ISSN 1440-771X

MONASH University

Australia

Department of Econometrics
and Business Statistics

http://www.buseco.monash.edu.au/depts/ebs/pubs/wpapers/

Some Results on the Identification and Estimation of Vector ARMAX Processes

D.S. Poskitt

Working Paper 12/04

Some Results on the Identification and Estimation of Vector ARMAX Processes

by

D.S. Poskitt

Department of Econometrics and Business Statistics

Monash University

Victoria 3800

Australia

Tel: +61 3 9905 9378

Fax: +61 3 9905 5474

E-mail: don.poskitt@Buseco.monash.edu.au

ABSTRACT

This paper addresses the problem of identifying echelon canonical forms for a vector autoregressive moving average model with exogenous variables using finite algorithms. For given values of the Kronecker indices a method for estimating the structural parameters of a model using ordinary least squares calculations is presented. These procedures give rise, rather naturally, to a technique for the determination of the structural indices based on the use of conventional model selection criteria. A detailed analysis of the statistical properties of the estimation and identification procedures is given and some evidence on the practical significance of the results obtained is also provided. Modifications designed to improve the performance of the methods are presented. Some discussion of the practical significance of the results obtained is also provided.

Keywords: ARMAX model, consistency, echelon canonical form, efficiency, estimation, identification, Kronecker invariants, least squares, selection criterion, structure determination, subspace algorithm.

JEL Subject Classifications: C32,C51

1. *Introduction*

This paper is concerned with the analysis of multivariate ARMAX systems of the form

$$\sum_{j=0}^{p} \mathbf{A}(j)\mathbf{y}(t-j) + \sum_{j=1}^{p} \mathbf{B}(j)\mathbf{x}(t-j) = \sum_{j=0}^{p} \mathbf{M}(j)\boldsymbol{\eta}(t-j) \ . \tag{1.1}$$

where $\mathbf{y}(t)$ is an observable v component vector of outputs, $\mathbf{x}(t)$ is an observable u component vector of input variables and $\boldsymbol{\eta}(t)$ is an unobservable vector of v elements that characterise the random disturbances, or noise, influencing the system. Interpreting z^{-1} as the unit lag operator, viz: $z^{-1}\mathbf{y}(t) = \mathbf{y}(t-1)$, (1.1) can be expressed more succinctly as

$$\mathbf{A}(z)\mathbf{y}(t) + \mathbf{B}(z)\mathbf{x}(t) = \mathbf{M}(z)\boldsymbol{\eta}(t) \ , \tag{1.1$'$}$$

where

$$\mathbf{A}(z) = \sum_{j=0}^{p} \mathbf{A}(j)z^{-j} \ , \quad \mathbf{B}(z) = \sum_{j=1}^{p} \mathbf{B}(j)z^{-j} \quad \text{and} \quad \mathbf{M}(z) = \sum_{j=0}^{p} \mathbf{M}(j)z^{-j} \ ,$$

and with regard to (1.1′) the following conditions will be assumed to hold.

(A1) The polynomial matrices $\mathbf{A}(z)$ and $\mathbf{M}(z)$ satisfy $\det \mathbf{A}(z) \neq 0$ and $\det \mathbf{M}(z) \neq 0$, $|z| \geq 1$. The triple $\big[\mathbf{A}(z) : \mathbf{B}(z) : \mathbf{M}(z)\big]$ is (left) coprime and in echelon canonical form.

Writing $a_{rc}(z)$ for the r, c th element of $\mathbf{A}(z)$ and similarly setting $\mathbf{B}(z) = \big[b_{rc}(z)\big]$ and $\mathbf{M}(z) = \big[m_{rc}(z)\big]$, the echelon canonical form is characterised by the following restrictions defining the row degrees and exclusion constraints on the polynomial operators;

$$a_{rr}(z) = 1 + \sum_{j=1}^{n_r} a_{rr}(j)z^{-j} \ , \tag{1.2a}$$

$$a_{rc}(z) = \sum_{j=n_r-n_{rc}+1}^{n_r} a_{rc}(j)z^{-j} \ , \tag{1.2b}$$

$$b_{rc}(z) = \sum_{j=1}^{n_r} b_{rc}(j)z^{-j} \ , \tag{1.2c}$$

$$m_{rr}(z) = \sum_{j=0}^{n_r} m_{rc}(j)z^{-j} \ , \text{ and} \tag{1.2d}$$

$$m_{rc}(\infty) = a_{rc}(\infty) \ . \tag{1.2e}$$

The integers $n_r, r = 1, \ldots, v$, are called the Kronecker indices and they define a multi-index $\nu = (n_1, \ldots, n_v)$ that determines the internal lag structure of the ARMAX process or model. In expression (1.2b) the index

$$n_{rc} \left.\begin{array}{ll} = \min(n_r + 1, n_c) & r \geq c \\ = \min(n_r, n_c) & r < c \end{array}\right\} \quad r, c = 1, \ldots, v \; . \tag{1.3}$$

See Hannan and Deistler (1988, §2.5) and Reinsel (1993, §3.1) for further details. Henceforth $ARMAX_E(\nu)$ will denote the set of all ARMAX structures in echelon form with structural index $\nu = \{n_1, \ldots, n_v\}$.

Now suppose that there exists an index $\nu_0 = \{n_{10}, \ldots, n_{v0}\}$, $n_{r0} < \infty$, $r = 1, \ldots, v$, with associated polynomial operators $\mathbf{A}_0(z)$, $\mathbf{B}_0(z)$ and $\mathbf{M}_0(z)$ such that

$$\mathbf{A}_0(z)\mathbf{y}(t) + \mathbf{B}_0(z)\mathbf{x}(t) = \mathbf{M}_0(z)\boldsymbol{\epsilon}(t) \quad t = 0, \pm 1, \ldots \tag{1.4}$$

where the appendage of a 0 is used to denote evaluation at the true parameter point and $\boldsymbol{\epsilon}(t)$ is the innovation process associated with $\mathbf{y}(t)$. In this case the $ARMAX_E(\nu_0)$ model is said to obtain or to hold. The connection between (1.4) and any model of the form (1.1) is derived by observing that the residual process $\boldsymbol{\eta}(t)$ is defined indirectly by inverting $\mathbf{M}(z)$ to give $\boldsymbol{\eta}(t) = \boldsymbol{\Psi}(z)\mathbf{y}(t) + \boldsymbol{\Phi}(z)\mathbf{x}(t)$ where $\boldsymbol{\Psi} = \mathbf{M}^{-1}\mathbf{A}$ and $\boldsymbol{\Phi} = \mathbf{M}^{-1}\mathbf{B}$, and $\boldsymbol{\eta}(t) = \boldsymbol{\epsilon}(t)$ whenever $[\boldsymbol{\Psi} : \boldsymbol{\Phi}] = \mathbf{M}_0^{-1}[\mathbf{A}_0 : \mathbf{B}_0]$. As here, the indeterminant z will often be omitted from polynomials and power series where this causes no confusion.

Let $[\mathbf{K} : \mathbf{L}] = \mathbf{A}^{-1}[-\mathbf{B} : \mathbf{M}]$. By Assumption A1 the squared norm of $\mathbf{K}(z) = \sum_{j=1}^{\infty} \mathbf{K}(j)z^{-j}$, $\|\mathbf{K}\|^2 = \sum_{j>1} \|\mathbf{K}(j)\|^2$, $\|\mathbf{K}(j)\|^2 = \operatorname{tr} \mathbf{K}(j)\mathbf{K}(j)'$, is bounded and similarly $\|\mathbf{L}\|^2 < \infty$. Assume also that the input processes $\mathbf{x}(t)$ and $\boldsymbol{\epsilon}(t)$ satisfy:

(A2) The process $\boldsymbol{\epsilon}(t) = \big(\epsilon_1(t), \ldots, \epsilon_v(t)\big)'$ is a stationary, ergodic, martingale difference sequence. Thus if F_t denotes the σ-algebra generated by $\boldsymbol{\epsilon}(s)$, $s \leq t$, then $\mathrm{E}\left[\boldsymbol{\epsilon}(t) \mid F_{t-1}\right] = 0$. Furthermore, $\mathrm{E}\left[\boldsymbol{\epsilon}(t)\boldsymbol{\epsilon}(t)' \mid F_{t-1}\right] = \boldsymbol{\Sigma} > 0$ and $\mathrm{E}\left[\epsilon_j(t)^4\right] < \infty$, $j = 1, \ldots, v$.

(A3) The input $\mathbf{x}(t)$ is a zero mean, stationary process independent of $\boldsymbol{\epsilon}(t)$ with finite fourth moment. Furthermore, for $H_T = (\log T)^c,\ 1 < c < \infty$

$$\sup_{0\le\tau\le H_T} \| \lim_{T\to\infty} T^{-1} \sum_{t=1}^{T} \mathbf{x}(t)\mathbf{x}(t+\tau)' - \boldsymbol{\Gamma}_{xx}(\tau)\| = O(\sqrt{\log\log T/T})$$

$$\boldsymbol{\Gamma}_{xx}(\tau) = \int_{-\pi}^{\pi} e^{i\omega\tau}\mathbf{H}_{xx}(e^{i\omega})\, d\omega$$

where $\mathbf{H}_{xx}(e^{i\omega})$ is a $u\times u$ Hermitian matrix-valued function satisfying $c_l\mathbf{I} \le \mathbf{H}_{xx}(e^{i\omega}) \le \mathbf{I}c_u,\ 0 < c_l \le c_u < \infty$, on the interval $[-\pi, \pi]$.

If the input-output system admits an ARMAX representation of the form (1.4) then there is a unique correspondence between $\mathbf{K}_0(z)$, $\mathbf{L}_0(z)$ and $\boldsymbol{\Sigma}_0$, and the second moment properties of the observable processes $\mathbf{y}(t)$ and $\mathbf{x}(t)$. In particular, if $\boldsymbol{\Gamma}_{yx}(z) = \sum \boldsymbol{\Gamma}_{yx}(s)z^{-s}$ denotes the cross-autocovariance generating function between $\mathbf{y}(t)$ and $\mathbf{x}(t)$ then

$$\boldsymbol{\Gamma}_{yy}(z) = \mathbf{K}_0(z)\boldsymbol{\Gamma}_{xx}(z)\mathbf{K}_0(z^{-1})' + \mathbf{L}_0(z)\boldsymbol{\Sigma}_0\mathbf{L}_0(z^{-1})' \qquad \text{and}$$

$$\boldsymbol{\Gamma}_{yx}(z) = \mathbf{K}_0(z)\boldsymbol{\Gamma}_{xx}(z)$$

and in principle $\mathbf{K}_0(z)$, $\mathbf{L}_0(z)$ and $\boldsymbol{\Sigma}_0$ can be determined directly from perfect knowledge of $\boldsymbol{\Gamma}_{yy}(z)$, $\boldsymbol{\Gamma}_{xx}(z)$ and $\boldsymbol{\Gamma}_{yx}(z)$. See Hannan and Deistler (1988). Such knowledge is generally not available however, and the statistical problem being addressed is that of identifying and estimating an $ARMAX_E(\nu)$ model using input-output data.

Multivariate time series models have, of course, been given considerable attention by research workers in the past and accounts of many of the methods and theoretical results currently available are given in Hannan and Deistler (1988), Lütkepohl (1991) and Reinsel (1993), for example. The question of how best to determine the internal structure of a multivariate model in a direct and straightforward manner has not, however, been completely resolved.

In the signal processing literature recent interest has focused on the so called subspace identification methods due to Van Overschee and De Moor (1994,1996). These techniques adapt ideas introduced in Akaike (1976) and use canonical correlations to estimate the

system matrices of an ARMAX model expressed in state-space form. The singular value decomposition underlying subspace algorithms can also be used to determine the order or McMillan degree $d = \sum_{r=1}^{v} n_r$ of the system. Subspace algorithms do not discriminate between members of the manifold $\mathcal{M}(m) = \{[\mathbf{K} : \mathbf{L}] : d = m\}$, the set of rational, proper and stable transfer functions of order m, however, since they do not use an explicit structural form. Here we consider the echelon canonical form since (i) it is simply expressed in terms of zero-one constraints, requiring fewer than $n(2v+u)$ parameters to describe the system, and (ii) the set $\{ARMAX_E(\nu) : d = m\}$ forms a disjoint cover of $\mathcal{M}(m)$, avoiding any difficulties associated with overlapping parameterisations, see Guidorzi (1981) and Hannan and Deistler (1988).

Earlier work in the statistics literature that builds on Akaike (1976) can be found in Tiao and Tsay (1989), and Nsiri and Roy (1992) and the references contained therein. The techniques discussed in Tiao and Tsay (1989) give rise to an approach to the examination of vector processes that is based on scalar-component models. An illuminating exposition of the similarities and differences between scalar-component models and echelon canonical forms and the orders of scalar-component representations and Kronecker indices in the context of ARMA processes is given in Tsay (1991). Nsiri and Roy (1992) deal directly with the Kronecker indices and their procedure is based on the detection of linear dependences implied by different structures. Both methods work in terms of the cross-autocovariances of the observed process and rely on the solution of different eigenvalue problems, solving the multiple decision problem via a sequence of hypothesis tests.

An alternative philosophical approach is taken in Hannan and Kavalieris (1984a) and Poskitt (1992), where the coefficients of an ARMA model expressed in echelon canonical form are estimated and the associated Kronecker indices determined using regression techniques and selection criteria that are structured in terms of the residual sums of squares and a penalty adjustment for the number of coefficients fitted, à la AIC (Akaike, 1974) or BIC (Schwarz, 1978). Reinsel (1993, §4.5) provides an interesting illustration of the use of

some of these different techniques and Lütkepohl and Poskitt (1996) present examples of the application of the regression based methodologies.

A basic purpose of the present paper is to indicate how the latter methodological approach can be employed to estimate and identify ARMAX systems using a new and simplified approach to the determination of the Kronecker invariants. A second objective is to fill a lacuna in the existing theory of multiple time series analysis. Lütkepohl and Poskitt (1996) observed that difficulties may arise when attempting to estimate Kronecker indices using regression methods because of singularities that are present when examining overparameterised models. A precise statement of the theoretical and practical consequences of fitting such models is given here and analytical results on the properties of the procedures that parallel those known to obtain in the context of scalar processes are developed.

Until recently little was known of the statistical properties of subspace algorithms but work by Peternell *et. al.* (1996) and Bauer *et. al.* (1999) has established the consistency and asymptotic normality of subspace-based systems parameter estimates under regularity conditions similar to those adopted here. These results require that the practitioner specify (backward and forward) truncation indices b and f that are bounded below by d but in general, of course, the true McMillan degree will not be known. Although it has been suggested that b can be chosen by reference to AIC applied to the Stage I (ARX) regression-autoregression of the following section, precise guidelines on how f should be selected have yet to be given. One off-shoot of the results presented in this paper is that the identification algorithm described in Section 3 may provide a natural choice of f.

The paper is organised as follows. In Section 2 a technique for estimating the structural parameters based on a two-stage least squares process is outlined and some statistical properties of the estimates are presented. The identification of the Kronecker invariants is then discussed in Section 3. A simple identification algorithm is advanced and theoretical results stating conditions under which strong convergence of the estimated values to the

true indices can be achieved are obtained. Results that provide an explanation of why the use of conventional model selection criterion such as AIC or BIC may lead to overparameterisations are also given. The fourth section of the paper presents some empirical evidence on the practical impact of the results obtained in Section 3. Section 5 then proposes a modification of the identification procedure that gives rise to a consistent model selection process that is governed by the law of the iterated logarithm. The sixth section of the paper presents some concluding remarks relating to the empirical import of the theoretical results obtained. Most proofs are assembled together in Section 7.

2. Two Stage Least Squares Estimation

The estimation process presented here is a single equation systems counterpart of an original proposal by Durbin (1960), that parallels the first two stages of the well known Hannan and Rissanen (1982) technique, see also Hannan and Deistler (1988, §6.5 & §6.7). To facilitate the presentation of the estimation method in terms of regular regression notation let

$$\mathbf{a}_r(z)\mathbf{y}(t) + \mathbf{b}_r(z)\mathbf{x}(t) = \mathbf{m}_r(z)\boldsymbol{\eta}(t) \tag{2.1}$$

denote the rth row of the system as defined in (1.1′). Set $\mathbf{a}_r(z) = \sum_{j=0}^{p} \mathbf{a}_r(j)z^{-j}$ and let $\boldsymbol{\lambda}_r$ and $\boldsymbol{\alpha}_r$ contain the freely varying parameters in $\mathbf{a}_r(0)$ and $\mathbf{a}_r(j),\ j = 1\ldots, p$, respectively, that are not restricted to be either zero or one by the indentification conditions in (1.2)-(1.3). Then

$$\mathbf{a}_r(z)\mathbf{y}(t) = \big[\mathbf{S}_a(r,\nu)(\boldsymbol{\zeta}_p \otimes \mathbf{y}(t))\big]'\boldsymbol{\alpha}_r + \big[\mathbf{S}_f(r,\nu)\mathbf{y}(t)\big]'\boldsymbol{\lambda}_r + \mathbf{y}(t)'\mathbf{e}_r \tag{2.2a}$$

where $\mathbf{S}_a(r,\nu)$ is a selection matrix that picks out appropriate lagged variables from $(\boldsymbol{\zeta}_p \otimes \mathbf{y}(t))$, $\boldsymbol{\zeta}_p' = (z^{-1}, \ldots, z^{-p})$, $\mathbf{S}_f(r,\nu)$ similarly selects appropriate components from $\mathbf{y}(t)$ and $\mathbf{e}_r = (0, \ldots, 0, 1, 0, \ldots, 0)'$ is the rth row of the $v \times v$ identity $\mathbf{I}_v$. Reexpressing $\mathbf{b}_r(z)\mathbf{x}(t)$ and $\mathbf{m}_r(z)\boldsymbol{\eta}(t)$ in a similar manner gives

$$\mathbf{b}_r(z)\mathbf{x}(t) = \big[\mathbf{S}_b(r,\nu)(\boldsymbol{\zeta}_p \otimes \mathbf{x}(t))\big]'\boldsymbol{\beta}_r \tag{2.2b}$$

and

$$\mathbf{m}_r(z)\boldsymbol{\eta}(t) = \big[\mathbf{S}_m(r,\nu)(\boldsymbol{\zeta}_p \otimes \boldsymbol{\eta}(t))\big]'\boldsymbol{\mu}_r + \big[\mathbf{S}_f(r,\nu)\boldsymbol{\eta}(t)\big]'\boldsymbol{\lambda}_r + \boldsymbol{\eta}(t)'\mathbf{e}_r \qquad (2.2c)$$

wherein an obvious notation has been employed for the different selection matrices and parameter vectors associated with $\mathbf{b}_r(z)$ and $\mathbf{m}_r(z)$. Now let $\boldsymbol{\theta}_r = (\boldsymbol{\alpha}_r' : \boldsymbol{\beta}_r' : \boldsymbol{\lambda}_r' : \boldsymbol{\mu}_r')'$ denote the vector of parameters that appear in the rth equation. Substituting (2.2a)-(2.2c) in to (2.1) and rearranging terms gives

$$y_r(t) = \mathcal{R}_{r,\nu}(t)'\boldsymbol{\theta}_r + \eta_r(t)$$

where the vector of regressors

$$\mathcal{R}_{r,\nu}(t) = \begin{bmatrix} -\mathbf{S}_a(r,\nu)(\boldsymbol{\zeta}_p \otimes \mathbf{y}(t)) \\ -\mathbf{S}_b(r,\nu)(\boldsymbol{\zeta}_p \otimes \mathbf{x}(t)) \\ \mathbf{S}_f(r,\nu)(\boldsymbol{\eta}(t) - \mathbf{y}(t)) \\ \mathbf{S}_m(r,\nu)(\boldsymbol{\zeta}_p \otimes \boldsymbol{\eta}(t)) \end{bmatrix}$$

is obtained by selecting from the vector of potential variables that occur in the system those that appear in the rth equation. Supposing that a realization of $T + H_T$ observations on $\mathbf{y}(t)$ and $\mathbf{x}(t)$, $t = 1 - H_T, \ldots, -1, 0, 1, \ldots, T$, is at hand, T effective observations with H_T initial values, this construction now allows the two stages of the estimation process to be presented in an uncomplicated manner.

STAGE 1: For $r = 1, \ldots, v$, regress $y_r(t)$ on $\mathbf{y}(t-j)$ and $\mathbf{x}(t-j)$, $j = 1, \ldots, h_T$, $t = 1, \ldots, T$, to obtain residuals

$$\hat{\varepsilon}_{r,T}(t) = y_r(t) - \sum_{j=1}^{h_T}\Big\{\sum_{c=1}^{v}\hat{\psi}_{rc}(j)y_c(t-j) + \sum_{c=1}^{u}\hat{\phi}_{rc}(j)x_c(t-j)\Big\}$$

where $h_T \to \infty$ as $T \to \infty$, $0 \le h_T \le H_T = (\log T)^c$, $1 < c < \infty$.

STAGE II: For $r = 1, \ldots, v$ determine the least squares estimates $[\hat{\mathbf{a}}_{r,T} : \hat{\mathbf{b}}_{r,T} : \hat{\mathbf{m}}_{r,T}]$, or equivalently $\hat{\boldsymbol{\theta}}_{r,T}$, by minimising the residual mean square

$$T^{-1}\sum_{t=1}^{T}\big(\mathbf{a}_r(z)\mathbf{y}(t) + \mathbf{b}_r(z)\mathbf{x}(t) - (\mathbf{m}_r(z) - \mathbf{e}_r')\hat{\boldsymbol{\epsilon}}_T(t)\big)^2$$
$$= T^{-1}\sum_{t=1}^{T}\big(y_r(t) - \hat{\mathbf{R}}_{r,\nu}(t)'\boldsymbol{\theta}_r\big)^2$$

with respect to the freely varying parameters in $[\mathbf{a}_r : \mathbf{b}_r : \mathbf{m}_r]$, or equivalently $\boldsymbol{\theta}_r$, where $\hat{\mathbf{R}}_{r,\nu}(t)$ is defined as for $\mathcal{R}_{r,\nu}(t)$ having replaced the unknown $\boldsymbol{\eta}(t-s),\ s = 0, \ldots, p$, by the corresponding innovation estimates obtained at Stage 1.

Stage I consists of the fitting of a regression-autoregression and the purpose of this stage is to provide estimates of $\boldsymbol{\epsilon}(t),\ \ t = 1, \ldots, T$, using the observed input and output. If h_T is sufficiently large we can expect $\hat{\boldsymbol{\epsilon}}_T(t) = (\varepsilon_{1,T}(t), \ldots, \varepsilon_{v,T}(t))'$ to approximate the innovation $\boldsymbol{\epsilon}(t)$ in a reasonable manner since under present assumptions the coefficients in $\boldsymbol{\Psi}$ and $\boldsymbol{\Phi}$ will decline at an exponential rate. In particular, if $ARMAX_E(\nu_0)$ obtains then $[\boldsymbol{\Psi}_0 : \boldsymbol{\Phi}_0] = \mathbf{M}_0^{-1}[\mathbf{A}_0 : \mathbf{B}_0]$ and a partial fractions expansion of $\mathbf{M}_0(z)^{-1}$ indicates that $|\psi_{rc}(j)|$ and $|\phi_{rc}(j)|$ are bounded by $c_0|z_0|^j,\ \ 0 < c_0 < \infty$, where z_0 is the zero of $\mathbf{M}_0(z)$ nearest $|z| = 1$. If h_T is appropriately prescribed the truncation effect in the regression-autoregression should therefore be asymptotically negligible. A precise statement concerning the approximation error obtained by substituting $\hat{\boldsymbol{\epsilon}}_T(t)$ for $\boldsymbol{\epsilon}(t)$ is given in the following lemma.

LEMMA 2.1. *Suppose that the input output system admits an ARMAX representation satisfying assumptions (A.1) to (A.4) and that* $h_{0T} < h_T \leq H_T,\ h_{0T} = \log T/(-2\log|z_0|)$. *Then uniformly in* h_T

$$T^{-1}\sum_{t=s+1}^{T}\big\{\hat{\boldsymbol{\epsilon}}_T(t) - \boldsymbol{\epsilon}(t)\big\}\big\{\hat{\boldsymbol{\epsilon}}_T(t-s) - \boldsymbol{\epsilon}(t-s)\big\}' = O(h_T Q_T^2) \quad a.s.$$
$$= \delta_{0,s} h_T T^{-1}(u+v)\boldsymbol{\Sigma} + o_p(h_T T^{-1})$$

and

$$T^{-1}\sum_{t=s+1}^{T}\boldsymbol{\epsilon}(t)\big\{\hat{\boldsymbol{\epsilon}}_T(t-s)-\boldsymbol{\epsilon}(t-s)\big\}' = O(h_T Q_T^2) \qquad a.s.$$
$$= -\delta_{0,s}h_T T^{-1}(u+v)\boldsymbol{\Sigma} + o_p(h_T T^{-1})$$

for $s \geq 0$ *where* $Q_T^2 = \log\log T/T$ *and* $\delta_{0,s} = 1$ *for* $s = 0$ *and is zero otherwise.*

This result is due to Hannan and Kavalieris (1984a), see also Hannan and Deistler (1988, §6.6).

In the second stage the freely varying parameters of the model are estimated equation by equation using least squares regressions for each $y_r(t),\ r = 1,\ldots,v$, with $\hat{\boldsymbol{\epsilon}}_T(t)$ substituted for the unobservable innovations or residual process. The stochastic properties of these estimates constitute the content of subsequent theoretical developments and are presented in the results that follow. In particular, if $\hat{\mathbf{g}}_T(r,\nu) = T^{-1}\Sigma y_r(t)\hat{\mathbf{R}}_{r,\nu}(t)'$, $\hat{\mathbf{G}}_T(r,\nu) = T^{-1}\Sigma\hat{\mathbf{R}}_{r,\nu}(t)\hat{\mathbf{R}}_{r,\nu}(t)'$ and $\hat{\sigma}^2_{r,T}(\nu) = \min_{\boldsymbol{\theta}_r} T^{-1}\sum_{t=1}^{T}\big(y_r(t)-\hat{\mathbf{R}}_{r,\nu}(t)'\boldsymbol{\theta}_r\big)^2$ then $\hat{\sigma}^2_{r,T}(\nu) = T^{-1}\sum_{t=1}^{T}\big(y_r(t)-\hat{\mathbf{R}}_{r,\nu}(t)'\hat{\boldsymbol{\theta}}_{r,T}\big)^2$ where by definition $\hat{\boldsymbol{\theta}}_{r,T}$ arises as a solution of the least squares normal equations $\hat{\mathbf{G}}_T(r,\nu)\boldsymbol{\theta}_r = \hat{\mathbf{g}}_T(r,\nu)$. Expanding the residual mean square as

$$T^{-1}\sum_{t=1}^{T} y_r(t)^2 - \hat{\boldsymbol{\theta}}'_{r,T}\hat{\mathbf{G}}_T(r,\nu)\hat{\boldsymbol{\theta}}_{r,T} \tag{2.3}$$

we see that the limiting behaviour of $\hat{\sigma}^2_r(\nu)$ and $\hat{\boldsymbol{\theta}}_{r,T}$ is governed by that of the second moment quantities in $\hat{\mathbf{g}}_T(r,\nu)$ and $\hat{\mathbf{G}}_T(r,\nu)$. These are described in the following complementary result.

LEMMA 2.2. *Set*

$$\mathbf{G}(r,\nu) = (2\pi)^{-1}\int_{-\pi}^{\pi}\mathbf{R}_{r,\nu}\mathbf{R}^*_{r,\nu}\,d\omega \qquad \text{and}$$
$$\mathbf{g}(r,\nu) = (2\pi)^{-1}\int_{-\pi}^{\pi} y_r\mathbf{R}^*_{r,\nu}\,d\omega$$

where

$$\mathbf{R}_{r,\nu}(z) = \begin{bmatrix} -\mathbf{S}_a(r,\nu)\big(\boldsymbol{\zeta}_p \otimes \big[\mathbf{K}(z)\mathbf{H}_{xx}(z)^{1/2} : \mathbf{L}(z)\boldsymbol{\Sigma}^{1/2}\big]\big) \\ -\mathbf{S}_b(r,\nu)\big(\boldsymbol{\zeta}_p \otimes [\mathbf{H}_{xx}(z)^{1/2} : \mathbf{0}]\big) \\ \mathbf{S}_f(r,\nu)\big(\big[\mathbf{K}(z)\mathbf{H}_{xx}(z)^{1/2} : (\mathbf{L}(z)-\mathbf{I})\boldsymbol{\Sigma}^{1/2}\big]\big) \\ \mathbf{S}_m(r,\nu)\big(\boldsymbol{\zeta}_p \otimes [\mathbf{0} : \boldsymbol{\Sigma}^{1/2}]\big) \end{bmatrix}$$

and $$y_r(z) = \mathbf{e}_r' \left[\mathbf{K}(z)\mathbf{H}_{xx}(z)^{1/2} : \mathbf{L}(z)\mathbf{\Sigma}^{1/2} \right] .$$

Then under the same conditions as for Lemma 2.1 $\hat{\mathbf{G}}_T(r,\nu) = \mathbf{G}(r,\nu) + O(Q_T)$ *and* $\hat{\mathbf{g}}_T(r,\nu) = \mathbf{g}(r,\nu) + O(Q_T)$.

In the statement of this lemma the argument $z = e^{i\omega}$ has been omitted and the asterisk denotes the complex conjugate transpose, conventions that will be adhered to throughout the paper. The proof of this and subsequent lemmas and theorems will be deferred to Section 6. A more detailed statement of the behaviour of $\hat{\sigma}^2_{r,T}(\nu)$ and $\hat{\boldsymbol{\theta}}_{r,T}$ when the structural index of the model $\nu = \nu_0 = \{n_{10}, \ldots, n_{v0}\}$ can now be given in the form of the following result.

LEMMA 2.3. *Suppose that assumptions (A1) to (A3) hold and* $ARMAX_E(\nu_0)$ *obtains. If the value* h_T *employed at Stage I is such that* $h_{0T} < h_T \leq H_T$ *and if for* $r = 1, \ldots, v$ $n_r = n_{r0}$, *then the Stage II estimate* $[\hat{\mathbf{a}}_{r,T} : \hat{\mathbf{b}}_{r,T} : \hat{\mathbf{m}}_{r,T}] = [\mathbf{a}_{r0} : \mathbf{b}_{r0} : \mathbf{m}_{r0}] + O(Q_T)$ *a.s. and* $\hat{\sigma}^2_{r,T}(\nu_0) = \sigma^2_{rr} + o(1)$ *a.s. for all* $r = 1, \ldots, v$.

At this point we observe that if ν_0 is known then Lemma 2.3 implies that $\hat{\boldsymbol{\theta}}_{r(q),T}$ provides a strongly consistent estimate of $\boldsymbol{\theta}_{r(q)0},\ q = 1, \ldots, v$, the parameter values associated with the unique representation of the system in terms of the Kronecker invariants. The two-stage least squares estimates will, however, be inefficient relative to those given by the Gaussian (maximum likelihood) estimator, see Poskitt and Salau (1994). Fully efficient estimates can be obtained by use of a full maximum likelihood procedure or by implementing Gauss–Newton type iterations using $\hat{\boldsymbol{\theta}}_{r(q),T},\ q = 1, \ldots, v$ as starting values, as described in Lütkepohl (1991, §7.2-7.4) or Reinsel (1993, §5.1-5.4) for example. Such calculations can present difficult computational burdens and complexities, however, and in general ν_0 will be unknown and a range of values for ν will have to be examined in order to estimate ν_0 and it would seem prudent to avoid the difficulties just alluded to before ν_0 has been identified, particularly in view of the *curse of dimensionality* implicit in the analysis of vector processes. We will therefore consider using the statistics derived in the

two stages presented above as the fundamental building blocks from which to construct an identification algorithm to determine the structural indices.

3. Identification of the Kronecker Invariants

The Kronecker indices are not invariant with respect to an arbitrary reordering of the elements of $\mathbf{y}(t)$ and to this extent the echelon canonical form is only unique modulo such rotations. However, the variables in $\mathbf{y}(t)$ can always be permuted so that the Kronecker indices are arranged in descending order, $n_{r(1)} \geq n_{r(2)} \geq \cdots \geq n_{r(v)}$, where $r(j),\ j = 1, \ldots, v$, denotes a rearrangement of $1, \ldots, v$ that induces the ordering. If $\mathbf{P}$ denotes a permutation matrix such that $\mathbf{P}(1, \ldots, v)' = (r(1), \ldots, r(v))'$ it is readily verified that $[\mathbf{PA}(z)\mathbf{P}^{-1} : \mathbf{PB}(z) : \mathbf{PM}(z)]$ provides an $ARMAX$ representation of $\mathbf{Py}_t$ and that the corresponding $ARMAX_E$ form has multi-index $(n_{r(1)}, \ldots, n_{r(v)})$. Note that the $r(j),\ j = 1, \ldots, v$, are unique modulo rotations of the indices that leave the ordering $n_{r(1)} \geq \cdots \geq n_{r(v)}$ unchanged. The $n_{r(j)}, j = 1, \ldots, v$, are referred to as the Kronecker invariants. When expressed in terms of the Kronecker invariants not only is the representation of the system in $ARMAX_E$ form canonical, but the individual variables $y_{r(j)}(t),\ j = 1, \ldots, v$ are uniquely represented.

In order to determine the Kronecker invariants let us begin by observing that $n_r = \delta_r[\mathbf{A} : \mathbf{B} : \mathbf{M}],\ \ r = 1, \ldots, v$, the row degrees of $[\mathbf{A} : \mathbf{B} : \mathbf{M}]$, and knowledge of the Kronecker index associated with $y_r(t)$ tells us the maximium lag of any variables appearing in the rth equation of the system. Knowing the ranking of n_r relative to the other indices, i.e. knowledge that $r = r(q)$, can be of no assistance, however, if the actual values $n_{r(j)},\ j = 1, \ldots, v$, and the associated permutation of the variables are not given, for otherwise any additional structure inherent in knowing that $n_r = n_{r(q)}$ cannot be exploited, unless that is $n_r = n_{r(v)} < n_{r(j)},\ j = 1, \ldots, v-1$, in which case the structure of the rth equation is determined solely by n_r. Since we wish to consider starting from a position of prior ignorance the approach that we shall adopt here is to determine the Kronecker indices by

searching through a collection of models for each $y_r(t)$ supposing that n_r coincides with the smallest Kronecker invariant. For any $n \geq 0$ let $\nu(n) = \{n, \ldots, n\}$. Formally the procedure is described in the following Identification Algorithm:

For each of $r = 1, \ldots, v$ perform steps (I) and (II):

(I) Calculate $\hat{\sigma}^2_{r,T}(\nu(n))$ for $n = 0, \ldots, N_T = O(\log T) \leq h_T$, the sequence of residual mean squares from the following regressions:

For $n = 0$ the regression of $y_r(t)$ on $(\varepsilon_{j,T}(t) - y_j(t)),\ j = 1, \ldots v,\ j \neq r$.

For $n = 1, \ldots, N_T$ the regression of $y_r(t)$ on $(\varepsilon_{j,T}(t) - y_j(t)),\ j = 1, \ldots v,\ j \neq r$, plus the regressors $-y_j(t-s),\ j = 1, \ldots, v,\ -x_c(t-s),\ c = 1, \ldots u$ and $\varepsilon_{j,T}(t-s),\ j = 1, \ldots, v,\ s = 1, \ldots, n$.

(II) Set the estimate of the rth Kronecker index equal to

$$\hat{n}_{r,T} = \arg \min_{0 \leq n \leq N_T} [\Lambda_{r,T}(\nu(n))].$$

where the criterion function

$$\Lambda_{r,T}(\nu(n)) = \log \hat{\sigma}^2_{r,T}(\nu(n)) + \kappa_T[(v-1) + n(2v+u)]/T$$

and κ_T is a nonnegative, nondecreasing function of T.

Note that as one cycles through the algorithm misspecified equations are being estimated for each $y_r(t),\ r = 1, \ldots, v$. This is because the $ARMAX_E$ form implies that the polynomial operators $\mathbf{a}_r(z)$, $\mathbf{b}_r(z)$ and $\mathbf{m}_r(z)$ that appear in the rth row of $[\mathbf{A} : \mathbf{B} : \mathbf{M}]$ exhibit additional zero restrictions that are governed by the values of the unknown Kronecker indices and such restrictions are not being explicitly accounted for. Thus, whenever $n < n_{r0}$ the rth equation will be misspecified since one or more lagged values required for a correct specification will be omitted. By adding additional lags we can therefore expect to reduce the magnitude of $\hat{\sigma}^2_{r,T}(\nu(n))$ until $n = n_{r0}$. At this point the maximum lag for the rth equation will be correctly specified but the equation will be potentially overparameterised in that some redundant variables may be included. When $n > n_{r0}$, however,

the rth equation will be incorrectly specified once again and overparameterised. Thus we might anticipate that if κ_T is prescribed appropriately the criterion function $\Lambda_{r,T}(\nu(n))$ will, asymptotically at least, possess a global minimum when $n = n_{r0}$ and that this is indeed the case is verified in Lemma 3.1.

LEMMA 3.1. *Suppose that* $\mathbf{x}(t),\ \mathbf{y}(t), t = 1, \ldots, T$, *is a realization of an input output process satisfying assumptions A1 and A3 and that the conditions stated in Lemma 2.1 hold. Let* $\hat{n}_{r,T},\ q = 1, \ldots, v$, *denote the estimated Kronecker indices obtained using the above identification algorithm. Then for all* $r = 1, \ldots, v$;

(i) $\Lambda_{r,T}(\nu(n_{r0})) < \Lambda_{r,T}(\nu(n))$ *whenever* $n < n_{r0}$ *and* $\liminf_{T\to\infty} \hat{n}_{r,T} \geq n_{r0}$ *a.s. if* $\kappa_T/T \to 0$, *and*

(ii) $\Lambda_{r,T}(\nu(n)) > \Lambda_{r,T}(\nu(n_{r0}))$ *for all* $n > n_{r0}$ *and* $\limsup_{T\to\infty} \hat{n}_{r,T} \leq n_{r(q)0}$ *a.s. if* $\log T/\kappa_T \to 0$.

It is perhaps worth pointing out that the determination of $\hat{n}_{r,T},\ r = 1, \ldots, v$, involves examining a total of $v(N_T + 1)$ different specifications. This is a considerable saving compared, for example, to the $(N_T + 1)^v$ specifications that would have to be examined if a full search over all $ARMAX$ structures in the set $\{ARMAX_E(\nu) : \nu \in \{\nu = (n_1, \ldots, n_v) : 0 \leq n_r \leq N_T,\ r = 1, \ldots, v\}\}$ were to be conducted. If $v = 4$ and $N_T = 6$ this gives only 28 different equations to be evaluated rather than 2401. Note also that the parameter correction term $\kappa_T[(v-1) + n(2v+u)]/T$ can be replaced by $C_{(v,u)}(T,n)/T$ where $C_{(v,u)}(T,n)$ is any function monotonically nondecreasing in T and n and such that $C_{(v,u)}(T,n)/T \to 0$ and $C_{(v,u)}(T,n)/\log T \to \infty$ as $T \to \infty$ without changing the basic result of the lemma.

Now let $\hat{n}_{\hat{r}(q),T},\ q = 1, \ldots, v$, denote the Kronecker invariants obtained by rearranging the $\hat{n}_{r,T},\ r = 1, \ldots, v$, into descending order and let $\hat{r}(q)_T,\ q = 1, \ldots, v$, denote the reordering of $r = 1, \ldots, v$ implied thereby. Note that identification of the Kronecker invariants involves the determination of both the value of the invariants themselves and the rearrangement $\mathbf{P}(1, \ldots, v)' = (r(1), \ldots, r(v))'$ since the order in which the variables

$y_r(t),\ r = 1, \ldots, v$, are presented is arbitrary. From Lemma 3.1 it is clear that $\hat{n}_{r(q)_0,T} \rightarrow n_{r(q)0}$ a.s. as $T \rightarrow \infty$ if κ_T increases with T such that $\kappa_T/T \rightarrow 0$ and $\log T/\kappa_T \rightarrow 0$. Suppose that this is the case and assume that $\hat{n}_{\hat{r}(j),T} = n_{r(j)0},\ j = 1, \ldots, q-1$. Then for T sufficiently large the ordering given by $\hat{n}_{\hat{r}(j),T},\ j = 1, \ldots, q$, will coincide with that given by $n_{r(j)0},\ j = 1, \ldots, q$ with probability one and hence, modulo invariant rotations, the estimate $\hat{r}(q)_T$ will converge to $r(q)_0$ a.s. if κ_T satisfies the requirements of parts (i) and (ii) of Lemma 3.1. Induction on $\hat{n}_{\hat{r}(j),T}$ and $\hat{r}(j)_T,\ j = 1, \ldots, v$, now yields the following Theorem.

THEOREM 3.2. *Assume that the conditions of Lemma 3.1 hold and the identification algorithm is implemented with the penalty term $\kappa_T \rightarrow \infty$ such that $\kappa_T/T \rightarrow 0$ and $\log T/\kappa_T \rightarrow 0$ as $T \rightarrow \infty$. Then modulo invariant rotations $\hat{r}(q)_T = r(q)_0$ a.s. for T sufficiently large and $Pr(\lim_{T\rightarrow\infty} \hat{n}_{\hat{r}(j),T} = n_{r(j)0}) = 1,\ j = 1, \ldots, v$.*

Lemma (3.1) indicates that use of the identification algorithm in conjunction with selection criteria which employ a value of κ_T that is at most $O(\log T)$, such as AIC or BIC, will most likely lead to the $\hat{n}_{r,T},\ r = 1, \ldots, v$, overestimating the true Kronecker indices and that we will have $\hat{n}_{r,T} \geq n_{r0}$, a.s., for T sufficiently large. A more detailed description of the extent of such overestimation is given in Theorem 3.4, which is a consequence of the following result characterising the properties of the two-stage estimation procedure when the Kronecker indices of the fitted model exceed those of the true process.

LEMMA 3.3. *Suppose that assumptions (A1) to (A3) hold, $ARMAX_E(\nu_0)$ obtains, and that the value h_T employed at Stage I is such that $h_{0T} < h_T \leq H_T$. Suppose also that $n_r \geq n_{r0},\ r = 1, \ldots, v$. Then for $r = 1, \ldots, v$*

$$[\hat{\mathbf{a}}_{r,T} : \hat{\mathbf{b}}_{r,T} : \hat{\mathbf{m}}_{r,T}] = \hat{\boldsymbol{\chi}}'_{r,T}[\mathbf{A}_0 : \mathbf{B}_0 : \mathbf{M}_0] + O(Q_T)$$

wherein $\hat{\boldsymbol{\chi}}_{r,T} = \boldsymbol{\chi}_{r0} + o_p(1)$ and $(\boldsymbol{\chi}_{r0}(z) - \mathbf{e}_r)'\mathbf{w}(t)$ is the minimum mean squared error predictor of $w_r(t)$ from $w_j(t-s),\ s = 0, \ldots, n_r - n_{j0},\ j \in \mathcal{K}_{r0} = \{q \in \{1, \ldots, v\} : n_{q0} \leq n_r\},\ \mathbf{w}(t) = (w_1(t), \ldots, w_v(t))' = \mathbf{M}_0(z)\boldsymbol{\epsilon}(t)$.

THEOREM 3.4. *Let* $\tau_r^2(\ell) = \mathrm{E}\left[(\boldsymbol{\chi}_{r0}(z)\mathbf{w}(t))^2\right]$ *be the minimum mean squared error from the projection of* $w_r(t)$ *on to the space spanned by* $w_j(t-s),\ s = 0, \ldots, n_r - n_{j0}$, $j \in \mathcal{K}_{r0},\ \ell = n_r - n_{r0},\ n_r \geq n_{r0}$. *If the conditions as stated in Lemma 3.3 obtain and* $\Lambda_{r,T}(\nu(n))$ *is employed at Stage II using the correction factor* κ_T *then, for* $r = 1, \ldots, v$, $\hat{n}_{r,T} = (n_{r0} + n_{r,T}^{\dagger})(1 + o_p(1))$ *where* $n_{r,T}^{\dagger}$ *denotes the non-negative integer* ℓ *that minimises*

$$\frac{h_T(u+v)}{\kappa_T}\Big(\frac{\tau_r^2(\ell)}{\sigma_{r,r}^2} - 1\Big) + \ell(2v+u).$$

Lemma 3.3 and Theorem 3.4 provide multivariate generalisations of the properties presented in Theorem 6.6.7 of Hannan and Deistler (1988), see also the comments following Theorem 6.7.2, Hannan and Deistler (1988, p.303). From the results given above it is apparent that the exact extent of the overestimation will depend not only on the structure of the system under investigation but also the values of h_T and κ_T employed in the analysis.

4. Empirical Illustrations

In order to implement the above procedures the practitioner will have to prescribe values for the design parameters $h_T,\ N_T$ and κ_T. The first of these can be chosen by using AIC to determine the order of the regression-autoregression at the first stage. If h_T^{AIC} denotes the value that minimises $T \log\det T^{-1} \sum \hat{\boldsymbol{\epsilon}}_T(t)\hat{\boldsymbol{\epsilon}}_T(t)' + 2h_T(v^2 + uv),\ 0 \leq h_T \leq H_T = (\log T)^{1.7}$ then by Theorem 6.6.3 of Hannan and Deistler (1988) and Proposition 4.3.1 of Lütkepohl (1991) $h_T = h_T^{AIC}$ will satisfy the conditions for application of the theoretical results presented above. Once h_T has been selected a natural choice for N_T is $h_T(u+v)/(2v+u)$. This value equates the number of freely varying coefficients in each equation of an $ARMAX_E(\nu(N_T))$ system with the number used in the regression-autoregression, recognising that the purpose of the $ARMAX$ model is to provide a more parsimonious but equally adequate representation of the observed multiple time series. The choice of κ_T is guided by Lemma 3.1 and Theorem 3.4, both of which show that any tendency to overestimate the Kronecker indices can be balanced by selecting κ_T to grow

at a rate at least as fast as $\log T$. Theorem 3.2 provides an asymptotic justification for a wide range of such values but if $\kappa_T = (\log T)^{1+\delta},\ \delta > 0$, then strongly consistent estimates will be generated. Given that δ can be arbitrarily small, to argue for anything other than the use of the limiting value $\kappa_T = \log T$ in most practical situations seems overly pedantic, particularly as time series folk-law suggests that the use of a parameter correction term that is too large in relation to T is likely to lead to underestimation, indicating that Theorem 3.4 may only be relevant when T is quite large.

In order to investigate the extent to which the predictions of asymptotic theory are reflected in finite sample behaviour and provide an indication of the practical significance of the results established above, simulation experiments have been used as a vehicle for analysing the sampling properties of the identification process. Realizations from bivariate, zero mean, (pseudo) Gaussian data generating mechanisms were generated and the sample sizes employed in the simulations were $T = 75(2^N),\ \ N = 0, 1, 2, 3, 4$ with the number of replications being given by $15(10^4)/T$ in each case. The latter rule is motivated by the notion that if T is large then large sample theory can be expected to provide a reasonable guide to the behaviour of the statistics of interest, whereas if T is small the adequacy of asymptotic approximations is questionable and it appears advisable to obtain more precise sampling information via an increase in the number of replications to be examined.

The results presented in Table 1 are derived from an $\text{ARMAX}_E(\nu_0)$ process with $\nu_0 = (2, 2)$,

$$\mathbf{A}_0(z) = \begin{bmatrix} 1.0 - 2.05z^{-1} + 0.615z^{-2} & 2.08z^{-1} - 0.85\ z^{-2} \\ -1.25z^{-1} + 0.613z^{-2} & 1.0 + 1.1z^{-1} - 0.938z^{-2} \end{bmatrix},$$

$$\mathbf{M}_0(z) = \begin{bmatrix} 1.0 - 4.75z^{-1} + 1.275z^{-2} & 4.95z^{-1} - 1.425z^{-2} \\ -3.9z^{-1} + 1.425z^{-2} & 1.0 + 4.0z^{-1} - 1.625z^{-2} \end{bmatrix},$$

$\mathbf{x}(t) \equiv 0$ and innovation variance-covariance matrix Σ_0 given by $\sigma_{11,0} = \sigma_{22,0} = 1.25$ and $\sigma_{12,0} = 1$. The values for the structural and scale parameters employed here are taken from Poskitt and Tremayne (1986), where they have been used for similar purposes and where some discussion of the experimental design considerations giving rise to such

values is provided. For each value of T the operational properties of the technique were summarised by dividing the realizations into mutually exclusive sets according to whether the structural index ν_0 was correctly identified or not. Denoting the first selection category by $\mathcal{C}$, its complement was further subdivided into those where the estimated Kronecker indices all exceeded $n_{i0},\ i = 1, 2$, denoted $\mathcal{E}$, or otherwise. In all the experiments reported here the first stage was implemented using AIC with $H_T = (\log T)^{1.7}$ and the average value of h_T^{AIC} for the regression-autoregression, $\bar{h}_T$, is also given.

TABLE 1

Results for Process I

T		75	150	300	600	1200
$\bar{h}_T$		4.6675	5.406	5.85	6.208	6.616
$\Lambda_{r,T}$	$\mathcal{C}$	0.54	0.69	0.496	0.24	0.04
$(\kappa_T = \log T)$	$\mathcal{E}$	0.05	0.18	0.46	0.73	0.95
$\Lambda_{r,T}$	$\mathcal{C}$	0.28	0.574	0.59	0.3	0.06
$(\kappa_T = \log T \log\log T)$	$\mathcal{E}$	0.01	0.06	0.33	0.69	0.94

Note: Entries in body of table give proportionate incidence of selection categories.

The propensity for the $\hat{n}_{r,T}$ to overestimate $n_{r0},\ r = 1, \ldots, v$, as T increases when $\kappa_T = \log T$ is clearly illustrated. Although the use of $\kappa_T = \log T \log\log T$ results in a decrease in the incidence of overestimation, as might be anticipated from the theoretical results presented above, the effect is small and transitory, and there is still a marked tendency to overestimate when T is large.

Table 2 provides a similar summary of the experimental outcomes from a data generating mechanism like the first except that $\mathbf{A}_0(z)$ and $\mathbf{M}_0(z)$ are replaced by

$$\begin{bmatrix} 1 - 1.002z^{-1} + 0.005z^{-2} & 2.993z^{-1} - 0.008z^{-2} \\ -1.99\ z^{-1} + 0.001z^{-2} & 1 + 0.55\ z^{-1} + 0.002z^{-2} \end{bmatrix}$$

and

$$\begin{bmatrix} 1 + 2z^{-1} & 4.333z^{-1} \\ -1.167z^{-1} & 1 - 2.5z^{-1} \end{bmatrix}$$

respectively. There is still some evidence of a propensity to overestimate the true Kronecker indices as the sample size increases for both choices of κ_T, but it seems that for this process T will have to considerably exceed 1200 before the experimental outcomes follow large sample dictates with a high degree of regularity.

TABLE 2

Results for Process II

T		75	150	300	600	1200
$\bar{h}_T$		2.525	2.998	3.282	3.948	4.6
$\Lambda_{r,T}$	$\mathcal{C}$	0.128	0.43	0.66	0.676	0.604
$(\kappa_T = \log T)$	$\mathcal{E}$	0.0015	0.006	0.036	0.02	0.384
$\Lambda_{r,T}$	$\mathcal{C}$	0.045	0.251	0.64	0.716	0.624
$(\kappa_T = \log T \log\log T)$	$\mathcal{E}$	0.0005	0.002	0.008	0.016	0.344

Note: Entries in body of table give proportionate incidence of selection categories.

A heuristic explanation for the differences observed with these two processes is not difficult to find. In both cases $\nu_0 = \{2, 2\}$, but unlike Process I, for Process II the elements $a_{ij0}(2), \quad i, j = 1, 2$, are all very small whilst the $m_{ij0}(2), \quad i, j = 1, 2$, are zero. This implies that the differences in the residual mean squares $\hat{\sigma}^2_{r,T}(\nu(n)) - \hat{\sigma}^2_{r,T}(\{2, 2\})$, $r = 1, 2$, $n = 0, 1$, are likely to be much smaller for the second process than for the first. Following the developments in Section 3 we can deduce that if $T = 1200$ and $\hat{\sigma}^2_{r,T}(\nu(n))$ for $n = 0, 1$ exceeds $\hat{\sigma}^2_{r,T}(\{2, 2\})$ by more than 1% then $\hat{n}_{r,T}$ will equal or exceed n_{r0}, but the number of observations needs to be increased eightfold in order for a difference between $\hat{\sigma}^2_{r,T}(\nu(n))$, $n = 0, 1$ and $\hat{\sigma}^2_{r,T}(\{2, 2\})$ of only 0.1% to lead to the same conclusion. Hence the higher frequency of occurrence of the event $\mathcal{E}$ for Process I than for Process II at these sample sizes.

The outcomes observed above serve to illustrate not only how the extent of overestimation can be influenced by the process under investigation, but also that overestimation can

present itself for values of T and the design parameters h_T and κ_T commonly encountered and employed in practice.

5. Second Phase Modifications

The properties presented in the previous section yield multivariate counterparts to phenomena first analysed by Hannan and Kavalieris (1984b) and subsequently described in Hannan and Deistler (1988) in the context of scalar models. Recognising that the overestimation implicit in Theorem 3.4 stems from the slow rate of convergence of the second moments of the regression-autoregression residuals to those of the innovation process these authors suggested that the residuals from the first stage be replaced by innovations estimates generated using the second stage coefficient values. A similar modification can be conducted here. The $\hat{\boldsymbol{\epsilon}}_T(t)$ determined at Stage I are replaced by $\tilde{\boldsymbol{\epsilon}}_T(t) = (\tilde{\epsilon}_{1,T}(t), \ldots, \tilde{\epsilon}_{v,T}(t))'$ where $\tilde{\boldsymbol{\epsilon}}_T(t)$ is generated recursively from

$$\sum_{j=0}^{\hat{p}_T} \tilde{\mathbf{M}}_T(j)\tilde{\boldsymbol{\epsilon}}_T(t-j) = \sum_{j=0}^{\hat{p}_T} \tilde{\mathbf{A}}_T(j)\mathbf{y}(t-j) + \sum_{j=1}^{\hat{p}_T} \tilde{\mathbf{B}}_T(j)\mathbf{x}(t-j) \ ,$$

for $t \geq 1 - H_T$ with initial values $\tilde{\boldsymbol{\epsilon}}_T(t) = \mathbf{0},\ t \leq -H_T$, an $ARMAX_E$ system wherein $\hat{p}_T = \max_{r=1,\ldots,v} \hat{n}_{r,T}$ and $[\tilde{\mathbf{A}}_T : \tilde{\mathbf{B}}_T : \tilde{\mathbf{M}}_T]$ denotes the system coefficient estimates obtained from the second stage, single equation least squares calculations carried out with those elements that are prescribed to zero by the echelon form based on the Kronecker invariants $\hat{n}_{\hat{r}(j),T},\ j = 1, \ldots, v$ set equal to zero. The second stage and identification algorithm are then repeated using $\tilde{\boldsymbol{\epsilon}}_T(t)$ in place of $\hat{\boldsymbol{\epsilon}}_T(t)$. The following lemma indicates the nature of the improvement obtained by using $\tilde{\boldsymbol{\epsilon}}_T(t)$ rather than $\hat{\boldsymbol{\epsilon}}_T(t)$ to estimate $\boldsymbol{\epsilon}(t)$.

LEMMA 5.1. *Suppose that the conditions of Lemma 3.1 obtain and that the identification algorithm is applied with κ_T of order $O(\log T)$ at most. Then for $0 \leq s \leq H_T$*

$$T^{-1} \sum_{t=s+1}^{T} \big\{\tilde{\boldsymbol{\epsilon}}_T(t) - \boldsymbol{\epsilon}(t)\big\}\big\{\tilde{\boldsymbol{\epsilon}}_T(t-s) - \boldsymbol{\epsilon}(t-s)\big\}' = O(Q_T^2) \qquad \textit{and}$$

$$T^{-1} \sum_{t=s+1}^{T} \boldsymbol{\epsilon}(t)\big\{\tilde{\boldsymbol{\epsilon}}_T(t-s) - \boldsymbol{\epsilon}(t-s)\big\}' = O(Q_T^2) \ .$$

The effect of the replacement is to improve the convergence rate of the residual mean square whilst leaving the properties of the coefficient estimates unchanged and simulation results presented in Hannan and Kavalieris (1984b) for the scalar case suggest that this enhances the performance of the model selection process.

More recently Kavalieris (1991) has pointed out that with univariate $ARMA$ models a repetition of the second stage is not necessary and that a marked improvement in performance can be achieved by basing the model selection criterion directly on the second stage innovations estimates. Reinsel (1993, §4.5) describes the application of a direct mutivariate generalisation of Kavalieris' procedure to vector $ARMA$ models expressed in normalised, simply identified form. In the current situation an immediate generalisation of this type is not possible because the generation of the $\tilde{\epsilon}_{r,T}(t),\ r = 1\ldots, v$, must be done simultaneously and this requires knowledge of the whole system. Such knowledge will not be forthcoming untill all the Kronecker invariants have been ascertained following a first pass through the identification algorithm. Nonetheless, an analogous procedure can be implemented by combining the innovations estimates $\tilde{\boldsymbol{\epsilon}}_T(t)$ with the different Stage II coefficient estimates $[\hat{\mathbf{a}}_{r,T} : \hat{\mathbf{b}}_{r,T} : \hat{\mathbf{m}}_{r,T}]$ that will become available once a first pass through the identification algorithm has been completed.

Let

$$\begin{aligned}\tilde{\hat{\sigma}}^2_{r,T}(\nu) &= T^{-1}\sum_{t=1}^{T}\left(\hat{\mathbf{a}}_{r,T}(z)\mathbf{y}(t) + \hat{\mathbf{b}}_{r,T}(z)\mathbf{x}(t) - (\hat{\mathbf{m}}_{r,T}(z) - \mathbf{e}_r')\tilde{\boldsymbol{\epsilon}}_T(t)\right)^2 \\ &= T^{-1}\sum_{t=1}^{T}\left(y_r(t) - \tilde{\mathbf{R}}_{r,\nu}(t)'\hat{\boldsymbol{\theta}}_{r,T}\right)^2\end{aligned}$$

where $[\hat{\mathbf{a}}_{r,T} : \hat{\mathbf{b}}_{r,T} : \hat{\mathbf{m}}_{r,T}]$, or equivalently $\hat{\boldsymbol{\theta}}_{r,T}$, are the second stage coefficient estimates obtained for the rth equation of an $ARMAX_E(\nu)$ system and $\tilde{\mathbf{R}}_{r,\nu}(t)$ is defined as for $\hat{\mathbf{R}}_{r,\nu}(t)$ except that $\tilde{\boldsymbol{\epsilon}}_T(t)$ replaces $\hat{\boldsymbol{\epsilon}}_T(t)$. Now set

$$\tilde{\Lambda}_{r,T}(\nu(n)) = \log\tilde{\hat{\sigma}}^2_{r,T}(\nu(n)) + \tilde{\kappa}_T[(v-1) + n(2v+u)]/T$$

where $\tilde{\kappa}_T$ is a nonnegative, nondecreasing function of T and define the estimate of the rth Kronecker index to be equal to

$$\tilde{n}_{r,T} = \arg \min_{0 \le n \le \hat{n}_{r,T}} [\tilde{\Lambda}_{r,T}(\nu(n))].$$

This modification determines the Kronecker indices as previously, but the new innovations estimates are combined with the second stage coefficient values to produce a different estimate of the residual variance and the $\tilde{n}_{r,T}$, $r = 1, \ldots, v$ are obtained by taking the previous estimates $\hat{n}_{r,T}$, $r = 1, \ldots, v$ as upper bounds and ascertaining whether the modification indicates the desirability of reducing these values. The following theorem and subsequent corollary show that strong consistency can be achieved via the modified identification algorithm with a lower bound to the rate of increase in $\tilde{\kappa}_T$ governed by the law of the iterated logarithm.

THEOREM 5.2. *Suppose that the assumptions of Lemma 5.1 hold. If the modified identification algorithm is applied using* $\tilde{\Lambda}_r(\nu(n))$ *with* $\tilde{\kappa}_T$ *such that* $\tilde{\kappa}_T / \log\log T \to \infty$ *and* $\tilde{\kappa}_T / T \to 0$ *as* $T \to \infty$ *then* $\lim_{T\to\infty} \tilde{n}_{r,T} = n_{r0}$, $r = 1, \ldots, v$ *with probability one.*

COROLLARY 5.3. *Let* $\tilde{n}_{\tilde{r}(q),T}$, $q = 1, \ldots, v$, *denote the Kronecker invariants obtained by rearranging the* $\tilde{n}_{r,T}$, $r = 1, \ldots, v$, *into descending order and let* $\tilde{r}(q)_T$, $q = 1, \ldots, v$, *denote the permutation of* $r = 1, \ldots, v$ *implied thereby. If* $\tilde{\kappa}_T \to \infty$ *such that* $\tilde{\kappa}_T / T \to 0$ *and* $\log\log T / \tilde{\kappa}_T \to 0$ *as* $T \to \infty$ *then modulo invariant rotations* $\tilde{r}(j)_T = r(j)_0$ *a.s. for* T *sufficiently large and* $Pr(\lim_{T\to\infty} \tilde{n}_{\tilde{r}(j),T} = n_{r(j)0}) = 1$, $j = 1, \ldots, v$.

Some idea of the impact of the above modification can be gained from Table 3. This table reports the outcomes observed when $\tilde{\Lambda}_r(\nu(n))$ with $\tilde{\kappa}_T = \log\log T$ is employed to identify the kronecker indices in the simulation experiments previously used to illustrate the sampling properties of the identification process in Section 4. Once again there is a higher frequency of occurrence of the event $\mathcal{E}$, indicating that the estimated Kronecker indices all exceeded n_{i0}, $i = 1, 2$, for Process I than for Process II, confirming that the sample

sizes needed before the asymptotic theory starts to bight may vary considerably from process to process. Nevertheless, the improvement in the performance of the identification methodology is clearly illustrated by the increased frequency with which the event $\mathcal{C}$ occurs for both processes, indicating that the structural index ν_0 is now being correctly identified much more regularly.

TABLE 3

Results for $\tilde{\Lambda}_r(\nu(n))$ with $\tilde{\kappa}_T = \log\log T$

T		75	150	300	600	1200
Process I	$\mathcal{C}$	0.59	0.73	0.88	0.94	0.98
	$\mathcal{E}$	0.06	0.16	0.16	0.002	0.001
Process II	$\mathcal{C}$	0.64	0.88	0.94	1.00	0.97
	$\mathcal{E}$	0.001	0.002	0.008	0.0	0.0

Note: Entries in body of table give proportionate incidence of selection categories.

6. Concluding Remarks

This paper has examined a generalisation of the Hannan and Rissanen (1982) technique of model selection to vector ARMAX processes expressed in echelon canonical form. It has shown that the phenomenon of order overestimation first analysed in the context of scalar models by Hannan and Kavalieris (1984b), and subsequently described in Hannan and Deistler (1988), carries over to the vector case, leading to the possible overestimation of the true Kronecker indices.

Such overestimation does not of itself constitute a condemnation of the identification algorithm since the purpose of the analysis may not be to determine n_{r0}, $r = 1, \ldots, v$, exactly. Indeed, as pointed out in the introduction, recent work on the statistical properties of subspace algorithms requires that the practitioner specify a truncation index $f \geq d_0 = \sum_{r=1}^{v} n_{r0}$, the true McMillan degree, see Bauer *et. al.* (1999). The results presented in this

paper suggest that the identification procedure of Section 3 may well provide a very sensible data directed method of selecting this truncation index, namely $f = \hat{d}_T = \sum_{r=1}^{v} \hat{n}_{r,T}$. The following result is an immediate consequence of the previous analytical developments.

COROLLARY 6.1. *Suppose that assumptions (A1) to (A3) hold whilst* $ARMAX_E(\nu_0)$ *obtains, that the value* h_T *employed at Stage I is such that* $h_{0T} < h_T \leq H_T$ *and the Identification Algorithm is implemented using the correction factor* κ_T *where* $\kappa_T/T \to 0$ *as* $T \to \infty$. *Then*

$$\begin{aligned} \hat{d}_T &\geq d_0 \qquad a.s. \\ &= (d_0 + d_T^{\dagger})(1 + o_p(1)) \end{aligned}$$

where $d_T^{\dagger} = n_{1,T}^{\dagger} + \cdots + n_{v,T}^{\dagger}$ *and the* $n_{r,T}^{\dagger}$, $r = 1, \ldots, v$, *are as defined in Theorem 3.4.*

If the relative-efficiency of subspace-based system parameter estimates is inversely related to the magnitude of f, as appears to be the case from the variance formulas given in Bauer *et. al.* (1999), then Corollary 5.1 intimates that $\hat{d}_T$ is likely to provide a value of f that will yield estimates with relatively good performance. Thus the choice $b = h_T^{AIC}$ determined at Stage I in conjunction with $f = \hat{d}_T$ as determined from the identification algorithm applied at Stage II with $\kappa_T = \log T$ suggests itself as a natural pairing for the two truncation indices input into subspace algorithms. Note also that if $n_r \geq n_{r0}$, $r = 1, \ldots, v$, and

$$[\hat{\mathbf{A}}_T : \hat{\mathbf{B}}_T : \hat{\mathbf{M}}_T] = \begin{bmatrix} \hat{\mathbf{a}}_{1,T} : \hat{\mathbf{b}}_{1,T} : \hat{\mathbf{m}}_{1,T} \\ \vdots \\ \hat{\mathbf{a}}_{v,T} : \hat{\mathbf{b}}_{v,T} : \hat{\mathbf{m}}_{v,T} \end{bmatrix}$$

then by Lemma 3.2 $[\hat{\mathbf{A}}_T : \hat{\mathbf{B}}_T : \hat{\mathbf{M}}_T] = \hat{\mathbf{X}}_T[\mathbf{A}_0 : \mathbf{B}_0 : \mathbf{M}_0] + O(Q_T)$ where the common factor matrix $\hat{\mathbf{X}}_T' = [\hat{\boldsymbol{\chi}}_{1,T} : \ldots : \hat{\boldsymbol{\chi}}_{v,T}]$. It follows that under appropriate regularity the transfer function $[\hat{\mathbf{K}}_T : \hat{\mathbf{L}}_T] = \hat{\mathbf{A}}_T^{-1}[-\hat{\mathbf{B}}_T : \hat{\mathbf{M}}_T]$ will satisfy $[\hat{\mathbf{K}}_T : \hat{\mathbf{L}}_T] = [\mathbf{K}_0 : \mathbf{L}_0] + O(Q_T)$. This suggests that the Stage II regression estimates upon which $\hat{d}_T$ are based might be usefully employed to construct initial values suitable for implementing the two-stage canonical correlation analysis subspace algorithm developed in Peternell et. al. (1996).

If the correct identification of n_{r0}, $r = 1, \ldots, v$, is important for the subsequent analysis, modifications of the Hannan and Rissanen (1982) technique due to Hannan and Kavalieris (1984b), and Kavalieris (1991), that are designed to circumvent the overestimation problem, have also been extended to cover the class of vector ARMAX processes expressed in echelon canonical form. These modifications give rise to order selection algorithms that yield strongly consistent estimates of the Kronecker indices and they have been shown to produce a marked improvement in the performance of the model selection process.

Finally, given that the overestimation is rooted in the use of the first stage regression-autoregression residuals as innovations estimates, it seems natural to consider the possibility of avoiding the use of the regression-autoregression residuals altogether rather than contemplating latter stage modifications. A method of parameter estimation and order determination for scalar $ARMA$ models that does not depend on estimating the innovations is presented in Poskitt and Chung (1996). The adaptation of their approach to the current situation is the subject of work in progress.

7. Proofs

7.1 Proof of Lemma 2.2 : Consider first $\hat{\mathbf{G}}_T(r,\nu)$. This symmetric matrix contains ten unique matrix blocks corresponding to the partitions induced by the mean squares and cross-products of the variables contained in the four-way partition of $\hat{\mathbf{R}}_{r,\nu}(t)$. The matrix in the second block of rows and third block of columns, for example, contains

$$\begin{aligned} &T^{-1}\sum_{t=1}^{T}\mathbf{S}_b(r,\nu)\big\{\boldsymbol{\zeta}_p\otimes\mathbf{x}(t)\big\}\big\{\mathbf{y}(t)-\hat{\boldsymbol{\epsilon}}_r(t)\big\}'\mathbf{S}_f(r,\nu)' \\ &=\mathbf{S}_b(r,\nu)T^{-1}\sum_{t=1}^{T}\begin{bmatrix}\mathbf{x}(t-1)\big\{\mathbf{y}(t)-\hat{\boldsymbol{\epsilon}}_T(t)\big\}'\\ \vdots\\ \mathbf{x}(t-p)\big\{\mathbf{y}(t)-\hat{\boldsymbol{\epsilon}}_T(t)\big\}'\end{bmatrix}\mathbf{S}_f(r,\nu)'\,. \end{aligned}$$

It is therefore necessary to determine the asymptotic properties of autocovariance estimates

of the form

$$\mathbf{C}_{x(y-\hat{\epsilon})}(j) = T^{-1}\sum_{t=j+1}^{T}\mathbf{x}(t-j)\big\{\mathbf{y}(t)-\hat{\boldsymbol{\epsilon}}_T(t)\big\}', \quad j=1,\ldots,p$$
$$= \mathbf{C}_{xy}(j) - \mathbf{C}_{x\epsilon}(j) + \mathbf{C}_{x(\epsilon-\hat{\epsilon})}(j)\ .$$

From the results presented in Hannan and Deistler (1988, §5.3) the first term is $\boldsymbol{\Gamma}_{xy}(j) + O(Q_T)$ and similarly the second is $O(Q_T)$ from the assumed independence of $\mathbf{x}(t)$ and $\boldsymbol{\epsilon}(t)$. The Cauchy–Schwarz inequality applied to the elements of $\mathbf{C}_{x(\epsilon-\hat{\epsilon})}(j)$ indicates that these are bounded by terms involving $\mathbf{C}_{(\epsilon-\hat{\epsilon})(\epsilon-\hat{\epsilon})}(0)$, which from Lemma 2.1 are $O(Q_T^2\log T)$. Bringing these results together gives

$$(2\pi)^{-1}\int_{-\pi}^{\pi}\mathbf{S}_b(r,\nu)(\boldsymbol{\zeta}_p\otimes\mathbf{H}_{xx}\mathbf{K}^*)\mathbf{S}_f(r,\nu)'\,d\omega + O(Q_T)\ .$$

Applying parallel arguments to the remaining subblocks of $\hat{\mathbf{G}}_T(r,\nu)$ and handling $\hat{\mathbf{g}}_T(r,\nu)$ analogously gives the desired result. □

7.2 Proof of Lemma 2.3 : Suppose that $\mathbf{G}(r,\nu_0)$ is singular for some $r=1,\ldots,v$. Then $\mathbf{G}(r,\nu_0)\bar{\boldsymbol{\theta}}_r = 0$ for some nonzero vector $\bar{\boldsymbol{\theta}}_r = (\bar{\boldsymbol{\alpha}}_r' : \bar{\boldsymbol{\beta}}_r' : \bar{\boldsymbol{\lambda}}_r' : \bar{\boldsymbol{\mu}}_r')'$ and

$$2\pi\bar{\boldsymbol{\theta}}_r'\mathbf{G}(r,\nu_0)\bar{\boldsymbol{\theta}}_r = \int_{-\pi}^{\pi}(\bar{\mathbf{a}}_r\mathbf{K}-\bar{\mathbf{b}}_r)\mathbf{H}_{xx}(\bar{\mathbf{a}}_r\mathbf{K}-\bar{\mathbf{b}}_r)^* +$$
$$(\bar{\mathbf{a}}_r\mathbf{L}-\bar{\mathbf{m}}_r)\boldsymbol{\Sigma}(\bar{\mathbf{a}}_r\mathbf{L}-\bar{\mathbf{m}}_r)^*\,d\omega$$
$$= 0.$$

Assumptions (A1) to (A3) imply that $\bar{\mathbf{a}}_r(z)\mathbf{K}(z) = \bar{\mathbf{b}}_r(z)$ and $\bar{\mathbf{a}}_r(z)\mathbf{L}(z) = \bar{\mathbf{m}}_r(z)$ a.e., $|z|=1$, and hence that $\{\mathbf{a}_{r0}(z)+\bar{\mathbf{a}}_{r(q)}(z)\}\mathbf{K}(z) = \{\mathbf{b}_{r0}(z)+\bar{\mathbf{b}}_{r(q)}(z)\}$ and $\{\mathbf{a}_{r(q)0}(a)+\bar{\mathbf{a}}_{r(q)}(z)\}\mathbf{L}(z) = \{\mathbf{m}_{r0}(z)+\bar{\mathbf{m}}_{r(q)}(z)\}$ a.e., $|z|=1$, giving a different representation of $\mathbf{K}$ and $\mathbf{L}$ within the same canonical form. By *reductio ad absurdum* it follows that $\mathbf{G}(r,\nu_0)$ is nonsingular. Since $\mathbf{G}(r,\nu_0)$ is nonsingular $\hat{\mathbf{G}}_T(r,\nu_0)$ will, with probability one, be nonsingular for T sufficiently large because the inequality $\lambda_{min}[\mathbf{A}+\mathbf{B}] \geq \lambda_{min}[\mathbf{A}] + \lambda_{min}[\mathbf{B}]$ where $\lambda_{min}[\cdot]$ denotes the smallest eigenvalue of the corresponding matrix implies that $\lambda_{min}\big[\hat{\mathbf{G}}_T(r,\nu_0)\big] \geq \lambda_{min}\big[\mathbf{G}(r,\nu_0)\big] + \lambda_{min}\big[\hat{\mathbf{G}}_T(r,\nu_0)-\mathbf{G}(r,\nu_0)\big]$. Hence from Lemma 2.2

$$\lambda_{min}\big[\hat{\mathbf{G}}_T(r,\nu_0)\big] \geq \lambda_{min}\big[\mathbf{G}(r,\nu_0)\big](1+O(Q_T))$$
$$\geq \lambda_{min}\big[\mathbf{G}(r,\nu_0)\big]/2$$

for T sufficiently large. Suppressing the arguments r and ν_0 for simplicity, from Lemma (2.2) and the inequality $\|\hat{\mathbf{G}}_T(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0)\| \leq \|\mathbf{G} - \hat{\mathbf{G}}_T\| \cdot \|\boldsymbol{\theta}_0\| + \|\hat{\mathbf{g}}_T - \mathbf{g}\|$ we can conclude that $\|\hat{\mathbf{G}}_T(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0)\| = O(Q_T)$ since $\|\boldsymbol{\theta}_0\| < \infty$ by assumption. It follows that $\hat{\boldsymbol{\theta}}_{r,T} = \boldsymbol{\theta}_{r0} + O(Q_T)$ and that $[\hat{\mathbf{a}}_{r,T} : \hat{\mathbf{b}}_{r,T} : \hat{\mathbf{m}}_{r,T}]$ converges to $[\mathbf{a}_{r0} : \mathbf{b}_{r0} : \mathbf{m}_{r0}]$ as stated.

A direct consequence of applying Lemma 2.2 and the immediately preceding result to equation (2.3) is that $\hat{\sigma}^2_{r,T}(\nu_0)$ converges a.s. to $\mathbf{e}_r' \boldsymbol{\Gamma}_{yy}(0) \mathbf{e}_r - \boldsymbol{\theta}_{r0}' \mathbf{G}(r, \nu_0) \boldsymbol{\theta}_{r0}$. Now let $\mathbf{R}_{r,\nu}(t)$ be defined as for $\hat{\mathbf{R}}_{r,\nu}(t)$ except that $\hat{\boldsymbol{\epsilon}}_T(t)$ is replaced by $\boldsymbol{\epsilon}(t)$ everywhere it occurs. Regarding $y_r(t)$ and the elements of $\mathbf{R}_{r,\nu}(t)$ as members of the Hilbert space of random variables and recalling the well known isomorphism between the time and frequency domains (Rozanov 1967) we see that $\mathbf{G}(r, \nu)$ and $\mathbf{g}(r, \nu)$ are the Grammians of these variables. Moreover $y_r(t) = f_{r0}(t) + \epsilon_r(t)$ where $f_{r0}(t) = \mathbf{R}_{r,\nu_0}(t)' \boldsymbol{\theta}_{r0}$, from which it follows that $\mathbf{G}(r, \nu_0) \boldsymbol{\theta}_{r0} = \mathbf{g}(r, \nu_0)$ and $\boldsymbol{\theta}_{r0}$ determines the projection of $y_r(t)$ on to the manifold spanned by $\mathbf{R}_{r,\nu_0}(t)$. The squared norm of the residual from that projection is then $\mathbf{e}_r' \boldsymbol{\Gamma}_{yy}(0) \mathbf{e}_r - \boldsymbol{\theta}_{r0}' \mathbf{G}(r, \nu_0) \boldsymbol{\theta}_{r0} = \sigma^2_{rr}$ and hence $\hat{\sigma}^2_{r,T}(\nu_0) = \sigma^2_{rr} + o(1)$ a.s. as required. □

7.3 Proof of Lemma 3.1 : The proof of the first part of Lemma 3.1 is modeled on that of Lemma 2.1(c) of Pőtscher(1989). To begin recall that $\hat{\sigma}^2_{r,T}(\nu)$ is the residual mean square from the regression of $y_r(t)$ on the regressors in $\hat{\mathbf{R}}_{r,\nu}(t)$. Now let $\nu_0(r, n)$ denote the multiindex $\{n_{10}, \ldots, n_{(r-1)0}, n, n_{(r+1)0}, \ldots, n_{v0}\}$, set $\mathbf{F}_{\nu_0(r,n)}(t)' = \left[f_{r0}(t) : \mathbf{R}_{r,\nu_0(r,n)}(t)'\right]$ and $\hat{\mathbf{F}}_{\nu_0(r,n)}(t) = \left[f_{r0}(t) : \hat{\mathbf{R}}_{r,\nu_0(r,n)}(t)\right]$. A similar logic to that employed in Lemma 2.3 shows that when $n < n_{r0}$ the variables in $\mathbf{F}_{\nu_0(r,n)}(t)$ are linearly independent almost surely for all t, from which we can conclude that $\liminf_{T\to\infty} \lambda_{min}\left[\sum_t \mathbf{F}_{\nu_0(r,n)}(t) \mathbf{F}_{\nu_0(r,n)}(t)'\right]/T > 0$, a.s.. Expanding $T^{-1}\sum_t (\hat{\mathbf{F}}_{\nu_0(r,n)}(t) \hat{\mathbf{F}}_{\nu_0(r,n)}(t)' - \mathbf{F}_{\nu_0(r,n)}(t) \mathbf{F}_{\nu_0(r,n)}(t)')$, in terms of $\hat{\mathbf{F}}_{\nu_0(r,n)}(t) - \mathbf{F}_{\nu_0(r,n)}(t)$ and $\mathbf{F}_{\nu_0(r,n)}(t)$, applying the Cauchy-Schwartz inequality, and using the bounds $T^{-1}\sum_t ||\hat{\mathbf{F}}_{\nu_0(r,n)}(t) - \mathbf{F}_{\nu_0(r,n)}(t)|| \leq T^{-1}\sum_t ||(\boldsymbol{\epsilon}(t) - \hat{\boldsymbol{\epsilon}}_T(t))||^2 = O(h_T Q_T^2)$ and $T^{-1}\sum_t ||\mathbf{F}_{\nu_0(r,n)}(t)||^2 = O(1)$, leads to the result $\lambda_{min}\left[\sum_t \hat{\mathbf{F}}_{\nu_0(r,n)}(t) \hat{\mathbf{F}}_{\nu_0(r,n)}(t)'\right]/T > 0$ with probability one as $T \to \infty$. Consequently, for T sufficiently large, $\hat{\mathbf{F}}_{\nu_0(r,n)}(t)$ consti-

tutes a set of linearly independent variables. From (1.6) of Lai and Wei (1982) it follows that the residual mean square from the projection of $f_{r0}(t)$ on to the space spanned by $\hat{\mathbf{R}}_{r,\nu_0(r,n)}(t)$ is bounded below by a constant $\rho_{r0}(n) > 0$. An argument similar to that employed by Pőtscher(1989, p1268) in proving his Lemma 2.1 therefore leads to the conclusion that $\hat{\sigma}^2_{r,T}\{\nu_0(r,n)\} > \rho_{r0}(n)(1+o(1)) + T^{-1}\sum_{t=1}^{T}\varepsilon_r(t)^2$. Since $\hat{\sigma}^2_{r,T}\{\nu_0\} = \sigma^2_{rr} + o(1)$ by Lemma 2.3 and $T^{-1}\sum_{t=1}^{T}\varepsilon_r(t)^2 \to \sigma^2_{rr}$ a.s. by assumption (A2) it follows that $\hat{\sigma}^2_{r,T}\{\nu_0(r,n)\} - \hat{\sigma}^2_{r,T}\{\nu_0\} > \rho_{r0}(n)(1+o(1))$ for all $n < n_{r0}$.

Now, the residual mean square from the regression of $y_r(t)$ on $\hat{\mathbf{R}}_{r,\nu(n)}(t)'$ equals $\hat{\sigma}^2_{r,T}\{\nu_0(r,n)\}$ minus the regression mean square obtained by regressing $y_r(t)$ on the component of $\hat{\mathbf{R}}_{r,\nu(n)}(t)$ orthogonal to $\hat{\mathbf{R}}_{r,\nu_0(r,n)}(t)$. If $n < n_{r(v)0}$ then $\hat{\mathbf{R}}_{r,\nu(n)}(t) = \hat{\mathbf{R}}_{r,\nu_0(r,n)}(t)$ and $\hat{\sigma}^2_{r,T}\{\nu(n)\} = \hat{\sigma}^2_{r,T}\{\nu_0(r,n)\}$ for all $r = 1,\ldots,v$. If, on the other hand, $n > n_{q0}$ for some q, $q \neq r$, then the equation $\mathbf{a}_{q0}(z)\mathbf{y}(t) + \mathbf{b}_{q0}(z)\mathbf{x}(t) = \mathbf{m}_{q0}(z)\boldsymbol{\epsilon}(t)$, the qth row of the $ARMAX_E(\nu_0)$ system defines an exact linear relationship between $\mathbf{y}(t-s)$, $\mathbf{x}(t-s)$ and $\boldsymbol{\epsilon}(t-s)$, $s = 0,\ldots,n_{j0}$, that implies that the additional variables $y_j(t-s)$, $s = 0,\ldots,n-n_{j0}$, appearing in $\mathbf{R}_{r,\nu(n)}(t)$ are linearly dependent on those already contained in $\mathbf{R}_{r,\nu_0(r,n)}(t)$. From Lemma 2.2 it follows that asymptotically $\hat{\mathbf{R}}_{r,\nu(n)}(t)$ and $\hat{\mathbf{R}}_{r,\nu_0(r,n)}(t)$ will span the same space and that $\hat{\sigma}^2_{r,T}\{\nu(n)\} = \hat{\sigma}^2_{r,T}\{\nu_0(r,n)\} + O(h_T^{1/2}Q_T)$. Hence

$$\log\left[\hat{\sigma}^2_{r,T}\{\nu(n)\}/\hat{\sigma}^2_{r,T}\{\nu(n_{r0})\}\right] = \log\left[\hat{\sigma}^2_{r,T}\{\nu_0(r,n)\}/\hat{\sigma}^2_{r,T}\{\nu_0\}\right] + o(1)$$

and consequently

$$\liminf_{T\to\infty} \log\left[\hat{\sigma}^2_{r,T}\{\nu(n)\}/\hat{\sigma}^2_{r,T}\{\nu(n_{r0})\}\right] > \log(1+\rho_{r0}(n)(1-\delta)/\sigma^2_{rr})$$

with probability one for any δ, $0 < \delta < 1$. The assumption $\kappa_T/T \to 0$ therefore implies that $\Lambda_{r,T}(\nu(n_{r0})) < \Lambda_{r,T}(\nu(n))$ a.s. if $n < n_{r0}$ because $\kappa_T/(T\log(1+\rho_{r0}(n)(1-\delta)/\sigma^2_{rr})) \to 0$ a.s. and $\liminf_{T\to\infty}\hat{n}_{r,T} \geq n_{r(q)0}$ as required.

In addition, for any n the inequality

$$\log^+(\sum_{t=1}^{T}||\hat{\mathbf{R}}_{r,\nu}(t)||^2) \leq O\big(\log^+\{\sum_{t=1}^{T}(\sum_{j=0}^{n}\|\mathbf{y}(t-j)\|^2+\|\hat{\boldsymbol{\epsilon}}_T(t-j)\|^2)+||\mathbf{x}(t-j)||^2\}\big)$$

$$= O(\log n) + O\big(\log\{O(\sum_{t=1}^{T}\|\mathbf{y}(t)\|^2+\|\hat{\boldsymbol{\epsilon}}_T(t)\|^2+||\mathbf{x}(t)||^2)\}\big)\ ,$$

where $\log^+(x)$ denotes the positive part of $\log(x)$, applies. The right hand side is $O(\log T)$ *a.s.* since $\sum_t \|\mathbf{y}(t)\|^2$ and $\sum_t ||\mathbf{x}(t)||^2$ are both $O(T)$ and $\sum_t \|\hat{\boldsymbol{\epsilon}}_T(t)\|^2 \leq \sum_t \|\boldsymbol{\epsilon}(t)\|^2+\sum_t \|(\boldsymbol{\epsilon}(t)-\hat{\boldsymbol{\epsilon}}_T(t))\|^2$ is at most $O(T)+O(H_T \log\log T)$ by Lemma 2.1. Thus sufficient conditions for the application of Lemma 2.2(b) of Pőtscher(1989) are satisfied because $\log(\sum_t ||\hat{\mathbf{R}}_{r,\nu}(t)||^2)$ is of smaller order than κ_T a.s. if $\log T/\kappa_T \to 0$ and therefore the results as stated in (ii) hold. □

7.4 Proof of Lemma 3.3 : Lemma 3.3 corresponds to Lemma 2.2 when $\nu = \nu_0$ with $\boldsymbol{\chi}_{r0}(z) = \mathbf{e}_r$. If $n_r \geq n_{r0}$, $r = 1,\ldots,v$, however, and at least one inequality is strict, $n_q > n_{q0}$ say, then the qth row of $[\mathbf{A}_0 : \mathbf{B}_0 : \mathbf{M}_0]$ defines an exact linear relationship between the elements of $\mathbf{y}(t-s)$, $\mathbf{x}(t-s)$ and $\boldsymbol{\epsilon}(t-s)$, $s = 0,\ldots,n_{q0}$, that implies that the variables in $\mathbf{R}_{r,\nu}(t)$ are linearly dependent. Therefore, see Poskitt (1992 p.17), there exists a nonzero vector that annihilates $\mathbf{G}(r,\nu)$. Thus the convergence argument used to establish Lemma 2.2 is not available as $\mathbf{G}(r,\nu)$ has less than full rank. In order to handle such singularities we employ an adaptation of a method outlined in Hannan and Deistler (1988, pp. 307-308). Since $[\mathbf{A}_0 : \mathbf{B}_0 : \mathbf{M}_0]$ is coprime it follows from the generalised Bezout identity (Kailath 1980, p.382) that $(u+v)\times v$, $(u+v)\times u$ and $(u+v)\times v$ matrix operators $\nabla\mathbf{A}_0$, $\nabla\mathbf{B}_0$ and $\nabla\mathbf{M}_0$ exist such that

$$\begin{bmatrix} \nabla\mathbf{A}_0 & \nabla\mathbf{B}_0 & \nabla\mathbf{M}_0 \\ \mathbf{A}_0 & \mathbf{B}_0 & \mathbf{M}_0 \end{bmatrix}$$

is unimodular and postmultiplying $[\mathbf{a}_r : \mathbf{b}_r : \mathbf{m}_r]$ by the inverse of this matrix yields a one-to-one coordinate transformation to a new parameter vector $(\boldsymbol{\xi}_r' : \boldsymbol{\chi}_r')$ such that

$$[\boldsymbol{\xi}_r' : \boldsymbol{\chi}_r']\begin{bmatrix} \nabla\mathbf{A}_0 & \nabla\mathbf{B}_0 & \nabla\mathbf{M}_0 \\ \mathbf{A}_0 & \mathbf{B}_0 & \mathbf{M}_0 \end{bmatrix} = [\mathbf{a}_r : \mathbf{b}_r : \mathbf{m}_r]\ .$$

From the properties of Bezout identity it follows that $\boldsymbol{\xi}_r(z)$ is strictly proper and $\mathbf{a}_r[\mathbf{A}_0^{-1}(\mathbf{B}_0 : \mathbf{M}_0)] - [-\mathbf{b}_r : \mathbf{m}_r] = \mathbf{0}$ if and only if $\boldsymbol{\xi}_r = \mathbf{0}$. In this case $[\mathbf{a}_r : \mathbf{b}_r : \mathbf{m}_r] = \boldsymbol{\chi}_r'[\mathbf{A}_0 : \mathbf{B}_0 : \mathbf{M}_0]$ and this vector must constitute the rth row of an echelon form whose Kronecker indices are $n_r,\ r = 1, \ldots, v$. That such a linear combination exists is verified by construction.

Let $\mathbf{E}_r(z)$ denote the elementary $v \times v$ row transformation matrix that simultaneously multiplies the rth row of $[\mathbf{A}_0 : \mathbf{B}_0 : \mathbf{M}_0]$ by $(1 + dz^{-1}),\ \ |d| < 1$ but otherwise arbitrary, and adds to it $-a_{rj0}(n_{r0} - n_{j0} + 1)z^{-(n_{r0}-n_{j0}+1)}$ times row j if $n_{j0} \leq n_{r0} + 1,\ \ j = 1, \ldots, v$, $j \neq r$. Then $\mathbf{E}_r[\mathbf{A}_0 : \mathbf{B}_0 : \mathbf{M}_0]$ is in echelon form with row degrees $n_{j0}\ \ j = 1, \ldots, v,\ j \neq r$, and $n_{r0}+1$. A similar transformation can then be applied to $[\mathbf{E}_r(\mathbf{A}_0 : \mathbf{B}_0 : \mathbf{M}_0)]$ to increase the degree of any other row. By repeated application of such transformations the degree of each row of $[\mathbf{A}_0 : \mathbf{B}_0 : \mathbf{M}_0]$ can be increased as required untill the resultant echelon form, $[\bar{\mathbf{A}} : \bar{\mathbf{B}} : \bar{\mathbf{M}}]$ say, has row degrees $n_r,\ r = 1, \ldots, v$. The rth row of $[\bar{\mathbf{A}} : \bar{\mathbf{B}} : \bar{\mathbf{M}}]$ equals $\boldsymbol{\chi}_r'[\mathbf{A}_0 : \mathbf{B}_0 : \mathbf{M}_0]$ where $\boldsymbol{\chi}_r'$ is the rth row of a product of elementary row transformation matrices and has elements

$$\begin{aligned} \chi_{r,j}(z) &= \sum_{s=0}^{n_r - n_{j0}} \chi_{r,j}(s) z^{-s},\ n_{j0} \leq n_r \\ &= 0 \qquad\qquad\qquad\qquad ,\ \text{otherwise} \end{aligned}$$

for $j = 1, \ldots, v,\ \ j \neq r$, and

$$\chi_{r,r}(z) = 1 + \sum_{s=1}^{n_r - n_{r0}} \chi_{r,r}(s) z^{-s}\ .$$

Thus the vector $\boldsymbol{\chi}_r$ contains polynomials of degree $n_r - n_{j0}$ in the jth location, $j \in \mathcal{K}_{r0}$, and is otherwise zero.

Transforming from $[\mathbf{a} : \mathbf{b} : \mathbf{m}]$ to $[\boldsymbol{\xi}' : \boldsymbol{\chi}']$ we find that

$$\begin{aligned} &\mathbf{a}_r(z)\mathbf{y}(t) + \mathbf{b}_r(z)\mathbf{x}(t) - \big(\mathbf{m}_r(z) - \mathbf{e}_r'\big)\hat{\boldsymbol{\epsilon}}_T(t) \\ &= \big[\boldsymbol{\chi}_r(z)'\mathbf{A}_0(z) + \boldsymbol{\xi}_r(z)'\nabla\mathbf{A}_0(z)\big]\mathbf{y}(t) + \big[\boldsymbol{\chi}_r(z)'\mathbf{B}_0(z) + \boldsymbol{\xi}_r(z)'\nabla\mathbf{B}_0(z)\big]\mathbf{x}(t) \\ &\qquad - \big[\boldsymbol{\chi}_r(z)'\mathbf{M}_0(z) + \boldsymbol{\xi}_r(z)'\nabla\mathbf{M}_0(z) - \mathbf{e}_r'\big]\hat{\boldsymbol{\epsilon}}_T(t) \\ &= \mathbf{e}_r'\{\boldsymbol{\epsilon}(t) - (\mathbf{M}_0(z) - \mathbf{I})(\hat{\boldsymbol{\epsilon}}_T(t) - \boldsymbol{\epsilon}(t))\} - (\boldsymbol{\chi}_r(z) - \mathbf{e}_r)'\mathbf{M}_0(z)(\hat{\boldsymbol{\epsilon}}_T(t) - \boldsymbol{\epsilon}(t)) \\ &\qquad + \boldsymbol{\xi}_r(z)'\big[\nabla\mathbf{A}_0(z)\mathbf{y}(t) + \nabla\mathbf{B}_0(z)\mathbf{x}(t) - \nabla\mathbf{M}_0(z)\hat{\epsilon}_T(t)\big] \end{aligned}$$

where the last line follows from its predecessor by adding and subtracting $(\boldsymbol{\chi}_r'\mathbf{M}_0(z) - \mathbf{e}_r')\boldsymbol{\epsilon}(t)$ and then rearranging and simplifying the expression using the fact that $\mathbf{A}_0(z)\mathbf{y}(t) + \mathbf{B}_0(z)\mathbf{x}(t) - \mathbf{M}_0(z)\boldsymbol{\epsilon}(t) = \mathbf{0}$. It follows that $\hat{\sigma}^2_{r,T}(\nu)$ equals the residual mean square from the regression of $\hat{w}_r(t) = \mathbf{e}_r'\{\boldsymbol{\epsilon}(t) - (\mathbf{M}_0(z) - \mathbf{I})(\hat{\boldsymbol{\epsilon}}_T(t) - \boldsymbol{\epsilon}(t))\}$ on appropriate lagged values of $\mathbf{M}_0(z)(\hat{\boldsymbol{\epsilon}}_T(t) - \boldsymbol{\epsilon}(t))$ and $\mathbf{z}(t) = \left[\nabla\mathbf{A}_0(z)\mathbf{y}(t) + \nabla\mathbf{B}_0(z)\mathbf{x}(t) - \nabla\mathbf{M}_0(z)\hat{\epsilon}_T(t)\right]$ with coefficients corresponding to the elements of $[\boldsymbol{\xi}_r(z)' : \boldsymbol{\chi}_r(z)']$ not known to be either zero or one. Argument by contradiction now shows that the least squares coefficient values must satisfy $[\hat{\mathbf{a}}_{r,T} : \hat{\mathbf{b}}_{r,T} : \hat{\mathbf{m}}_{r,T}] = \hat{\boldsymbol{\chi}}_{r,T}'[\mathbf{A}_0 : \mathbf{B}_0 : \mathbf{M}_0] + \hat{\boldsymbol{\xi}}_{r,T}'[\nabla\mathbf{A}_0 : \nabla\mathbf{B}_0 : \nabla\mathbf{M}_0]$.

We will now establish that $\hat{\boldsymbol{\xi}}_{r,T} = O(Q_T)$. To this end note that, by definition,

$$\mathbf{z}(t) = \sum_{j=0}^{h_T+h_0} \bar{\boldsymbol{\Psi}}(j)\mathbf{y}(t-j) - \bar{\boldsymbol{\Phi}}(j)\mathbf{x}(t-j)$$

for some fixed integer $h_0 > 0$ where $[\bar{\boldsymbol{\Psi}} : \bar{\boldsymbol{\Phi}}] = [\nabla\mathbf{A}_0 - \nabla\mathbf{M}_0\hat{\boldsymbol{\Psi}}_T : -\nabla\mathbf{B}_0 + \nabla\mathbf{M}_0\hat{\boldsymbol{\Phi}}_T]$ and from Hannan and Deistler (1988, Theorem 6.6.10) it follows that the matrix of mean squares and cross products with $T^{-1}\sum\mathbf{z}(t-s)\mathbf{z}(t-r)'$ in the $(s,r)th$ block, $s, r = 1, \ldots, \delta(\boldsymbol{\xi}_r)$, converges to a nonsingular limit. From Theorem 5.3.1 of Hannan and Deistler (1988) we also conclude that $T^{-1}\sum\mathbf{z}(t-s)\boldsymbol{\epsilon}(t)' = O(Q_T),\ \ s = 1, \ldots, \delta(\boldsymbol{\xi}_r)$. Similarly $T^{-1}\sum_t\mathbf{y}(t-s)\{\hat{\boldsymbol{\epsilon}}_T(t-r) - \boldsymbol{\epsilon}(t-r)\}' = O(Q_T)$ for any s and r. That this equality obtains can be seen by substituting for $\hat{\boldsymbol{\epsilon}}_T(t) - \boldsymbol{\epsilon}(t)$ and rewriting the transpose of the quantity of interest as

$$T^{-1}\sum_t\Big\{\sum_{j=1}^{h_T}\left[\hat{\boldsymbol{\Psi}}_T(j) - \boldsymbol{\Psi}_0(j)\right]\mathbf{y}(t-r-j)\mathbf{y}(t-s)' - \left[\hat{\boldsymbol{\Phi}}_T(j) - \boldsymbol{\Phi}_0(j)\right]\mathbf{x}(t-r-j)\mathbf{y}(t-s)'\Big\}$$

plus a remainder term that is $o(T^{-1/2})$, this order of magnitude arising from the fact that $\sum_{j>h_T}\|\boldsymbol{\Psi}_0(j) : \boldsymbol{\Phi}_0(j)\|$ is dominated by $const \cdot |z_0|^{h_T},\ \ h_T > h_{0T}$ and $\sum\|\boldsymbol{\Gamma}_{yy}(k) : \boldsymbol{\Gamma}_{xy}(k)\| < \infty$. Since $T^{-1}\sum[\mathbf{y}(t-s) : \mathbf{x}(t-s)]\mathbf{y}(t-r)' = [\boldsymbol{\Gamma}_{yy}(s-r) : \boldsymbol{\Gamma}_{xy}(s-r)] + O(Q_T)$ and $sup_{1\le j\le h_T}\|\hat{\boldsymbol{\Psi}}_T(j) - \boldsymbol{\Psi}_0(j) : \hat{\boldsymbol{\Phi}}_T(j) - \boldsymbol{\Phi}_0(j)\| = O(Q_T)$, see Hannan and Deistler (1988, Corollary 6.6.2), the stated result follows. The same argument can now be used to show that $T^{-1}\sum_t\mathbf{x}(t-s)\{\hat{\boldsymbol{\epsilon}}_T(t-r) - \boldsymbol{\epsilon}(t-r)\}' = O(Q_T)$ by simply replacing $\mathbf{y}(t-s)$ by $\mathbf{x}(t-s)$.

This implies not only that $T^{-1}\sum_t \mathbf{z}(t-s)w_r(t) = O(Q_T)$ but also that the sample cross-covariance matrix between the regressors in $\mathbf{z}(t-s)$ and $\mathbf{M}_0(z)(\hat{\boldsymbol{\epsilon}}_T(t-r) - \boldsymbol{\epsilon}(t-r))$ is $O(Q_T)$ and thus we can conclude that $\hat{\boldsymbol{\xi}}_{r,T} = O(Q_T)$, as claimed.

Finally, using the immediately preceding results we see that the coefficients in $\hat{\boldsymbol{\chi}}_{r,T}$ are obtained from the regression of $\hat{w}_r(t)$ on $\mathbf{e}_j'\mathbf{M}_0(z)(\hat{\boldsymbol{\epsilon}}_T(t-s)-\boldsymbol{\epsilon}(t-s))$, $s = 0,\ldots,n_r - n_{j0},\ j \in \mathcal{K}_{r0}$, plus a correction term $O(Q_T)$. From Lemma 2.1, however,

$$T^{-1}\sum_{t=1}^{T}\mathbf{M}_0(z)(\hat{\boldsymbol{\epsilon}}_T(t-s)-\boldsymbol{\epsilon}(t-s))(\hat{\boldsymbol{\epsilon}}_T(t-r)-\boldsymbol{\epsilon}(t-r))'\mathbf{M}_0(z)' = (u+v)h_T(2\pi T)^{-1}\int_{-\pi}^{\pi}\mathbf{M}_0\boldsymbol{\Sigma}\mathbf{M}_0^* e^{i\omega(s-r)}d\omega + o_p(h_T T^{-1})$$

and

$$T^{-1}\sum_{t=1}^{T}\mathbf{M}_0(z)(\hat{\boldsymbol{\epsilon}}_T(t-s)-\boldsymbol{\epsilon}(t-s))\hat{w}_r(t) = (u+v)h_T(2\pi T)^{-1}\int_{-\pi}^{\pi}\mathbf{M}_0\Sigma\mathbf{M}_0^*\mathbf{e}_r e^{i\omega s}d\omega + o_p(h_T T^{-1})\ .$$

Neglecting terms that are $o_p(1)$ it is readily verified that $\hat{\boldsymbol{\chi}}_{r,T}$ corresponds to the solution of the Toeplitz equations associated with the minimum mean squared error prediction of $w_r(t)$ from $w_j(t-s)$, $s = 0,\ldots,n_r - n_{j0},\ j \in \mathcal{K}_{r0}$. This completes the proof. □

7.5 Proof of Theorem 3.4 : Recall from the above that when $n \geq n_{r0}$ $\hat{\sigma}^2_{r,T}(\nu(n))$ equals the residual mean square from the regression of $\hat{w}_r(t) = \mathbf{e}_r'\{\boldsymbol{\epsilon}(t) - (\mathbf{M}_0(z)-\mathbf{I})(\hat{\boldsymbol{\epsilon}}_T(t)-\boldsymbol{\epsilon}(t))\}$ on $\mathbf{e}_j'\mathbf{M}_0(z)(\hat{\boldsymbol{\epsilon}}_T(t-s)-\boldsymbol{\epsilon}(t-s))$ $s = 0,\ldots,n_r - n_{j0},\ j \in \mathcal{K}_{r0}$, plus $O(Q_T^2)$. Using Lemma 2.1 once more we find that

$$T^{-1}\sum_{t=1}^{T}\hat{w}_r(t)^2 = T^{-1}\sum_{t=1}^{T}\epsilon_r(t)^2 + (u+v)h_T(2\pi T)^{-1}\mathbf{e}_r'\big[\int_{-\pi}^{\pi}(\mathbf{M}_0\Sigma_0\mathbf{M}_0^* - \boldsymbol{\Sigma}_0)d\omega\big]\mathbf{e}_r$$

plus $o_p(h_T T^{-1})$. Subtracting the regression mean square, which has already been shown to equal that from the optimal prediction of $w_r(t)$ from $w_j(t-s)$, $s = 0,\ldots,n_r - n_{j0},\ j \in \mathcal{K}_{r0}$, yields the result that

$$\hat{\sigma}^2_{r,T}(\nu(n)) = T^{-1}\sum_{t=1}^{T}\epsilon_r(t)^2 + (u+v)h_T T^{-1}(\tau_r^2(\ell) - \sigma_{rr}^2) + o_p(1)\ ,$$

where $\ell = n - n_{r0}$. Substituting into $\Lambda_{r,T}(\nu(n))$, replacing $T^{-1}\sum_{t=1}^{T}\epsilon_r(t)^2$ by $\sigma_{rr}^2 + o(1)$ and rearranging terms gives

$$\Lambda_{r,T}(\nu(n)) = \log\sigma_{rr}^2 + (u+v)h_T T^{-1}(\tau_r^2(\ell)/\sigma_{rr}^2 - 1) + \big[(v-1) + n(2v+u)\big]\kappa_T T^{-1} + o_p(1)\ ,$$

from which the result presented in the theorem follows directly. □

7.6 Proof of Lemma 5.1: To prove Lemma 5.1 let $[\tilde{\Delta}\boldsymbol{\Psi}_{0T} : \tilde{\Delta}\boldsymbol{\Phi}_{0T}] = [\tilde{\boldsymbol{\Psi}}_T : \tilde{\boldsymbol{\Phi}}_T] - [\boldsymbol{\Psi}_0 : \boldsymbol{\Phi}_0]$ where the coefficients of $[\tilde{\boldsymbol{\Psi}}_T : \tilde{\boldsymbol{\Phi}}_T]$ are generated from the recursions

$$\sum_{j=0}^{i}\tilde{\mathbf{M}}_T(j)[\tilde{\boldsymbol{\Psi}}_T(i-j) : \tilde{\boldsymbol{\Phi}}_T(i-j)] = [\tilde{\mathbf{A}}_T(i) : \tilde{\mathbf{B}}_T(i)]\ ,\ i = 1,\ldots,\hat{p}_T,$$
$$= \mathbf{0}\ , \qquad i > \hat{p}_T,$$

with initial conditions $[\tilde{\boldsymbol{\Psi}}_T(0) : \tilde{\boldsymbol{\Phi}}_T(0)] = [\mathbf{I}_v : \mathbf{0}]$. Expanding $\tilde{\boldsymbol{\epsilon}}_T(t-s) - \boldsymbol{\epsilon}(t-s)$ in terms of $[\tilde{\Delta}\boldsymbol{\Psi}_{0T} : \tilde{\Delta}\boldsymbol{\Phi}_{0T}]$ and substituting into $T^{-1}\sum_{t=s+1}^{T}(\tilde{\boldsymbol{\epsilon}}_T(t-s) - \boldsymbol{\epsilon}(t-s))\boldsymbol{\epsilon}(t)'$, for example, we obtain

$$T^{-1}\sum_{t=s+1}^{T}\Big\{\sum_{j=1}^{H_T}[\tilde{\Delta}\boldsymbol{\Psi}_{0T}(j)\mathbf{y}(t-s-j) : \tilde{\Delta}\boldsymbol{\Phi}_{0T}(j)\mathbf{x}(t-s-j)]\boldsymbol{\epsilon}(t)'\Big\}$$
$$+ T^{-1}\sum_{t=s+1}^{T}\Big\{\sum_{j=H_T+1}^{t-s+H_T-1}[\tilde{\Delta}\boldsymbol{\Psi}_{0T}(j)\mathbf{y}(t-s-j) : \tilde{\Delta}\boldsymbol{\Phi}_{0T}(j)\mathbf{x}(t-s-j)]\boldsymbol{\epsilon}(t)'\Big\}$$
$$- T^{-1}\sum_{t=s+1}^{T}\Big\{\sum_{j=t-s+H_T}^{\infty}[\boldsymbol{\Psi}_0(j)\mathbf{y}(t-s-j) : \boldsymbol{\Phi}_0(j)\mathbf{x}(t-s-j)]\boldsymbol{\epsilon}(t)'\Big\}\ .$$

The first term is $O(Q_T^2)$ because $T^{-1}\sum_{t=s+1}^{T}\boldsymbol{\epsilon}(t)[\mathbf{y}(t-s)' : \mathbf{x}(t-s)'] = O(Q_T),\ 0 < s \leq (\log T)^c,\ c > 1$, by Theorem 5.3.1 of Hannan and Deistler (1988) and, as will be shown below, $\|\tilde{\Delta}\boldsymbol{\Psi}_{0T} : \tilde{\Delta}\boldsymbol{\Phi}_{0T}\| = O(Q_T)$. The norm of the second term is bounded by

$$T^{-1}\sum_{t=s+1}^{T}\|\boldsymbol{\epsilon}(t)\|\sum_{j=H_T+1}^{t-s+H_T-1}\|\tilde{\Delta}\boldsymbol{\Psi}_{0T}(j)\mathbf{y}(t-s-j) : \tilde{\Delta}\boldsymbol{\Phi}_{0T}(j)\mathbf{x}(t-s-j)\|\ ,$$

which is less than

$$T^{1/4}o(1)\sum_{j=H_T+1}^{T+H_T-1}\|\tilde{\Delta}\boldsymbol{\Psi}_{0T}(j) : \tilde{\Delta}\boldsymbol{\Phi}_{0T}(j)\|T^{-1}\sum_{t=s+1}^{T}\|\mathbf{y}(t-s-j)' : \mathbf{x}(t-s-j)'\|$$

since $\|\boldsymbol{\epsilon}(t)\| = o(t^{1/4})$ by finiteness of the fourth moment. By stationarity and ergodicity $T^{-1}\sum_{t=s+1}^{T}\|\mathbf{y}(t-s-j)' : \mathbf{x}(t-s-j)'\|$ converges to $\mathrm{E}\left[\|\mathbf{y}(t)' : \mathbf{x}(t)'\|\right] < \infty$ and the second term is therefore $o(T^{1/4}Q_T)$. Similarly, the third term is of smaller order than $T^{-1/2}$ since $\sum_{j>H_T}\|\boldsymbol{\Psi}_0(j) : \boldsymbol{\Phi}_0(j)\| = o(T^{-1/2})$. Hence $T^{-1}\sum_{t=s+1}^{T}\boldsymbol{\epsilon}(t)(\tilde{\boldsymbol{\epsilon}}_T(t-s)-\boldsymbol{\epsilon}(t-s))' = O(Q_T^2)$ as required. The first part of Lemma 5.1 is established in a similar manner.

To show that $\limsup_{T\to\infty}\|\tilde{\Delta}\boldsymbol{\Psi}_{0T} : \tilde{\Delta}\boldsymbol{\Phi}_{0T}\| = O(Q_T)$ suppose that this statement is false. Then there exists a subsequence, still denoted by T, and a constant $C < \infty$ such that $\|\tilde{\Delta}\boldsymbol{\Psi}_{0T} : \tilde{\Delta}\boldsymbol{\Phi}_{0T}\| > CQ_T + \delta,\ \delta > 0$. But from Lemma 3.1 $\hat{n}_{r,T} \geq n_{r0},\ r = 1,\ldots,v$, a.s. as $T \to \infty$ and, therefore, by Lemma 3.3 $[\tilde{\mathbf{A}}_T : \tilde{\mathbf{B}}_T : \tilde{\mathbf{M}}_T] = \hat{\mathbf{X}}_T[\mathbf{A}_0 : \mathbf{B}_0 : \mathbf{M}_0] + O(Q_T)$ and

$$\tilde{\mathbf{M}}_T[\tilde{\boldsymbol{\Psi}}_T : \tilde{\boldsymbol{\Phi}}_T] - \hat{\mathbf{X}}_T\mathbf{M}_0[\boldsymbol{\Psi}_0 : \boldsymbol{\Phi}_0] = [\tilde{\mathbf{A}}_T : \tilde{\mathbf{B}}_T] - \hat{\mathbf{X}}_T[\mathbf{A}_0 : \mathbf{B}_0] = O(Q_T)\ .$$

Also, by the same lemma, $\hat{\mathbf{X}}_T' = [\hat{\boldsymbol{\chi}}_{1,T},\ldots,\hat{\boldsymbol{\chi}}_{v,T}]$ converges in probability to the matrix $\mathbf{X}_0' = [\boldsymbol{\chi}_{10},\ldots,\boldsymbol{\chi}_{v0}]$ of minimum mean squared error filter coefficients. $\mathbf{X}_0(z)$ is non-singular, $|z| \geq 1$, because otherwise there would exist a linear combination of the stationary processes $\boldsymbol{\chi}_{r0}'(z)\mathbf{w}(t),\ r = 1,\ldots,v$ that would equal zero $\forall\ t$ a.s., implying that $\mathbf{w}(t) = \mathbf{M}_0(z)\boldsymbol{\epsilon}(t)$ has less than maximal rank, contradicting assumptions (A1) and (A4), see Rozanov(1967, §2.6 & 2.9). Thus there exists a sub-subsequence such that $\hat{\mathbf{X}}_T$ is non-singular a.s. and hence $\tilde{\mathbf{M}}_T^{-1} = \mathbf{M}_0^{-1}\hat{\mathbf{X}}_T^{-1} + O(Q_T)$. From this we conclude that there exists a $C' < \infty$ such that $\|\tilde{\Delta}\boldsymbol{\Psi}_{0T} : \tilde{\Delta}\boldsymbol{\Phi}_{0T}\| \leq C'Q_T = CQ_T + (C'-C)Q_T$ and we have arrived at a contradiction to $\|\tilde{\Delta}\boldsymbol{\Psi}_{0T} : \tilde{\Delta}\boldsymbol{\Phi}_{0T}\| > CQ_T + \delta$. □

7.6 Proof of Theorem 5.2: To prove Theorem 5.2 note that

$$\tilde{\hat{\sigma}}_{r,T}(\nu) \geq \min_{\boldsymbol{\theta}_r} T^{-1}\sum_{t=1}^{T}\left(y_r(t) - \tilde{\mathbf{R}}_{r,\nu}(t)'\boldsymbol{\theta}_r\right)^2$$

and, as will be shown below,

$$\begin{aligned}\tilde{\hat{\sigma}}_{r,T}(\nu_0) &= T^{-1}\sum_{t=1}^{T}\left(y_r(t) - f_{r0}(t) + (\mathbf{R}_{r,\nu_0}(t) - \tilde{\mathbf{R}}_{r,\nu_0}(t))'\hat{\boldsymbol{\theta}}_{rT} + \mathbf{R}_{r,\nu_0}(t)'(\boldsymbol{\theta}_{r0} - \hat{\boldsymbol{\theta}}_{r,T})\right)^2 \\ &= T^{-1}\sum_{t=1}^{T}\varepsilon_r(t)^2 + O(Q_T^2)\ ,\end{aligned}$$

the last line arising directly from Lemmas 2.4 and 5.1. Following the proof of Lemma 3.3(i) substituting $\tilde{\mathbf{R}}_{r,\nu(n)}(t)$ for $\hat{\mathbf{R}}_{r,\nu(n)}(t)$ therefore leads to the conclusion that

$$\liminf_{T\to\infty} \log\left[\tilde{\hat{\sigma}}^2_{r,T}\{\nu(n)\}/\tilde{\hat{\sigma}}^2_{r,T}\{\nu(n_{r0})\}\right] > \log(1+\rho_{r0}(n)(1-\delta)/\sigma^2_{rr})$$

with probability one for any $\delta,\ 0<\delta<1$, if $n<n_{r0}$ and hence that $\liminf_{T\to\infty}\tilde{n}_{r,T}\geq n_{r0}$ a.s. if $\tilde{\kappa}_T/T\to 0$.

When $n\geq n_{r0}$ we can expand $\tilde{\hat{\sigma}}_{r,T}(\nu(n))$ by replacing $\mathbf{y}(t)$ by $\mathbf{K}(z)\mathbf{x}(t)+\mathbf{L}(z)\boldsymbol{\epsilon}(t)$ and rearranging terms to give

$$\begin{aligned}\tilde{\hat{\sigma}}^2_{r,T}\{\nu(n)\} &= T^{-1}\sum_{t=1}^{T}\left(\hat{\mathbf{a}}_{r,T}(z)\mathbf{y}(t)+\hat{\mathbf{b}}_{r,T}(z)\mathbf{x}(t)-(\hat{\mathbf{m}}_{r,T}(z)-\mathbf{e}_r')\tilde{\boldsymbol{\epsilon}}_T(t)\right)^2\\ &= T^{-1}\sum_{t=1}^{T}\left(\mathbf{e}_r'\boldsymbol{\epsilon}(t)-(\hat{\mathbf{m}}_{r,T}(z)-\mathbf{e}_{r(q)}')\{\tilde{\boldsymbol{\epsilon}}_T(t)-\boldsymbol{\epsilon}(t)\}\right.\\ &\qquad \left.+(\hat{\mathbf{a}}_{r,T}(z)\mathbf{K}(z)+\hat{\mathbf{b}}_{r,T}(z))\mathbf{x}(t)+(\hat{\mathbf{a}}_{r,T}(z)\mathbf{L}(z)-\hat{\mathbf{m}}_{r,T}(z))\boldsymbol{\epsilon}(t)\right)^2.\end{aligned}$$

But both

$$\hat{\mathbf{a}}_{r,T}\mathbf{K}+\hat{\mathbf{b}}_{r,T}=(\hat{\mathbf{a}}_{r,T}-\hat{\boldsymbol{\chi}}_{r,T}'\mathbf{A}_0)\mathbf{K}+(\hat{\mathbf{b}}_{r,T}-\hat{\boldsymbol{\chi}}_{r,T}'\mathbf{B}_0)$$

and

$$\hat{\mathbf{a}}_{r,T}\mathbf{L}-\hat{\mathbf{m}}_{r,T}=(\hat{\mathbf{a}}_{r,T}-\hat{\boldsymbol{\chi}}_{r,T}'\mathbf{A}_0)\mathbf{L}-(\hat{\mathbf{m}}_{r,T}-\hat{\boldsymbol{\chi}}_{r,T}'\mathbf{M}_0)$$

are of order $O(Q_T)$ by Lemma 3.3 and both are strictly proper. Truncating $\hat{\mathbf{a}}_{r,T}\mathbf{K}+\hat{\mathbf{b}}_{r,T}$ and $\hat{\mathbf{a}}_{r,T}\mathbf{L}-\hat{\mathbf{m}}_{r,T}$ after H_T terms and using Theorem 5.3.1 of Hannan and Deistler (1988) once more in conjunction with Lemma 5.1 now yields the result that

$$\tilde{\hat{\sigma}}^2_{r,T}\{\nu(n)\}=T^{-1}\sum_{t=1}^{T}\epsilon_r(t)^2+O(Q_T^2)\ .$$

Therefore, the probability that $\tilde{\Lambda}_{r,T}\{\nu(n)\}-\tilde{\Lambda}_{r,T}\{\nu(n_{r0})\}$ will be negative or zero for $T>T'$ will be arbitrarily small as $T'\to\infty$ if $\tilde{\kappa}_T/\log\log T\to\infty$ and this completes the proof of the theorem. □

REFERENCES

Akaike, H. (1974) A new look at statistical model identification, *IEEE Transactions on Automatic Control*, AC-19, 716–723.

Akaike, H. (1976) Canonical correlation analysis of time series and the use of an information criterion. In *System Identification: Advances and Case Studies*, ed. R.K. Mehra and D.G. Lainiotis, Academic Press, New York. 27–96.

Bauer, D., Deistler, M. and Scherrer, W. (1999) Consistency and asymptotic normality of some subspace algorithms for systems without observed inputs, *Automatica*, 35, 1243-1254.

Durbin, J. (1960) The Fitting of Time Series Models. *International Statistical Review*, 28, 233-244.

Guidorzi, R. (1981) Invariants and canonical forms for systems structural and parameteric identification. *Automatica*, 17, 117-133.

Hannan, E.J. and Deistler, M. (1988) *The Statistical Theory of Linear Systems.* John Wiley, New York.

Hannan, E.J. and Kavalieris, L. (1984a) Multivariate linear time series models, *Advances in Applied Probability*, 16, 492–561.

Hannan, E. J. and Kavalieris, L. (1984b) A method for autoregressive-moving average estimation, *Biometrika*, 72, 273-280.

Kailath, T. (1980) *Linear Systems.* Prentice-Hall, Englewood Cliffs.

Kavalieris, L. (1991) A Note on Estimating Autoregressive-Moving Average Order. *Biometrika*, 78, 920-922.

Lai, T. L. and Wei, C. Z. (1982) Asymptotic Properties of Projections with Applications to Stochastic Regression Problems. *Journal of Multivariate Analysis*, 12, 346-370.

Lütkepohl, H. (1991) *Introduction to Multiple Time Series Analysis.* Springer-Verlag, Berlin.

Lütkepohl, H. and Poskitt, D.S. (1995a) Specification of Echelon Form VARMA Models. *Journal of Business and Economic Statistics*, 14, 69-79.

Nsiri, S. and Roy, R. (1992) On the identification of ARMA echelon form models, *Canadian Journal of Statistics*, 20, 369-386.

Peternell, K., Scherrer, W. and Deistler, M. (1996) Statistical analysis of noval subspace identification methods, *Signal Processing*, 52, 161-177.

Poskitt, D.S. (1992) Identification of echelon canonical forms for vector linear processes using least squares, *Annals of Statistics*, 20, 195–215.

Poskitt, D. S. and Salau, M. O. (1994) On the asymptotic relative efficiency of Gaussian and least squares estimators for vector ARMA models, *Journal of Multivariate Analysis*, 51, 294-317.

Poskitt, D.S. and Chung, S.H. (1996) Markov Chain Models, Time Series Analysis and Extreme Value Theory . *Advances in Applied Probability*, 28, 405-425.

Poskitt, D.S. and Tremayne, A.R. (1986) Some aspects of the performance of diagnostic checks in bivariate time series models, *Journal of Time Series Analysis*, 7, 217–233.

Pötscher, B. M. (1989), Model selection under nonstationarity: Autoregressive models and stochastic linear regression models. *Annals of Statistics*, 17, 1257-1274.

Reinsel, G. C. (1993) *Elements of Multivariate Time Series Analysis*. Springer-Verlag, New York.

Rozanov, Yu A. (1967) *Stationary Random Processes.* Holden-Day, San Francisco.

Schwarz, G. (1978) Estimating the dimension of a model, *Annals of Statistics*, 6, 461–464.

Tiao, G.C. and Tsay, R.S. (1989) Model specification in multivariate time series, *Journal of the Royal Statistical Society*, B-51, 157–213.

Tsay, R.S. (1991) Two canonical forms for vector ARMA processes, *Statistica Sinica*, 1, 247-269.

Van Overschee, P. and De Moor, B. (1994) N4SID: Subspace algorithms for the identification of combined deterministic-stochastic systems. *Automatica*, 30, 75-93.

Van Overschee, P. and De Moor, B. (1996) *Subspace Identification for Linear Systems: Theory, Implementation, Application*. Kluwer, Dordrecht.