



Suggested Running Head:

Partially Nonstationary ARMAX Systems

Address for correspondence:

D. S. Poskitt

Department of Econometrics and Business Statistics

Monash University

Victoria 3800

Australia

### **Abstract**

This paper extends current theory on the identification and estimation of vector time series models to nonstationary processes. It examines the structure of dynamic simultaneous equations systems or ARMAX processes that start from a given set of initial conditions and evolve over a given, possibly infinite, future time horizon. The analysis proceeds by deriving the echelon canonical form for such processes. The results are obtained by amalgamating ideas from the theory of stochastic difference equations with adaptations of the Kronecker index theory of dynamic systems. An extension of these results to the analysis of unit-root, partially nonstationary (cointegrated) time series models is also presented, leading to straightforward identification conditions for the error correction, echelon canonical form. An innovations algorithm for the evaluation of the exact Gaussian likelihood is given and the asymptotic properties of the approximate Gaussian estimator and the exact maximum likelihood estimator based upon the algorithm are derived. Examples illustrating the theory are discussed and some experimental evidence is also presented. ( JEL CLASSIFICATION C10, C32 )

## 1 Introduction

The concept of cointegration, due to Granger (1981), has proven to be an extremely useful tool in the analysis of many economic and financial time series and it has given rise to an extensive literature. Lucid surveys of this development can be found in Banerjee, Dolado, Galbraith, and Hendry (1993) and Hatanaka (1996). Following the seminal papers by Engle and Granger (1987) and Johansen (1991), much of the empirical and theoretical work on cointegration has been conducted in the context of vector autoregressive, *AR*, processes but more recently interest has been shown in extending the ideas to more general models. Yap and Reinsel (1995) and Lütkepohl and Claessen (1997), for example, consider estimation problems associated with cointegrated autoregressive moving-average, *ARMA*, processes, as does Dhrymes (1998). In such treatise it is commonly supposed that results on the identification of stationary vector time series models associated with the work of E. J. Hannan and M. Deistler (Hannan (1974, 1976), Deistler (1983, 1985) and Hannan and Deistler (1988)), can be readily extended to the analysis of unit-root nonstationary and partially nonstationary (cointegrated) processes without modification. As the results presented below will show, such an assumption significantly understates the theoretical issues associated with such an extension and is only partially correct.

This paper will provide a detailed discussion of the structure of nonstationary dynamic simultaneous equations systems of the form

$$\mathbf{A}(\mathcal{L})\mathbf{y}_t + \mathbf{B}(\mathcal{L})\mathbf{x}_t = \boldsymbol{\xi}_t, \quad t = 1, \dots, T. \quad (1.1)$$

In equation (1.1) the vector  $\mathbf{y}_t = (y_{1t}, \dots, y_{vt})'$  denotes a  $v$  component observable output process and  $\mathbf{x}_t = (x_{1t}, \dots, x_{ut})'$ , if present, is a  $u$  component vector of observable exogenous input variables. The  $v \times v$  and  $v \times u$  matrix operators  $\mathbf{A}(z) = \mathbf{A}_0 + \mathbf{A}_1 z^1 + \dots + \mathbf{A}_p z^p$  and  $\mathbf{B}(z) = \mathbf{B}_0 + \mathbf{B}_1 z^1 + \dots + \mathbf{B}_p z^p$  in the unit-delay or lag operator  $\mathcal{L}$ , *viz.*  $\mathcal{L}\mathbf{y}_t = \mathbf{y}_{t-1}$ , determine the basic evolutionary properties of  $\mathbf{y}_t$  and the stochastic disturbance,  $\boldsymbol{\xi}_t = (\xi_{1t}, \dots, \xi_{vt})'$ , which is unobserved, determines how chance or random influences enter the system. The endogenous process  $\mathbf{y}_t$  is assumed to evolve over the time period  $t = 1, \dots, T$ , according to the specification given in (1.1) starting from initial values given by  $\mathbf{y}_t$  and  $\mathbf{x}_t$  for  $t = 1 - p, \dots, 0$ .

With economic and financial phenomena it will rarely if ever be appropriate to think of the process as having evolved unchanged from the infinite past. Conditioning on initial values, which is what the current paradigm implies and which corresponds to common current practice in the analysis of nonstationary time series, is therefore only natural. Thus we are faced with the task of analysing a discrete time, time invariant and causal dynamic system where time, following a finite sequence of initial values, is explicitly confined to the positive integers.

It will be assumed that  $\boldsymbol{\xi}_t$  is a full rank, zero mean, stationary process with covariance  $E[\boldsymbol{\xi}_t \boldsymbol{\xi}_{t+\tau}'] = \boldsymbol{\Gamma}_\xi(\tau) = \boldsymbol{\Gamma}_\xi(-\tau)'$ ,  $\tau = 0, \pm 1, \pm 2, \dots$ ,  $\boldsymbol{\Gamma}_\xi(\tau) = \mathbf{0}$  for  $|\tau| > p$ . It is well known (see Hannan, 1971, Theorem 10' and the associated discussion) that this implies the existence of a sequence of zero mean, uncorrelated random variates  $\boldsymbol{\varepsilon}_t$ ,  $t = 1 - p, \dots, 0, 1, \dots, T$ , defined on the

same probability space as  $\boldsymbol{\xi}_t$  such that  $\boldsymbol{\xi}_t = \mathbf{M}(\mathcal{L})\boldsymbol{\varepsilon}_t$ ,  $t = 1, \dots, T$ , where  $E[\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t'] = \boldsymbol{\Sigma}_\varepsilon$  is positive definite and, without loss of generality, the  $v \times v$  matrix operator  $\mathbf{M}(z) = \mathbf{M}_0 + \mathbf{M}_1z^1 + \dots + \mathbf{M}_pz^p$  satisfies  $\det(\mathbf{M}(z)) \neq 0$ ,  $|z| < 1$ .<sup>1</sup> Expressed in the form

$$\mathbf{A}(\mathcal{L})\mathbf{y}_t + \mathbf{B}(\mathcal{L})\mathbf{x}_t = \mathbf{M}(\mathcal{L})\boldsymbol{\varepsilon}_t, \quad t = 1, \dots, T, \quad (1.2)$$

equation (1.1) gives us an autoregressive moving-average model with exogenous variables, commonly referred to as an *ARMAX* system.

An *ARMAX* process of the type described above clearly violates the standard conditions for  $\mathbf{y}_t$  to be stationary. More importantly, we wish to explicitly allow for unit root, partially nonstationary behaviour and hence we will assume that  $\det \mathbf{A}(z)$  has  $\zeta \leq v$  roots of unity, all other zeroes of  $\det \mathbf{A}(z)$  lie outside the unit circle, and that the individual series  $y_{i,t}$ ,  $i = 1, \dots, s$ , are asymptotically-stationary after first differencing, i.e.,  $\Delta \mathbf{y}_t = (1 - \mathcal{L})\mathbf{y}_t = \mathbf{y}_t - \mathbf{y}_{t-1}$ ,  $t = 1, \dots, T$ , is  $I(0)$ .<sup>2</sup> The process  $\mathbf{y}_t$  is said to be integrated of order one,  $I(1)$ , although the possibility of individual, but not all, elements  $y_{i,t}$  being  $I(0)$  without differencing is not excluded. If  $\zeta$  is strictly less than  $v$  then it can be shown (see Section 3.5 of Dhrymes (1998) for example) that there are  $\varrho = v - \zeta$  linear combinations of the  $y_{i,t}$ ,  $i = 1, \dots, v$ , that are asymptotically-stationary even though  $\mathbf{y}_t$  is integrated. It is this feature, of course, that is referred to as cointegration and  $\mathbf{y}_t$  is said to be cointegrated with cointegrating rank  $\varrho$ .

A specification often assumed in the analysis of *ARMAX* systems is the reduced form simply identified  $ARMAX(p_a, p_b, p_m)$  structure in which the normalisation  $\mathbf{A}_0 = \mathbf{M}_0 = \mathbf{I}_v$  is imposed and the degrees  $p_a = \delta[\mathbf{A}(z)]$ ,  $p_b = \delta[\mathbf{B}(z)]$  and  $p_m = \delta[\mathbf{M}(z)]$  are prescribed where  $\delta[\mathbf{A}(z)]$  equals the degree of  $[\mathbf{A}(z)]$  and so on. The coefficient matrices  $\mathbf{A}_u$ ,  $1 \leq u \leq p_a$ ,  $\mathbf{B}_u$ ,  $1 \leq u \leq p_b$ , and  $\mathbf{M}_u$ ,  $1 \leq u \leq p_m$ , are then free to vary subject to the identifiability conditions that  $[\mathbf{A}(z) : \mathbf{B}(z) : \mathbf{M}(z)]$  is left coprime and the matrix  $[\mathbf{A}_{p_a} : \mathbf{B}_{p_b} : \mathbf{M}_{p_m}]$  has full row rank, see Hannan (1971, 1976). This is the structure considered in Dhrymes (1998, Section 1.4.2) and Hsiao (1997, Section 3), for example. Questions relating to the choice of model structure and identification are not straightforward however, even in the stationary case. As pointed out by Hannan, the reduced form simply identified  $ARMAX(p_a, p_b, p_m)$  specification is actually over-identifying in the sense that it excludes particular structures from consideration. In fact, from Theorem 5.1 of Gevers (1986) we know that a simply identified *ARMAX* model can only represent a system in which the McMillan degree is a multiple of  $v$ , and it is not canonical. As pointed out by Lütkepohl and Poskitt (1996), such features can lead to serious practical problems when investigating observed time series. Reduced form simply identified *ARMAX* structures will generate similar problems for the current class of processes *a fortiori*.

In the analysis of cointegrated systems it has proved to be advantageous for both theoretical and practical purposes to separate the long-run behaviour of the system from the more transient dynamics by using the error correction, *EC*, form of the model due to Engle and Granger (1987). It seems sensible, therefore, to contemplate combining the advantages of the *EC* specification with the merits of the echelon canonical form of *ARMAX* processes, and consider modelling

an observed time series using an error correction - autoregressive moving-average with exogenous variables - expressed in echelon form, a specification that we will henceforth denote by the acronym  $ECARMAX_E$ .  $ECARMAX_E$  specifications offer a flexible class of models with which to capture the dynamics of a system under study and recent work of Lütkepohl and Claessen (1997) and Poskitt (2003) indicates that in practice it is possible to construct a more parsimonious but equally adequate representation of an observed multiple time series using an  $ECARMAX_E$  model rather than a more conventional  $ARX$  model whilst incurring little increase in either numerical or analytic complexity.

The structure of an  $ECARMAX_E$  model is characterized by a set of  $v + 1$  nonnegative integers, namely,  $v$  Kronecker indices, that specify the polynomial degrees of the rows of  $\mathbf{A}(z)$ ,  $\mathbf{B}(z)$  and  $\mathbf{M}(z)$ , and the cointegrating rank. The nature of this structure is examined in Sections 3 and 4 following Section 2 in which the general framework of the discussion is outlined whilst establishing additional definitions and notational conventions. In Section 3 the Kronecker index theory is extended to nonstationary  $ARMAX$  systems as described in (1.2). In Section 4 this extension is adapted and further expanded to cover partially nonstationary processes and the canonical form of  $ECARMAX_E$  specifications is established, thereby providing a rigorous foundation for the empirical implementation of these models.

As a prelude to the discussion in Sections 3 and 4 note that approaches to the Kronecker index theory that parallel the development in Akaike (1974) and which depend on Hilbert space ideas involving prediction spaces stretching back into the infinite past are commonly used in the analysis of stationary processes, as are derivations that construct the indices from the stationary autocovariance sequence. See for example Hannan and Deistler (1988), or Reinsel (1993), and the references contained therein. Such approaches are obviously not available here.

Motivated by results of Phillips (1991) indicating that the best way to proceed when analysing cointegrated systems is via maximum likelihood incorporating all prior knowledge about the presence of unit roots and the short run dynamics, Section 5 provides an algorithm for evaluating the Gaussian likelihood of an  $ECARMAX_E$  model and presents the asymptotic distribution of the approximate Gaussian estimator and the exact maximum likelihood estimator. Section 5 also presents some simulation results. The paper ends in Section 6 with some brief remarks. Most proofs are provided in the Appendix.

## 2 The Model, Assumptions and Preliminary Results

Isolating the input variables  $\mathbf{x}_t$  and  $\boldsymbol{\varepsilon}_t$  on the right hand side of (1.2) gives

$$\mathbf{A}(\mathcal{L})\mathbf{y}_t = \mathbf{N}(\mathcal{L})\mathbf{w}_t, \quad t = 1, \dots, T, \quad (2.1)$$

where  $\mathbf{w}'_t = (\boldsymbol{\varepsilon}'_t : -\mathbf{x}'_t)'$  and  $\mathbf{N}(z) = (\mathbf{M}(z) : \mathbf{B}(z))$ . For fixed values of  $v$  and  $u$  let  $[\mathbf{A} : \mathbf{N}]$  denote the set of pairs  $[\mathbf{A}(z) : \mathbf{N}(z)]$  such that  $\mathbf{M}_0 = \mathbf{A}_0$  and  $\det \mathbf{A}(0) \neq 0$ . Now set  $\delta[\mathbf{A}(z) : \mathbf{N}(z)]$  equal to the degree of  $[\mathbf{A}(z) : \mathbf{N}(z)]$ , defined as  $\max_{1 \leq i \leq v} \delta_i[\mathbf{A}(z) : \mathbf{N}(z)]$ ,  $i =$

$1, \dots, v$  denotes the polynomial degree of the  $i$ th row of  $[\mathbf{A}(z) : \mathbf{N}(z)]$ . Let  $\{[\mathbf{A} : \mathbf{N}]\}_p$  denote the set  $\{[\mathbf{A} : \mathbf{N}] : \delta[\mathbf{A}(z) : \mathbf{N}(z)] = p\}$ , with  $p$  finite. For  $[\mathbf{A}(z) : \mathbf{N}(z)] \in \{[\mathbf{A} : \mathbf{N}]\}_p$  define the coefficient sequence  $\{\Phi_0, \Phi_1, \Phi_2, \dots, \Phi_{T+p-1}\}$  via the recursive relationships

$$\sum_{j=0}^i \mathbf{A}_j \Phi_{i-j} = \mathbf{N}_i, \quad i = 0, \dots, p, \quad \text{and} \quad \sum_{j=0}^p \mathbf{A}_j \Phi_{i-j} = \mathbf{0}, \quad i = p+1, \dots, T+p-1. \quad (2.2)$$

Using standard nomenclature,  $\{\Phi_0, \Phi_1, \Phi_2, \dots, \Phi_{T+p-1}\}$  will be referred to as the impulse response sequence. Note that by construction  $\|\Phi_i\| < \infty$ ,  $i = 0, 1, \dots, T+p-1$ , where  $\|\cdot\|$  denotes the Euclidean norm, and the condition that  $\det \mathbf{A}(0) \neq 0$  implies that the power series  $\Phi(z) = \lim_{T \rightarrow \infty} \sum_0^{T+p-1} \Phi_i z^i$  will be convergent for  $|z| < c$  for some  $c > 0$ . If  $\det \mathbf{A}(z) \neq 0$ ,  $|z| \leq 1$ , then  $\|\Phi_i\| \rightarrow 0$  at an exponential rate as  $i \rightarrow \infty$ ,  $c = 1$ , and  $\mathbf{A}(z)$  is said to be stable.

**Assumption 2.1** : *The series  $\mathbf{y}_t$  is an  $I(1)$  process that admits an ARMAX representation as in (1.2), or equivalently (2.1), with  $[\mathbf{A}(z) : \mathbf{N}(z)] \in \{[\mathbf{A} : \mathbf{N}]\}_p$  where:*  
*(i)  $\det \mathbf{A}(z) = (1-z)^\zeta d(z)$ ,  $\zeta \leq v$ , and  $d(z)$  is stable, (ii)  $\det(\mathbf{M}(z)) \neq 0$ ,  $|z| < 1$ , (iii) the normalisation  $\mathbf{M}_0 = \mathbf{A}_0$ ,  $\det \mathbf{A}_0 \neq 0$ , is imposed, and (iv)  $\Sigma_\varepsilon > 0$ .*

The equations in (2.2) define a mapping from  $[\mathbf{A}(z) : \mathbf{N}(z)]$  to  $\Phi(z)$  which is sufficient to determine the characteristics of the data generating mechanism in the stationary case by virtue of the Wold representation theorem, see Lemma 1 of Deistler (1983). The properties of a process satisfying Assumption 2.1 also depend on the homogenous solution to equation (2.1) when viewed as a stochastic difference equation, however, and it is this feature, amongst others, that serves to distinguish the current situation from the stationary case.

## 2.1 A Realization Theorem

For completeness let us briefly review the structure of the solutions to (2.1). It is well known that the solution to a difference equation can be expressed as the sum of a particular solution and a homogeneous solution and this is reflected in the following theorem relating the specification in (2.1) to the representation of  $\mathbf{y}_t$  in input-output final form.

**Theorem 2.1** *The process  $\mathbf{y}_t$  admits an ARMAX representation of the form*

$$\mathbf{A}(\mathcal{L})\mathbf{y}_t = \mathbf{N}(\mathcal{L})\mathbf{w}_t, \quad t = 1, \dots, T,$$

*with  $[\mathbf{A}(z) : \mathbf{N}(z)] \in \{[\mathbf{A} : \mathbf{N}]\}_p$  and initial conditions given by  $(\mathbf{y}'_t : \mathbf{w}'_t)'$ ,  $t = 1 - p, \dots, 0$ , if and only if  $\mathbf{y}_t$  admits a linear input-output representation*

$$\mathbf{y}_t = \sum_{s=0}^{t+p-1} \Phi_s \mathbf{w}_{t-s} + \mathbf{m}_t, \quad t = 1 - p, \dots, 0, 1, \dots, T, \quad (2.3)$$

*in which the conditions  $\sum_{j=0}^p \mathbf{A}_j \Phi_{i-j} = \mathbf{0}$ ,  $i = p+1, \dots, T+p-1$ , and  $\sum_{j=0}^p \mathbf{A}_j \mathbf{m}_{t-j} = \mathbf{0}$ ,  $t = 1, \dots, T$ , are satisfied.*

The final form in (2.3) expresses  $\mathbf{y}_t$  as a function of current and past values of the input, the initial values and the coefficients. The system is therefore described as being causal and time invariant, time invariance meaning that the coefficient values are held constant over time.

Re-expressing the final form using the original partition of  $\mathbf{w}_t$  into  $\boldsymbol{\varepsilon}_t$  and  $\mathbf{x}_t$  and employing an obvious notation for the corresponding partition of the impulse response sequence we obtain the following mean value for  $\mathbf{y}_t$  conditional on the initial values and the exogenous inputs,

$$\boldsymbol{\mu}_y(t) = \mathbf{m}_t - \sum_{s=0}^{t+p-1} \boldsymbol{\Phi}_{x,s} \mathbf{x}_{t-s}, \quad t = 1, \dots, T, \quad (2.4)$$

where  $\mathbf{m}_t$  is calculated deterministically from the initial values. The covariance function is

$$\begin{aligned} \boldsymbol{\Gamma}_y(t, s) &= \sum_{r=0}^{s+p-1} \boldsymbol{\Phi}_{\varepsilon, r+(t-s)} \boldsymbol{\Sigma}_\varepsilon \boldsymbol{\Phi}'_{\varepsilon, r}, \quad t \geq s, \\ &= \sum_{r=0}^{t+p-1} \boldsymbol{\Phi}_{\varepsilon, r} \boldsymbol{\Sigma}_\varepsilon \boldsymbol{\Phi}'_{\varepsilon, r+(s-t)}, \quad t < s, \\ &= \boldsymbol{\Gamma}_y(s, t)', \quad t, s = 1, \dots, T. \end{aligned} \quad (2.5)$$

Since a Gaussian process with mean (2.4) and covariance (2.5) can be readily constructed we will, for ease of exposition, assume that  $\mathbf{y}_t$  is Gaussian.

## 2.2 Identification

Suppose then that  $\boldsymbol{\varepsilon}_t$  is a zero mean Gaussian process with variance  $\boldsymbol{\Sigma}_\varepsilon$  for  $t = 1-p, \dots, 0, 1, \dots, T$  and let  $\boldsymbol{\Lambda}_\xi = [\boldsymbol{\Gamma}_\xi(t-s)]$ ,  $t, s = 1, \dots, T$ , denote the  $Tv \times Tv$  block Toeplitz covariance matrix of  $\boldsymbol{\xi}_t$ ,  $t = 1, \dots, T$ , where  $\boldsymbol{\Gamma}_\xi(\tau) = \boldsymbol{\Gamma}_\xi(-\tau)' = \sum_{j=0}^{p-\tau} \mathbf{M}_j \boldsymbol{\Sigma}_\varepsilon \mathbf{M}'_{j+\tau}$ ,  $\tau = 0, 1, \dots, p$ , and is otherwise zero. Then the function

$$\begin{aligned} f(\mathbf{y}_1^T | \mathbf{y}_{1-p}^0, \mathbf{x}_{1-p}^T; \boldsymbol{\lambda}) &= \prod_{t=1}^T f(\mathbf{y}_t | \mathbf{y}_{1-p}^{t-1}, \mathbf{x}_{1-p}^t; \boldsymbol{\lambda}) \\ &= (2\pi)^{-Tv/2} (\det \boldsymbol{\Lambda}_\xi)^{-\frac{1}{2}} \exp(-(\boldsymbol{\xi}_1^T)' \boldsymbol{\Lambda}_\xi^{-1} \boldsymbol{\xi}_1^T / 2), \end{aligned} \quad (2.6)$$

where  $\mathbf{y}_1^T = (\mathbf{y}'_1, \dots, \mathbf{y}'_T)'$ ,  $\mathbf{y}_{1-p}^0 = (\mathbf{y}'_{1-p}, \dots, \mathbf{y}'_0)'$  and so on, defines the partial likelihood for the parameter vector  $\boldsymbol{\lambda}' = (\boldsymbol{\beta}' : \boldsymbol{\sigma}')$  where  $\boldsymbol{\beta} = \text{vec}[\mathbf{A}_0 : \dots : \mathbf{A}_p : \mathbf{B}_0 : \dots : \mathbf{B}_p : \mathbf{M}_0 : \dots : \mathbf{M}_p]$  contains the structural coefficients and  $\boldsymbol{\sigma} = \text{vech}[\boldsymbol{\Sigma}_\varepsilon]$  the scale parameters, Cox (1975). Note that the density of the endogenous variable is conditional on both the initial values and the exogenous input.

**Assumption 2.2** : *The statistic  $\mathbf{y}_{1-p}^0$  is, using statistical parlance, ancillary for  $\boldsymbol{\lambda}$  and (2.6) defines the partial likelihood for  $\boldsymbol{\lambda}$  where the exogenous process satisfies Assumption 2.3. There are no restrictions on  $\boldsymbol{\sigma}$  other than those that ensure  $\boldsymbol{\Sigma}_\varepsilon > 0$  and there are no joint restrictions linking the elements of  $\boldsymbol{\sigma}$  to those of  $\boldsymbol{\beta}$ .*

**Assumption 2.3** : Let  $\langle \Delta \mathbf{x} \rangle_{t|t-1} = \sum_{j=1}^{t+p-2} \mathbf{L}_j \Delta \mathbf{z}_{t-j}$  denote the projection of  $\Delta \mathbf{x}_t$  on to the space spanned by  $\Delta \mathbf{z}_{2-p}^{t-1}$  where  $\Delta \mathbf{z}_t = (\Delta \mathbf{y}'_t, \Delta \mathbf{x}'_t)'$ ,  $t = 2-p, \dots, T$ . Then there exist constants  $\Phi$  and  $\lambda$ ,  $0 < \Phi < \infty$  and  $0 \leq \lambda < 1$ , such that  $\|\mathbf{L}_s\| < \Phi \lambda^s$  as  $s \rightarrow \infty$ . The exogenous disturbance  $\Delta \boldsymbol{\eta}_t = \Delta \mathbf{x}_t - \langle \Delta \mathbf{x} \rangle_{t|t-1}$  is a Gaussian process that is independent of  $\boldsymbol{\varepsilon}_t$  and  $\lim_{T \rightarrow \infty} T^{-1} \sum_{t=1}^T \Delta \boldsymbol{\eta}_t \Delta \boldsymbol{\eta}'_{t-\tau} = \boldsymbol{\Gamma}_{\Delta \eta}(\tau)$  a.s. where  $\boldsymbol{\Gamma}_{\Delta \eta}(\tau)$ ,  $\tau = 0, 1, 2, \dots$  is a positive definite sequence.

Presuming that Assumptions 2.1, 2.2 and 2.3 hold, identification can now be defined by the standard requirement that equality between the likelihood values  $f(\mathbf{y}_1^T | \mathbf{y}_{1-p}^0, \mathbf{x}_{1-p}^T; \boldsymbol{\lambda})$  and  $f(\mathbf{y}_1^T | \mathbf{y}_{1-p}^0, \mathbf{x}_{1-p}^T; \boldsymbol{\lambda}^*)$  for all  $\mathbf{y}_1^T$ ,  $\mathbf{y}_{1-p}^0$  and  $\mathbf{x}_{1-p}^T$  implies that  $\boldsymbol{\lambda} = \boldsymbol{\lambda}^*$ , otherwise there would exist parameter values  $\boldsymbol{\lambda}$  and  $\boldsymbol{\lambda}^*$  with  $\boldsymbol{\lambda} \neq \boldsymbol{\lambda}^*$  such that  $\boldsymbol{\lambda}$  and  $\boldsymbol{\lambda}^*$  are observationally equivalent. Given that there are no constraints linking the elements of  $\boldsymbol{\sigma}$  to those of  $\boldsymbol{\beta}$ , and  $\mathbf{M}_0 = \mathbf{A}_0$ ,  $\boldsymbol{\Sigma}_\varepsilon$  is determined uniquely by  $\boldsymbol{\sigma}$  without further ado. Additional constraints must be placed upon  $\boldsymbol{\beta}$ , however, in order to select a characteristic element from within each observational equivalence class. Theorem 2.1 implies that this can be done by constructing a one-to-one correspondence between the initial values and  $[\mathbf{A}(z) : \mathbf{N}(z)]$ , on the one hand, and  $\boldsymbol{\Phi}_s$ ,  $s = 0, \dots, T+p-1$ , and  $\mathbf{m}_s$ ,  $s = 1-p, \dots, T$ , on the other, so that  $\mathbf{y}_1^T$  and  $f(\mathbf{y}_1^T | \mathbf{y}_{1-p}^0, \mathbf{x}_{1-p}^T; \boldsymbol{\lambda})$  are uniquely determined once  $\mathbf{y}_{1-p}^0$  and  $\mathbf{x}_{1-p}^T$  are known and the value of  $\boldsymbol{\lambda}$  has been given.

### 3 The Kronecker Index Theory for Nonstationary ARMAX Processes

Consider re-couching Theorem 2.1 in terms of a sequence of block Hankel matrices of finite dimension derived from the input-output representation. To this end, set  $\mathbf{K}_\tau = [\boldsymbol{\Phi}_\tau : \mathbf{m}_{\tau-p+1}]$   $\tau = 0, 1, \dots, T+p-1$ , and define  $\mathbf{H}_{R,T}$  to be the  $Rv \times (T+p-R)(v+u+1)$  block Hankel matrix with  $\mathbf{K}_{i+j-1}$  in the  $(i, j)$ th  $(v \times (v+u+1))$  block,  $i = 1, \dots, R$ ,  $j = 1, \dots, T+p-R$ . That is,

$$\mathbf{H}_{R,T} = \begin{bmatrix} \mathbf{K}_1 & \mathbf{K}_2 & \cdots & \mathbf{K}_{T+p-R} \\ \mathbf{K}_2 & \mathbf{K}_3 & \cdots & \mathbf{K}_{T+p-R+1} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{K}_R & \mathbf{K}_{R+1} & \cdots & \mathbf{K}_{T+p-1} \end{bmatrix}.$$

Let  $\mathbf{h}_{R,T}(i, r)$  denote row  $(i-1)v+r$  of  $\mathbf{H}_{R,T}$ ,  $i = 1, \dots, R$ ,  $r = 1, \dots, v$ . From the Hankel structure of  $\mathbf{H}_{R,T}$  it follows that if  $\mathbf{h}_{R,T}(i, r)$  lies in the linear span of  $\mathbf{h}_{R,T}(i_1, r_1), \dots, \mathbf{h}_{R,T}(i_L, r_L)$  where  $i_j < i$ ,  $j = 1, \dots, L$ , then row  $\mathbf{h}_{R+1,T}(i+1, r)$  of  $\mathbf{H}_{R+1,T}$  lies in the linear span of  $\mathbf{h}_{R+1,T}(i_1+1, r_1), \dots, \mathbf{h}_{R+1,T}(i_L+1, r_L)$ . Thus the block Hankel matrix sequence  $\mathbf{H}_{R,T}$ ,  $R = 1, \dots, T+p-1$  exhibits similar linear dependence properties to those found in the infinite dimensional block Hankel matrix conventionally analysed in the stationary case (*cf.* Hannan and Deistler (1988, expression 2.3.5), for example, or Reinsel (1993, expression 3.2)). Arguments that

parallel those employed in the stationary case can therefore be used to establish corresponding results.

Our interest centers on how the properties of  $\mathbf{H}_{R,T}$ ,  $R = 1, \dots, T + p - 1$ , can be used to determine the structure of the input-output system.<sup>4</sup> If we define the rank of the sequence as  $\sup_{1 \leq R \leq T+p-1} [\rho(\mathbf{H}_{R,T})]$  where  $\rho(\mathbf{H}_{R,T})$  denotes the rank of  $\mathbf{H}_{R,T}$  then the following corollary to Theorem 2.1 indicates that the rank provides a partial characterisation of the system equivalent to the McMillan degree.

**Corollary 3.1 :** *The process  $\mathbf{y}_t$  admits an ARMAX representation of the form  $\mathbf{A}(\mathcal{L})\mathbf{y}_t = \mathbf{N}(\mathcal{L})\mathbf{w}_t$ ,  $t = 1, \dots, T$ , for all  $T > vp$ , with initial conditions given by  $(\mathbf{y}'_t : \mathbf{w}'_t)'$ ,  $t = 1 - p, \dots, 0$ , if and only if  $\sup_{1 \leq R \leq T+p-1} [\rho(\mathbf{H}_{R,T})] = \rho(\mathbf{H}_{p,T}) \leq vp$ .*

The characterisation is completed by selecting a basis for the row space of  $\mathbf{H}_{p+1,T}$ , thereby obtaining a unique parameterisation for the operator pair  $[\mathbf{A}(z) : \mathbf{N}(z)]$  in which

- (i)  $a_{rc,0} = n_{rc,0}$ ,  $r, c = 1, \dots, v$ ,
  - (ii)  $a_{rr}(z) = 1 + a_{rr,1}z + \dots + a_{rr,n_r}z^{n_r}$ ,  
 $a_{rc}(z) = a_{rc,n_r-n_{rc}+1}z^{n_r-n_{rc}+1} + \dots + a_{rc,n_r}z^{n_r}$  and
  - (iii)  $n_{rc}(z) = n_{rc,0} + n_{rc,1}z + \dots + n_{rc,n_r}z^{n_r}$ ,  $r = 1, \dots, v$ ,  $c = 1, \dots, u + v$ ,
- where

$$n_{rc} = \begin{cases} \min(n_r + 1, n_c) & \text{for } c < r \\ \min(n_r, n_c) & \text{for } c \geq r . \end{cases}$$

A pair  $[\mathbf{A}(z) : \mathbf{N}(z)]$  satisfying (i)–(iii) is said to be in echelon form and the nonnegative integers  $n_i$ ,  $i = 1, \dots, v$  are called the Kronecker indices. Such a pair defines a canonical structure called an echelon canonical form.

**Theorem 3.1** *For all  $T > vp$  a nonstationary ARMAX process  $\mathbf{y}_t$  is uniquely defined via the initial conditions  $(\mathbf{y}'_t : \mathbf{w}'_t)'$ ,  $t = 1 - p, \dots, 0$ , and the representation*

$$\sum_{j=0}^p \mathbf{A}_j \mathbf{y}_{t-j} = \sum_{j=0}^p \mathbf{N}_j \mathbf{w}_{t-j}, \quad t = 1, \dots, T,$$

when the pair  $[\mathbf{A}(z) : \mathbf{N}(z)] \in \{[\mathbf{A} : \mathbf{N}]\}_p$  are in echelon canonical form.

Since by assumption  $p$  is finite  $\rho(\mathbf{H}_{p,T}) = n_1 + \dots + n_v \leq vp$  is bounded and the echelon form depends on fixed, finite values of the Kronecker indices, as it does in the stationary case. It is clear, however, that the values  $n_r$ ,  $r = 1, \dots, v$ , are not invariant with respect to a reordering of the elements of  $\mathbf{y}_t$ , for if  $\mathbf{P}$  denotes an arbitrary permutation matrix then

$$\mathbf{P}\mathbf{y}_t = \sum_{s=0}^{t+p-1} \mathbf{P}\Phi_s \mathbf{w}_{t-s} + \mathbf{P}\mathbf{m}_t, \quad t = 1 - p, \dots, 0, 1, \dots, T,$$

and the linear dependences in the Hankel matrix sequence that previously generated the Kro-

necker indices may no longer hold for the permuted sequence

$$(\mathbf{P} \otimes \mathbf{I}_R)\mathbf{H}_{R,T} = \begin{bmatrix} \mathbf{PK}_1 & \mathbf{PK}_2 & \dots & \mathbf{PK}_{T+p-R} \\ \mathbf{PK}_2 & \mathbf{PK}_3 & \dots & \mathbf{PK}_{T+p-R+1} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{PK}_R & \mathbf{PK}_{R+1} & \dots & \mathbf{PK}_{T+p-1} \end{bmatrix},$$

$R = 1, \dots, T + p - 1$ . To this extent the echelon canonical form is only unique modulo such rotations. Following an argument that exactly parallels the development of Gevers (1986, pp. 1750-1751), however, we can establish the following version of Gevers' Lemma 2.4, which yields a unique invariant form.

**Lemma 3.1** : *The variables in  $\mathbf{y}_t = (y_{1t}, \dots, y_{vt})'$  can be permuted such that the Kronecker indices of  $(y_{r(1)t}, \dots, y_{r(v)t})'$  are arranged in descending order,  $n_{r(1)} \geq n_{r(2)} \geq \dots \geq n_{r(v)}$ , where  $r(j), j = 1, \dots, v$ , denotes a permutation of  $1, \dots, v$  that induces the ordering. The  $n_{r(j)}, j = 1, \dots, v$ , are unique and are referred to as the Kronecker invariants.*

Thus if  $\mathbf{P}$  now denotes a permutation matrix such that  $\mathbf{P}(1, \dots, v)' = (r(1), \dots, r(v))'$  then  $\mathbf{P}\mathbf{y}_t = (y_{r(1)t}, \dots, y_{r(v)t})'$  has an echelon form ARMAX representation with Kronecker indices  $(n_{r(1)}, \dots, n_{r(v)})$  equal to the Kronecker invariants. When expressed in terms of the Kronecker invariants the coefficient matrix  $\mathbf{A}_0 = \mathbf{M}_0$  is lower triangular, the representation of the system is canonical and the ordered variables  $y_{r(j)t}, j = 1, \dots, v$  possess unique characterisations.

**Example:(i)** Suppose that  $\mathbf{y}_t$  is a  $v$  component process generated by the reduced form structure

$$\mathbf{y}_t + \mathbf{A}\mathbf{y}_{t-1} = \boldsymbol{\varepsilon}_t + \mathbf{M}\boldsymbol{\varepsilon}_{t-1}, \quad t = 1, \dots, T,$$

with initial values  $\mathbf{y}_0$  and  $\boldsymbol{\varepsilon}_0$ . Then  $\boldsymbol{\Phi}_0 = \mathbf{I}$ ,  $\boldsymbol{\Phi}_1 = \mathbf{M} - \mathbf{A} = \mathbf{D}$ ,  $\boldsymbol{\Phi}_j = (-\mathbf{A})^{j-1}\mathbf{D}$ ,  $j = 2, 3, \dots, T$ , and  $\mathbf{m}_t = (-\mathbf{A})^t\mathbf{d}$ ,  $t = 0, 1, \dots, T$ , where  $\mathbf{d} = \mathbf{y}_0 - \boldsymbol{\varepsilon}_0$ . Hence, from Theorem 2.1,

$$\mathbf{y}_t = \boldsymbol{\varepsilon}_t + \sum_{s=1}^t (-\mathbf{A})^{s-1}\mathbf{D}\boldsymbol{\varepsilon}_{t-s} + (-\mathbf{A})^t\mathbf{d}, \quad t = 1, \dots, T,$$

as can be verified by a direct sequence of successive substitutions, and

$$\mathbf{H}_{R,T} = \begin{bmatrix} (\mathbf{D} : \mathbf{d}) & (-\mathbf{A})(\mathbf{D} : \mathbf{d}) & \dots & (-\mathbf{A})^{T-R+1}(\mathbf{D} : \mathbf{d}) \\ (-\mathbf{A})(\mathbf{D} : \mathbf{d}) & (-\mathbf{A})^2(\mathbf{D} : \mathbf{d}) & \dots & (-\mathbf{A})^{T-R+2}(\mathbf{D} : \mathbf{d}) \\ \vdots & \vdots & \ddots & \vdots \\ (-\mathbf{A})^{R-1}(\mathbf{D} : \mathbf{d}) & (-\mathbf{A})^R(\mathbf{D} : \mathbf{d}) & \dots & (-\mathbf{A})^T(\mathbf{D} : \mathbf{d}) \end{bmatrix}.$$

From  $\mathbf{H}_{R,T}$ ,  $R = 1, \dots, T$ , it is clear that  $\sup_{1 \leq R \leq T+p-1} [\rho(\mathbf{H}_{R,T})] = \rho(\mathbf{H}_{1,T}) \leq v$  and the Kronecker indices  $n_r \leq 1$ ,  $r = 1, \dots, v$ . Assume that  $\rho(\mathbf{H}_{1,T}) = k < v$ . Then the Kronecker

invariants are  $n_{r(j)} = 1$ ,  $j = 1, \dots, k$ ,  $n_{r(j)} = 0$ ,  $j = k + 1, \dots, v$ , and the echelon form of  $(y_{r(1)t}, \dots, y_{r(k)t}, y_{r(k+1)t}, \dots, y_{r(v)t})'$  is given by

$$\mathbf{A}_0 = \begin{bmatrix} \mathbf{I}_k & \mathbf{0} \\ * & \mathbf{I}_{v-k} \end{bmatrix} \quad \text{and} \quad \mathbf{A}_1 = \begin{bmatrix} * & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$$

with  $\mathbf{M}_0 = \mathbf{A}_0$  and  $\mathbf{M}_1 = \mathbf{D} + \mathbf{A}_1$  where  $*$  indicates a block of the  $(k : (v - k))' \times (k : (v - k))$  partitioned matrix whose entries are not restricted by the canonical form .  $\square$

### 3.1 Special Features

Although the current situation shares several features in common with the stationary case certain critical differences do arise. In particular, in the stationary case boundedness of the rank of the infinite block Hankel matrix  $[\Phi(i - j + 1)]_{\{i,j=1,\dots,\infty\}}$  is associated with rationality and leads to the conclusion that the canonical representation will be coprime, see Hannan and Deistler (1988) for a self-contained discussion of rational transfer functions, coprimeness and other related issues. For a nonstationary process of the type being considered here the Hankel matrices in the sequence  $\mathbf{H}_{R,T}$ ,  $R = 1, \dots, T + p - 1$ , all have finite size and hence finite rank, but the echelon form need not be coprime.

**Example:(i')** Observe that the canonical structure given in Example:(i) is applicable whatever the values of  $\mathbf{A}$  and  $\mathbf{M}$ . Thus, if  $\mathbf{A} = \mathbf{M} = -\mathbf{I}$  and  $\mathbf{d} = \mathbf{y}_0 - \boldsymbol{\varepsilon}_0 \neq \mathbf{0}$  where, without loss of generality,  $d_1 \neq 0$ , then  $\sup_{1 \leq R \leq T+p-1} [\rho(\mathbf{H}_{R,T})] = 1$  and a simple calculation shows that

$$\mathbf{A}_0 = \mathbf{M}_0 = \begin{bmatrix} 1 & 0 & \dots & 0 \\ -d_2/d_1 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ -d_v/d_1 & 0 & \dots & 1 \end{bmatrix}$$

and  $\mathbf{A}_1 = \mathbf{M}_1$  has first row  $(-1, 0, \dots, 0)$  and is otherwise zero. Expressed in terms of the individual components we have  $y_{1,t} - y_{1,(t-1)} = \varepsilon_{1,t} - \varepsilon_{1,(t-1)}$  with initial values  $y_{1,0}$  and  $\varepsilon_{1,0}$ , and  $y_{i,t} - \varepsilon_{i,t} = (d_i/d_1)(y_{1,t} - \varepsilon_{1,t})$  for  $i = 2, \dots, v$ ,  $t = 1, \dots, T$ . Solving the echelon form leads to the representation  $y_{i,t} - d_i = \varepsilon_{i,t}$ ,  $i = 1, \dots, v$ ,  $t = 1, \dots, T$ . Note that the constants  $d_i = y_{i,0} - \varepsilon_{i,0}$ ,  $i = 1, \dots, v$ , are definitive whereas the stationary solutions to  $\mathbf{y}_t - \mathbf{y}_{t-1} = \boldsymbol{\varepsilon}_t - \boldsymbol{\varepsilon}_{t-1}$ , for  $t \in \mathcal{Z} = \{0, \pm 1, \pm 2, \dots\}$  are given by  $\mathbf{y}_t = \mathbf{d}^* + \boldsymbol{\varepsilon}_t$  where  $\mathbf{d}^*$  is arbitrary.  $\square$

This example illustrates that in a nonstationary world with fixed starting points coprimeness is not a generic property of the canonical form. In order to examine the non-coprime situation in a little more depth and assess the practical implications, let us extend the above example.

**Example:(ii)** Presume that  $\mathbf{y}_t$  is observed for  $t = 0, 1, \dots, T$ , and it is known that  $n_r = 1$ ,

$r = 1, \dots, v$ , but the coefficients  $\mathbf{A}$  and  $\mathbf{M}$  in the echelon form

$$\mathbf{y}_t + \mathbf{A}\mathbf{y}_{t-1} = \boldsymbol{\varepsilon}_t + \mathbf{M}\boldsymbol{\varepsilon}_{t-1}, \quad t = 1, \dots, T,$$

are unknown. Suppose also that there exist first stage estimates  $\bar{\boldsymbol{\varepsilon}}_t$  of  $\boldsymbol{\varepsilon}_t$  for  $t = 0, 1, \dots, T$ , such that  $T^{-1} \sum_{t=0}^T \|\bar{\boldsymbol{\varepsilon}}_t - \boldsymbol{\varepsilon}_t\|^2 = O(H_T/T)$  a.s. where  $H_T$  is an increasing sequence of values such that  $H_T/T \rightarrow 0$  as  $T \rightarrow \infty$ . At least two versions of such estimates are available. The first is given by the residuals from an AR of order  $h_T = [c(\log T)^a]$ ,  $c > 0$ ,  $a > 1$ , fitted to  $\mathbf{y}_t$ ,  $t = 1, \dots, T$ , supposing the initial values  $\mathbf{y}_t$ ,  $t = -h_T + 1, \dots, 0$ , are available. Proposition 3.1 of Poskitt (2003) tells us that  $H_T = O(\log T)$  at most for this estimate. The second possibility is to construct residuals using instrumental variable estimates of  $\mathbf{A}$  and  $\mathbf{M}$ , see *inter alia* Yap and Reinsel (1995). Substituting  $\bar{\boldsymbol{\varepsilon}}_{t-1}$  for  $\boldsymbol{\varepsilon}_{t-1}$  in  $\mathbf{y}_t = -\mathbf{A}\mathbf{y}_{t-1} + \mathbf{M}\boldsymbol{\varepsilon}_{t-1} + \boldsymbol{\varepsilon}_t$  it can be seen that second stage least squares estimates of  $\mathbf{A}$  and  $\mathbf{M}$  can be obtained by solving the normal equations

$$[-\bar{\mathbf{A}}_T : \bar{\mathbf{M}}_T] \sum_{t=1}^T \begin{bmatrix} \mathbf{y}_{t-1}\mathbf{y}'_{t-1} & \mathbf{y}_{t-1}\bar{\boldsymbol{\varepsilon}}'_{t-1} \\ \bar{\boldsymbol{\varepsilon}}_{t-1}\mathbf{y}'_{t-1} & \bar{\boldsymbol{\varepsilon}}_{t-1}\bar{\boldsymbol{\varepsilon}}'_{t-1} \end{bmatrix} = \sum_{t=1}^T [\mathbf{y}_t\mathbf{y}'_{t-1} : \mathbf{y}_t\bar{\boldsymbol{\varepsilon}}'_{t-1}]. \quad (3.1)$$

Now assume that  $\mathbf{A} = \mathbf{M} = \mathbf{C}$  where  $\mathbf{C}$  has singular values on the interval  $(0, 1]$  and  $\mathbf{y}_0 \neq \boldsymbol{\varepsilon}_0$ . Examining the components of 3.1 we find that

$$\begin{aligned} T^{-1} \sum_{t=1}^T \mathbf{y}_{t-1}\mathbf{y}'_{t-1} &= T^{-1} \sum_{t=1}^{T-1} \boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}'_t + T^{-1} \sum_{t=1}^{T-1} \{\boldsymbol{\varepsilon}_t\mathbf{d}'(-\mathbf{C}')^t + (-\mathbf{C})^t\mathbf{d}\boldsymbol{\varepsilon}'_t\} + \\ &\quad T^{-1} \sum_{t=1}^{T-1} (-\mathbf{C})^t\mathbf{d}\mathbf{d}'(-\mathbf{C}')^t + \mathbf{y}_0\mathbf{y}'_0/T \\ &= \boldsymbol{\Sigma}_\varepsilon + \mathbf{R}_T + O(l_2(T)) \text{ a.s.} \end{aligned}$$

where  $l_2(T) = (\log \log T/T)^{\frac{1}{2}}$  and  $\mathbf{R}_T = T^{-1} \sum_{t=1}^{T-1} (-\mathbf{C})^t\mathbf{d}\mathbf{d}'(-\mathbf{C}')^t > 0$ . The second and third terms are  $O(l_2(T))$  with probability one because the process  $\mathbf{u}_t = \text{vec}(\boldsymbol{\varepsilon}_t\mathbf{d}'(-\mathbf{C}')^t) = ((-\mathbf{C})^t\mathbf{d} \times \mathbf{I}_v)\boldsymbol{\varepsilon}_t$  is a martingale difference sequence and therefore  $\text{vec}(T^{-1} \sum_{t=1}^{T-1} \boldsymbol{\varepsilon}_t\mathbf{d}'(-\mathbf{C}')^t) = T^{-1} \sum_{t=1}^{T-1} \mathbf{u}_t = O(l_2(T))$ . By the same arguments

$$\begin{aligned} T^{-1} \sum_{t=1}^T \mathbf{y}_t\mathbf{y}'_{t-1} &= T^{-1} \sum_{t=2}^T \boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}'_{t-1} + T^{-1} \sum_{t=2}^T \{\boldsymbol{\varepsilon}_t\mathbf{d}'(-\mathbf{C}')^{t-1} + (-\mathbf{C})^t\mathbf{d}\boldsymbol{\varepsilon}'_{t-1}\} + \\ &\quad T^{-1} \sum_{t=2}^T (-\mathbf{C})^t\mathbf{d}\mathbf{d}'(-\mathbf{C}')^{t-1} + \mathbf{y}_1\mathbf{y}'_0/T \\ &= -\mathbf{C}\mathbf{R}_T + O(l_2(T)) \text{ a.s.} \end{aligned}$$

Now, trivially,

$$T^{-1} \sum_{t=1}^T \bar{\boldsymbol{\varepsilon}}_{t-1}\mathbf{y}'_{t-1} = T^{-1} \sum_{t=1}^T \boldsymbol{\varepsilon}_{t-1}\mathbf{y}'_{t-1} + T^{-1} \sum_{t=1}^T (\bar{\boldsymbol{\varepsilon}}_{t-1} - \boldsymbol{\varepsilon}_{t-1})\mathbf{y}'_{t-1}.$$

The second term on the right hand side is bounded in norm by

$$T^{-1} \left\{ \left( \sum_t \|\mathbf{y}_{t-1}\|^2 \right) \left( \sum_t \|\bar{\boldsymbol{\varepsilon}}_{t-1} - \boldsymbol{\varepsilon}_{t-1}\|^2 \right) \right\}^{1/2} = O(\{H_T/T\}^{1/2})$$

and the first term

$$\begin{aligned} T^{-1} \sum_{t=1}^T \boldsymbol{\varepsilon}_{t-1} \mathbf{y}'_{t-1} &= T^{-1} \sum_{t=2}^T \boldsymbol{\varepsilon}_{t-1} \boldsymbol{\varepsilon}'_{t-1} + T^{-1} \sum_{t=2}^T \boldsymbol{\varepsilon}_{t-1} \mathbf{d}' (-\mathbf{C}')^{t-1} + \boldsymbol{\varepsilon}_0 \mathbf{y}'_0 / T \\ &= \boldsymbol{\Sigma}_\varepsilon + O(l_2(T)) \text{ a.s. .} \end{aligned}$$

Thus we can conclude that  $T^{-1} \sum_{t=1}^T \bar{\boldsymbol{\varepsilon}}_{t-1} \mathbf{y}'_{t-1} = \boldsymbol{\Sigma}_\varepsilon + O(l_2(T)) + O(\{H_T/T\}^{1/2})$ . Similarly

$$\begin{aligned} T^{-1} \sum_{t=1}^T \mathbf{y}_t \bar{\boldsymbol{\varepsilon}}'_{t-1} &= T^{-1} \sum_{t=2}^T \boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}'_{t-1} + T^{-1} \sum_{t=2}^T (-\mathbf{C})^t \mathbf{d} \boldsymbol{\varepsilon}'_{t-1} + \mathbf{y}_1 \boldsymbol{\varepsilon}'_0 / T + O(\{H_T/T\}^{1/2}) \\ &= O(l_2(T)) + O(\{H_T/T\}^{1/2}) \text{ a.s.} \end{aligned}$$

and the equality

$$\begin{aligned} T^{-1} \sum_{t=1}^T (\bar{\boldsymbol{\varepsilon}}_{t-1} \bar{\boldsymbol{\varepsilon}}'_{t-1} - \boldsymbol{\varepsilon}_{t-1} \boldsymbol{\varepsilon}'_{t-1}) &= T^{-1} \sum_{t=1}^T (\bar{\boldsymbol{\varepsilon}}_{t-1} - \boldsymbol{\varepsilon}_{t-1})(\bar{\boldsymbol{\varepsilon}}_{t-1} - \boldsymbol{\varepsilon}_{t-1})' + \\ &\quad T^{-1} \sum_{t=1}^T (\boldsymbol{\varepsilon}_{t-1}(\bar{\boldsymbol{\varepsilon}}_{t-1} - \boldsymbol{\varepsilon}_{t-1})' + (\bar{\boldsymbol{\varepsilon}}_{t-1} - \boldsymbol{\varepsilon}_{t-1}) \boldsymbol{\varepsilon}'_{t-1}) \end{aligned}$$

implies that  $T^{-1} \sum_{t=1}^T \bar{\boldsymbol{\varepsilon}}_{t-1} \bar{\boldsymbol{\varepsilon}}'_{t-1}$  equals  $\boldsymbol{\Sigma}_\varepsilon + O(l_2(T)) + O(\{H_T/T\}^{1/2})$  with probability one as  $T \rightarrow \infty$ .

If  $\mathbf{d} \neq \mathbf{0}$  and the largest singular value of  $\mathbf{C}$  is bounded away from one then  $\|(-\mathbf{C})^t \mathbf{d}\| < C\lambda^t$  for some  $C > 0$ ,  $\lambda < 1$ ,  $\mathbf{R}_T = O(T^{-1})$ , and the large sample behaviour of  $[\bar{\mathbf{A}}_T : \bar{\mathbf{M}}_T]$  will reflect that  $\mathbf{y}_t$  is asymptotically stationary. If, however,  $\mathbf{d} \neq \mathbf{0}$  and  $\mathbf{C}$  has at least one singular value of unity, then the magnitude of  $\mathbf{R}_T$  as  $T$  increases will be such that it will dominate the  $O(l_2(T))$  and  $O(\{H_T/T\}^{1/2})$  remainder terms that appear in the sums of squares and cross products that appear in 3.1. From Lemma A.2 of Poskitt (2000) it follows that

$$[-\bar{\mathbf{A}}_T : \bar{\mathbf{M}}_T] = [(-\mathbf{C}\mathbf{R}_T : \mathbf{0}) + o(1)] \left[ \begin{pmatrix} \boldsymbol{\Sigma}_\varepsilon + \mathbf{R}_T & \boldsymbol{\Sigma}_\varepsilon \\ \boldsymbol{\Sigma}_\varepsilon & \boldsymbol{\Sigma}_\varepsilon \end{pmatrix}^{-1} + o(1) \right] \text{ a.s. .}$$

Using standard formulae for partitioned inversion we can therefore conclude that  $[\bar{\mathbf{A}}_T : \bar{\mathbf{M}}_T]$  will converge to the value  $[\mathbf{C} : \mathbf{C}]$  as the time horizon increases. The consistency observed in the nonstationary case stems from the fact that  $\mathbf{m}_t = (-\mathbf{C})^t \mathbf{d}$  plays an important part in determining the evolution and structure of the process and feeds information about the parameters through to the observed statistics, information that is inevitably lost in the asymptotically stationary case because  $\mathbf{m}_t$  converges to zero at an exponential rate.  $\square$

## 4 Cointegration and the Error Correction Echelon Form

We now specialise the results of the previous sections to unit-root nonstationary cointegrated systems and investigate the consequences of the identification conditions presented above for the analysis of partially nonstationary *ARMAX* structures.

First, consider interchanging the roles of  $\mathbf{y}_t$  and  $\boldsymbol{\varepsilon}_t$  in (1.2). Making the replacements  $\mathbf{y}_t \mapsto \boldsymbol{\varepsilon}_t$ ,  $\mathbf{M}(z) \leftrightarrow \mathbf{A}(z)$ , and  $\mathbf{w}_t \mapsto \mathbf{z}_t$  where  $\mathbf{z}_t = (\mathbf{y}'_t : \mathbf{x}'_t)'$  we obtain

$$\mathbf{M}(\mathcal{L})\boldsymbol{\varepsilon}_t = \mathbf{N}(\mathcal{L})\mathbf{z}_t, \quad t = 1, \dots, T,$$

with initial values  $(\boldsymbol{\varepsilon}'_t : \mathbf{z}'_t)'$ ,  $t = 1 - p, \dots, 0$ . The arguments employed in the previous section can now be repeated to produce the following parallel to Theorem (3.1).

**Theorem 4.1** *The values of the stochastic disturbance  $\boldsymbol{\varepsilon}_t$  are given uniquely via the initial values  $(\boldsymbol{\varepsilon}'_t : \mathbf{y}'_t : \mathbf{x}'_t)'$ ,  $t = 1 - p, \dots, 0$ , and the expression*

$$\sum_{j=0}^p \mathbf{M}_j \boldsymbol{\varepsilon}_{t-j} = \sum_{j=0}^p \mathbf{A}_j \mathbf{y}_{t-j} + \sum_{j=0}^p \mathbf{B}_j \mathbf{x}_{t-j}, \quad t = 1, \dots, T,$$

where  $[\mathbf{A}(z) : \mathbf{B}(z) : \mathbf{M}(z)]$  satisfy the conditions

- (i')  $a_{rc,0} = m_{rc,0}$ ,  $r, c = 1, \dots, v$ ,
- (ii')  $m_{rr}(z) = 1 + m_{rr,1}z + \dots + m_{rr,n_r}z^{n_r}$ ,  
 $m_{rc}(z) = m_{rc,n_r-n_{rc}+1}z^{n_r-n_{rc}+1} + \dots + m_{rc,n_r}z^{n_r}$ ,  $r, c = 1, \dots, v$ ,
- (iii')  $a_{rc}(z) = a_{rc,0} + a_{rc,1}z + \dots + a_{rc,n_r}z^{n_r}$ ,  $r, c = 1, \dots, v$  and  
 $b_{rc}(z) = b_{rc,0} + b_{rc,1}z + \dots + b_{rc,n_r}z^{n_r}$ ,  $r = 1, \dots, v$ ,  $c = 1, \dots, u$ .

Note that the canonical form in (i')–(iii') employs the same normalisation as previously, namely that  $\mathbf{A}_0 = \mathbf{M}_0$  with unit diagonal elements, and  $\delta_r[\mathbf{A}(z) : \mathbf{B}(z) : \mathbf{M}(z)] = n_r$ ,  $r = 1, \dots, s$ , as before, but additional exclusion constraints are placed on the elements of  $\mathbf{M}(z)$ , rather than  $\mathbf{A}(z)$  as in (i)–(iii). Those elements not so restricted are freely varying. This differs from what is commonly found in the literature on echelon forms and for clarity we will therefore call this structure the inverse echelon canonical form. The terminology is based on the fact that the derivations leading to Theorem (3.1) parallel the manipulations used to invert the *ARMAX* system in order to represent the stochastic disturbance in terms of the model parameters and the observables when constructing the likelihood function, see Proposition 5.1. The alternative identification convention used in the inverse echelon canonical form has no bearing on the uniqueness properties of the representation but the modification turns out to be particularly convenient when discussing cointegration.

#### 4.1 Error Correction and the Kronecker Indices

In order to facilitate discussion in a general framework let  $\mathbf{z}_t = (\mathbf{y}'_t : \mathbf{x}'_t)'$ , as above, and consider embedding equation (1.2) in the *ARMA* system

$$\Psi(\mathcal{L})\mathbf{z}_t = \Theta(\mathcal{L})\mathbf{e}_t, \quad t = 1, \dots, T, \quad (4.1)$$

with initial conditions  $(\mathbf{z}'_t, \mathbf{e}'_t)'$ ,  $t = 1 - p, \dots, 0$ , where  $\Psi(z) = \sum_{j=0}^p \Psi_j z^j$ ,  $\Theta(z) = \sum_{j=0}^p \Theta_j z^j$  and  $\mathbf{e}_t = (\boldsymbol{\varepsilon}'_t : \boldsymbol{\eta}'_t)'$  is an  $s = v + u$  component white noise process with covariance matrix  $\Sigma_{(\varepsilon, \eta)}$ . Assume that  $\mathbf{z}_t$  satisfies Assumption 2.1 and has cointegrating rank  $\rho$ . To isolate the integrated components of the process we can apply the following variant of the Beveridge-Nelson decomposition:

**Proposition 4.1** : *Let  $\mathbf{K}(z) = \sum_{j \geq 0} \mathbf{K}_j z^j$ . Then  $\mathbf{K}(z) = \mathbf{K}(1)z + (1 - z)\mathbf{L}(z)$  where  $\mathbf{L}(z) = \sum_{j \geq 0} \mathbf{L}_j z^j$ ,  $\mathbf{L}_0 = \mathbf{K}_0$ ,  $\mathbf{L}_j = -\sum_{i \geq j+1} \mathbf{K}_i$ ,  $j = 1, \dots$ . Furthermore, if  $\sum_{j \geq 0} j^p \|\mathbf{K}_j\| \leq \infty$  then  $\sum_{j \geq 0} j^{p-1} \|\mathbf{L}_j\| \leq \infty$ .*

The result is obtained by writing the identity  $\mathbf{K}(z) \equiv \mathbf{K}(1)z + \sum_{j \geq 0} \mathbf{K}_j(z^j - z)$  and then rearranging terms using the telescoping sum  $z^j - z = \sum_{i=1}^{j-1} (z^{i+1} - z^i)$ ,  $j \geq 2$ . See Phillips and Solo (1992), who provide an interesting example of the use of this decomposition in a rather different context. Applied to (4.1) the proposition leads to the *EC* representation

$$\tilde{\Psi}(\mathcal{L})\Delta \mathbf{z}_t + \mathbf{\Pi} \mathbf{z}_{t-1} = \Theta(\mathcal{L})\boldsymbol{\varepsilon}_t, \quad t = 1, \dots, T, \quad (4.2)$$

with the same initial conditions where  $\mathbf{\Pi} = \Psi_0 + \Psi_1 + \dots + \Psi_p$  and  $\tilde{\Psi}(z) = \tilde{\Psi}_0 + \tilde{\Psi}_1 z + \dots + \tilde{\Psi}_{p-1} z^{p-1}$  with  $\tilde{\Psi}_0 = \Psi_0$  and  $\tilde{\Psi}_i = -(\Psi_{i+1} + \dots + \Psi_p)$ ,  $i = 1, \dots, p-1$ . The only term that involves potentially integrated variables in levels is  $\mathbf{\Pi} \mathbf{z}_{t-1}$  and it is the coefficient matrix  $\mathbf{\Pi}$  that summarises the cointegrating relations.

Consider now the identification of (4.2). Suppose that the original *ARMA* system in (4.1) is expressed in inverse echelon canonical form and let  $\boldsymbol{\alpha}_p = \text{vec}[\Psi_0, \dots, \Psi_p, \Theta_0, \dots, \Theta_p]$ . Then conditions (i')–(iii') of Theorem (3.1) can be expressed in the form of the imposition of linear constraints  $\mathbf{R}_{\{n_1, \dots, n_s\}} \boldsymbol{\alpha}_p = \mathbf{r}_{\{n_1, \dots, n_s\}}$  where  $\mathbf{R}_{\{n_1, \dots, n_s\}}$ ,  $d \times s^2 p$ , and  $\mathbf{r}_{\{n_1, \dots, n_s\}}$ ,  $d \times 1$ ,  $d = 2s^2(p+1) - (\sum \sum_{i < j} \{\min(n_i, n_j) + \min(n_i, n_j + 1)\} + (s+1) \sum_i n_i)$ , are known. In addition, the *EC* representation is obtained via the parametric transformation

$$[\tilde{\Psi}_0, \dots, \tilde{\Psi}_{p-1}, \mathbf{\Pi}, \Theta_0, \dots, \Theta_p] = [\Psi_0, \dots, \Psi_p, \Theta_0, \dots, \Theta_p] \begin{bmatrix} \mathbf{S} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{s(p+1)} \end{bmatrix}$$

where  $\mathbf{S}$  is the  $s(p+1) \times s(p+1)$  matrix

$$\begin{bmatrix} \mathbf{I}_s & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \cdots & \cdots & \mathbf{I}_s \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \cdots & \cdots & \mathbf{I}_s \\ \mathbf{0} & -\mathbf{I}_s & \mathbf{0} & \mathbf{0} & \cdots & \cdots & \cdots & \mathbf{I}_s \\ \mathbf{0} & -\mathbf{I}_s & -\mathbf{I}_s & \mathbf{0} & \cdots & \cdots & \cdots & \mathbf{I}_s \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \mathbf{0} & -\mathbf{I}_s & -\mathbf{I}_s & -\mathbf{I}_s & \cdots & -\mathbf{I}_s & \mathbf{0} & \mathbf{I}_s \\ \mathbf{0} & -\mathbf{I}_s & -\mathbf{I}_s & -\mathbf{I}_s & \cdots & -\mathbf{I}_s & -\mathbf{I}_s & \mathbf{I}_s \end{bmatrix}.$$

This implies that

$$\tilde{\boldsymbol{\alpha}}_p = \text{vec}[\tilde{\boldsymbol{\Psi}}_0, \dots, \tilde{\boldsymbol{\Psi}}_{p-1}, \mathbf{\Pi}, \boldsymbol{\Theta}_0, \dots, \boldsymbol{\Theta}_p] = \left( \begin{bmatrix} \mathbf{S}' & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{s(p+1)} \end{bmatrix} \otimes \mathbf{I}_s \right) \boldsymbol{\alpha}_p$$

and hence that  $\tilde{\mathbf{R}}_{\{n_1, \dots, n_s\}} \tilde{\boldsymbol{\alpha}}_p = \mathbf{r}_{\{n_1, \dots, n_s\}}$  where

$$\tilde{\mathbf{R}}_{\{n_1, \dots, n_s\}} = \mathbf{R}_{\{n_1, \dots, n_s\}} \left( \begin{bmatrix} \mathbf{S}' & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{s(p+1)} \end{bmatrix} \otimes \mathbf{I}_s \right).$$

Since the mapping from  $\boldsymbol{\alpha}_p$  to  $\tilde{\boldsymbol{\alpha}}_p$  is one-to-one it follows that the original *ARMA* system in (4.1) satisfies the constraints  $\mathbf{R}_{\{n_1, \dots, n_s\}} \boldsymbol{\alpha}_p = \mathbf{r}_{\{n_1, \dots, n_s\}}$  if and only if the *EC* representation in (4.2) satisfies  $\tilde{\mathbf{R}}_{\{n_1, \dots, n_s\}} \tilde{\boldsymbol{\alpha}}_p = \mathbf{r}_{\{n_1, \dots, n_s\}}$ , and the identification of one implies that of the other.

Examining these restrictions in detail we find that for any nonzero Kronecker index  $n_r < p, r = 1, \dots, s$ ,  $\psi_{rc,j}$  is freely varying for  $j = 1, \dots, n_r$ , but  $\psi_{rc,j} = 0, j = n_r + 1, \dots, p, c = 1, \dots, s$ . It follows that  $\tilde{\psi}_{rc,j}$  is freely varying for  $j = 1, \dots, n_r - 1$  but  $\tilde{\psi}_{rc,j} = 0, j = n_r, \dots, p - 1, c = 1, \dots, s$ . If  $n_r = p$  then the  $\psi_{rc,j}, j = 1, \dots, n_r, c = 1, \dots, s$ , are all freely varying and the same is true for  $\tilde{\psi}_{rc,j}, j = 1, \dots, n_r - 1, c = 1, \dots, s$ . Thus the restrictions  $\tilde{\mathbf{R}}_{\{n_1, \dots, n_s\}} \tilde{\boldsymbol{\alpha}}_p = \mathbf{r}_{\{n_1, \dots, n_s\}}$  incorporate the conditions that  $\delta_r[\tilde{\boldsymbol{\Psi}}(z)] = n_r - 1, r = 1, \dots, s$ . We also have that  $\tilde{\boldsymbol{\Psi}}_0 = \boldsymbol{\Theta}_0$  and  $\boldsymbol{\Theta}(z)$  is subject to the same restrictions in  $\tilde{\mathbf{R}}_{\{n_1, \dots, n_s\}} \tilde{\boldsymbol{\alpha}}_p = \mathbf{r}_{\{n_1, \dots, n_s\}}$  as in  $\mathbf{R}_{\{n_1, \dots, n_s\}} \boldsymbol{\alpha}_p = \mathbf{r}_{\{n_1, \dots, n_s\}}$ . Thus  $\boldsymbol{\Theta}(z)$  presents as the *MA* operator of an inverse echelon canonical form with Kronecker indices  $n_r, r = 1, \dots, s$  in both (4.1) and (4.2).<sup>3</sup>

There are  $s^2$  fewer parameters in  $\tilde{\boldsymbol{\Psi}}(z)$  and  $\boldsymbol{\Theta}(z)$  than in the original pair  $[\boldsymbol{\Psi}(z) : \boldsymbol{\Theta}(z)]$  and the degrees of freedom so released are taken up by the elements of  $\mathbf{\Pi}$ . Thus far  $\mathbf{\Pi}$  remains unrestricted. By assumption, however,  $\det \boldsymbol{\Psi}(z) = \psi(z)(1-z)^\zeta$  where  $\psi(z)$  is stable and  $\zeta < s$  and it is well known (see, *inter alia*, Yap and Reinsel, 1995, Section 2) that the rank of  $\mathbf{\Pi} = \boldsymbol{\Psi}(1)$  equals  $s - \zeta$ . Consequently, additional restrictions must be applied over and above those given by  $\tilde{\mathbf{R}}_{\{n_1, \dots, n_s\}} \tilde{\boldsymbol{\alpha}}_p = \mathbf{r}_{\{n_1, \dots, n_s\}}$  if we require a particular cointegrating rank  $\varrho = s - \zeta$  to hold.

In order to link the identification of  $\mathbf{\Pi}$  to that of the original specification note that if  $\text{rank}(\mathbf{\Pi}) = \varrho$  then  $\mathbf{\Pi} = \mathbf{F}\mathbf{G}'$ , where  $\mathbf{F}$  and  $\mathbf{G}$  are  $(s \times \varrho)$  matrices with full column rank. To

ensure a one-to-one correspondence between  $\mathbf{\Pi}$  and its reduced rank factorisation we follow a standard procedure. Since the columns of  $\mathbf{G}$  are linearly independent there exists a nonsingular  $(\varrho \times \varrho)$  matrix  $\mathbf{E}$ , constructed from a sequence of elementary column transformations, such that  $\mathbf{\Gamma} = \mathbf{G}\mathbf{E}$  is column equivalent to  $\mathbf{G}$  and in reduced column-echelon form. Post-multiplying  $\mathbf{F}$  by  $\mathbf{E}^{-1}$  leads to an identified pair  $[\mathbf{\Upsilon} : \mathbf{\Gamma}]$  with  $\mathbf{\Upsilon} = \mathbf{F}\mathbf{E}^{-1}$ . The condition that  $\text{rank}(\mathbf{\Pi}) = \varrho$  is thereby obtained by imposing an additional  $\varrho^2$  constraints on  $\mathbf{\Gamma}$  and leaving the elements of  $\mathbf{\Upsilon}$  unconstrained.

Let (4.2) denote an *EC* representation in which the operator pair  $[\tilde{\Psi}(z) : \Theta(z)]$  are in inverse echelon canonical form but with the added restrictions  $\delta_r[\tilde{\Psi}(z)] = n_r - 1$ , where  $n_r > 0$ ,  $r = 1, \dots, s$ , imposed, and  $\mathbf{\Pi} = \mathbf{\Upsilon}\mathbf{\Gamma}'$  where  $\mathbf{\Upsilon}$  and  $\mathbf{\Gamma}$  are  $(s \times \varrho)$  matrices with full column rank and  $\mathbf{\Gamma}$  is in reduced column-echelon form. Then the upshot of the preceding argument is that the structure in (4.2) is identified and is equivalent to an inverse echelon canonical form *ARMA* representation (4.1) in which the cointegrating rank  $\varrho$  has been imposed.

If any of the Kronecker indices are zero,  $n_q = 0$ ,  $1 \leq q \leq s$ , say, then the  $q$ th row of  $\tilde{\Psi}(z)$  and  $\mathbf{\Pi}$  are equal and hence the  $q$ th row of (4.1) and (4.2) are identical. From the structure of the echelon form this implies that the variable  $z_{q,t}$  can be expressed as a contemporaneous linear combination of the remaining variables in the system, and the innovations, and therefore it will inherit all its dynamics from these other variables. Hence  $z_{q,t}$  must be either  $I(0)$  or cointegrated with some of the other variables so that  $\varrho \geq 1$ . More generally, assume that the variables have been permuted such that the system is represented in terms of the Kronecker invariants. If  $n_{r(s)} = \dots = n_{r(s-q+1)} = 0$  and  $n_{r(j)} \geq 1$ ,  $j = 1, \dots, s - q$ , then some relatively straightforward manipulations indicate that

$$\mathbf{\Pi} = \begin{bmatrix} \mathbf{\Pi}_{11} & \mathbf{\Pi}_{10} \\ \mathbf{\Pi}_{01} & \mathbf{I}_q \end{bmatrix}$$

where  $\mathbf{\Pi}_{11}$ ,  $\mathbf{\Pi}_{10}$  and  $\mathbf{\Pi}_{01}$  are  $((s - q) \times (s - q))$ ,  $((s - q) \times q)$  and  $(q \times (s - q))$  coefficient matrices respectively. Obviously the rank of this matrix is at least  $q$ , so  $\varrho \geq q$ .

There are two basic conclusions to be drawn from the above analysis: First, that the conditions for identifying the short-run dynamics in an *EC* inverse echelon canonical form, conditions (*EC i'*)–(*EC iii'*) of Theorem 4.2 below, can be separated from those that identify the long-run relationships, (*EC iv'*)–(*EC v'*) of Theorem 4.2. Moreover, the long-run relationships can be present amongst any of the variables in  $\mathbf{z}_t$ . Second, that specifying an original dynamic system in which one or more of the Kronecker indices are zero amounts to a presumption that cointegration is present, the static equations corresponding to the zero indices representing the associated long-run equilibrium relationships assumed. On the whole it seems unlikely that preconditions of the latter type will be imposed *a-priori* in a pure time series setting as the data is usually left to speak for itself and such precise information is rarely available. On the other hand, for some relatively simple economic models the short-run dynamics and long-run relationships can be written down directly, see Wickens and Breusch (1988) for example.

More generally, if a dynamic simultaneous equations perspective is taken then conclusions about the cointegrating structure are often implicit in the model formulation and the assumptions made. In the notation of the current section the dynamic structural equation (1.2) corresponds to the specialisation  $\Psi_{11}(z) = \mathbf{A}(z)$ ,  $\Psi_{12}(z) = \mathbf{B}(z)$ ,  $\Theta_{11}(z) = \mathbf{M}(z)$  and  $\Theta_{12}(z) = \mathbf{0}$  where  $\Psi_{ij}(z)$  and  $\Theta_{ij}(z)$   $i, j = 1, 2$  denote partitions of the  $s \times s$  operators  $\Psi(z)$  and  $\Theta(z)$  into the first  $v$  and last  $u$  rows and columns. Adding the restrictions that  $\Psi_{21}(z) = \Theta_{21}(z) = \mathbf{0}$  and  $\Sigma_{(\varepsilon, \eta)} = \text{diag}(\Sigma_\varepsilon, \Sigma_\eta)$  amounts to the imposition of the assumption that  $\mathbf{x}_t$  is strictly exogenous, see Engle, Hendry, and Richard (1983). If we also suppose that  $\Psi_{22}(z) = \mathbf{D}(z)(1 - z)$  where  $\mathbf{D}(z)$  is stable then

$$\Pi = \begin{bmatrix} \mathbf{A}(1) & \mathbf{B}(1) \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$$

and  $\varrho \leq v$ . But following the argument employed by (Hsiao, 1997, p. 653) we find that the possibility that  $\varrho < v$  is ruled out, for otherwise there would exist a nonzero vector  $\mathbf{v}$  such that  $\mathbf{v}'\mathbf{A}(1) = \mathbf{0}$ , implying that  $\mathbf{v}'\mathbf{B}(1)\mathbf{x}_t$  is asymptotically-stationary, contradicting the assumption that  $\mathbf{x}_t = \sum_{s=1}^t \mathbf{u}_s + \mathbf{x}_0$ ,  $t = 1, \dots, T$ , where  $\mathbf{D}(\mathcal{L})\mathbf{u}_t = \Theta_{22}(\mathcal{L})\boldsymbol{\eta}_t$ . Thus  $\varrho = v$ ,

$$\Pi = \begin{bmatrix} \mathbf{A}(1) \\ \mathbf{0} \end{bmatrix} [\mathbf{I}_v : \mathbf{A}(1)^{-1}\mathbf{B}(1)] ,$$

and, following Wickens (1996), the variables in  $\mathbf{x}_t$  may be regarded as the common trends of Stock and Watson (1988).

The strength of the previous conclusion, which corresponds to that drawn by Hsiao (1997), obviously depends critically on the stringency of the assumptions. If the previous coefficient conditions are relaxed by replacing  $\Psi_{21}(z) = \mathbf{0}$  by  $\Psi_{21,0} = \mathbf{0}$  so that  $\Psi_{21}(z) = \Psi_{21,1}z^1 + \dots + \Psi_{21,p}z^p$  and the requirement that  $\Psi_{22}(z) = \mathbf{D}(z)(1 - z)$  is dropped then a range of different possibilities for the cointegrating structure are possible. In particular, the dichotomy between the endogenous and exogenous variables and that between the short-run dynamics and long-run equilibrium relationships need no longer coincide. Such a system, in which  $\mathbf{x}_t$  exhibits feedback but is weakly exogenous, might be appropriate where  $\mathbf{x}_t$  contains policy instruments determined via a partial adjustment process or control variates determined via a linear-quadratic control rule, for example. For a discussion of causality and feedback in the context of non-stationary processes and exogeneity see Hosoya (1977) and Geweke (1984) respectively.

## 4.2 The Error Correction Echelon Form

In the light of the previous discussion, the most general result concerning the canonical structure of our original partially nonstationary ARMAX model is as follows:

**Theorem 4.2** *Let*

$$\sum_{j=0}^{p-1} \tilde{\mathbf{D}}_j \Delta \mathbf{z}_{t-j} + \Pi \mathbf{z}_{t-1} = \sum_{j=0}^p \mathbf{M}_j \varepsilon_{t-j}, \quad t = 1, \dots, T, \quad (4.3)$$

denote an EC representation in which  $\mathbf{z}_t = [\mathbf{y}'_t : \mathbf{x}'_t]'$ ,  $\tilde{\mathbf{D}}_0 = \mathbf{D}_0$ ,  $\tilde{\mathbf{D}}_i = -(\mathbf{D}_{i+1} + \dots + \mathbf{D}_p)$ ,  $i = 1, \dots, p-1$ , and  $\mathbf{\Pi} = \mathbf{D}_0 + \mathbf{D}_1 + \dots + \mathbf{D}_p$ , where  $\mathbf{D}(z) = [\mathbf{A}(z) : \mathbf{B}(z)]$  and the polynomial operators in  $[\tilde{\mathbf{D}}(z) : \mathbf{\Pi} : \mathbf{M}(z)]$  satisfy the conditions

$$\begin{aligned} (EC \text{ i}') \quad & \tilde{a}_{rc,0} = m_{rc,0}, \quad r, c = 1, \dots, v, \\ (EC \text{ ii}') \quad & m_{rr}(z) = 1 + m_{rr,1}z + \dots + m_{rr,n_r}z^{n_r}, \\ & m_{rc}(z) = m_{rc,n_r-n_{rc}+1}z^{n_r-n_{rc}+1} + \dots + m_{rc,n_r}z^{n_r} \quad r, c = 1, \dots, v, \\ (EC \text{ iii}') \quad & \tilde{a}_{rc}(z) = \tilde{a}_{rc,0} + \tilde{a}_{rc,1}z + \dots + \tilde{a}_{rc,n_r}z^{n_r-1} \quad r, c = 1, \dots, v \text{ and} \\ & \tilde{b}_{rc}(z) = \tilde{b}_{rc,0} + \tilde{b}_{rc,1}z + \dots + \tilde{b}_{rc,n_r}z^{n_r-1} \quad r = 1, \dots, v \quad c = 1, \dots, u. \end{aligned}$$

where  $n_r > 0$ ,  $r = 1, \dots, v$ , and  $\mathbf{\Pi} = \mathbf{\Upsilon}\mathbf{\Gamma}'$  where

$$\begin{aligned} (EC \text{ iv}') \quad & \mathbf{\Upsilon} \text{ and } \mathbf{\Gamma} \text{ are } (v \times \varrho) \text{ and } ((v+u) \times \varrho) \text{ matrices with full column rank and} \\ (EC \text{ v}') \quad & \mathbf{\Gamma} \text{ is in reduced column-echelon form.} \end{aligned}$$

Then under Assumptions 2.1, 2.2 and 2.3 the structure in (4.3) is identified and is equivalent to an inverse echelon canonical form ARMAX representation in which the cointegrating rank  $\varrho \leq v$  has been imposed.

It is the system in Theorem (4.2) that has previously been christened an  $ECARMAX_E$  form.

In order to prevent a proliferation of notation the same symbolism is employed in Theorem (4.2) for the cointegrating relationships as was adopted above, namely  $\mathbf{\Pi} = \mathbf{F}\mathbf{G}' = \mathbf{\Upsilon}\mathbf{\Gamma}'$ , only now  $\mathbf{\Upsilon} = \mathbf{F}\mathbf{E}'$  and  $\mathbf{\Gamma} = \mathbf{G}\mathbf{E}$  are  $(v \times \varrho)$  and  $((v+u) \times \varrho)$  matrices with full column rank. The reduced column-echelon form is a mathematical artifact often employed in matrix algebra that serves here to solve the statistical identification problem. In such a matrix the first nonzero entry in any column is unity and appears below the first nonzero entry in the preceding column. All other entries in the same row as the first nonzero entry in any column are zero. If, after suitable permutation denoted by  $\mathbf{R}\mathbf{G}$ , the first  $\varrho$  rows of  $\mathbf{G}$  are linearly independent then the reduced column-echelon form becomes

$$\mathbf{R}\mathbf{G}\mathbf{E} = \mathbf{R}\mathbf{\Gamma} = \begin{bmatrix} \mathbf{I}_\varrho \\ \mathbf{\Gamma}_\varrho \end{bmatrix}.$$

This gives the triangular structure  $\mathbf{\Gamma}'\mathbf{z}_{t-1} = [\mathbf{I}_\varrho : \mathbf{\Gamma}'_\varrho]\mathbf{R}'\mathbf{z}_{t-1}$  introduced by Phillips (1991). For the reduced column-echelon form  $\mathbf{\Gamma}$  any arrangement of the variables is permitted and this allows the system to be ordered according to the permutation induced by the Kronecker invariants, which need not coincide with the reordering implicit in the triangular structure.

## 5 Parameter Estimation

Given that the  $ECARMAX_E$  form is identified we are now interested in estimating the unknown parameters in  $\boldsymbol{\lambda}$ . Suppose that  $\varrho$  and  $n_{r(i)}$ ,  $i = 1, \dots, v$ , are given and that the variables and equation system have been ordered according to the permutation  $r(1), \dots, r(v)$  induced by the Kronecker invariants to give the unique invariant form. Let  $\{\boldsymbol{\lambda} : \varrho, (n_{r(1)}, \dots, n_{r(v)})\}$  denote the

set of parameter values  $\boldsymbol{\lambda} = (\boldsymbol{\beta}' : \boldsymbol{\sigma}')'$  such that  $\boldsymbol{\beta}$  satisfies conditions (EC i')–(EC v') and let  $\tilde{\boldsymbol{\lambda}}_T$  denote the maximum likelihood estimator. By definition  $\tilde{\boldsymbol{\lambda}}_T$  is the value of  $\boldsymbol{\lambda}$  that maximises  $L_T(\boldsymbol{\lambda}) = \log\{f(\mathbf{y}_1^T | \mathbf{y}_{1-p}^0, \mathbf{x}_{1-p}^T; \boldsymbol{\lambda})\}$  over  $\{\boldsymbol{\lambda} : \varrho, (n_{r(1)}, \dots, n_{r(v)})\}$ ,

$$\tilde{\boldsymbol{\lambda}}_T = \arg \max_{\boldsymbol{\lambda} \in \{\boldsymbol{\lambda} : \varrho, (n_{r(1)}, \dots, n_{r(v)})\}} L_T(\boldsymbol{\lambda}) .$$

The determination of  $\tilde{\boldsymbol{\lambda}}_T$  will necessitate the use of numerical optimisation techniques and the function evaluations required to implement such numerical methods can be readily computed in practice using the following algorithm.

**Proposition 5.1** *Suppose that  $\mathbf{y}_t$  admits an ARMAX representation as in Theorem 4.2. Then for  $t = 2, \dots, T$  and  $q = \min\{t - 1, p\}$  set*

$$\mathbf{M}_{\langle t, q \rangle} = \boldsymbol{\Gamma}_\xi(q) [\boldsymbol{\Sigma}_{t-q|t-q-1} \mathbf{M}'_0]^{-1} \quad (5.1)$$

and for  $j = q - 1, \dots, 1$ ,

$$\mathbf{M}_{\langle t, j \rangle} = [\boldsymbol{\Gamma}_\xi(j) - \sum_{r=j+1}^q \mathbf{P}_{\langle t, r \rangle} \mathbf{P}'_{\langle t-j, r-j \rangle}] [\boldsymbol{\Sigma}_{t-j|t-j-1} \mathbf{M}'_0]^{-1} \quad (5.2)$$

where  $\mathbf{P}_{\langle t, j \rangle} = \mathbf{M}_{\langle t, j \rangle} \boldsymbol{\Sigma}_{t-j|t-j-1}^{\frac{1}{2}}$  and  $\boldsymbol{\Sigma}_{t|t-1} = \boldsymbol{\Sigma}_{t|t-1}^{\frac{1}{2}} (\boldsymbol{\Sigma}_{t|t-1}^{\frac{1}{2}})'$ ,

$$\boldsymbol{\Sigma}_{t|t-1}^{\frac{1}{2}} = \mathbf{M}_0^{-1} \left[ \boldsymbol{\Gamma}_\xi(0) - \sum_{r=1}^q \mathbf{M}_{\langle t, r \rangle} \boldsymbol{\Sigma}_{t-r|t-r-1} \mathbf{M}'_{\langle t, r \rangle} \right]^{\frac{1}{2}} \quad (5.3)$$

with initial value  $\boldsymbol{\Sigma}_{1|0}^{\frac{1}{2}} = \mathbf{M}_0^{-1} \boldsymbol{\Gamma}_\xi(0)^{\frac{1}{2}}$ . Then the partial log-likelihood function

$$\begin{aligned} L_T(\boldsymbol{\lambda}) &= \log \left( \prod_{t=1}^T f(\mathbf{y}_t | \mathbf{y}_{1-p}^{t-1}, \mathbf{x}_{1-p}^t; \boldsymbol{\lambda}) \right) \\ &= - \sum_{t=1}^T \frac{1}{2} (v \log(2\pi) + \log(\det(\boldsymbol{\Sigma}_{t|t-1})) + \boldsymbol{\varepsilon}'_{\langle t|t-1 \rangle} \boldsymbol{\Sigma}_{t|t-1}^{-1} \boldsymbol{\varepsilon}_{\langle t|t-1 \rangle}) \end{aligned}$$

where

$$\boldsymbol{\varepsilon}_{\langle 1|0 \rangle} = \mathbf{M}_0^{-1} \left[ \sum_{j=0}^{p-1} \tilde{\mathbf{D}}_j \Delta \mathbf{z}_{1-j} + \boldsymbol{\Pi} \mathbf{z}_0 \right]$$

and

$$\boldsymbol{\varepsilon}_{\langle t|t-1 \rangle} = \mathbf{M}_0^{-1} \left[ \sum_{j=0}^{p-1} \tilde{\mathbf{D}}_j \Delta \mathbf{z}_{t-j} + \boldsymbol{\Pi} \mathbf{z}_{t-1} - \sum_{j=1}^q \mathbf{M}_{\langle t, j \rangle} \boldsymbol{\varepsilon}_{\langle t-j|t-j-1 \rangle} \right], \quad t = 2, \dots, T.$$

The recursive calculations given in (5.1)–(5.3) are derived from the finite span Wiener-Hopf equations due to Rissanen and Barbosa (1969) applied to the  $MA(p)$  process  $\boldsymbol{\xi}_t = \sum_{j=0}^p \mathbf{M}_j \boldsymbol{\varepsilon}_{t-j}$ .

They are constructed from the Cholesky factorisation of the covariance matrix of  $(\boldsymbol{\xi}'_1, \dots, \boldsymbol{\xi}'_T)'$  via a Gram-Schmidt orthonormalisation based on  $\boldsymbol{\xi}_{\langle t|t-1 \rangle}$ , the projection of  $\boldsymbol{\xi}_t$  on to the space spanned by  $\boldsymbol{\xi}_{t-1}, \dots, \boldsymbol{\xi}_1$ . This gives a square-root, orthonormal version of what is often referred to as the innovations algorithm. Proposition 5.1 follows by noting that  $\mathbf{M}_{\langle t,0 \rangle} \equiv \mathbf{M}_0$  and  $\boldsymbol{\xi}_{\langle t|t-1 \rangle} = \sum_{j=1}^q \mathbf{M}_{\langle t,j \rangle} \boldsymbol{\varepsilon}_{\langle t-j|t-j-1 \rangle}$ ,  $t = 2, \dots, T$ , and  $\sum_{j=0}^q \mathbf{M}_{\langle t,j \rangle} \boldsymbol{\varepsilon}_{\langle t-j|t-j-1 \rangle} = \sum_{j=0}^p \mathbf{M}_j \boldsymbol{\varepsilon}_{t-j} = \sum_{j=0}^{p-1} \tilde{\mathbf{D}}_j \Delta \mathbf{z}_{t-j} + \mathbf{\Pi} \mathbf{z}_{t-1}$  for  $t = 1, \dots, T$ . The detailed steps of the argument, which can be deduced by consulting the manipulations presented in Rissanen and Barbosa *op. cit.*, are omitted.<sup>1</sup>

It is of interest to observe that if  $\mathbf{M}(z) = \mathbf{M}_0$  then the model in equation (4.3) reduces to

$$\sum_{j=0}^{p-1} \tilde{\mathbf{D}}_j \Delta \mathbf{z}_{t-j} + \mathbf{\Pi} \mathbf{z}_{t-1} = \mathbf{u}_t, \quad t = 1, \dots, T, \quad (5.4)$$

where  $\mathbf{u}_t = \mathbf{M}_0 \boldsymbol{\varepsilon}_t$ , a cointegrated *ARX* structure with row degrees  $n_{r(i)}$ ,  $i = 1, \dots, v$ . In this case it is relatively straightforward to verify that the recursions in (5.1)–(5.3) yield  $\mathbf{M}_{\langle t,j \rangle} = \mathbf{0}$ ,  $1 \leq j \leq q$ , and  $\boldsymbol{\Sigma}_{t|t-1} = \boldsymbol{\Sigma}_\varepsilon$  for all  $t = 1, \dots, T$ . We are thereby lead to the conclusion that  $\boldsymbol{\varepsilon}_{\langle t|t-1 \rangle} = \boldsymbol{\varepsilon}_t$  and that the partial log-likelihood function can be reexpressed as

$$L_T(\boldsymbol{\lambda}) = -\frac{Tv}{2} \log(2\pi) - \frac{T}{2} \log(\det \boldsymbol{\Sigma}_u) - \frac{1}{2} \sum_{t=1}^T \mathbf{u}'_t \boldsymbol{\Sigma}_u^{-1} \mathbf{u}_t \quad (5.5)$$

where  $\boldsymbol{\Sigma}_u = \mathbf{M}_0 \boldsymbol{\Sigma}_\varepsilon \mathbf{M}'_0$ , because  $\mathbf{M}_0$  is lower triangular with leading diagonal equal to the identity and hence  $\det \mathbf{M}_0 = 1$ . Expressions (5.4) and (5.5) coincide with those commonly considered in the analysis of cointegrated autoregressive systems, following Johansen (1991).

To investigate the statistical properties of  $\tilde{\boldsymbol{\lambda}}_T$  consider the estimator  $\hat{\boldsymbol{\lambda}}_T$  obtained by maximising the following approximation to the partial log-likelihood function,

$$L_T^a(\boldsymbol{\lambda}) = -\frac{Tv}{2} \log(2\pi) - \frac{T}{2} \log(\det \boldsymbol{\Sigma}_\varepsilon) - \frac{1}{2} \sum_{t=1}^T \boldsymbol{\varepsilon}'_t \boldsymbol{\Sigma}_\varepsilon^{-1} \boldsymbol{\varepsilon}_t \quad (5.6)$$

where  $\boldsymbol{\varepsilon}_t$  for  $t = 1, \dots, T$  are calculated from the recursion

$$\boldsymbol{\varepsilon}_t = \mathbf{M}_0^{-1} \left[ \sum_{j=0}^{p-1} \tilde{\mathbf{D}}_j \Delta \mathbf{z}_{t-j} + \mathbf{\Pi} \mathbf{z}_{t-1} - \sum_{j=1}^p \mathbf{M}_j \boldsymbol{\varepsilon}_{t-j} \right],$$

starting from the initial values  $\boldsymbol{\varepsilon}_t = \mathbf{0}$ ,  $t = 1 - p, \dots, 0$ . Then  $\hat{\boldsymbol{\lambda}}_T = (\hat{\boldsymbol{\beta}}'_T, \hat{\sigma}'_T)'$  where  $\hat{\sigma}_T = \text{vech}[\bar{\boldsymbol{\Sigma}}_T] |_{\hat{\boldsymbol{\beta}}_T}$ ,  $\bar{\boldsymbol{\Sigma}}_T = T^{-1} \sum_{t=1}^T \boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}'_t$ , and  $\hat{\boldsymbol{\beta}}_T$  is given by the solution to the score equations  $\sum_{t=1}^T (\partial \boldsymbol{\varepsilon}'_t / \partial \boldsymbol{\beta}) \bar{\boldsymbol{\Sigma}}_T^{-1} \boldsymbol{\varepsilon}_t = \mathbf{0}$ . These equations can be solved numerically, leading to the iterative Newton-Raphson approximation

$$\hat{\boldsymbol{\beta}}_T^{(i+1)} = \hat{\boldsymbol{\beta}}_T^{(i)} - \left( \sum_{t=1}^T \frac{\partial \boldsymbol{\varepsilon}'_t}{\partial \boldsymbol{\beta}} \bar{\boldsymbol{\Sigma}}_T^{-1} \frac{\partial \boldsymbol{\varepsilon}_t}{\partial \boldsymbol{\beta}'} \right)^{-1} \left( \sum_{t=1}^T \frac{\partial \boldsymbol{\varepsilon}'_t}{\partial \boldsymbol{\beta}} \bar{\boldsymbol{\Sigma}}_T^{-1} \boldsymbol{\varepsilon}_t \right) |_{\hat{\boldsymbol{\beta}}_T^{(i)}} \quad (5.7)$$

where, by definition,  $\hat{\beta}_T^{(i)}$  is the  $i$ th iterate. It is  $\hat{\lambda}_T^{(i)}$  that is commonly referred to as the Gaussian estimator and it is this estimator that is considered by Yap and Reinsel (1995) and Dhrymes (1998). We will now specialise (5.7) to the  $ECARMAX_E$  form.

Rearranging equation (4.3) so as to isolate the common parameters in  $\mathbf{D}_0 = [\mathbf{A}_0 : \mathbf{B}_0]$  and  $\mathbf{M}_0 = \mathbf{A}_0$  gives

$$\mathbf{A}_0(\mathbf{y}_t - \varepsilon_t) + \mathbf{B}_0\mathbf{x}_t + \sum_{j=1}^{p-1} \tilde{\mathbf{D}}_j \Delta \mathbf{z}_{t-j} + \Upsilon \Gamma' \mathbf{z}_{t-1} - \sum_{j=1}^p \mathbf{M}_j \varepsilon_{t-j} = \mathbf{0}. \quad (5.8)$$

Following Poskitt (1992) let us now rewrite the left hand side by vectorising each term. The first two terms become  $(\mathbf{y}_t - \varepsilon_t) + ((\mathbf{y}_t - \varepsilon_t)' : \mathbf{x}_t') \otimes \mathbf{I}_v \zeta$  where  $\zeta = \text{vec}([\mathbf{A}_0 - \mathbf{I} : \mathbf{B}_0])$  and the third is  $-(\xi_{p-1}(\mathcal{L})' \otimes \Delta \mathbf{z}_t' \otimes \mathbf{I}_v) \delta$  where  $\xi_{p-1}(z)' = (z, z^2, \dots, z^{p-1})$  and  $\delta = \text{vec}(\mathbf{D}_1 : \dots : \mathbf{D}_{p-1})$ . The fourth term becomes  $-(\mathbf{z}_{t-1}' \Gamma \otimes \mathbf{I}_v) \mathbf{v}$  where the parameter vector  $\mathbf{v} = \text{vec}(\Upsilon)$  and the fifth term gives  $(\xi_p(\mathcal{L})' \otimes \mathbf{e}_t' \otimes \mathbf{I}_v) \mu$ , where  $\mu = \text{vec}(\mathbf{M}_1 : \dots : \mathbf{M}_p)$ . In the derivation of these expressions the well known rule  $\text{vec}(\mathbf{ABC}) = (\mathbf{C}' \otimes \mathbf{A}) \text{vec} \mathbf{B}$  has been employed. Set  $\theta = (\delta' : \zeta' : \mu' : \mathbf{v})'$ . The vector  $\theta$  contains the parameters of  $\mathbf{D}(z)$ ,  $\mathbf{M}(z)$  and  $\Upsilon$  not restricted to be unity. For any  $ECARMAX_E$  form with  $\sum_{i=1}^v n_{r(i)} \leq vp$  the exclusion constraints implicit in (*EC i'*)–(*EC iii'*) are simply incorporated by deleting the corresponding elements of  $\theta$ . To complete the parameterisation of the model let  $\gamma = \text{vec}(\Gamma_\rho)$  denote the coefficients in  $(\mathbf{R}\Gamma)' = [\mathbf{I}_\rho : \Gamma'_\rho]$  not restricted to be zero or one in the echelon form. This gives us a freely varying parameter vector  $\beta = (\theta', \gamma')'$  of dimension  $d_\beta = d_\theta + d_\gamma$ , where  $d_\theta = (v + u + 1) \sum_{i=1}^v n_{r(i)} - (v + u)v + \sum_{i \neq j} n_{r(i)r(j)} + v\rho$  and  $d_\gamma = \rho(v + u - \rho)$ , such that  $\lambda = (\beta', \sigma')' \in \{\lambda : \rho, (n_{r(1)}, \dots, n_{r(v)})\}$ .

Treating (5.8) as an implicit function of the stochastic disturbance, the data and  $\beta$  and differentiating with respect to  $\beta$  we find that  $\partial \varepsilon_t / \partial \beta' = -\mathbf{W}_t$  where  $\mathbf{W}_t = [\mathbf{W}_{\theta t} : \mathbf{W}_{\gamma t}]$  equals the  $v \times d_\beta$  matrix obtained by selecting the appropriate columns of

$$[\xi_{p-1}(\mathcal{L})' \otimes \mathbf{W}_{\delta t} : (\mathbf{W}_{\delta t} - [\mathbf{W}_{\mu t} : \mathbf{0}]) : -\xi_p(\mathcal{L})' \otimes \mathbf{W}_{\mu t} : \mathbf{W}_{v t} : \mathbf{W}_{\gamma t}]$$

where the  $v \times v(v + u)$ ,  $v \times v^2$ ,  $v \times \rho v$  and  $v \times \rho(v + u - \rho)$  derivative processes  $\mathbf{W}_{\delta t}$ ,  $\mathbf{W}_{\mu t}$ ,  $\mathbf{W}_{v t}$  and  $\mathbf{W}_{\gamma t}$  are generated from

$$\sum_{j=0}^p \mathbf{M}_j [\mathbf{W}_{\delta(t-j)} : \mathbf{W}_{\mu(t-j)} : \mathbf{W}_{v(t-j)} : \mathbf{W}_{\gamma(t-j)}] = [\Delta \mathbf{z}_t' \otimes \mathbf{I}_v : \mathbf{e}_t' \otimes \mathbf{I}_v : \mathbf{z}_{t-1}' \Gamma \otimes \mathbf{I}_v : \Upsilon \otimes \mathbf{z}_{t-1}' \mathbf{H}],$$

$\mathbf{H}' = [\mathbf{0} : \mathbf{I}_{(v+u-\rho)}] \mathbf{R}'$ . The Gaussian iterations of (5.7) can then be expressed as

$$\hat{\beta}_T^{(i+1)} = \left( \sum_{t=1}^T \mathbf{W}_t' \bar{\Sigma}_T^{-1} \mathbf{W}_t \right)^{-1} \left( \sum_{t=1}^T \mathbf{W}_t' \bar{\Sigma}_T^{-1} [\mathbf{W}_t \beta + \varepsilon_t] \right) \Big|_{\hat{\beta}_T^{(i)}}. \quad (5.9)$$

Now let  $\tilde{\beta}_T^{(i)}$  denote the iterate obtained by substituting  $\varepsilon_{\langle t|t-1 \rangle}$  for  $\varepsilon_t$ ,  $\partial \varepsilon_{\langle t|t-1 \rangle} / \partial \beta' = -\mathbf{W}_{\langle t|t-1 \rangle}$  for  $-\mathbf{W}_t$  and  $\Sigma_{\langle t|t-1 \rangle}$  for  $\bar{\Sigma}_T$  in (5.9). On convergence  $\tilde{\beta}_T^{(i)}$  will yield a critical point of the partial likelihood, but to ensure that the iterates  $\tilde{\beta}_T^{(i)}$  will converge to  $\tilde{\beta}_T$  when  $T$  is

sufficiently large the iterations must be initiated using a consistent estimator. Suppose that the iterations commence at  $\tilde{\beta}_T^{(0)} = \hat{\beta}_T^{(0)} = \bar{\beta}_T$  where  $\bar{\beta}_T$  denotes a preliminary estimator chosen such that  $\mathbf{N}_T(\bar{\beta}_T - \beta) = O_p(1)$ ,  $\mathbf{N}_T = \text{diag}[T^{\frac{1}{2}}\mathbf{I}_{d_\theta} : T\mathbf{I}_{d_\gamma}]$ . Then we have the following result.

**Theorem 5.1** *If Assumptions 2.1, 2.2 and 2.3 hold and, in addition, the operator  $\mathbf{M}(z)$  is invertible, then the iterates  $\tilde{\beta}_T^{(i)}$  and  $\hat{\beta}_T^{(i)}$  obtained using the initial value  $\bar{\beta}_T$  are asymptotically equivalent in the sense that  $\|\mathbf{N}_T(\tilde{\beta}_T^{(i)} - \hat{\beta}_T^{(i)})\| = o_p(1)$  for all  $i \geq 1$  as  $T \rightarrow \infty$ . Moreover, both estimators converge in distribution to  $\beta_T^a = (\theta_T^a, \gamma_T^a)'$  where:*

$$\sqrt{T}(\theta_T^a - \theta) \xrightarrow{\mathcal{D}} \mathbf{N}(\mathbf{0}, \mathbf{V}_\theta^{-1})$$

where

$$\mathbf{V}_\theta = \lim_{T \rightarrow \infty} T^{-1} E \left[ \sum_{t=1}^T \mathbf{W}'_{\theta t} \Sigma_\varepsilon^{-1} \mathbf{W}_{\theta t} \right],$$

the asymptotic information matrix for  $\theta$ ; the vectors  $\sqrt{T}(\theta_T^a - \theta)$  and  $T(\gamma_T^a - \gamma)$  are asymptotically mutually uncorrelated; and  $T(\gamma_T^a - \gamma) = T \text{vec}(\bar{\Gamma}_{\rho, T}^a - \Gamma_\rho)$  where the components of  $(\bar{\Gamma}_{\rho, T}^a - \Gamma_\rho)$  satisfy the asymptotic mixed-normality result

$$\text{vec} \left( \left[ \sum_{t=1}^T \mathbf{H}' \mathbf{z}_{t-1} \mathbf{z}'_{t-1} \mathbf{H} \right]^{1/2} [\bar{\Gamma}_{\rho, T}^a - \Gamma_\rho] \right) \xrightarrow{\mathcal{D}} \mathbf{N}(\mathbf{0}, \mathbf{V}_\gamma)$$

where

$$\mathbf{V}_\gamma = ((\Upsilon' [\mathbf{M}(1) \Sigma_\varepsilon \mathbf{M}(1)']^{-1} \Upsilon)^{-1} \otimes \mathbf{I}_{(v+u-\rho)}).$$

When applying Theorem 5.1  $\mathbf{V}_\theta$  and  $\mathbf{V}_\gamma$  can be consistently estimated by replacing the unknown parameters in  $\beta$  by their Gaussian maximum likelihood values and substituting a consistent estimate for  $\Sigma_\varepsilon$ , whilst dropping the expectation from  $\mathbf{V}_\theta$ . Two simple estimates of  $\Sigma_\varepsilon$  that can be constructed as a by-product of the output from the algorithm in Proposition 5.1 are  $\tilde{\Sigma}_{1, T} = T^{-1} \sum_{t=1}^T \varepsilon_{\langle t|t-1 \rangle} \varepsilon'_{\langle t|t-1 \rangle}$  and  $\tilde{\Sigma}_{2, T} = T^{-1} \sum_{t=1}^T \Sigma_{t|t-1}$ .

**Corollary 5.1** *Both  $\tilde{\Sigma}_{1, T}$  and  $\tilde{\Sigma}_{2, T}$  are consistent estimators of  $\Sigma_\varepsilon$ . Moreover,  $\sqrt{T} \text{vech}(\tilde{\Sigma}_{1, T} - \Sigma_\varepsilon)$  and  $\sqrt{T} \text{vech}(\tilde{\Sigma}_{2, T} - \Sigma_\varepsilon)$  have the same limiting distribution as  $\sqrt{T} \text{vech}(T^{-1} \sum_{t=1}^T \varepsilon_t \varepsilon'_t - \Sigma_\varepsilon)$ .*

Finally, the initial estimator  $\bar{\beta}_T$  must be chosen. One such estimate can be obtained by first using instrumental variables with instruments chosen from  $\mathbf{y}_{t-\tau}$ ,  $\mathbf{x}_{t-\tau}$  and  $\mathbf{z}_{t-\tau-1}$  plus  $\Delta \mathbf{z}_{t-\tau-1}, \dots, \Delta \mathbf{z}_{t-\tau-p}$ , for  $\tau > p$ , to calculate  $\bar{\zeta}_T$ ,  $\bar{\nu}_T$  and  $\bar{\gamma}_T$ , and  $\bar{\delta}_T$ , respectively. From these estimates  $\bar{\xi}_{T, t}$ , to use an obvious notation, can be generated and then  $\bar{\mu}_T$  can be evaluated using the spectral factorization method of Tunnicliffe-Wilson (1972), as suggested by Yap and Reinsel (1995, Section 3.3) and Dhrymes (1998, pp. 325-326). Alternatively,  $\bar{\mu}_T$  can be obtained by implementing the technique proposed by Brockwell and Davis (1988) following a preliminary pass through Proposition 5.1 with  $\Gamma_\xi(\tau)$  replaced by  $\bar{\Gamma}_{\xi, T}(\tau) = T^{-1} \sum_{t=1}^{T-\tau} \bar{\xi}_{T, t} \bar{\xi}'_{T, t+\tau}$ .

**Example:(ii')** In order to provide some indication of the possible impact of the results outlined above a sequence of Monte-Carlo experiments have been conducted using the data generating mechanism  $\mathbf{y}_t + \mathbf{A}\mathbf{y}_{t-1} = \boldsymbol{\varepsilon}_t + \mathbf{M}\boldsymbol{\varepsilon}_{t-1}$ ,  $t = 1, \dots, T$ , where the pair  $[\mathbf{A} : \mathbf{M}]$  accord with the structure considered in Example (ii), namely  $\mathbf{A} = \mathbf{M} = \mathbf{C}$ . The behaviour of the maximum likelihood and Gaussian estimates  $[\tilde{\mathbf{A}}_T : \tilde{\mathbf{M}}_T]$  and  $[\hat{\mathbf{A}}_T : \hat{\mathbf{M}}_T]$  constructed from a single iteration of the Newton-Raphson (Scoring) algorithm initiated at  $[\bar{\mathbf{A}}_T : \bar{\mathbf{M}}_T]$  has been monitored. The behaviour of  $[\tilde{\mathbf{A}}_T : \tilde{\mathbf{M}}_T]$  and  $[\hat{\mathbf{A}}_T : \hat{\mathbf{M}}_T]$  was examined by calculating the observed mean-squared error, the results of which are summarised by presenting the empirical mean and the average squared Euclidean distance between the estimate and the true value  $[\mathbf{C} : \mathbf{C}]$ , labelled *M.S.E.* in the tables that follow.

In all cases  $v = 2$ ,  $n_1 = n_2 = 1$  and the coefficient values were given by either

$$\mathbf{C} = \begin{bmatrix} -0.5 & 0.5 \\ -0.4 & 0.6 \end{bmatrix} \text{ or } \mathbf{C} = \begin{bmatrix} -0.6 & 0.4 \\ 1.4 & 0.4 \end{bmatrix}.$$

In the first case, process *P1*, the zeroes of  $\det(\mathbf{I} + \mathbf{C}z)$  are  $1.1 \pm i0.8888$  and in the second, process *P2*,  $\zeta_1 = -1.25$  and  $\zeta_2 = 1.0$ . The covariance matrix

$$\boldsymbol{\Sigma}_\varepsilon = \begin{bmatrix} 1.0 & \rho_\varepsilon \\ \rho_\varepsilon & 1.0 \end{bmatrix} \text{ where } \rho_\varepsilon = 0.8 \text{ or } 0.2$$

and the sample sizes considered were  $T = 50, 100, 150$  and  $250$ . All simulation results listed here were based on 1000 replications.

The influence of the initial conditions was controlled by setting  $\mathbf{d} = (1, 1)'d$  where the scalar  $d$  was chosen so as to make  $\rho_{y,m} = \det(\mathbf{R}_T) / \det(\mathbf{R}_T + \boldsymbol{\Sigma}_\varepsilon)$  equal to  $0.1Q_T$ ,  $Q_T = (\log(T))^{1/2}l_2(T)$ , for *P1*, and  $0.1$  for *P2*. The rationale for this follows from observing that for both processes  $\rho_{y,m}$  is the (asymptotic) coefficient of vector correlation (Hotelling, 1936) between  $\mathbf{y}_t$  and  $\mathbf{m}_t$ . Process *P1* is asymptotically-stationary and multiplication by the factor  $Q_T$  ensures that the influence of the initial conditions does not die away too quickly, so that  $\mathbf{R}_T \rightarrow \mathbf{0}$  as  $T \rightarrow \infty$  but  $\mathbf{R}_T/l_2(T) = O((\log(T))^{1/2})$ . For *P2*, of course,  $\mathbf{R}_T$  is  $O(1)$  and such re-scaling is not necessary. For process *P1*  $\rho_{y,m} = 0.1Q_T$  takes the values  $0.0327, 0.0265, 0.0232$  and  $0.0194$  for  $T = 50, 100, 150$  and  $250$ , respectively, indicating that fluctuations in  $\mathbf{m}_t$  account for well below 5% of the observed variation in  $\mathbf{y}_t$  for both processes.

The values presented in Table 1 are typical of those obtained using different parameterisations of process *P1*. The figures indicate a somewhat superior performance for  $[\tilde{\mathbf{A}}_T : \tilde{\mathbf{M}}_T]$ , with the observed relative *M.S.E.*  $\|\tilde{\mathbf{A}}_T - \mathbf{C} : \tilde{\mathbf{M}}_T - \mathbf{C}\|^2/2\|\mathbf{C}\|^2$  decreasing from  $0.3319$  when  $T = 50$  to  $0.0849$  when  $T = 250$  whereas  $\|\hat{\mathbf{A}}_T - \mathbf{C} : \hat{\mathbf{M}}_T - \mathbf{C}\|^2/2\|\mathbf{C}\|^2$  decreases from  $1.2133$  to  $0.255$ . Equivalent results for *P2* are given in Table 2. A striking feature of the figures given in this second table is the relatively poor performance of  $[\hat{\mathbf{A}}_T : \hat{\mathbf{M}}_T]$ . This presumably reflects aspects of the Gaussian approximation that work less well in the presence of a unit root and further emphasises the benefits of determining the maximum likelihood estimate.

Although both  $P1$  and  $P2$  are in a sense pathological processes this is not obvious from the figures reported in Tables 1 and 2. Mean-squared error does not capture the distributional properties of the estimators however. Figure (1) presents kernel density estimates of the distribution of the Mahalanobis distance  $Q_T = (\hat{\beta}_T - \beta)' \left( \sum_{t=1}^T \mathbf{W}_t' \Sigma_\varepsilon^{-1} \mathbf{W}_t \right)^\dagger (\hat{\beta}_T - \beta)$  compared to theoretical  $\chi^2$  densities for process  $P1$ . Process  $P1$  is asymptotically stationary and the specification  $\mathbf{A} = \mathbf{M}$  means that it will also be asymptotically unidentified. Such a lack of identification is well known to manifest itself in  $\sum_{t=1}^T \mathbf{W}_t' \Sigma_\varepsilon^{-1} \mathbf{W}_t$  having less than full rank and  $Q_T$  having fewer degrees of freedom than might be anticipated on the basis of conventional asymptotic theory, see Poskitt and Tremayne (1982) for example. Figure (1) lends clear testimony to this feature. For  $P2$  both  $\mathbf{A}$  and  $\mathbf{M}$  are identified and the estimates  $\tilde{\beta}_T^{(i)}$  and  $\hat{\beta}_T^{(i)}$  can be readily calculated but Theorem 5.1 is not applicable since the invertibility condition is violated. It seems reasonable to conjecture that a similar theorem could be established that would allow for the singularity present under the less restrictive condition  $\det(\mathbf{M}(z)) \neq 0, |z| < 1$ , of Assumption 2.1, but that avenue will not be pursued here. See Tanaka (1996) for a discussion of the issues associated with lack of invertibility.  $\square$

## 6 Conclusion

This paper has filled an important gap in the identification theory of nonstationarity vector  $ARMAX$  systems by showing that  $ECARMAX_E$  models provide a canonical form for partially nonstationary (cointegrated)  $ARMAX$  processes. It has established the asymptotic equivalence of the Gaussian estimator  $\hat{\beta}_T$  and the maximum likelihood estimator  $\tilde{\beta}_T$  constructed using an innovations algorithm. It has also established the large sample distribution of both estimators in such models. Examples illustrating the theory and some experimental evidence on the empirical impact of the results have been presented.

The normality assumption underlying the analysis conducted in this paper is commonly adopted in the literature on cointegration. Normality does not play a key role beyond motivating the estimators, however, and it seems likely that the asymptotic properties of  $\hat{\beta}_T$  and  $\tilde{\beta}_T$  can be extended to more general processes under much weaker regularity conditions.

In closing it is worth emphasizing that the identification conditions for an  $ECARMAX_E$  model depend on fixed, finite values of the Kronecker indices,  $n_r, r = 1, \dots, v$ , and are applicable at any sample size  $T > vp$ . This latter point is of significance for any future development of exact finite sample distribution theory for  $\hat{\beta}_T$  and  $\tilde{\beta}_T$  or, perhaps more importantly, Bootstrapping methodology.

### NOTES

<sup>1</sup>Observe that the possibility that  $\mathbf{M}(z)$  is noninvertible is not ruled out. For such a process the minimum mean squared error predictor of  $\xi_t$  based on  $\xi_{t-1}, \dots, \xi_1, \xi_{\lfloor t/2 \rfloor}$ , is well defined and can be evaluated using the algorithm in Proposition 5.1. Since the algorithm is structured in terms of the covariances  $E[\xi_t \xi_{t+\tau}'] = \mathbf{\Gamma}_\xi(\tau)$  it is independent of any assumptions concerning the invertibility of  $\mathbf{M}(z)$ . See Hannan (1974, Chapter III. 2) for a

discussion of the synthesis of  $\xi_{(t|t-1)}$  when  $\det(\mathbf{M}(z)) = 0$  on the unit circle.

<sup>2</sup>A process is said to be asymptotically-stationary, denoted  $I(0)$ , if it admits a representation as in (1.1) or (1.2) where  $\mathbf{A}(z)$  is stable (see Section 2) and the exogenous input is also  $I(0)$ .

<sup>3</sup>The term “freely varying” is used here to indicate that the echelon form imposes no restrictions on the value of the parameter. Stability and minimum phase conditions will, of course, impose separate constraints that will limit the admissible parameter values.

<sup>4</sup>A discussion of a closely related issue, known in the engineering literature as the partial realization problem, can be found in Hanzon (1989).

## References

- Akaike, H. (1974) Stochastic theory of minimal realization. *IEEE Transactions on Automatic Control* AC-19, 667–674.
- Banerjee, M., J. Dolado, J. Galbraith, & D. F. Hendry (1993) *Co-integration, Error-correction and the Econometric Analysis of Non-stationary Data*. Oxford University Press, Oxford.
- Brockwell, P. J. & R. A. Davis (1988) Simple consistent estimation of the coefficients of a linear filter. *Stochastic Processes and Their Applications* 22, 47–59.
- Cox, D. R. (1975) Partial likelihood. *Biometrika* 62, 269–276.
- Deistler, M. (1983) The properties of the parameterisation of ARMAX systems and their relevance for structural estimation and dynamic specification. *Econometrica* 51, 1187–1206.
- Deistler, M. (1985) The general structure and parameterisation of ARMA and State-Space systems and its relation to statistical problems. In E. J. Hannan, P. R. Krishnaiah, and M. M. Rao (eds.), *Handbook of Statistics, Volume 5: Time Series in the Time Domain*, pp. 257–278, North-Holland, Amsterdam.
- Dhrymes, P. J. (1998) *Time Series, Unit Roots, and Cointegration*. Academic Press, San Diego.
- Engle, R. F. & C. W. J. Granger (1987) Cointegration and error correction: Representation, estimation and testing. *Econometrica* 55, 251–276.
- Engle, R. F., D. F. Hendry, & J. F. Richard (1983) Exogeneity. *Econometrica* 51, 277–304.
- Gevers, M. (1986) ARMA models, their Kronecker indices and their McMillan degree. *International Journal of Control* 43, 1745–1761.
- Geweke, J. (1984) Measures of conditional linear independence and feedback between time series. *Journal of the American Statistical Association* 79, 907–915.
- Granger, C. W. J. (1981) Some properties of time series data and their use in econometric model specification. *Journal of Econometrics* 16, 121–130.
- Hannan, E. J. (1971) *Multiple Time Series*. J. Wiley, New York.

- Hannan, E. J. (1974) The identification problem for multiple equation systems with moving average errors. *Econometrica* 39, 751–765.
- Hannan, E. J. (1976) The identification and parameterisation of ARMA and State Space forms. *Econometrica* 44, 713–723.
- Hannan, E. J. & M. Deistler (1988) *The Statistical Theory of Linear Systems*. Wiley, New York.
- Hanzon, B. (1989) *Identifiability, Recursive Identification and Spaces of Linear Dynamical Systems: Part I*. CWI Tract 63, Center for Mathematics and Computer Science, Amsterdam.
- Hatanaka, M. (1996) *Time-Series-Based Econometrics: Unit Roots and Co-integration*. Oxford University Press, Oxford.
- Hosoya, Y. (1977) On the Granger condition for non-causality. *Econometrica* 45, 1735–1736.
- Hotelling, H. (1936) Relations between two sets of variates. *Biometrika* 28, 321–377.
- Hsiao, C. (1997) Cointegration and dynamic simultaneous equations model. *Econometrica* 65, 647–670.
- Johansen, S. (1991) Estimation and hypothesis testing of cointegration vectors in gaussian vector autoregressive models. *Econometrica* 59, 1551–1580.
- Lütkepohl, H. & H. Claessen (1997) Analysis of cointegrated VARMA processes. *Journal of Econometrics* 80, 223–239.
- Lütkepohl, H. & D. Poskitt (1996) Specification of echelon form VARMA models. *Journal of Business and Economic Statistics* 14, 69–79.
- Phillips, P. C. B. (1991) Optimal inference in cointegrated systems. *Econometrica* 59, 283–306.
- Phillips, P. C. B. & V. Solo (1992) Asymptotics for linear processes. *Annals of Statistics* 20, 971–1001.
- 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. (2000) Strongly consistent determination of cointegrating rank via canonical correlations. *Journal of Business and Economic Statistics* 18, 71–90.
- Poskitt, D. S. (2003) On the specification of cointegrated autoregressive moving-average forecasting systems. *International Journal of Forecasting* 19, 503–519.
- Poskitt, D. S. & A. R. Tremayne (1982) Diagnostic tests for multiple time series models. *Annals of Statistics* 10, 114–120.
- Reinsel, G. C. S. (1993) *Elements of Multivariate Time Series Analysis*. Springer-Verlag, Berlin.

- Rissanen, J. & L. Barbosa (1969) A factorization problem and the problem of predicting non-stationary vector-valued stochastic processes. *Z. Wahrsch. Verw. Gebiete* 12, 255–266.
- Rissanen, J. & P. E. Caines (1979) The strong consistency of maximum likelihood estimators for arma processes. *Annals of Statistics* pp. 297–315.
- Saikkonen, P. (1996) Problems with the asymptotic theory of maximum likelihood estimation in integrated and cointegrated systems. *Econometric Theory* 11, 888–911.
- Stock, J. H. & M. W. Watson (1988) Testing for common trends. *Journal of The American Statistical Association* 83, 1097–1107.
- Tanaka, K. (1996) *Time Series Analysis: Nonstationary and Noninvertible Distribution Theory*. John Wiley, New York.
- Tunnicliffe-Wilson, G. (1972) The factorization of matricial spectral densities. *SIAM Journal of Applied Mathematics* 23, 420–426.
- Wickens, M. R. (1996) Interpreting cointegrating vectors and common stochastic trends. *Journal of Econometrics* 74, 255–271.
- Wickens, M. R. & T. S. Breusch (1988) Dynamic specification, the long-run and the estimation of transformed regression models. *The Economic Journal* 98, 189–205.
- Yap, S. & G. Reinsel (1995) Estimation and testing for unit roots in a partially nonstationary vector autoregressive moving average model. *Journal of the American Statistical Association* 90, 253–267.

## Appendix: Proofs

**Proof of Theorem 2.1:** Let  $\mathbf{y}_{P,t} = \sum_{s=0}^{t+p-1} \Phi_s \mathbf{w}_{t-s}$ ,  $t = 1-p, \dots, 0, 1, \dots, T$ , where the impulse response sequence is as given in (2.2). By construction  $\sum_{j=0}^{\min(\tau-1,p)} \mathbf{A}_j z^j \{ \sum_{k=0}^{\tau-j-1} \Phi_k z^k \} = \sum_{l=0}^{\min(\tau-1,p)} \mathbf{N}_l z^l$  for  $\tau = 1, 2, \dots, T+p-1$ , which when evaluated at  $\tau = t+p$  gives

$$\sum_{j=0}^p \mathbf{A}_j \mathbf{y}_{P,t-j} = \sum_{j=0}^p \mathbf{A}_j \left\{ \sum_{s=0}^{t-j+p-1} \Phi_s \mathbf{w}_{t-j-s} \right\} = \sum_{l=0}^p \mathbf{N}_l \mathbf{w}_{t-l}$$

for  $t = 1, \dots, T$ . Hence  $\mathbf{y}_{P,t}$  provides a particular solution to (1.2).

Now let  $\mathbf{m}_t = \mathbf{y}_t - \mathbf{y}_{P,t}$ ,  $t = 1-p, \dots, 0$ , and set  $\mathbf{m}_t = -\mathbf{A}_0^{-1} (\sum_{j=1}^p \mathbf{A}_j \mathbf{m}_{t-j})$ ,  $t = 1, \dots, T$ . Then  $\mathbf{m}_t$  defines an appropriate complementary function and  $\mathbf{y}_{P,t} + \mathbf{m}_t$  gives the general solution to (1.2) since by construction  $\mathbf{y}_t = \mathbf{y}_{P,t} + \mathbf{m}_t$ ,  $t = 1-p, \dots, 0$ , and

$$\mathbf{A}(\mathcal{L})\mathbf{y}_t = \sum_{j=0}^p \mathbf{A}_j (\mathbf{y}_{P,t-j} + \mathbf{m}_{t-j}) = \mathbf{A}(\mathcal{L})\mathbf{y}_{P,t} = \mathbf{N}(\mathcal{L})\mathbf{w}_t, \quad t = 1, \dots, T.$$

Conversely, suppose that the process  $\mathbf{y}_t$  admits an input-output representation as in (2.3) and that  $v \times v$  coefficient values  $\mathbf{A}_j$ ,  $j = 0, \dots, p$ , exist such that the impulse response sequence satisfies the difference equation  $\sum_{j=0}^p \mathbf{A}_j \Phi_{i-j} = \mathbf{0}$ ,  $i > p$  and, similarly,  $\mathbf{m}_t$  solves  $\sum_{j=0}^p \mathbf{A}_j \mathbf{m}_{t-j} = \mathbf{0}$ ,  $t = 1, \dots, T$ . Then for  $t = 1, \dots, T$ .

$$\sum_{j=0}^p \mathbf{A}_j \mathbf{y}_{t-j} = \sum_{j=0}^p \mathbf{A}_j \left\{ \sum_{s=0}^{t-j+p-1} \Phi_s \mathbf{w}_{t-j-s} \right\} + \sum_{j=0}^p \mathbf{A}_j \mathbf{m}_{t-j} = \sum_{s=0}^p \mathbf{N}_s \mathbf{w}_{t-s} . \quad \blacksquare$$

**Proof of Corollary 3.1:** Necessity follows by noting that the conditions  $\sum_{j=0}^p \mathbf{A}_j \Phi_{i-j} = \mathbf{0}$ ,  $i = p+1, \dots, T+p-1$ , and  $\sum_{j=0}^p \mathbf{A}_j \mathbf{m}_{t-j} = \mathbf{0}$ ,  $t = 1, \dots, T$  imply that

$$\sum_{j=0}^p \mathbf{A}_j [\mathbf{K}_{p+1-j} : \dots : \mathbf{K}_{T+p-1-j}] = [\mathbf{0} : \dots : \mathbf{0}].$$

From the nonsingularity of  $\mathbf{A}_0$  it follows that rows  $pv + r$ ,  $r = 1, \dots, v$ , of  $\mathbf{H}_{p+1,T}$  can be expressed as linear combinations of rows  $(i-1)v + r$ ,  $i = 1, \dots, p$ ,  $r = 1, \dots, v$  and hence that  $\rho(\mathbf{H}_{p+1,T}) \leq vp$ . Corollary 3.1.2.3–37 of Hanzon (1989) and the linear dependence properties of  $\mathbf{H}_{R,T}$ ,  $R = 1, \dots, T+p-1$ , now imply that  $\rho(\mathbf{H}_{R,T}) \leq vp$ ,  $R = p+1, \dots, T+p-1$ , and thus  $\sup_{1 \leq R \leq T+p-1} [\rho(\mathbf{H}_{R,T})] = \rho(\mathbf{H}_{p,T}) \leq vp$ .

To establish sufficiency suppose that  $\sup_{1 \leq R \leq T+p-1} [\rho(\mathbf{H}_{R,T})] = \rho(\mathbf{H}_{p,T}) \leq vp$ . Then the rows of  $\mathbf{H}_{p+1,T}$  are linearly dependent and each of the last  $v$  rows can be expressed as a linear combination of the rows that precede it. These row combinations generate a sequence of  $v \times v$  coefficient values  $\mathbf{A}_j$ ,  $j = 0, \dots, p$ , such that

$$[\mathbf{A}_p : \dots : \mathbf{A}_0] \begin{bmatrix} \mathbf{K}_1 & \mathbf{K}_2 & \dots & \mathbf{K}_{T-1} \\ \vdots & \vdots & & \vdots \\ \mathbf{K}_{p+1} & \mathbf{K}_{p+2} & \dots & \mathbf{K}_{T+p-1} \end{bmatrix} = [\mathbf{0} : \dots : \mathbf{0}], \quad (\text{A.1})$$

where  $\mathbf{A}_0$  is lower triangular and nonsingular, with leading diagonal equal to the identity, and all of  $\mathbf{A}_0, \mathbf{A}_1, \dots, \mathbf{A}_p$  cannot be zero. That is,  $\sum_{j=0}^p \mathbf{A}_j \Phi_{i-j} = \mathbf{0}$ ,  $i = p+1, \dots, T+p-1$ , and  $\sum_{j=0}^p \mathbf{A}_j \mathbf{m}_{t-j} = \mathbf{0}$ ,  $t = 1, \dots, T$ . Appeal to Theorem 2.1 completes the proof.  $\blacksquare$

**Proof of Theorem 3.1:** Selecting the first basis rows of  $\mathbf{H}_{p+1,T}$  in natural order produces  $v$  integers  $n_r$ ,  $r = 1, \dots, v$ , such that  $n_1 + \dots + n_v = \rho(\mathbf{H}_{p+1,T})$  and  $\mathbf{h}_{p+1,T}(1, 1), \dots, \mathbf{h}_{p+1,T}(n_1, 1)$  through to  $\mathbf{h}_{p+1,T}(1, v), \dots, \mathbf{h}_{p+1,T}(n_v, v)$  form a basis for the rows of  $\mathbf{H}_{p+1,T}$ . Expressing row  $n_r v + r$  of  $\mathbf{H}_{p+1,T}$  as a linear combination of its linearly independent antecedents results in an equation system analogous to (A.1) that may be solved uniquely for the coefficient values  $a_{rj, n_r - n_{rj} + 1}, \dots, a_{rj, n_r}$ ,  $r, j = 1, \dots, v$ . Equating  $a_{rj, n_r - s + 1}$  to the  $(r, j)$ 'th element of  $\mathbf{A}_{n_r - s + 1}$ ,  $s = 1, \dots, n_{rj}$ , with  $a_{rr, 0} = 1$ ,  $r = 1, \dots, v$ , and all other elements equal to zero, yields the autoregressive operator  $\mathbf{A}(z) = \mathbf{A}_0 + \mathbf{A}_1 z^1 + \dots + \mathbf{A}_p z^p$  where  $p = \max_{1 \leq r \leq v} (n_r)$ . Given  $\mathbf{A}(z)$  the coefficients of  $\mathbf{N}(z)$  are obtained by evaluating  $\sum_{j=0}^i \mathbf{A}_j \mathbf{K}_{i-j} = \sum_{j=0}^i \mathbf{A}_j [\Phi_{i-j} : \mathbf{m}_{i-p-j+1}] = [\mathbf{N}_i : *]$ ,  $i = 0, \dots, p$ . By construction the row degrees of  $[\mathbf{A}(z) : \mathbf{N}(z)]$  equal the Kronecker

indices, that is,  $\delta_r [\mathbf{A}(z) : \mathbf{N}(z)] = n_r$ ,  $r = 1, \dots, v$ . The detailed steps in the above argument follow those used by Hannan and Deistler (1988) in establishing their Theorem 2.5.1.

Thus, for any given ARMAX system  $\mathbf{H}_{R,T}$ ,  $R = 1, \dots, T + p - 1$ , can be readily determined and from these the rank  $\rho(\mathbf{H}_{p,T})$ , the Kronecker indices  $n_i$ ,  $i = 1, \dots, v$ , and the echelon form can subsequently be constructed. Conversely, every ARMAX system such that  $[\mathbf{A}(z) : \mathbf{N}(z)]$  satisfies (i), (ii) and (iii) obviously defines an echelon form representation with Kronecker indices  $n_r$ ,  $r = 1, \dots, v$ . ■

**Proof of Theorem 5.1:** To show that  $\|\mathbf{N}_T(\hat{\boldsymbol{\beta}}_T^{(i)} - \tilde{\boldsymbol{\beta}}_T^{(i)})\| = o_p(1)$  we use the principle of induction. Assume that  $\mathbf{N}_T(\hat{\boldsymbol{\beta}}_T^{(i-1)} - \boldsymbol{\beta})$  and  $\mathbf{N}_T(\tilde{\boldsymbol{\beta}}_T^{(i-1)} - \boldsymbol{\beta})$  are  $O_p(1)$ . Substituting  $\hat{\boldsymbol{\beta}}_T^{(i-1)} = \boldsymbol{\beta} + \mathbf{N}_T^{-1}O_p(1)$  in (5.9) and using the stochastic equicontinuity results given in Saikkonen (1996) we obtain the asymptotic representation

$$\mathbf{N}_T(\hat{\boldsymbol{\beta}}_T^{(i)} - \boldsymbol{\beta}) = \left( \mathbf{N}_T^{-1} \sum_{t=1}^T \mathbf{W}'_t \bar{\boldsymbol{\Sigma}}_T^{-1} \mathbf{W}_t \mathbf{N}_T^{-1} \right)^{-1} \mathbf{N}_T^{-1} \sum_{t=1}^T \mathbf{W}'_t \bar{\boldsymbol{\Sigma}}_T^{-1} \boldsymbol{\varepsilon}_t |_{\boldsymbol{\beta}} + o_p(1). \quad (\text{A.2})$$

A corresponding expression with  $\hat{\boldsymbol{\beta}}_T^{(i)}$  replaced by  $\tilde{\boldsymbol{\beta}}_T^{(i)}$ , and  $\mathbf{W}_{\langle t|t-1 \rangle}$  substituted for  $\mathbf{W}_t$ ,  $\boldsymbol{\Sigma}_{t|t-1}$  for  $\bar{\boldsymbol{\Sigma}}_T$ , and  $\boldsymbol{\varepsilon}_{\langle t|t-1 \rangle}$  for  $\boldsymbol{\varepsilon}_t$ , also obtains. It follows that the probability limit of  $\|\mathbf{N}_T(\hat{\boldsymbol{\beta}}_T^{(i)} - \tilde{\boldsymbol{\beta}}_T^{(i)})\|$  will be zero if

$$\mathbf{N}_T^{-1} \sum_{t=1}^T \left\{ \mathbf{W}'_t \bar{\boldsymbol{\Sigma}}_T^{-1} \mathbf{W}_t - \mathbf{W}'_{\langle t|t-1 \rangle} \boldsymbol{\Sigma}_{t|t-1}^{-1} \mathbf{W}_{\langle t|t-1 \rangle} \right\} \mathbf{N}_T^{-1} |_{\boldsymbol{\beta}} = o_p(1) \quad (\text{A.3})$$

and

$$\mathbf{N}_T^{-1} \sum_{t=1}^T \left\{ \mathbf{W}'_t \bar{\boldsymbol{\Sigma}}_T^{-1} \boldsymbol{\varepsilon}_t - \mathbf{W}'_{\langle t|t-1 \rangle} \boldsymbol{\Sigma}_{t|t-1}^{-1} \boldsymbol{\varepsilon}_{\langle t|t-1 \rangle} \right\} |_{\boldsymbol{\beta}} = o_p(1). \quad (\text{A.4})$$

Suppressing explicit evaluation at the point  $\boldsymbol{\beta}$  for notational convenience, both (A.3) and (A.4) are obtained by expressing the left hand side in terms of the differences  $\nabla \boldsymbol{\varepsilon}_t = \boldsymbol{\varepsilon}_t - \boldsymbol{\varepsilon}_{\langle t|t-1 \rangle}$ ,  $\partial \nabla \boldsymbol{\varepsilon}_t / \partial \boldsymbol{\beta}' = \mathbf{W}_{\langle t|t-1 \rangle} - \mathbf{W}_t = -\nabla \mathbf{W}_t = -[\nabla \mathbf{W}_{\theta t} : \nabla \mathbf{W}_{\gamma t}]$  and  $\nabla \boldsymbol{\Sigma}_{Tt} = \bar{\boldsymbol{\Sigma}}_T - \boldsymbol{\Sigma}_{t|t-1}$ .

Consider first (A.3). Expanding each term in the summation as

$$\begin{aligned} \mathbf{W}'_t \bar{\boldsymbol{\Sigma}}_T^{-1} \mathbf{W}_t - \mathbf{W}'_{\langle t|t-1 \rangle} \boldsymbol{\Sigma}_{t|t-1}^{-1} \mathbf{W}_{\langle t|t-1 \rangle} &= \nabla \mathbf{W}'_t \boldsymbol{\Sigma}_{t|t-1}^{-1} \mathbf{W}_t + \mathbf{W}'_t \boldsymbol{\Sigma}_{t|t-1}^{-1} \nabla \mathbf{W}_t \\ &\quad - \nabla \mathbf{W}'_t \boldsymbol{\Sigma}_{t|t-1}^{-1} \nabla \mathbf{W}_t - \mathbf{W}'_t \bar{\boldsymbol{\Sigma}}_T^{-1} \nabla \boldsymbol{\Sigma}_{Tt} \boldsymbol{\Sigma}_{t|t-1}^{-1} \mathbf{W}_t \end{aligned}$$

we can see that the left hand side of (A.3) can be decomposed into four series. The fourth of these,  $-\sum_{t=1}^T \mathbf{N}_T^{-1} \mathbf{W}'_t \bar{\boldsymbol{\Sigma}}_T^{-1} \nabla \boldsymbol{\Sigma}_{Tt} \boldsymbol{\Sigma}_{t|t-1}^{-1} \mathbf{W}_t \mathbf{N}_T^{-1}$ , is bounded in norm by

$$\sum_{t=1}^T \|\bar{\boldsymbol{\Sigma}}_T^{-1}\| \cdot \|\nabla \boldsymbol{\Sigma}_{Tt}\| \cdot \|\boldsymbol{\Sigma}_{t|t-1}^{-1}\| \cdot \|\mathbf{N}_T^{-1} \mathbf{W}'_t\|^2. \quad (\text{A.5})$$

But  $\bar{\boldsymbol{\Sigma}}_T = \boldsymbol{\Sigma}_\varepsilon + o(1)$  by ergodicity and a direct application of Lemma (1) of Rissanen and Caines (1979) tells us that  $\boldsymbol{\Sigma}_{t|t-1}$  converges to  $\boldsymbol{\Sigma}_\varepsilon$  at a geometric rate. The latter implies that there exist constants  $K_\sigma$  and  $\lambda_\sigma$  with  $0 \leq K_\sigma < \infty$  and  $0 < \lambda_\sigma < 1$ ,  $K_\sigma \lambda_\sigma^T > \|\bar{\boldsymbol{\Sigma}}_T - \boldsymbol{\Sigma}_\varepsilon\|$ , such that

$\|\boldsymbol{\Sigma}_{t|t-1}^{-1}\| \leq \|\boldsymbol{\Sigma}_\varepsilon^{-1}\| + K_\sigma \lambda_\sigma^t$  and  $\|\nabla \boldsymbol{\Sigma}_{Tt}\| \leq \|\bar{\boldsymbol{\Sigma}}_T - \boldsymbol{\Sigma}_\varepsilon\| + \|\boldsymbol{\Sigma}_\varepsilon - \boldsymbol{\Sigma}_{t|t-1}\| \leq 2K_\sigma \lambda_\sigma^t$ . It follows that

$$\|\bar{\boldsymbol{\Sigma}}_T^{-1}\| \cdot \|\nabla \boldsymbol{\Sigma}_{Tt}\| \cdot \|\boldsymbol{\Sigma}_{t|t-1}^{-1}\| \leq 2\|\boldsymbol{\Sigma}_\varepsilon^{-1}\|^2 \{1 + o(1)\} K_\sigma \lambda_\sigma^t$$

and therefore A.5 is majorised by

$$2\|\boldsymbol{\Sigma}_\varepsilon^{-1}\|^2 \{1 + o(1)\} \sum_{t=1}^T K_\sigma \lambda_\sigma^t \|\mathbf{N}_T^{-1} \mathbf{W}'_t\|^2.$$

Now let  $\ell_T = -2 \log(T) / \log(\lambda_\sigma)$ . Then

$$\begin{aligned} \sum_{t=1}^T K_\sigma \lambda_\sigma^t \|\mathbf{N}_T^{-1} \mathbf{W}'_t\|^2 &\leq K_\sigma \left\{ T^{-1} \sum_{t=1}^{\ell_T} \|\mathbf{W}_{\theta t}\|^2 + T^{-2} \sum_{t=1}^{\ell_T} \|\mathbf{W}_{\gamma t}\|^2 + \lambda_\sigma^{\ell_T} \sum_{t=\ell_T+1}^T \|\mathbf{N}_T^{-1} \mathbf{W}'_t\|^2 \right\} \\ &= K_\sigma \left\{ \frac{\ell_T}{T} (O(1) + O_p(\ell_T/T)) + \lambda_\sigma^{\ell_T} O_p(1) \right\} \\ &= o_p(1) \end{aligned}$$

because

$$\sum_{t=1}^T \|\mathbf{N}_T^{-1} \mathbf{W}'_t\|^2 = T^{-1} \sum_{t=1}^T \|\mathbf{W}_{\theta t}\|^2 + T^{-2} \sum_{t=1}^T \|\mathbf{W}_{\gamma t}\|^2 = O_p(1).$$

and  $\ell_T/T \rightarrow 0$  as  $T \rightarrow \infty$  and  $\lambda_\sigma^{\ell_T} = T^{-2}$  by construction. Thus we can conclude that A.5 is  $o_p(1)$  and hence  $-\sum_{t=1}^T \mathbf{N}_T^{-1} \mathbf{W}'_t \bar{\boldsymbol{\Sigma}}_T^{-1} \nabla \boldsymbol{\Sigma}_{Tt} \boldsymbol{\Sigma}_{t|t-1}^{-1} \mathbf{W}_t \mathbf{N}_T^{-1} = o_p(1)$ .

The first two series in the decomposition of A.3 are bounded in norm by

$$\sum_{t=1}^T \|\mathbf{N}_T^{-1} \nabla \mathbf{W}'_t\| \cdot \|\boldsymbol{\Sigma}_{t|t-1}^{-1}\| \cdot \|\mathbf{N}_T^{-1} \mathbf{W}'_t\| \leq (\|\boldsymbol{\Sigma}_\varepsilon^{-1}\| + K_\sigma) \left( \sum_{t=1}^T \|\mathbf{N}_T^{-1} \nabla \mathbf{W}'_t\|^2 \cdot \sum_{t=1}^T \|\mathbf{N}_T^{-1} \mathbf{W}'_t\|^2 \right)^{\frac{1}{2}}$$

and the norm of the third is bounded by

$$\sum_{t=1}^T \|\boldsymbol{\Sigma}_{t|t-1}^{-1}\| \cdot \|\mathbf{N}_T^{-1} \nabla \mathbf{W}'_t\|^2 \leq (\|\boldsymbol{\Sigma}_\varepsilon^{-1}\| + K_\sigma) \sum_{t=1}^T \|\mathbf{N}_T^{-1} \nabla \mathbf{W}'_t\|^2.$$

Thus these three series will converge to zero in probability if we can show that

$$\sum_{t=1}^T \|\mathbf{N}_T^{-1} \nabla \mathbf{W}'_t\|^2 = o_p(1). \quad (\text{A.6})$$

To establish A.6 note that simple manipulation of the equality between  $\sum_{j=0}^q \mathbf{M}_{\langle t,j \rangle} \boldsymbol{\varepsilon}_{\langle t-j|t-j-1 \rangle}$  and  $\sum_{j=0}^p \mathbf{M}_j \boldsymbol{\varepsilon}_{t-j}$  gives

$$\sum_{j=0}^q \mathbf{M}_{\langle t,j \rangle} \nabla \boldsymbol{\varepsilon}_{t-j} = \sum_{j=0}^p (\mathbf{M}_{\langle t,j \rangle} - \mathbf{M}_j) \boldsymbol{\varepsilon}_{t-j} \quad (\text{A.7})$$

and therefore the proof given by Rissanen and Caines (1979, pp. 312-314 ) that

$$T^{-1} \sum_{t=1}^T \|\nabla \varepsilon_t\|^2 \rightarrow 0 \text{ a.s. as } T \rightarrow \infty, \quad (\text{A.8})$$

cf. Rissanen and Caines (1979, Equation A3.4), is directly applicable. Differentiating (A.7) with respect to  $\beta$  and rearranging gives

$$\sum_{j=0}^q \mathbf{M}_{(t,j)} \nabla \mathbf{W}_{t-j} = \sum_{j=0}^p (\mathbf{M}_{(t,j)} - \mathbf{M}_j) \mathbf{W}_{t-j} - \sum_{j=0}^p \frac{\partial (\mathbf{M}_{(t,j)} - \mathbf{M}_j)}{\partial \beta'} \varepsilon_{t-j} + \sum_{j=0}^q \frac{\partial \mathbf{M}_{(t,j)}}{\partial \beta'} \nabla \varepsilon_{t-j}.$$

By Lemma (1) of Rissanen and Caines (1979) there exist constants  $K_\mu$  and  $\lambda_\mu$ ,  $0 \leq K_\mu < \infty$ ,  $0 < \lambda_\mu < 1$ , such that  $\|\mathbf{M}_{(t,j)} - \mathbf{M}_j\| < K_\mu \lambda_\mu^t$  uniformly in  $\beta$  and therefore the same is true for  $\|\partial(\mathbf{M}_{(t,j)} - \mathbf{M}_j)/\partial \beta'\|$ . A straightforward adaptation of the argument that gives (A.8) therefore leads to the conclusion that  $T^{-1} \sum_{t=1}^T \|\nabla \mathbf{W}_{\theta t}\|^2 = o(1)$  and  $T^{-2} \sum_{t=1}^T \|\nabla \mathbf{W}_{\gamma t}\|^2 = o(1)$  as  $T \rightarrow \infty$ , giving A.6 as required. Equation (A.3) now follows.

Now consider equation (A.4). First we bound the norm of the left hand side by

$$\begin{aligned} & \sum_{t=1}^T \|\mathbf{N}_T^{-1} \nabla \mathbf{W}'_t \Sigma_{t|t-1}^{-1} \varepsilon_t\| + \sum_{t=1}^T \|\mathbf{N}_T^{-1} \mathbf{W}'_t \Sigma_{t|t-1}^{-1} \nabla \varepsilon_t\| \\ & - \sum_{t=1}^T \|\mathbf{N}_T^{-1} \nabla \mathbf{W}'_t \Sigma_{t|t-1}^{-1} \nabla \varepsilon_t\| - \sum_{t=1}^T \|\mathbf{N}_T^{-1} \mathbf{W}'_t \bar{\Sigma}_T^{-1} \nabla \Sigma_{Tt} \Sigma_{t|t-1}^{-1} \varepsilon_t\|. \end{aligned}$$

The fourth term

$$\begin{aligned} \sum_{t=1}^T \|\mathbf{N}_T^{-1} \mathbf{W}'_t \bar{\Sigma}_T^{-1} \nabla \Sigma_{Tt} \Sigma_{t|t-1}^{-1} \varepsilon_t\| & \leq 2 \|\Sigma_\varepsilon^{-1}\|^2 \{1 + o(1)\} \sum_{t=1}^T K_\sigma \lambda_\sigma^t \|\mathbf{N}_T^{-1} \mathbf{W}'_t\| \cdot \|\varepsilon_t\| \\ & \leq 2 \|\Sigma_\varepsilon^{-1}\|^2 \{1 + o(1)\} K_\sigma \left( \sum_{t=1}^T \lambda_\sigma^t \|\mathbf{N}_T^{-1} \mathbf{W}'_t\|^2 \cdot \sum_{t=1}^T \lambda_\sigma^t \|\varepsilon_t\|^2 \right)^{\frac{1}{2}}. \end{aligned}$$

But

$$\sum_{t=1}^T \lambda_\sigma^t \|\mathbf{N}_T^{-1} \mathbf{W}'_t\|^2 \cdot \sum_{t=1}^T \lambda_\sigma^t \|\varepsilon_t\|^2 = \sum_{t=1}^T \lambda_\sigma^t \left( \frac{\|\mathbf{W}_{\theta t}\|^2}{T^{1/2}} + \frac{\|\mathbf{W}_{\gamma t}\|^2}{T^{3/2}} \right) \cdot \sum_{t=1}^T \lambda_\sigma^t \frac{\|\varepsilon_t\|^2}{T^{1/2}}.$$

Now observe that for any  $\lambda$ ,  $0 \leq \lambda < 1$ ,

$$\begin{aligned} \sum_{t=1}^T \lambda^t \frac{\|\mathbf{W}_{\theta t}\|^2}{T^{1/2}} & \leq \frac{\ell_T}{T^{1/2}} \sum_{t=1}^{\ell_T} \frac{\|\mathbf{W}_{\theta t}\|^2}{\ell_T} + \lambda^{\ell_T} T^{1/2} \sum_{t=\ell_T+1}^T \frac{\|\mathbf{W}_{\theta t}\|^2}{T} \\ & = O(\log(T)/T^{1/2}) + O(T^{-3/2}) \end{aligned} \quad (\text{A.9})$$

where  $\ell_T = -2 \log(T)/\log(\lambda)$ . An analogous derivation also gives

$$\sum_{t=1}^T \lambda^t \|\varepsilon_t\|^2 / T^{1/2} \leq O(\log(T)/T^{1/2}) + O(T^{-3/2}) \quad (\text{A.10})$$

and

$$\begin{aligned} \sum_{t=1}^T \lambda^t \frac{\|\mathbf{W}_{\gamma t}\|^2}{T^{3/2}} &\leq \frac{\ell_T^2}{T^{3/2}} \sum_{t=1}^{\ell_T} \frac{\|\mathbf{W}_{\gamma t}\|^2}{\ell_T^2} + \lambda^{\ell_T} T^{1/2} \sum_{t=\ell_T+1}^T \frac{\|\mathbf{W}_{\gamma t}\|^2}{T^2} \\ &= O_p(\log(T)^2/T^{3/2}) + O_p(T^{-3/2}). \end{aligned} \quad (\text{A.11})$$

Substituting these bounds into the above gives  $\sum_{t=1}^T \|\mathbf{N}_T^{-1} \mathbf{W}'_t \bar{\Sigma}_T^{-1} \nabla \Sigma_{Tt} \Sigma_{t|t-1}^{-1} \boldsymbol{\varepsilon}_t\| = o_p(1)$ .

Equivalent bounds on the order of magnitude of the first three terms are derived using the fact that both  $\|\nabla \boldsymbol{\varepsilon}_t\|$  and  $\|\nabla \mathbf{W}'_t\|$  converge to zero at a geometric rate in  $t$ . Confirmation of this result is obtained by noting that the conditions  $\|\mathbf{M}_{(t,j)} - \mathbf{M}_j\| < K_\mu \lambda_\mu^t$  and  $\det(\mathbf{M}(z)) \neq 0$ ,  $|z| \leq 1$  when applied to (A.7) imply that

$$\|\nabla \boldsymbol{\varepsilon}_t\| \leq K_\varepsilon \left( \sum_{s=1}^{t-1} \lambda_\varepsilon^{2s} \sum_{j=1}^p \lambda_\varepsilon^{2t} \|\boldsymbol{\varepsilon}_{t-j-s}\| + \lambda_\varepsilon^{2t} \|\boldsymbol{\eta}_\varepsilon\| \right) \leq K_\varepsilon \lambda_\varepsilon^{2t} t \left( \sum_{j=1}^p \sum_{s=1}^{t-1} \frac{\|\boldsymbol{\varepsilon}_{t-j-s}\|}{t} + \frac{\|\boldsymbol{\eta}_\varepsilon\|}{t} \right)$$

for some  $K_\varepsilon, K_\mu \leq K_\varepsilon < \infty$  and  $\lambda_\varepsilon, \lambda_\mu \leq \lambda_\varepsilon^2 < 1$ , where  $\|\boldsymbol{\eta}_\varepsilon\| < \infty$  bounds the influence of the initial conditions. Hence  $\|\nabla \boldsymbol{\varepsilon}_t\| \leq \lambda_\varepsilon^t M_\varepsilon$  where  $M_\varepsilon < \infty$  but exceeds

$$K_\varepsilon \lambda_\varepsilon^t t \left( \sum_{j=1}^p \sum_{s=1}^{t-1} \frac{\|\boldsymbol{\varepsilon}_{t-j-s}\|}{t} + \frac{\|\boldsymbol{\eta}_\varepsilon\|}{t} \right) \rightarrow 0 \text{ as } t \rightarrow \infty.$$

Using the inequality  $\|\nabla \boldsymbol{\varepsilon}_t\| \leq \lambda_\varepsilon^t M_\varepsilon$  we also have

$$\begin{aligned} \|\nabla \mathbf{W}_t\| &\leq K_W \left( \sum_{s=1}^{t-1} \lambda_W^{2s} \sum_{j=1}^p \lambda_W^{2t} \|\mathbf{W}_{t-j-s}\| + \lambda_W^{2t} \|\boldsymbol{\varepsilon}_{t-j-s}\| + \|\nabla \boldsymbol{\varepsilon}_{t-j-s}\| + \lambda_W^{2t} \|\boldsymbol{\eta}_W\| \right) \\ &\leq K_W \lambda_W^{2t} \left( \sum_{j=1}^p \sum_{s=1}^{t-1} \lambda_W^{2s} (\|\mathbf{W}_{t-j-s}\| + \|\boldsymbol{\varepsilon}_{t-j-s}\|) + K_W \lambda_W^{-j} + \|\boldsymbol{\eta}_W\| \right) \\ &\leq K_W \lambda_W^{2t} t^{3/2} \left( \sum_{j=1}^p \sum_{s=1}^{t-1} \frac{\|\mathbf{W}_{t-j-s}\|}{t^{3/2}} + \frac{\|\boldsymbol{\varepsilon}_{t-j-s}\|}{t^{3/2}} + \frac{K_W \lambda_W^{-p}}{t^{3/2}} + \frac{\|\boldsymbol{\eta}_W\|}{t^{3/2}} \right), \end{aligned}$$

$\max\{K_\varepsilon, M_\varepsilon\} < K_W < \infty$ ,  $\sqrt{\lambda_\varepsilon} < \lambda_W < 1$ , leading to the conclusion that  $\|\nabla \mathbf{W}_t\| \leq \lambda_W^t M_W$ .

Applying these geometric bounds yields the inequalities

$$\begin{aligned} \sum_{t=1}^T \|\mathbf{N}_T^{-1} \nabla \mathbf{W}'_t \Sigma_{t|t-1}^{-1} \boldsymbol{\varepsilon}_t\| &\leq \sum_{t=1}^T \|\mathbf{N}_T^{-1} \nabla \mathbf{W}'_t\| (\|\Sigma_\varepsilon^{-1}\| + K_\sigma \lambda_\sigma^t) \|\boldsymbol{\varepsilon}_t\| \\ &\leq (\|\Sigma_\varepsilon^{-1}\| + K_\sigma) \sum_{t=1}^T \|\mathbf{N}_T^{-1} \nabla \mathbf{W}'_t\| \cdot \|\boldsymbol{\varepsilon}_t\| \\ &\leq (\|\Sigma_\varepsilon^{-1}\| + K_\sigma) M_W d_\beta \sum_{t=1}^T \lambda_W^t \|\boldsymbol{\varepsilon}_t\| / T^{1/2} \end{aligned}$$

for the first term,

$$\begin{aligned} \sum_{t=1}^T \|\mathbf{N}_T^{-1} \mathbf{W}'_t \boldsymbol{\Sigma}_{t|t-1}^{-1} \nabla \boldsymbol{\varepsilon}_t\| &\leq \sum_{t=1}^T \|\mathbf{N}_T^{-1} \mathbf{W}'_t\| (\|\boldsymbol{\Sigma}_\varepsilon^{-1}\| + K_\sigma \lambda_\sigma^t) \|\nabla \boldsymbol{\varepsilon}_t\| \\ &\leq (\|\boldsymbol{\Sigma}_\varepsilon^{-1}\| + K_\sigma) \sum_{t=1}^T \|\mathbf{N}_T^{-1} \mathbf{W}'_t\| \cdot \|\nabla \boldsymbol{\varepsilon}_t\| \\ &\leq (\|\boldsymbol{\Sigma}_\varepsilon^{-1}\| + K_\sigma) M_\varepsilon \sum_{t=1}^T \lambda_\varepsilon^t \|\mathbf{N}_T^{-1} \mathbf{W}'_t\| \end{aligned}$$

for the second and, similarly,

$$\sum_{t=1}^T \|\mathbf{N}_T^{-1} \nabla \mathbf{W}'_t \boldsymbol{\Sigma}_{t|t-1}^{-1} \nabla \boldsymbol{\varepsilon}_t\| \leq (\|\boldsymbol{\Sigma}_\varepsilon^{-1}\| + K_\sigma) M_\varepsilon M_W d_\beta \sum_{t=1}^T \lambda_W^t \lambda_\varepsilon^t / T^{1/2}$$

for the third. But

$$\sum_{t=1}^T \lambda_W^t \|\boldsymbol{\varepsilon}_t\| / T^{1/2} \leq (T^{-1} \lambda_W / (1 - \lambda_W)) \sum_{t=1}^T \lambda_W^t \|\boldsymbol{\varepsilon}_t\|^2)^{1/2} = o_p(1)$$

by the Cauchy-Schwartz inequality and (A.10) and

$$\sum_{t=1}^T \lambda_\varepsilon^t \|\mathbf{N}_T^{-1} \mathbf{W}'_t\| \leq (\lambda_\varepsilon / (1 - \lambda_\varepsilon)) \sum_{t=1}^T \lambda_\varepsilon^t \|\mathbf{N}_T^{-1} \mathbf{W}'_t\|^2)^{1/2} = o_p(1)$$

since (A.9) and (A.11) imply that  $\sum_{t=1}^T \lambda_\varepsilon^t \|\mathbf{N}_T^{-1} \mathbf{W}'_t\|^2 = o_p(1)$ . Finally,

$$\sum_{t=1}^T \lambda_W^t \lambda_\varepsilon^t / T^{1/2} \leq (T^{-1} \lambda_W^2 / (1 - \lambda_W^2)) \lambda_\varepsilon^2 / (1 - \lambda_\varepsilon^2)^{1/2} = o_p(1).$$

Thus we have established (A.4).

The induction to show that  $\|\mathbf{N}_T(\hat{\boldsymbol{\beta}}_T^{(i)} - \tilde{\boldsymbol{\beta}}_T^{(i)})\| = o_p(1)$  for all  $i \geq 1$ , as well as the convergence in distribution to  $\overset{a}{\boldsymbol{\beta}}_T$ , is now completed by verifying that the components of  $\mathbf{N}_T(\hat{\boldsymbol{\beta}}_T^{(i)} - \boldsymbol{\beta})$  converge in distribution as stated. The proof of the latter follows along lines that parallel the developments in Yap and Reinsel (1995) and Dhrymes (1998). Recall that for an  $ECARMAX_E$  model the constraints of the echelon form are incorporated by simply deleting appropriate elements of  $\boldsymbol{\theta}$ . This implies that corresponding rows and columns in previous expressions involving  $\mathbf{W}_t$  are similarly removed and the same is true of (A.12)–(A.13) below. Hence the arguments of Yap and Reinsel (1995 *cf.* § 4, see in particular the comment in § 4.1) and Dhrymes (1998 *cf.* § 6.4) can be applied at this stage with very little modification. The argument proceeds by showing that  $\mathbf{N}_T^{-1} \sum_{t=1}^T \mathbf{W}'_t \boldsymbol{\Sigma}_T^{-1} \mathbf{W}_t \mathbf{N}_T^{-1}$  equals

$$\begin{bmatrix} T^{-1} \sum_{t=1}^T \mathbf{W}'_{\theta t} \boldsymbol{\Sigma}_\varepsilon^{-1} \mathbf{W}_{\theta t} & \mathbf{0} \\ \mathbf{0} & T^{-2} \sum_{t=1}^T \boldsymbol{\Upsilon}' [\mathbf{M}(1) \boldsymbol{\Sigma}_\varepsilon \mathbf{M}(1)']^{-1} \boldsymbol{\Upsilon} \otimes \mathbf{H}' \mathbf{z}_{t-1} \mathbf{z}'_{t-1} \mathbf{H} \end{bmatrix} + o_p(1) \quad (\text{A.12})$$

and that  $\mathbf{N}_T^{-1} \sum_{t=1}^T \mathbf{W}'_t \bar{\boldsymbol{\Sigma}}_T^{-1} \boldsymbol{\varepsilon}_t$  can be rewritten as

$$\begin{bmatrix} T^{-\frac{1}{2}} \sum_{t=1}^T \mathbf{W}'_{\theta t} \boldsymbol{\Sigma}_\varepsilon^{-1} \boldsymbol{\varepsilon}_t + o_p(1) \\ T^{-1} \sum_{t=1}^T (\boldsymbol{\Upsilon}' \mathbf{M}(1)^{-1} \boldsymbol{\Sigma}_\varepsilon^{-1} \otimes \mathbf{H}' \mathbf{z}_{t-1}) \boldsymbol{\varepsilon}_t + o_p(1) \end{bmatrix}. \quad (\text{A.13})$$

Both (A.12) and (A.13) are established by applying Proposition (4.1) to  $\mathbf{M}(z)$  to give

$$\mathbf{M}(1) \mathbf{W}_{\gamma t} = (\boldsymbol{\Upsilon} \otimes \mathbf{z}'_{t-1} \mathbf{H}) - \sum_{j=0}^{p-1} \tilde{\mathbf{M}}_j \Delta \mathbf{W}_{\gamma(t-j)}.$$

The sums of squares and cross-products involving  $\mathbf{W}_{\gamma t}$  will therefore be dominated by the components in  $(\mathbf{M}(1)^{-1} \boldsymbol{\Upsilon} \otimes \mathbf{z}'_{t-1} \mathbf{H})$  since the process  $\mathbf{H}' \mathbf{z}_t$  is integrated and  $\Delta \mathbf{W}_{\gamma t}$  and  $\mathbf{W}_{\theta t}$  are asymptotically-stationary processes. Substituting (A.12) and (A.13) into (A.2) and applying standard central limit theorems to (A.13) yields the required result.  $\blacksquare$

**Proof of Corollary 5.1:** We have already shown that  $\|\nabla \boldsymbol{\varepsilon}_t\| \leq \lambda_\varepsilon^t M_\varepsilon$  where  $M_\varepsilon < \infty$  and  $0 < \lambda_\varepsilon < 1$ . From the inequality  $\|\boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}'_t - \boldsymbol{\varepsilon}_{\langle t|t-1} \boldsymbol{\varepsilon}'_{\langle t|t-1}\rangle\| \leq \|\nabla \boldsymbol{\varepsilon}_t\|^2 + 2\|\boldsymbol{\varepsilon}_t\| \cdot \|\nabla \boldsymbol{\varepsilon}_t\|$  we therefore obtain the upper bound

$$\begin{aligned} \|\bar{\boldsymbol{\Sigma}}_T - \tilde{\boldsymbol{\Sigma}}_{1,T}\| &\leq T^{-1} M_\varepsilon^2 \sum_{t=1}^T \lambda_\varepsilon^{2t} + T^{-1} M_\varepsilon \sum_{t=1}^T \|\boldsymbol{\varepsilon}_t\| \lambda_\varepsilon^t \\ &\leq \frac{M_\varepsilon^2}{T(1-\lambda_\varepsilon^2)} + \frac{M_\varepsilon}{(1-\lambda_\varepsilon)^{\frac{1}{2}}} \sqrt{\frac{\sum_{t=1}^T \|\boldsymbol{\varepsilon}_t\|^2 \lambda_\varepsilon^t}{T^2}} \end{aligned}$$

and from (A.10) it follows that  $T^{\frac{1}{2}} \|\bar{\boldsymbol{\Sigma}}_T - \tilde{\boldsymbol{\Sigma}}_{1,T}\| = o(1)$ . Similarly, the inequality  $\|\nabla \boldsymbol{\Sigma}_{Tt}\| \leq 2K_\sigma \lambda_\sigma^t$  implies that  $\|\bar{\boldsymbol{\Sigma}}_T - \tilde{\boldsymbol{\Sigma}}_{2,T}\| \leq 2K_\sigma/T(1-\lambda_\sigma)$  and hence that  $T^{\frac{1}{2}} \|\bar{\boldsymbol{\Sigma}}_T - \tilde{\boldsymbol{\Sigma}}_{2,T}\| = o(1)$ .  $\blacksquare$

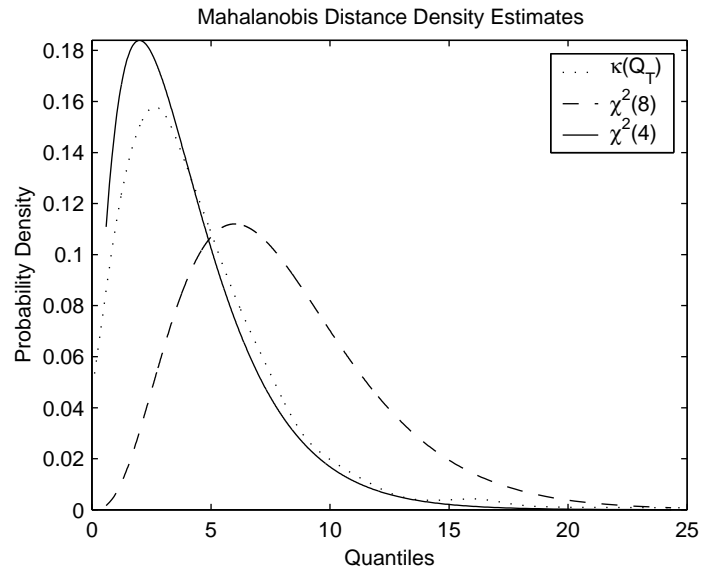


Figure 1: Empirical Distribution of Mahalanobis Distance. Processes  $P1$  with  $T = 150$ ,  $\rho_\epsilon = 0.8$  and  $\rho_{y,m} = 0.1Q_T$ . Kernel density estimate  $\kappa(Q_T)$  obtained using Gaussian kernel with optimal bandwidth.

**Table 1** Experimental Outcomes for Process 1

$\rho_{y,m} = 0.1Q_T$ and $\rho_\epsilon = 0.8$								
	$T = 50$				$T = 100$			
$[\tilde{\mathbf{A}}_T : \tilde{\mathbf{M}}_T]$	-0.4879	0.4998	-0.4999	0.4827	-0.4991	0.4981	-0.5082	0.4933
	-0.3896	-0.6019	-0.3854	-0.6329	-0.3985	-0.6027	-0.4019	-0.6143
	$M.S.E. = 0.7675$				$M.S.E. = 0.4325$			
$[\hat{\mathbf{A}}_T : \hat{\mathbf{M}}_T]$	-0.5919	0.6221	-0.6203	0.3936	-0.5880	0.5768	-0.5738	0.4461
	-0.5740	-0.5581	-0.5248	-0.7848	-0.5179	-0.5560	-0.4813	-0.7049
	$M.S.E. = 2.8421$				$M.S.E. = 1.3445$			
	$T = 150$				$T = 250$			
$[\tilde{\mathbf{A}}_T : \tilde{\mathbf{M}}_T]$	-0.5099	0.4974	-0.5178	0.5010	-0.5040	0.5021	-0.5063	0.5003
	-0.4076	-0.6044	-0.4070	-0.6091	-0.4047	-0.5986	-0.4031	-0.6049
	$M.S.E. = 0.2899$				$M.S.E. = 0.1989$			
$[\hat{\mathbf{A}}_T : \hat{\mathbf{M}}_T]$	-0.5529	0.5407	-0.5597	0.4072	-0.5412	0.5381	-0.5352	0.4780
	-0.4818	-0.5770	-0.4573	-0.6706	-0.4758	-0.5662	-0.4419	-0.6504
	$M.S.E. = 0.8454$				$M.S.E. = 0.5973$			

**Table 2** Experimental Outcomes for Process 2

$\rho_{y,m} = 0.1$ and $\rho_\epsilon = 0.8$								
	$T = 50$				$T = 100$			
$[\tilde{\mathbf{A}}_T : \tilde{\mathbf{M}}_T]$	-0.6043	0.3953	-0.5919	0.3813	-0.5996	0.4010	-0.6003	0.3989
	1.3912	0.3947	1.4078	0.3793	1.4007	0.4031	1.3989	0.3999
	$M.S.E. = 0.5257$				$M.S.E. = 0.2318$			
$[\hat{\mathbf{A}}_T : \hat{\mathbf{M}}_T]$	-0.5166	0.3062	-0.4732	0.4677	-0.6162	0.3080	-0.5387	0.4537
	1.4587	0.3763	1.4872	0.7091	1.7816	0.4398	1.5289	0.5939
	$M.S.E. = 31.8905$				$M.S.E. = 16.1529$			
	$T = 150$				$T = 250$			
$[\tilde{\mathbf{A}}_T : \tilde{\mathbf{M}}_T]$	-0.6034	0.3976	-0.6102	0.4019	-0.6021	0.3985	-0.5964	0.3951
	1.4004	0.4019	1.3970	0.4013	1.3959	0.3961	1.4009	0.3923
	$M.S.E. = 0.1646$				$M.S.E. = 0.0957$			
$[\hat{\mathbf{A}}_T : \hat{\mathbf{M}}_T]$	-0.5985	0.3327	-0.5812	0.4108	-0.5946	0.3728	-0.5753	0.4121
	1.6700	0.4040	1.4764	0.5178	1.5292	0.3899	1.4553	0.4504
	$M.S.E. = 3.1711$				$M.S.E. = 0.8459$			