2_{r(s)}\{\nu(s,n)\} > \sigma^2_{r(s)}\{\nu(s,n+1)\}$ whenever $n < n_{r(s)0}$, as required. □

5. Approximation of nonrational transfer functions. When $K(z)$ is not rational, that is, it is not possible to find $n_r < \infty$, $r = 1, \ldots, v$, and a pair $[A:M]$ in (reversed) echelon form such that $A(z)K(z) = M(z)$ a.s., $|z| = 1$, finite order models may be used to approximate the structure of the data-generating mechanism. If the identification procedures defined in Section 3 and 4 are employed to estimate and determine such an approximation, then in order to analyse these methods it is necessary to alter the previous prescription concerning the true stochastic structure. Assumption (A1) is modified to (A1)′ by including the additional condition $\sum_{j\geq 0} j^{1/2}\|k(j)\| < \infty$ and (A2) is applied directly to the innovation process, rather than indirectly via $\eta(t)$ and the notion of a true model to give (A2)′. The alternations are required because the model class being entertained no longer obtains. With these modifications to the assumptions about $x(t)$, Proposition (3.2) and Lemma (3.3) still apply with Q_T replaced by $Q'_T = (\log T/T)^{1/2}$ and $h_T = O(T/\log T)^{1/2}$ [Hannan and Kavalieris (1986), An, Chen and Hannan (1982)]. Thus, Theorem (3.1) is also applicable with Q'_T exchanged for Q_T. Furthermore, since $x(t)$ does not admit an ARMA representation, the same argument as previously employed shows that Lemma 4.2(i) holds true for any fixed n, the $\text{ARMA}_E\{\nu(n)\}$ approximant being uniquely identified.

Referring now to the two steps of the structure determination process, let $\sigma^2_r\{\nu(n)\}$ be [as before] the residual variance from the projection of $x_r(t)$ on to the regressors $x_j(t-l)$, $\varepsilon_j(t-l)$, $j = 1, \ldots, v$, $l = 1, \ldots, n$, and $x_j(t) - \varepsilon_j(t)$, $j = 1, \ldots, r-1$, in $\mathscr{L}_2(P)$. A repetition of the logic employed in Section 4 shows that $\sigma^2_r\{\nu(n)\} > \sigma^2_r\{\nu(n+1)\}$, $r = 1, \ldots, v$. Recall that $\sigma^2_r(\nu)$ is given by the rth diagonal element of $\Omega + \Psi(\nu)$, $\omega_{rr} + \psi_{rr}(\nu)$ and put $\bar{\Lambda}_r\{\nu(n)\} =$

$\log(1 + \psi_{rr}\{\nu(n)\}/\omega_{rr}) + (2vn + r - 1)C(T)/T$. From the modifications indicated in the previous paragraph we find, in a manner similar to the derivation of (4.3) and (4.4), that $\hat{\sigma}_r^2\{\nu(n)\} = \sigma_r^2\{\nu(n)\} + o(1)$ a.s. and hence that $|\Lambda_r\{\nu(n)\} - \log \omega_{rr} - \bar{\Lambda}_r\{\nu(n)\}| = o(1)$ with probability 1. For any fixed $T < \infty$ $\bar{\Lambda}_r\{\nu(n)\}$ attains a global minimum for n finite for $\psi_{rr}\{\nu(n)\}$ declines monotonically in n, ω_{rr} is independent of n and the correction term increases linearly with n. Combining this argument in an obvious way with those used to establish Theorem (4.1) shows that the $\hat{n}_{r(j)}$, $j = 1, \ldots, v$, derived from the identification process will converge to the ordered set of $\bar{n}_{rT}$, $r = 1, \ldots, v$, $\bar{n}_{r(1)T} \leq \bar{n}_{r(2)T} \leq \cdots \leq \bar{n}_{r(v)T}$, respectively. Nevertheless, for any integer $N > 0$, $\bar{\Lambda}_r\{\nu(n)\} > \bar{\Lambda}_r\{\nu(n + 1)\}$ for $0 \leq n < N$ a.s. since, by assumption, $C(T)/T \to 0$ a.s. as $T \to \infty$. Therefore the minimum of $\bar{\Lambda}_r\{\nu(n)\}$ occurs at $\bar{n}_{rT} \geq N$ for T sufficiently large and since N is arbitrary, this implies $\bar{n}_{rT} \to \infty$ as $T \to \infty$. Summarising these results we have the following.

THEOREM 5.1. *Suppose that $x(t)$ does not admit an ARMA representation but satisfies* (A1)′ *and* (A2)′. *If $\bar{n}_{rT}$, $r = 1, \ldots, v$, denotes a sequence of positive integers at each of which the minimum of $\bar{\Lambda}_r\{\nu(n)\}$ with respect to n is attained, $r = 1, \ldots, v$, then $\bar{n}_{rT} \to \infty$, $r = 1, \ldots, v$, a.s. This implies that $\hat{n}_{r(j)} \to \infty$ a.s., $j = 1, \ldots, v$.*

Because $\bar{n}_{rT}$ yields the minimum of $\bar{\Lambda}_r\{\nu(n)\}$, $r = 1, \ldots, v$, the intuitive interpretation that the identification procedure selects the model that appears nearest the generating mechanism of the process is still available. As an illustration of the application of this notion, suppose that for large n, $\psi_{rr}\{\nu(n)\} = \rho_{rr}(n)\{1 + o(1)\}$, which we write as $\psi_{rr}\{\nu(n)\} \sim \rho_{rr}(n)$, where $\rho_{rr}(n)$ is a twice continuously differentiable function satisfying $d\rho_{rr}(n)/dn < 0$, $d^2\rho_{rr}(n)/dn^2 > 0$ and $\rho_{rr} > 0$. Using the fact that for y small, $\log(1 + y) \sim y$, we see that $\bar{n}_{rT}$ is governed by the behavior of $\rho_{rr}(n) + \omega_{rr}(2vn + r - 1)C(T)/T$. If $\rho_{rr}(n) = c_0 n^{-\beta_0}$, $\beta_0 > 1$, then $\bar{n}_{rT} \sim (C_0(T))^{1/(1+\beta_0)}$, where $C_0(T) = c_0 T/2vC(T)\omega_{rr}$; if $\rho_{rr}(n) = c_0\beta_0^n$, $0 < \beta_0 < 1$, then $\bar{n}_{rT} \sim (C_0(T)/(-\log \beta_0))$. Observe that $1 + \psi_{rr}\{\nu(n)\}/\omega_{rr}$ represents the ratio of the one-step ahead prediction error variance obtained from predicting $x_r(t)$ from the model over the innovation variance. More generally, $\Sigma_r \psi_{rr}\{\nu(n)\}$ can be taken as a measure of the goodness-of-fit of the rational approximation $\Phi_\infty(z) = A_\infty(z)^{-1}M_\infty(z)$ to $K(z)$. If these quantities are viewed as being indicative of the relative merits or efficiencies of alternative specifications, then it appears from the inverse relationship between $\bar{n}_{rT}$ and $C(T)$ that consistency and efficiency are not compatible, a state of affairs already known to exist in the context of autoregressive approximation [Shibata (1980) and Hannan and Kavalieris (1986)].

6. Some practical considerations. In order to implement the above identification procedures, numerical algorithms for solving the least squares problems described in Section 3 must be chosen. One such choice corresponds to the methods outlined in Hannan and Kavalieris [(1984), Section 3.1]. These

are based on Levison–Whittle recursions and it is well known [Makhoul (1981), Tjöstheim and Paulsen (1983)] that the pre- and post-sample windowing implicit in the Toeplitz assumption underlying this method can produce substantial small sample bias. This problem might be circumvented by use of the class of lattice algorithms; see Friedlander (1982) for a survey and further references. Alternatively, the required least squares calculations could be performed directly via the QR algorithm. This procedure is numerically robust, and for the purposes of Section 4, readily lends itself to simple recursive calculations for the introduction and deletion of regressors [Seber (1977)]. Asymptotically, of course, these different methods are all equivalent and ultimately the choice made by the applied worker may be governed by questions of availability and convenience. Nevertheless, the need for empirical experience and experimental evidence on the properties of the different algorithms in finite samples is indicated.

Whatever method is employed in practice it will be necessary to monitor the calculations involved in the determination of $\hat{n}_{r(j)}$, $j = 1, \ldots, v$. Consider, for example, the first pass through Steps 1 and 2 of Section 4. The submatrix of $\hat{G}_T$, corresponding to the rth equation $\hat{G}_{rT}$, converges to G_r, where G_r is composed of the mean squares and cross-products of $x_j(t-l)$, $\varepsilon_j(t-l)$, $l = 1, \ldots, n$, $j = 1, \ldots, v$, and $x_j(t) - \varepsilon_j(t)$, $j = 1, \ldots, r-1$. Examination of (2.6) reveals, however, that if $\text{ARMA}_E(\nu_0)$ obtains, then

$$\sum_{j=1}^{v} \sum_{s=s(1,j)_0}^{n_{r(1)0}} a_{r(1)c(j)0}(s) x_{c(j)}(t-s) = \sum_{j=1}^{v} \sum_{s=1}^{n_{r(1)0}} m_{r(1)c(j)0} \varepsilon_j(t-s).$$

For $l = n - n_{r(1)0} > 0$, this defines an exact linear dependence between $x_{r(1)}(t-l)$ and $x_j(t-l-1), \ldots, x_j(t-n)$, $\varepsilon_j(t-l-1), \ldots, \varepsilon_j(t-n)$, $j = 1, \ldots, v$, and $x_j(t-l) - \varepsilon_j(t-l)$, $j = 1, \ldots, r(1)-1$, with weights given by the coefficients of the $r(1)$th row of $[A_0 : M_0]$. When $n > n_{r(1)0}$, these variables are a subset of those appearing in row $r(1)$ of $R_{v(n)}(t)'$ and the singularity of $G_{r(1)}$ and hence G follows. Lemma (3.3) therefore implies that when $n = n_{r(1)0} + 1$ $\det \hat{G}_{r(1)T} = O(Q_T)$ a.s. and it is this approach to singularity that will have to be checked as n is increased. Similar but notationally more complex derivations indicate the presence of equivalent properties whenever n in $\nu(u, n)$ exceeds $n_{r(u)0}$, $u = 2, \ldots, v$, and the need to scrutinize the least squares evaluations is apparent.

The singularities discussed immediately above clearly provide additional information on the Kronecker indices and structure of $[A : M]$ that can be constructively exploited. Indeed, if $n_{r(u)0}$ is known, the parameters of the $r(u)$th equation can be estimated by appropriate scaling of the elements of the eigenvector of $\hat{G}_{r(u)T}$ corresponding to the zero eigenvalue obtained when $n = n_{r(u)0} + 1$. The index $n_{r(u)}$ can itself be determined by developing a method for assessing when, as n is increased, an eigenvalue of $\hat{G}_{r(u)T}$ first appears to be not significantly different from zero. Such an approach is closely related to the singular value method used in canonical correlation analysis of Akaike (1976); see Tsay and Tiao (1985) and Tiao and Tsay (1989). Moreover,

as suggested by a referee, examination of the singularities present could indicate alternative constraints on $[A:M]$ to those given in (2.4) and this is related to the concept of redundant parameters discussed in Tiao and Tsay (1989). In the present context of using closed form least squares calculations, a more direct way of exploiting this information would be to equate $C(T)$ in $\Lambda_r(\nu)$ with $(1 + \operatorname{tr} \hat{G}_{rT}^{-1} \log T)$. Some justification for this value can be obtained from the arguments presented in Poskitt (1987a). This assignment generates a stochastic parameter adjustment term satisfying the conditions of Theorem (4.1) as $C(T)$ is $O(\log T)$ when the equation is underparameterised but $O(Q_T^{-1} \log T)$ if the specification is too profligate. Other choices for $C(T)$ consistent with the requirements of Theorem (4.1) can be made, $C(T) = (\log T)^{1+\delta_T}$, where $\delta_T > 0$ with $\delta_T \to 0$ as $T \to \infty$, so that high-dimensional models are penalised less as sample size increases, seems reasonable for example. One possible advantage that the former suggestion has over the latter one, however, is that it provides a straightforward data-oriented method of deciding on the magnitude of the parameter adjustment term; see Atkinson (1980) for a discussion of this point.

Finally, in the same vein, a value for $H_T = (\log T)^a$ will have to be chosen in practice and it is clear that such a choice could be critical. If $K(z)$ has a zero near the unit circle, for example, a and hence H_T will need to be reasonably large in order for $\hat{\varepsilon}_T(t)$ to provide a good approximation to the unobserved innovations. Note that $[\log T]$ only increases from 4 to 9 between $T = 100$ and $T = 10{,}000$. A common practice when fitting autoregressions is to set $H_T = qT/v^2$, $0 < q < 1$, so that the total number of coefficients estimated is bounded by a fixed proportion of the realisation length. The use of such a rule implies that $a = \log T / \log\log T + o(1)$ and seems quite reasonable, but the need for empirical experience and experimental evidence is once again apparent.

Acknowledgments. I would like to thank two anonymous referees for constructive and incisive comments on preliminary versions of this paper. These have led to the correction of errors, the clarification of ideas and, I believe, significant improvements in presentation.

REFERENCES

AKAIKE, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automat. Control* **AC-19** 716–723.

AKAIKE, H. (1976). Canonical correlation analysis of time series and use of an information criterion. In *Systems Identification: Advances and Case Studies* (R. K. Metra and D. G. Lainiotis, eds.) 27–96. Academic, New York.

AN, H.-Z., CHEN, Z.-G. and HANNAN, E. J. (1982). Autocorrelation, autoregression and autoregressive approximation. *Ann. Statist.* **10** 926–936.

ATKINSON, A. C. (1980). A note on the generalised information criterion for choice of a model. *Biometrika* **67** 413–418.

BOX, G. E. P. and JENKINS, G. M. (1976). *Time Series Analysis: Forecasting and Control*, 2nd ed. Holden-Day, San Francisco.

BOX, G. E. P. and TIAO, G. C. (1977). A canonical analysis of multiple time series. *Biometrika* **64** 355–365.

CHAN, N. H. and WALLIS, K. F. (1978). Multiple time series modelling: Another look at the mink-muskrat interaction. *J. Roy. Statist. Soc. Ser. C* **27** 168–175.

COOPER, D. M. and WOOD, E. F. (1982). Indentifying multivariate time series models. *J. Time Ser. Anal.* **3** 153–164.

DEISTLER, M. (1985). General structure and parametrization of ARMA and state space systems and its relation to statistical problems. In *Handbook of Statistics* (E. J. Hannan, P. R. Krishnaiah and M. M. Rao, eds.) **5** 257–277. North-Holland, Amsterdam.

DUNSMUIR, W. and HANNAN, E. J. (1976). Vector linear time series models. *Adv. in Appl. Probab.* **8** 339–364.

FRIEDLANDER, B. (1982). Lattice filters for adaptive processing. *Proceedings of the IEEE* **70** 829–867.

HANNAN, E. J. (1987). Rational transfer function approximation (with discussion). *Statist. Sci.* **2** 135–161.

HANNAN, E. J. and DEISTLER, M. (1988). *The Statistical Theory of Linear Systems*. Wiley, New York.

HANNAN, E. J. and KAVALIERIS, L. (1983). The convergence of autocorrelations and autoregressions. *Austral. J. Statist.* **25** 287–297.

HANNAN, E. J. and KAVALIERIS, L. (1984). Multivariate linear time series models. *Adv. in Appl. Probab.* **16** 492–561.

HANNAN, E. J. and KAVALIERIS, L. (1986). Regression autoregression models. *J. Time Ser. Anal.* **7** 27–49.

HANNAN, E. J. and POSKITT, D. S. (1988). Unit canonical correlations between future and past. *Ann. Statist.* **16** 784–790.

HOSKING, J. R. M. (1980). The multivariate portmanteau statistic. *J. Amer. Statist. Assoc.* **75** 602–608.

KAILATH, T. (1980). *Linear Systems*. Prentice-Hall, Englewood Cliffs, N.J.

KOHN, R. (1978). Asymptotic properties of time domain Gaussian estimators. *Adv. in Appl. Probab.* **10** 339–359.

MAKHOUL, J. (1981). Lattice methods in spectral estimation. In *Applied Time Series Analysis II* (D. F. Findley ed.) 301–326. Academic, New York.

MCDOUGALL, A. (1988). *Algorithms in Time Series*. Doctoral dissertation, Australian National Univ.

NEUDECKER, H. (1969). Some theorems on matrix differentiation with special reference to Kronecker matrix products. *J. Amer. Statist. Assoc.* **64** 953–963.

NICHOLLS, D. F. (1977). A comparison of estimation methods for vector linear time series models. *Biometrika* **64** 85–90.

POSKITT, D. S. (1987a). Precision, complexity and Bayesian model determination. *J. Roy. Statist. Soc. Ser. B* **49** 199–208.

POSKITT, D. S. (1987b). A modified Hannan–Rissaren strategy for mixed autoregressive moving-average order determination. *Biometrika* **74** 781–790.

POSKITT, D. S. and TREMAYNE, A. R. (1982). Diagnostic tests for multiple time series models. *Ann. Statist.* **10** 114–120.

PÖTSCHER, B. M. (1990). Estimation of autoregressive moving average order given an infinite number of models and approximation of spectral densities. *J. Time Ser. Anal.* **11** 165–179.

RAO, C. R. and MITRA, S. K. (1971). *Generalised Inverse of Matrices and Its Applications*. Wiley, New York.

RISSANEN, J. (1978). Modelling by shortest data description. *Automatica* **14** 465–471.

ROSENBERG, M. (1963). The square-integrability of matrix-valued functions with respect to a nonnegative Hermitian measure. *Duke Math. J.* **31** 291–298.

ROZANOV, YU, A. (1967). *Stationary Random Processes*. Holden-Day, San Francisco.

SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464.

SEBER, G. A. F. (1977). *Linear Regression Analysis*. Wiley, New York.

SHIBATA, R. (1980). Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. *Ann. Statist.* **8** 147–164.

TIAO, G. C. and BOX, G. E. P. (1981). Modelling multiple time series with applications. *J. Amer. Statist. Assoc.* **76** 802–816.

TIAO, G. C. and TSAY, R. S. (1989). Model specification in multivariate time series. *J. Roy. Statist. Soc. Ser. B* **51** 157–213.

TJÖSTHEIM, D. and PAULSEN, J. (1983). Bias in some commonly-used time series estimates. *Biometrika* **70** 389–399.

TSAY, R. S. and TIAO, G. C. (1985). Use of canonical analysis in time series indentification. *Biometrika* **72** 299–316.

TUNNICLIFFE-WILSON, G. (1973). The estimation of parameters in multivariate time series. *J. Roy. Statist. Soc. Ser. B* **35** 76–85.

VAINBERG, M. M. (1964). *Variational Methods for the Study of Nonlinear Operators*. Holden-Day, San Francisco.

DEPARTMENT OF STATISTICS
AUSTRALIAN NATIONAL UNIVERSITY
GPO BOX 4, CANBERRA ACT 2601
AUSTRALIA