COWLES FOUNDATION FOR RESEARCH IN ECONOMICS
AT YALE UNIVERSITY

Box 2125, Yale University
New Haven, Connecticut 06520

COWLES FOUNDATION DISCUSSION PAPER NO. 866R

OPTIMAL INFERENCE IN COINTEGRATED SYSTEMS

by

Peter C. B. Phillips

August 9, 1989

OPTIMAL INFERENCE IN COINTEGRATED SYSTEMS

by

P. C. B. Phillips*

Cowles Foundation for Research in Economics
Yale University

First draft: February, 1988

Revision: August, 1989

---

*This paper was written in January 1988 while the author was living on Waiheke Island and visiting the University of Auckland in New Zealand. I am grateful to P. Jeganathan, Lars Hansen and a referee for helpful comments on the first version of the paper, which is available as a Cowles Foundation Discussion Paper No. 866. My thanks as always go to Glena Ames for her skill and effort in keyboarding the manuscript of this paper and to the NSF for support under grant numbers SES 8519595 and SES 8821180.

0. ABSTRACT

This paper studies the properties of maximum likelihood estimates of cointegrated systems. Alternative formulations of such models are considered including a new triangular system error correction mechanism. It is shown that full system maximum likelihood brings the problem of inference within the family that is covered by the locally asymptotically mixed normal asymptotic theory provided that all unit roots in the system have been eliminated by specification and data transformation. This result has far reaching consequences. It means that cointegrating coefficient estimates are symmetrically distributed and median unbiased asymptotically, that an optimal asymptotic theory of inference applies and that hypothesis tests may be conducted using standard asymptotic chi–squared tests. In short, this solves problems of specification and inference in cointegrated systems that have recently troubled many investigators.

Methodological issues are also addressed and these provide the major focus of the paper. Our results favor the use of full system estimation in error correction mechanisms or subsystem methods that are asymptotically equivalent. They also point to disadvantages in the use of unrestricted VAR's that are formulated in levels and in certain single equation approaches to the estimation of error correction mechanisms. Unrestricted VAR's implicitly estimate unit roots that are present in the system and the relevant asymptotic theory for the VAR estimates of the cointegrating subspace inevitably involves unit root asymptotics. Single equation error correction mechanisms generally suffer from similar disadvantages through the neglect of additional equations in the system. Both examples point to the importance of the proper use of information in the estimation of cointegrated systems. In classical estimation theory the neglect of information typically results in a loss of statistical efficiency. In cointegrated systems deeper consequences occur. Single equation and VAR approaches sacrifice asymptotic median unbiasedness as well as optimality and they run into inferential difficulties through the presence of nuisance parameters in the limit distributions. The advantages of the use of fully specified systems techniques are shown to be all the more compelling in the light of these alternatives.

Key words: Cointegration; Error correction mechanisms; LAMN family; Maximum likelihood; SUR systems; Unit roots.

1. INTRODUCTION

Cointegration systems have recently been attracting the attention of both macroeconomists and econometricians. The field is unusually active with theoretical and empirical research going forward together. It has proved particularly interesting that well defined links exist between cointegrated systems, vector autoregressions (VAR's) and error correction models (ECM's). These links have served to bring different econometric methodologies closer together. But there is still little agreement amongst researchers about how best to proceed in empirical research. Is it appropriate to continue to use unrestricted VAR's in estimation and if so what theory of inference applies? Is it better to estimate a model in ECM format rather than as an unrestricted VAR? If so, can one improve further on the ECM methodology?

This paper attempts to address some of the questions above. Our approach is to compare the properties of full information estimation of ECM systems with alternatives such as unrestricted VAR's and direct estimation of cointegrating regressions. The critical differences between these procedures have not come to light in the existing literature. But it turns out that they are easily understood. In some cases, such as unrestricted VAR estimation, unit roots are implicitly or explicitly estimated along with other parameters. In other cases such as properly formulated ECM's they are not. This difference, which is rather obvious from the formulation of the two systems once it is pointed out, has a critical effect on the relevant asymptotic behavior of the likelihood function. In the former case one cannot avoid a unit root theory in the characterization of the likelihood. This puts us in the class of models which I have described elsewhere in Phillips (1989) as a limiting Gaussian functional (LGF) family. In the latter case, however, the problem turns out to belong to the locally asymptotically mixed normal (LAMN) family. The distinction is critical because in the latter case an optimal theory of inference exists (see Jeganathan (1980, 1982, 1988), Basawa and Scott (1983, 1984), Davies (1986) and LeCam (1986)). Whereas

in the former this is not so. Moreover, in the LAMN case, conventional asymptotic theory which relies on tabulations of the chi–squared distribution forms a valid asymptotic basis of inference. In the LGF case this is again not so and tabulations of nonstandard distributions are required as well as elimination of surplus nuisance parameters.

The present paper is related to a recent study by Johansen (1988). Johansen considers a nonstationary Gaussian VAR with some unit roots. He obtains the limit distribution of the maximum likelihood estimator (MLE) of the cointegrating vectors and the limit distributions of likelihood ratio tests of the dimension of the cointegrating space and of linear hypotheses about the coefficients. We also deal with full system maximum likelihood (ML) estimation of cointegrated systems and derive an asymptotic theory for our estimators and tests. But we distinguish between those cases where information about the presence of unit roots is used in estimation and those where it is not. This enables us to compare structural equation methods like FIML (which impose no unit roots) and full system ML estimation of ECM models (which impose a certain number of unit roots by virtue of their construction). These comparisons are facilitated by the use of a triangular system ECM representation which is quite different from the Engle–Granger (1987) representation that is employed by Johansen. Our system is linear in the parameters that define the cointegration space, whereas in the Engle–Granger representation the same parameters appear non linearly. This simplification means that explicit formulae for the estimators are usually available in our set up and the limit distribution theory is easy to derive. More general parametric and nonparametric models for the errors are also easily accommodated in our approach and, as we shall see, involve few complications over the simple case of iid errors. Finally, the triangular structure provides important insights concerning the special conditions under which different estimators are related, in particular when systems estimators are equivalent or asymptotically equivalent to certain subsystem estimators. This helps to furnish a link between the models and methods that we discuss here and the single equation ECM models that are common in empirical research.

The paper is organized as follows. All of our results are given in Section 2. This section sets up and motivates the triangular system ECM representation referred to above. A prototypical model with iid errors is used to demonstrate the properties of full system estimation of the ECM under a Gaussian likelihood. Theorem 1 gives the asymptotic distribution of the MLE of the cointegrating matrix and the parameters on which it depends, in this simple environment. The remainder, and the bulk, of Section 2 is organized as a series of remarks on this theorem. These serve to relate the results to other approaches like structural equation methods, unrestricted VAR's, nonlinear least squares, and subsystem and single equation approaches. We also show how the simplifying structure of the prototypical model and the conclusions of Theorem 1 continue to apply in the general context of a cointegrated system with linear process errors. Links with simultaneous equation methods and empirical ECM methodology are also explored. Many of the remarks emphasize heuristics and these are intended to help in understanding the similarities and the differences between conventional structural equation econometric theory and cointegrated systems theory. Some conclusions and recommendations for empirical research that emerge from the study are given in Section 3. Proofs are given in the Appendix.

A word on notation. We use $\text{vec}(A)$ to stack the rows of a matrix A into a column vector, A^* to represent the complex conjugate transpose of A, P_A to represent the orthogonal projection operator onto the range space of A, $\|A\|$ to signify the matrix norm $(\text{tr}(A'A))^{1/2}$, $[x]$ to denote the smallest integer $\leq x$ and $(x)_{-\infty}^{t}$ to represent the collection $(x_t\, x_{t-1}, \ldots)$. We use the symbol " $\Rightarrow$ " to signify weak convergence, the symbol " $\equiv$ " to signify equality in distribution and the inequality " > 0 " to signify positive definite when applied to matrices. Stochastic processes such as the Brownian motion $W(r)$ on $[0,1]$ are frequently written as W to achieve notational economy. Similarly, we write integrals with respect to Lebesgue measure such as $\int_0^1 W(s)ds$ more simply as $\int_0^1 W$. Vector Brownian motion with covariance matrix Ω is written " $BM(\Omega)$ ". We use $P(\cdot)$ to signify the probability measure of its argument, $\tilde{E}$ to denote wide sense conditional

expectation and I(1) to signify a time series that is integrated of order one. Finally, all limits given in the paper are taken as the sample size $T \to \infty$.

2. COINTEGRATED MODELS, THE TRIANGULAR SYSTEM ECM REPRESENTATION, ESTIMATION AND INFERENCE

Let y_t be an n–vector I(1) process and u_t be an n–vector stationary time series. We partition these vectors into subvectors of dimension n_1 and n_2 with $n = n_1 + n_2$ and assume that the generating mechanism for y_t is the cointegrated system

$$y_{1t} = By_{2t} + u_{1t} \tag{1}$$

$$\Delta y_{2t} = u_{2t} . \tag{2}$$

Here B is an $n_1 \times n_2$ matrix of coefficients and (1) may be thought of as a stochastic version of the linear long run equilibrium relationship $y_{1t} = By_{2t}$, with u_{1t} representing stationary deviations from equilibrum.

The ECM system arising from (1) and (2) can be written in triangular system format as follows:

$$\Delta y_t = -EAy_{t-1} + v_t \tag{3}$$

where

$$E = \begin{bmatrix} I_{n_1} \\ 0 \end{bmatrix}, \quad A = [I, -B], \quad v_t = \begin{bmatrix} I & B \\ 0 & I \end{bmatrix} u_t .$$

Equation (3) is a very convenient representation of the ECM and has several advantages over the autoregressive ECM representation that is used in Engle and Granger (1987) and Johansen (1988). First, the block triangular format of (3) ensures that generalized least squares (GLS) procedures are asymptotically equivalent to full maximum likelihood estimates. This is already a well known result of simultaneous equations theory in stationary

models with iid errors (e.g. Lahiri and Schmidt (1978)). When v_t in (3) is stationary rather than iid its serial covariance properties need to be attended to. This can be achieved by parametric maximum likelihood, by semiparametric corrections (see Phillips and Hansen (1989)) or by generalized least squares in the frequency domain (see Phillips (1988c)). The latter method is especially appealing since finite Fourier transforms preserve the triangular structure of (3) and enable us to deal with rather general stationary errors u_t on the original system. Second, the cointegrating coefficient matrix A and submatrix B appear linearly as the coefficients of y_{t-1} in (3). This is a great advantage because it simplifies estimation and makes the asymptotic theory much easier to follow. Third, all short–run dynamic behavior is absorbed in the residual v_t of (3). Again this simplifies the theory because questions of optimal inference about the long–run coefficients B are formally the same when v_t is a general stationary process as they are when v_t is iid. We shall discuss this more fully in Remarks (j)–(ℓ) below, which deal with models with stationary time series errors and show how an approximate pseudo–model with iid errors may be constructed when v_t is a stationary linear process.

For the reasons just given let us now assume that (3) is a prototypical system whose error vector $v_t \equiv$ iid $N(0,\Omega)$ with $\Omega > 0$. The normality theory is, as usual, needed for the development of the optimality theory but it is not necessary for the development of the asymptotics. The Gaussian log likelihood of (3) is

$$L(B,\Omega) = -(T/2)\ell n|\Omega| - (1/2)\Sigma_1^T(\Delta y_t + EAy_{t-1})'\Omega^{-1}(\Delta y_t + EAy_{t-1}) . \tag{4}$$

Partition Ω conformably with y and define $\Omega_{11\cdot 2} = \Omega_{11} - \Omega_{12}\Omega_{22}^{-1}\Omega_{21}$. Then $L(B,\Omega)$ may be written as the sum of the conditional log likelihood

$$\begin{aligned} &-(T/2)\ell n|\Omega_{11\cdot 2}| - (1/2)\Sigma_1^T(y_{1t} - By_{2t-1} - \Omega_{12}\Omega_{22}^{-1}\Delta y_{2t})' \\ &\quad \cdot \Omega_{11\cdot 2}^{-1}(y_{1t} - By_{2t-1} - \Omega_{12}\Omega_{22}^{-1}\Delta y_{2t}) \end{aligned} \tag{5}$$

and the marginal likelihood

$$(6) \qquad -(T/2)\ell n|\Omega_{22}| - (1/2)\Sigma_1^T \Delta y_{2t}' \Omega_{22}^{-1} \Delta y_{2t} .$$

Of course, the latter does not depend on the matrix B because of the triangular structure of (3). Moreover, provided B is unrestricted it is apparent from (5) that the maximum likelihood estimate of B is equivalent to the ordinary least squares (OLS) estimate from the linear model

$$(7) \qquad y_{1t} = By_{2t-1} + C\Delta y_{2t} + v_{1\cdot 2t}$$

where $C = \Omega_{12}\Omega_{22}^{-1}$ and $v_{1\cdot 2t} = v_{1t} - \Omega_{12}\Omega_{22}^{-1}v_{2t}$. Partitioned regression on (7) now yields in an obvious notation the formula

$$(8) \qquad T(\hat{B}-B) = (T^{-1}V_{1\cdot 2}'Q_\Delta \underline{Y}_2)\left[T^{-2}\underline{Y}_2'Q_\Delta \underline{Y}_2\right]^{-1}$$

where $\underline{Y}_2$ is the matrix of observations of y_{2t-1}. If there are restrictions on B, which lead, let us say to the form $\text{vec } B = J\alpha$ for some p–vector α and known matrix J of rank p, then optimal estimation requires the use of a consistent estimate $\hat{\Omega}_{11\cdot 2}$ of the error covariance matrix in (7). Such an estimate may be obtained by a preliminary unrestricted least squares regression on (7). We then have

$$\hat{\alpha} = [J'(\hat{\Omega}_{11\cdot 2} \otimes \underline{Y}_2'Q_\Delta Y_2)J]^{-1}[J'(\hat{\Omega}_{11\cdot 2} \otimes Y_2'Q_\Delta)\text{vec}(Y_1')] .$$

To extract the relevant asymptotics we use the fact that the innovations v_t in (3) satisfy the invariance principle

$$(9) \qquad T^{-1/2}\Sigma_1^{[Tr]} v_t \Rightarrow S(r) \equiv BM(\Omega) .$$

This will certainly be true when v_t is $\text{iid}(0,\Omega)$ or a strictly stationary and ergodic sequence of martingale differences with conditional variance matrix Ω —see Billingsley

(1968, Theorem 23.1). It also holds for much more general stationary processes, as discussed for example in Phillips and Durlauf (1986). We partition the limit process S conformably with Ω as $S' = (S_1', S_2')$ and define the component process $S_{1.2} = S_1 - \Omega_{12}\Omega_{22}^{-1}S_2 \equiv BM(\Omega_{11.2})$, which is independent of S_2. Using arguments analogous to those developed in Phillips (1986, 1987) we obtain the following asymptotics:

THEOREM 1.

$$(10) \qquad T(\hat{B}-B) \Rightarrow \left(\int_0^1 dS_{1.2}S_2'\right)\left[\int_0^1 S_2S_2'\right]^{-1} \equiv \int_{G>0} N(0, \Omega_{11.2} \otimes G)dP(G)$$

where $G = \left(\int_0^1 S_2S_2'\right)^{-1}$ *and* P *is its associated probability measure. When* $\text{vec } B = J\alpha$ *for some* p*–vector* α *and matrix* J *of rank* p *we have*

$$(11) \qquad T(\hat{\alpha}-\alpha) \Rightarrow \left[J'(\Omega_{11.2}^{-1} \otimes \int_0^1 S_2S_2')J\right]^{-1}[J'(\Omega_{11.2}^{-1} \otimes I)\int_0^1 dS_{1.2} \otimes S_2]$$

$$\equiv \int_{G>0} N\left[0, \left[J'(\Omega_{11.2}^{-1} \otimes G)J\right]^{-1}\right]dP(G) .$$

REMARK (a). The mixture representation of the limit distribution given in (10) is a simple consequence of the independence of the Brownian motions $S_{1.2}$ and S_2. The mixing variate may be a matrix as in (10) or a scalar as in the following representation established in Phillips (1989, Theorem 3.2):

$$\int_{g>0} N(0, g\Omega_{11.2} \otimes \Omega_{22}^{-1})dP(g) , \quad g = e'\left[\int_0^1 W_2W_2'\right]^{-1}e$$

where $W_2 \equiv BM(I_m)$ and e is any unit vector (with unity in one coordinate position and zeroes elsewhere).

REMARK (b). The asymptotics of Theorem 1 fall within the LAMN theory for the likelihood ratio as developed by Jeganathan (1980, 1982), LeCam (1986) and Davies (1986). This theory tells us that the likelihood ratio may be locally approximated by a quadratic in which the Hessian has a random limit. This leads to a random information matrix in the

limit and mixed normal asymptotics. It is worth showing the details in the present case. Let (H_B, H_Ω) be matrices of deviations for the parameter matrices (B,Ω). Set $h_B = \mathrm{vec}(H_B)$, $h_\Omega = D^+ \mathrm{vec}(H_\Omega)$ and $h' = (h_B', h_\Omega')$ where $D^+ = (D'D)^{-1}D'$ is the Moore Penrose inverse of the duplication matrix D. We expand the likelihood ratio that is based on (4) to the second order as follows:

$$\begin{aligned}
&\Lambda_T(h) = L(B + T^{-1}H_B, \Omega + T^{-1/2}H_\Omega) - L(B,\Omega) \\
&= \left[(1/2)\mathrm{tr}\{\Omega^{-1}T^{1/2}(M_{vv} - \Omega)\Omega^{-1}H_\Omega\} - \mathrm{tr}\{H_B(T^{-1}\underline{Y}_2'V)\Omega^{-1}E\}\right] \\
&\quad + (1/2)\left[-(1/2)\mathrm{tr}(\Omega^{-1}H_\Omega\Omega^{-1}H_\Omega) - \mathrm{tr}\{\Omega^{-1}EH_B(T^{-2}\underline{Y}_2'\underline{Y}_2)H_B'E'\}\right] + o_p(1) \\
(12) \quad &= h'w_T - (1/2)h'Q_Th + o_p(1)
\end{aligned}$$

where $M_{vv} = T^{-1}\Sigma_1^T v_t v_t'$, $\underline{Y}_2' = [y_{20}, \cdots, y_{2T-1}]$,

$$w_T = \begin{bmatrix} w_{1T} \\ w_{2T} \end{bmatrix} = \begin{bmatrix} -(E'\Omega^{-1} \otimes I)T^{-1}\Sigma_1^T v_t \otimes y_{2t-1} \\ (1/2)D'(\Omega^{-1} \otimes \Omega^{-1})\mathrm{vec}\{T^{1/2}(M_{vv} - \Omega)\} \end{bmatrix},$$

$$Q_T = \begin{bmatrix} E'\Omega^{-1}E \otimes T^{-2}\underline{Y}_2'\underline{Y}_2 & 0 \\ 0 & (1/2)D'(\Omega^{-1} \otimes \Omega^{-1})D \end{bmatrix}.$$

Now

$$(13) \qquad (w_T, Q_T) \Rightarrow (w,Q)$$

where

$$w = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} (E'\Omega^{-1} \otimes I)\int_0^1 dS \otimes S_2 \\ (1/2)D'(\Omega^{-1} \otimes \Omega^{-1})N(0, 2P_D(\Omega \otimes \Omega)) \end{bmatrix}$$

and

$$Q = \begin{bmatrix} E'\Omega^{-1}E \otimes \int_0^1 S_2 S_2' & 0 \\ 0 & (1/2)D'(\Omega^{-1} \otimes \Omega^{-1})D \end{bmatrix} .$$

The approximation (12) and the limit behavior given in (13) ensure that the log likelihood ratio belongs to the LAMN family of Jeganathan (1980).

Note that if we let $\mathcal{F}_t$ denote the σ–field generated by $\{S(r) : r \leq t\}$ and $\mathrm{var}(\cdot\,|\mathcal{F}_t)$ signifies the conditional variance relative to $\mathcal{F}_t$ then we have

$$(14) \qquad \int_0^1 \mathrm{var}(E'\Omega^{-1}dS \otimes S_2|\mathcal{F}_t) = E'\Omega^{-1}E \otimes \int_0^1 S_2(t)S_2(t)'dt .$$

This follows because the increments in a Brownian motion are independent of its past history. But note that $E'\Omega^{-1}S = \Omega_{11}^{-1}{}_{\cdot 2}(S_1 - \Omega_{12}\Omega_{22}^{-1}S_2) = \Omega_{11}^{-1}{}_{\cdot 2}S_{1\cdot 2} = \mathrm{BM}(\Omega_{11}^{-1}{}_{\cdot 2})$ and this process is independent of S_2. The random matrix (14) is the leading submatrix of Q. It is also a finite dimensional element of the quadratic variation process of $\int_0^t \Omega_{11}^{-1}{}_{\cdot 2}dS_{1\cdot 2} \otimes S_2$. Thus, in the notation of Metivier (1982) we have the square bracketed process

$$\left[\int_0^t \Omega_{11}^{-1}{}_{\cdot 2}dS_{1\cdot 2} \otimes S_2\right]_t = \Omega_{11}^{-1}{}_{\cdot 2} \otimes \int_0^t S_2 S_2' .$$

With this interpretation, the leading submatrix of Q is a natural candidate as a random variance for the limit process w_1.

In addition we have

$$\mathrm{var}(w_2) = (1/4)D'(\Omega^{-1} \otimes \Omega^{-1})[2P_D(\Omega \otimes \Omega)](\Omega^{-1} \otimes \Omega^{-1})D = (1/2)D'(\Omega^{-1} \otimes \Omega^{-1})D$$

corresponding to the lower diagonal submatrix of Q.

Finally, the inverse of Q is the information matrix

$$Q^{-1} = \begin{bmatrix} \Omega_{11\cdot 2} \otimes (\int_0^1 S_2 S_2')^{-1} & 0 \\ 0 & 2D^+(\Omega \otimes \Omega)D^{+\prime} \end{bmatrix} .$$

The leading submatrix of Q^{-1} is random and signifies random information in the limit for the maximum likelihood estimates of the cointegrating matrix B. This corresponds with the normal mixture given in (10). The lower diagonal submatrix gives the asymptotic variance matrix of the maximum likelihood estimates of the non redundant elements of Ω. If $\hat{\Omega}$ is the corresponding element of Ω we have

$$\sqrt{T}(\hat{\Omega}-\Omega) \Rightarrow N(0, 2P_D(\Omega \otimes \Omega)) .$$

This final result refers to the model (3) with error vector $v_t \equiv iid(0,\Omega)$. In the general case of stationary v_t, Ω is the long–run variance of v_t and kernel methods are usually employed in its estimation to deal with the fact that Ω depends on the entire serial covariance structure of v_t. This naturally affects the asymptotics for estimates of Ω. But the discussion above continues to apply in this case for the estimation of B.

The sense in which the estimator $\hat{B}$ is optimal under Gaussian assumptions is quite precise, just as in traditional ML estimation with a non random information matrix. A theory of optimality for inference from stochastic processes that is suitable in the present context has been developed by Sweeting (1983) and is discussed by Prakasa Rao (1987). We shall rely on their treatment here. We first observe that from the proof of Theorem 1 it is apparent that convergence to the limit distribution in (10) is uniform in B since the weak convergence results that are used there are independent of and hence uniform in B. If R_B denotes the limit probability measure in (10), $\mathcal{M}$ is the class of sets in $R^{n_1 \times n_2}$ that are convex and symmetric about the origin and $M \in \mathcal{M}$ then

$$P(T(\hat{B}-B) \in M) \rightarrow_u P_B(M)$$

where "$\rightarrow_u$" signifies uniform convergence on compact subsets of $R^{n_1 \times n_2}$. Now let $\mathcal{T}$ be a class of estimators B_T of B for which

$$T(B_T - B) \Rightarrow_u \tau_B$$

where τ_B is a limit variate with probability measure Q_B on $R^{n_1 \times n_2}$ and "$\Rightarrow_u$" signifies uniform weak convergence (with respect to $B \in R^{n_1 \times n_2}$). Under Gaussian assumptions the MLE $\hat{B}$ is optimal asymptotically in the class $\mathcal{T}$ in the sense that for any alternative estimate B_T whose limit variate is τ_B we have the inequality

$$Q_B(M) \leq R_B(M)$$

$\forall M \in \mathcal{M}$ and $\forall B \in R^{n_1 \times n_2}$. This implies that the MLE is efficient in the usual sense of having an asymptotic maximum concentration probability for all estimators in the class $\mathcal{T}$. When v_t is not Gaussian Theorem 1 still holds provided partial sums of v_t satisfy the invariance principle (9). But the Gaussian estimator $\hat{B}$ is no longer optimal. In this event the possibility of adaptive estimation exists. It has been explored recently in a deep and extensive study by Jeganathan (1988).

REMARK (c) Note that the coefficient matrix E in (3) is known and the ECM is just another algebraic representation of the original cointegrated system (1) and (2). The MLE $\hat{B}$ may therefore be obtained by applying ML directly to this original system rather than (3). ML estimation requires full specification of the model that generates u_t and the system must be estimated as specified with the n_2 unit roots eliminated as they are in (3). If the unit roots are estimated, either explicitly or implicitly, then the asymptotic distribution of the maximum likelihood estimator of B is different from that of $\hat{B}$ and, with one important exception that will be discussed below, no longer belongs to the LAMN family.

To see this it is simplest to write (1) and (2) in simultaneous equations format as

$$(15) \qquad \begin{bmatrix} I & -B \\ 0 & I \end{bmatrix} y_t = \begin{bmatrix} 0 \\ \Pi \end{bmatrix} y_{2t-1} + u_t \quad \text{with} \quad \Pi = I_{n_2}.$$

It is also convenient for the purposes of this demonstration to continue to assume serially independent errors and to set $u_t = \text{iid}(0,\Sigma)$. Then (15) is a conventional simultaneous

system with predetermined variables y_{2t-1}. Note that (15), like (3), is in triangular format and the second block is in reduced form. Assuming that there are no restrictions on Π or Σ, the full information maximum likelihood estimator (FIML) of B in (15) is simply the subsystem limited information maximum likelihood (LIML) estimator of B from the first n_1 equations. We shall derive the asymptotic distribution of this estimator.

As in the stationary simultaneous equations case, subsystem LIML is asymptotically equivalent to subsystem three stage least squares (3SLS)—the proof of this statement follows the same lines as the proof given by Sargan (1988, Theorem 5, p. 120) for the usual stationary case with some minor changes to the standardization factors for sample moment matrices. Furthermore, when there are no restrictions on the matrix B, subsystem 3SLS is equivalent to equation by equation two stage least squares (2SLS). The 2SLS estimator of B can be written quite simply as the matrix quotient $B^{\dagger} = Y_1'P_{-1}Y_2(Y_2'P_{-1}Y_2)^{-1}$ where P_{-1} is the orthogonal projector onto the range of $\underset{\sim}{Y}_2$. The asymptotic distribution theory for this estimator is straightforward and leads directly to the following result.

THEOREM 2. *If* $\tilde{B}$ *is the* FIML *estimator of* B *in the simultaneous system* (15) *then*

$$T(\tilde{B}-B) \Rightarrow (A\int_0^1 dSS_2')\left[\int_0^1 S_2S_2'\right]^{-1}$$

$$\equiv (\int_0^1 dS_{1.2}S_2')\left[\int_0^1 S_2S_2'\right]^{-1} + \Sigma_{12}\Sigma_{22}^{-1}(\int_0^1 dS_2S_2')\left[\int_0^1 S_2S_2'\right]^{-1}. \tag{16}$$

The FIML *estimator of* B *in* (15) *is asymptotically equivalent to that of the* MLE *in* (3) *iff* $\Sigma_{12} = 0$ *i.e. iff* y_{2t} *is strictly exogenous in the first block of* (15).

Note that, in general, the limit distribution (16) is a linear combination of the "unit root" distribution given by $(\int_0^1 dS_2S_2')\left[\int_0^1 S_2S_2'\right]^{-1}$ and the compound normal distribution $(\int_0^1 dS_{1.2}S_2')\left[\int_0^1 S_2S_2'\right]^{-1}$. This limit distribution falls within the LAMN family iff $\Sigma_{12} = 0$ i.e. iff y_{2t} is strictly exogenous in (15). The presence of the "unit root" component in the limit distribution is the consequence of the fact that FIML applied to (15)

(or equivalently subsystem LIML, 3SLS or 2SLS) involves the (implicit) estimation of the reduced form and, thereby, the unit roots that occur in the model. This inevitably means a breakdown in the LAMN theory, evidenced here by the form of (16). Only in the special case where y_{2t} is exogenous does the LAMN theory apply.

REMARK (d) It is of interest to observe that the special case above in which $\Sigma_{12} = 0$ is precisely the case when FIML and subsystem LIML reduce to ordinary least squares (OLS) on the first n_1 equations of (15). This is explained by the fact that when $\Sigma_{12} = 0$ (15) becomes a triangular system in which the covariance matrix Σ is block diagonal. The stated reduction of FIML to OLS is then well known from traditional econometric theory when $n_1 = 1$. When $n_1 > 1$ the reduction continues to apply provided the matrix B is unrestricted. Note that the equivalence of LIML and OLS on the first block of (15) means that the unit roots in the second block of (15) are not estimated either implicitly or explicitly and therefore the LAMN theory goes through.

REMARK (e) When $\Sigma_{12} \neq 0$, subsystem LIML and OLS on the first block of (15) are not equivalent. In this event the OLS estimator B^* has the following asymptotics:

$$T(B^* - B) \sim (A\int_0^1 dSS_2' + \Sigma_{12})\left[\int_0^1 S_2 S_2'\right]^{-1}$$

which differ from (16) by the additional bias term Σ_{12} in the numerator of the matrix quotient. Thus, in the general case, the use of simultaneous equations methods like LIML would seem to reduce the second order bias effects that occur with OLS but not to eliminate them entirely.

Theorem 1 shows that maximum likelihood estimation eliminates all bias effects asymptotically. This is of particular interest when we compare the asymptotic distributions of the MLE $\hat{B}$ and the FIML estimator $\tilde{B}$ in (15). Note that the usual effect in asymptotic statistical theory from employing more information is greater statistical efficiency. Here the extra information is the knowledge that the submatrix of the reduced

form coefficient matrix $\Pi = I$ in (15). Use of this information is all that distinguishes $\hat{B}$ from $\tilde{B}$. The effect on the asymptotic distribution of the use of this information is dramatic. All second order bias effects are removed, the asymptotic distribution becomes symmetric about B, it belongs to the LAMN family and an optimal theory of inference applies. None of these advantages apply if the information is not used, except when $\Sigma_{12} = 0$ and y_{2t} is strictly exogenous.

REMARK (f) The comments just made apply equally well in time series models to the comparison between unrestricted VAR estimation and maximum likelihood estimation of the full system ECM. In the former case unit roots are implicitly estimated unless, of course, the system is formulated in differences, which is not the approach followed in most empirical implementations of VAR's. It follows that the asymptotic theory for VAR based estimates of cointegrating vectors involves "unit root" type asymptotics, as in the case of the conventional FIML estimator discussed in Remark (c) above. These asymptotics have been studied elsewhere (see Park and Phillips (1988, 1989) and Phillips (1988a)) and we will not go into details here. It is sufficient to remark that the VAR estimates of the cointegrating subspace (i.e. the space spanned by the rows of A) involve nuisance parameters asymptotically and the relevant asymptotic theory is LGF, in the terminology of Phillips (1989), not LAMN. This means that nonstandard limit distributions are needed for inference, tabulations of these distributions need to allow for nuisance parameters, which have to be estimated, and no optimal asymptotic theory of inference is applicable. None of these drawbacks applies to full system ECM estimation by maximum likelihood and it would seem that the latter is preferable for empirical applications.

REMARK (g) As discussed in (e), knowledge of the presence of the n_2 unit roots in (2) has major statistical effects. The methodological aspects of this information are also interesting. From the form of the conditional Gaussian likelihood (5) we observed earlier that the MLE of B is just OLS on the linear model (7). By adding and subtracting Bu_{2t} to

the right side of (7) it is easy to see that this model may be written in the equivalent form

$$(7)' \qquad y_{1t} = By_{2t} + D\Delta y_{2t} + u_{1\cdot 2t}$$

where

$$D = \Omega_{12}\Omega_{22}^{-1} - B = \Sigma_{12}\Sigma_{22}^{-1}$$

$$u_{1\cdot 2t} = u_{1t} - \Sigma_{12}\Sigma_{22}^{-1}u_{2t} = u_{1t} - (\Omega_{12}\Omega_{22}^{-1} - B)u_{2t} = v_{1\cdot 2t}\,.$$

Of course $(7)'$ is just the original equation (1) with the error corrected for its conditional mean given $\Delta y_{2y} = u_{2t}$. Note that $(7)'$ is specified in levels (unlike the ECM) but it involves differences as additional regressors (whereas the ECM has levels as additional regressors). In the present case, the role of the difference Δy_{2t} in $(7)'$ as an additional regressor is simply to adjust the conditional mean and thereby remove the second order asymptotic bias effects that are present when OLS is applied directly to (1).

It should now be clear that what is important in estimation and inference in cointegrated systems, at least as far as ensuring the applicability of the LAMN theory, is not the precise form of the specification but the information concerning the presence of unit roots that is employed in estimation. If unit roots are known to be present, then our results argue strongly that they should be directly incorporated in model specification. It is perhaps one of the central advantages of the ECM formulation that it does this in a constructive way as part of the overall specification.

REMARK (h) The above remark should *not* be construed to mean that ECM formulations as they are presently used in econometric research automatically embody the advantages of the LAMN asymptotic theory. Virtually all ECM empirical work is conducted on a single equation basis and this is generally insufficient for the LAMN theory to apply. Our own analysis, and Theorem 1 in particular, is based on full system maximum likelihood estimation of (3). Since (3) is block triangular it is tempting to focus attention on the first block

of (3). However, neglect of the second block of equations in estimation involves more than a loss of efficiency, as we have seen. In most cases single equation estimation leads to a second order asymptotic bias of the type discussed earlier and complicates inference through the presence of nuisance parameters.

When the error vector $u_t \equiv$ iid $N(0,\Sigma)$ (or $v_t \equiv$ iid $N(0,\Omega)$) there is a simple way of incorporating the information that is necessary for efficient estimation into the first block of (3). In this case we have seen that full system maximum likelihood is equivalent to OLS on the regression equation (7)—i.e. the first block of (3) augmented by the regressor $u_{2t} = \Delta y_{2t}$. Thus, subsystem estimation is optimal on the augmented equation (7) or (7)′. When the error vector u_t is serially dependent the situation is more complex because there are feedbacks among the errors and the minimal information set for efficient estimation depends on the serial covariance structure of the errors. This issue, together with the link between ECM formulations and optimal estimation of cointegrated systems, is explored in Phillips (1988d). It is shown there that typical ECM specifications that include the present and past history of Δy_{2t} in the regressor set lead to optimal estimation by OLS when $u_2 = \Delta y_2$ is *strongly exogenous* in the sense of Engle *et al.* (1983). In addition to weak exogeneity (viz. that the marginal distribution of $(u_2)_1^T$ carries no information about the cointegrating coefficient matrix B) this requires that u_1 does not Granger cause u_2 (see Definition 2.6 of Engle *et al.* (1983)). When this applies we have the equivalence of the wide sense conditional expectations

$$\tilde{E}(u_{1t} | (u_2)_{-\infty}^{t}) = \tilde{E}(u_{1t} | (u_2)_{-\infty}^{t}, (u_2)_{t+1}^{\infty}) . \tag{17}$$

Obviously (17) is true when $u_t \equiv$ iid$(0,\Sigma)$. But when (17) does not hold and Δy_t is not strongly exogenous for B , it is necessary to augment the regression further by the inclusion of leads as well as lags of Δy_2 . Clearly, such augmentation reduces the advantages of working with single equation ECM formulations. An alternative semiparametric single

equation (or subsystem) method that avoids this problem is developed in Phillips and Hansen (1989).

We observe that the nonlinear least squares (NLS) procedure studied by Stock (1987) falls into the single equation category just described. This procedure involves a single equation NLS applied to an autoregressive version of the first equation of (3). In general, this approach has the same disadvantages of bias and nuisance parameter dependencies that have been discussed above. In fact, the simulation evidence reported in Stock (1987) indicates that the bias in the NLS cointegrating coefficient estimates can be substantial even in large samples. Stock's experimental study is based on the following two variable system (formulated with Stock's notation for the parameters)

$$(18) \qquad (1-\rho L)\Delta y_t = -\begin{bmatrix} \gamma_1 \\ \gamma_2 \end{bmatrix} \alpha' y_{t-1} + \epsilon_t\,, \quad \alpha' = (1, -\theta)$$

where $\epsilon_t \equiv$ iid $N(0, I_2)$. Stock reports large biases in the estimation of θ when $\gamma_2 \neq 0$ and ρ is small. On the other hand, a careful study of Stock's simulation results shows that the bias in the estimation of θ seems negligible when $\gamma_2 = 0$ and, in this case, the sampling distribution of the estimate is nearly symmetric about the true coefficient. Interestingly, $\gamma_2 = 0$ is a special case in which the asymptotic distribution of the NLS estimate of θ is the same as that of full system maximum likelihood and in this special case the LAMN theory applies.

It is easy to see why this is true. Since $\text{var}(\epsilon_t) = I$ and ρ is scalar it is clear that when $\gamma_2 = 0$ there is no information about θ in the second equation of (18). Moreover, with the autoregressive operator in (18) being diagonal there is no feedback from ϵ_1 to Δy_2. Thus, full system maximum likelihood estimation of θ in (18) is asymptotically equivalent to NLS on the first equation when $\gamma_2 = 0$, thereby explaining the good simulation performance of NLS in this case. In general, this asymptotic equivalence does not hold. In order to bring the NLS procedure within the realm of the LAMN theory and to

remove the second order asymptotic bias, it is generally necessary to do systems estimation. For Stock's procedure this amounts to seemingly unrelated systems NLS.

This example provides another illustration of the far reaching effects of prior information in regressions with nonstationary series. In a stationary time series or classical estimation context the neglect of information typically results in a loss of statistical efficiency. Thus, the use of single equation least squares instead of systems seemingly unrelated regression (SUR) techniques involves efficiency considerations alone in the classical regression setting. In the present context, deeper asymptotic issues come into play. The Stock example (18) shows that single equation approaches sacrifice asymptotic median unbiasedness as well as optimality and they run into major inferential difficulties through the presence of nuisance parameters in the limit distributions. The advantages of systems methods or SUR procedures are well known in classical regression. In the present context the advantages seem to be even more compelling.

REMARK (i) When Theorem 1 applies statistical testing may be conducted in the usual fashion as for asymptotic chi–squared criteria. This is a direct consequence of the mixed normal limit theory. For example, suppose we wish to test the hypotheses $H_0 : h(B) = 0$ where $h(\cdot)$ is a q–vector of twice continuously differentiable functions of the elements of B and $H = \partial h(B)/\partial \operatorname{vec} B'$ has full rank q. Then the Wald statistic for H_0 is $M_T = h(\hat{B})'(\hat{H}\hat{V}_T^{-1}\hat{H}')^{-1}h(\hat{B})$ where $\hat{H} = H(\hat{B})$ and $V_T = E'\hat{\Omega}^{-1}E \otimes \underline{Y}_2'\underline{Y}_2$. When $\hat{B}$ satisfies Theorem 1 and $\hat{\Omega}$ is any consistent estimator of Ω we have $M_T \Rightarrow \chi_q^2$. This theory continues to apply when the model has serially dependent errors but then $\Omega = 2\pi f_{uu}(0)$ is the long–run rather than the short–run covariance matrix and it must be estimated accordingly. The same result also holds for LR and LM tests of H_0 in the present context. Indeed, as in the classical setting, these tests are asymptotically equivalent with the same asymptotic χ_q^2 distribution as the Wald test M_T under the null. A closely related result has been given by Johansen (1988), who considers a Gaussian VAR

with cointegrated variates. Johansen proves that the likelihood ratio test of a linear hypothesis about the cointegrating vector is asymptotically distributed as chi–squared. For the reasons given here his theory applies also to more general hypotheses about the cointegrating coefficients and to other tests.

REMARK (j) Theorem 1 and the discussion contained in the preceding remarks refer to the prototypical model (3) with $v_t \equiv \text{iid}(0,\Omega)$. The time series case where v_t is stationary would seem *prima facie* to be much more complex. Surprisingly, this is not the case. All of the above ideas and results, especially our remarks concerning systems estimation and prior information about unit roots, continue to apply. What is required for the continued validity of Theorem 1 is the use of full systems estimation on (3) or at least an asymptotically equivalent subsystem procedure. If v_t is driven by a parametric scheme such as a vector ARMA model, then full system estimation by MLE involves the simultaneous estimation of the parameters of the stationary ARMA system and the coefficient matrix B of the long–run equilibrium relationship. Obviously this involves the construction of the likelihood function for general ARMA systems. An alternative approach that is developed in Phillips (1988c) is to deal with the time series properties of v_t nonparametrically by the use of systems spectral regression procedures on (3). The latter approach turns out to be most convenient because a discrete Fourier transform (dft) of (3) retains the basic form of this equation, including its triangular structure and the linearity of the coefficients. Moreover, for Fourier frequencies $\omega_j = 2\pi j/T$ that converge to zero as $T \to \infty$, the dft's of v_t are approximately distributed as iid $N(0,\underline{\Omega})$ with $\underline{\Omega} = 2\pi f(0)$, where $f(\omega)$ is the spectral density of v_t. Thus, for frequencies in the neighborhood of the origin, the dft of (3) is just a frequency domain version of our prototypical model. Spectral regression methods on (3) therefore have the same asymptotic properties for general stationary errors v_t as those of the MLE in Theorem 1 for $v_t \equiv \text{iid } N(0,\Omega)$. All that is needed in adjusting the results is to replace the contemporaneous (or short–run) covariance

matrix Ω by the long–run covariance matrix $\underline{\Omega}$. Since this approach is explored in detail in the cited paper (1988c) and in related work (1988e) by the author on continuous time systems estimation we shall say no more about it here.

It is worthwhile to look further at the parametric likelihood approach. Suppose, for example, that v_t in (3) is generated by the parametric linear process

$$(19) \qquad v_t = \Sigma_{j=0}^{\infty} C_j(\theta)\epsilon_{t-j}$$

where $\epsilon_t \equiv \text{iid}(0, \Sigma_\epsilon(\theta))$, $C_0 = I$ and the coefficient matrices $C_j(\cdot)$ depend on a q–vector of parameters θ and satisfy the summability condition

$$(20) \qquad \Sigma_{j=0}^{\infty} j^{1/2}\|C_j(\theta)\| < \infty$$

for all θ in a prescribed parameter space Θ. The model (19) includes AR models of the type considered by Johansen (1988), general ARMA systems and many other parametric linear time series models. For observable processes v_t, estimation of θ in (19) has been extensively studied in the stationary time series literature. In particular, Dunsmuir and Hannan (1976) and Dunsmuir (1979) establish strong laws and central limit theorems for Gaussian estimates of θ in (19) under quite general conditions using frequency domain approximations to the Gaussian likelihood—the so–called Whittle likelihood. This approach may also be applied in the context of the ECM (3) with linear process errors as in (19). In this case the Whittle likelihood that is to be minimized is given by

$$(21) \qquad L_T(B,\theta) = \ell n|\Sigma_\epsilon(\theta)| + T^{-1}\Sigma_j \text{tr}\{f(\omega_j;\theta)^{-1}I(\omega_j)\}, \quad -T/2 < j \le [T/2]$$

In this formula

$$f(\omega;\theta) = (1/2\pi)D(e^{i\omega};\theta)\Sigma_\epsilon(\theta)D(e^{i\omega};\theta)^*, \quad D(z;\theta) = \Sigma_0^{\infty} C_j(\theta)z^j$$

is the spectral density matrix of v_t, $I(\omega) = w(\omega)w(\omega)^*$ is the periodogram at frequency

$\omega \in (-\pi,\pi]$, $w(\omega) = (2\pi T)^{-1/2}\Sigma_1^T(\Delta y_t + EAy_{t-1})e^{it\omega}$ is a dft and $\omega_j = 2\pi j/T$ are the fundamental Fourier frequencies for $-T/2 < j \leq [T/2]$.

Now let $\hat{B}$ and $\hat{\theta}$ be the full system MLE's obtained by minimizing (21). Assuming that the regularity conditions used by Dunsmuir (1979) are satisfied we now have the following simple extension of Theorem 1 to the general time series case.

THEOREM 1′. *If* $\underline{\Omega} = 2\pi f(0) > 0$

$$T(\hat{B}-B) \Rightarrow (\int_0^1 dS_{1\cdot 2}S_2')\left[\int_0^1 S_2 S_2'\right]^{-1} \tag{22}$$

where $S \equiv BM(\underline{\Omega})$, $S_{1\cdot 2} \equiv BM(\underline{\Omega}_{11\cdot 2})$, $\underline{\Omega}_{11\cdot 2} = \underline{\Omega}_{11} - \underline{\Omega}_{12}\underline{\Omega}_{22}^{-1}\underline{\Omega}_{21}$ *and* S *and* $\underline{\Omega}$ *are partitioned conformably with* y_t .

REMARK (k) There is another, conceptually simpler way of looking at the time series case. The idea is to find an approximate pseudo–model that leads to the same asymptotics as Theorem 1′ but avoids the complications of explicit time series modeling. This is possible because the I(1) character of y_t is determined by partial sums of the errors that enter the ECM (3) period by period and these may be approximated by a suitable martingale. Thus, back substitution in (3) and initialization at $y_0 = 0$ gives rise to the representation

$$y_t = -E\Sigma_{j=1}^{t-1}Ay_{t-j} + \Sigma_{j=1}^{t}v_j . \tag{23}$$

The partial sum process $\Sigma_{j=1}^{t}v_j$ in (23) can be replaced by the martingale $Y_t = \Sigma_1^t V_j$ with an error that can be neglected in the asymptotics. When v_t is generated by (19) and (20) we may use $V_t = (\Sigma_{j=0}^{\infty}C_j)\epsilon_t$ as the approximating martingale difference sequence, just as we do in the martingale approach to central limit theory for a linear process (e.g. see Hall and Heyde (1980), Corollary 5.2, p. 135). Since $\epsilon_t \equiv iid(0, \Sigma_\epsilon)$ we have $V_t \equiv iid(0,\underline{\Omega})$ with $\underline{\Omega} = (\Sigma_{j=0}^{\infty}C_j)\Sigma_\epsilon(\Sigma_{j=1}^{\infty}C_j') = 2\pi f(0)$, as in Theorem 1′. The

approximating pseudo–model for (3) is obtained simply by replacing v_t with V_t giving

$$(3)' \qquad \Delta y_t = -EAy_{t-1} + V_t .$$

The Gaussian likelihood for (3)′ is identical with that of our earlier prototypical model (viz. (4)) upon replacement of the short run covariance matrix Ω with $\underline{\Omega}$. The asymptotic behavior of the full system MLE $\hat{B}$ may now be obtained by working from the pseudo model (3)′ with iid errors V_t, just as in Theorem 1.

Remark (ℓ) The simple heuristics of the last remark point to another interesting feature of optimal estimates of B. Such estimates rely only on consistent estimates of the covariance matrix—here the long run covariance matrix $\underline{\Omega}$. It is not necessary for optimal estimation of B that $\underline{\Omega}$ be jointly estimated. This is true even when $\underline{\Omega}$ is restricted as it may be, for instance, in the linear process case where $\underline{\Omega} = \underline{\Omega}(\theta)$. Interestingly, even in the prototypical model where $v_t \equiv \text{iid}(0,\Omega)$ and $\Omega = \sigma^2\Omega_0$ with Ω_0 a known matrix there is no information loss asymptotically for the estimation of B in estimating the full matrix Ω. Thus, if Ω_0 is known the coefficient matrix $C = \Omega_{12}\Omega_{22}^{-1}$ in (7) is also known and may be used in estimating the contracted system

$$(7)'' \qquad y_{at} = By_{2t-1} + v_{1\cdot 2t}\,,\ y_{at} = y_{1t} - C\Delta y_{2t}$$

rather than (7), where C is estimated. However, least squares on (7)" has the same asymptotic distribution as the estimate of B derived from (7) and is the same as that given in Theorem 1. Thus, in contrast to conventional simultaneous equations theory where there are efficiency gains in coefficient estimation from restrictions on the covariance matrix, there are no such gains in cointegrated systems estimation. The situation is analogous to SUR systems, where the regressors are exogenous and the information matrix is block diagonal. In cointegrated systems the regressors are not exogenous but they may be treated as such when Ω (or $\underline{\Omega}$ as appropriate) is consistently estimated. The pseudo-

model (3)′ where y_{t-1} and V_t are independent helps to explain this in the general time series case.

3. CONCLUSIONS

This paper started with two main objectives. The first was to study the asymptotic properties of maximum likelihood estimates of cointegrated systems. It has been shown that full system estimation by maximum likelihood brings the problem within the family that is covered by the LAMN theory of inference, provided all unit roots have been eliminated by specification and data transformation. This condition is crucial. If maximum likelihood does involve the estimation of unit roots, then the likelihood no longer belongs to the LAMN family. Instead it involves unit root asymptotics in terms of Gaussian functionals. These asymptotics import a bias and asymmetry into the cointegrating coefficient estimates and they carry nuisance parameter dependencies into the limit theory which inhibit inference.

The second and more important objective of the paper was to address the general question of how best to proceed in empirical research with cointegrated systems. Fortunately, the answer seems unambiguous. Full system estimation by maximum likelihood or asymptotically equivalent subsystem techniques that incorporate all prior knowledge about the presence of unit roots are most desirable. This approach ensures that coefficient estimates are symmetrically distributed and median unbiased, that an optimal theory of inference applies under Gaussian assumptions and that hypothesis tests may be conducted using standard asymptotic chi–squared tests. These are major advantages. The simplest approach in practice is to perform systems estimation of a fully specified ECM. Single equation estimation of an ECM is generally not sufficient unless the variables in the regressor set are strongly exogenous for the cointegrating coefficients. In *stationary* time series regression single equation estimation usually leads to a loss of statistical efficiency, as

in the seemingly unrelated regression context. But in cointegrated systems the use of single equation techniques imports bias, nuisance parameter dependencies and loses optimality. As a result the arguments for the use of systems methods in cointegrated systems seem more compelling than they are in a classical regression context.

We remark that in the cases where the system falls within the VAR framework unrestricted estimation of the VAR in levels does not bring the likelihood within the LAMN family. This is because in an unrestricted estimation in levels, unit roots are implicitly estimated in the regression. In consequence, the use of VAR's for inferential purposes about the cointegrating subspace suffers drawbacks relative to systems ECM estimation. However, as we stressed in Remark (g), the formulation of the model is less important than the information that it incorporates. If unit roots are known to be present, then our results indicate that it is best to incorporate them directly in the model specification. This can be done in VAR's, just as it is done constructively in ECM's. It might even be argued that suitably chosen Bayesian priors in VAR's go some way towards achieving the same end.

APPENDIX

Proof of Theorem 1. Continuing the partitioned regression notation in (8) we have

$$T^{-2}\underline{Y}_2'Q_\Delta\underline{Y}_2 = T^{-2}\underline{Y}_2'\underline{Y}_2 - T^{-2}(T^{-1}\underline{Y}_2'\Delta Y_2)\left[T^{-1}\Delta Y_2'\Delta Y_2\right]^{-1}(T^{-1}\Delta Y_2'\underline{Y}_2)$$
$$\Rightarrow \int_0^1 S_2 S_2'$$

and

$$T^{-1}V_{1\cdot 2}'Q_\Delta\underline{Y}_2 = T^{-1}V_{1\cdot 2}'\underline{Y}_2 - (T^{-1}V_{1\cdot 2}'\Delta Y_2)\left[T^{-1}\Delta Y_2'\Delta Y_2\right]^{-1}(T^{-1}\Delta Y_2'\underline{Y}_2)$$
$$\Rightarrow \int_0^1 dS_{1\cdot 2}S_2',$$

with both limits following by conventional weak convergence arguments (see Phillips (1988a, 1988b) for the required theory). Since joint weak convergence applies and $\hat{\Sigma}_{11\cdot 2} \to_p \Sigma_{11\cdot 2}$, both (10) and (11) follow directly.

Proof of Theorem 2. Since $\tilde{B}$ and $B^\dagger$ are asymptotically equivalent we need only consider $T(B^\dagger - B) = (T^{-1}U_1'P_{-1}Y_2)\left[T^{-2}Y_2'P_{-1}Y_2\right]^{-1}$. But

$$T^{-2}Y_2'P_{-1}Y_2 = (T^{-2}Y_2'\underline{Y}_2)\left[T^{-2}\underline{Y}_2'\underline{Y}_2\right]^{-1}(T^{-2}\underline{Y}_2'Y_2) \Rightarrow \int_0^1 S_2 S_2',$$

and

$$T^{-1}U_1'P_{-1}Y_2 = A(T^{-1}V'\underline{Y}_2)\left[T^{-2}\underline{Y}_2'\underline{Y}_2\right]^{-1}(T^{-2}\underline{Y}_2'Y_2) \Rightarrow A\int_0^1 dSS_2'.$$

Now note that

$$S_\sigma = \begin{bmatrix} AS \\ S_2 \end{bmatrix} \equiv BM\begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}.$$

Decompose $S_a = AS$ as follows

$$S_a = S_{a\cdot 2} + \Sigma_{12}\Sigma_{22}^{-1}S_2$$

where $S_{a\cdot 2} = BM(\Sigma_{11\cdot 2})$ and is independent of S_2. Notice that $\Sigma_{11\cdot 2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} = A\Omega A' - (\Omega_{12} - B\Omega_{22})\Omega_{22}^{-1}(\Omega_{21} - \Omega_{22}B') = \Omega_{11} - \Omega_{12}\Omega_{22}^{-1}\Omega_{21} = \Omega_{11\cdot 2}$. Thus $S_{a\cdot 2} \equiv S_{1\cdot 2} \equiv BM(\Omega_{11\cdot 2})$ and the stated result follows.

Proof of Theorem 1′. The first order conditions for $\hat{B}$ from the Whittle likelihood take the form

$$\Sigma_j E' f(\omega_j; \hat{\theta})^{-1} E(\hat{B}-B) w_2(\omega_j) w_2(\omega_j)^* = \Sigma_j E' f(\omega_j; \hat{\theta})^{-1} w_v(\omega_j) w_2(\omega_j)^*$$

where $w_v(\,)$ and $w_2(\,)$ are the dft's of v_t and y_{2t-1}, respectively. Under the regularity conditions in Dunsmuir (1979), $\hat{\theta}$ and $\hat{B}$ are consistent. Then, using the same lines of argument as those in the proof of Theorem 3.1 of (1988c) we find that

$$T^{-2}\Sigma_j E' f(\omega_j, \hat{\theta})^{-1} E \otimes w_2(\omega_j) w_2(\omega_j)^* \Rightarrow E'\underline{\Omega}^{-1}E \otimes \int_0^1 S_2 S_2' = \underline{\Omega}_{11\cdot 2}^{-1} \otimes \int_0^1 S_2 S_2'$$

and

$$T^{-1}\Sigma_j E' f(\omega_j, \hat{\theta})^{-1} w_v(\omega_j) w_2(\omega_j)^* \Rightarrow E'\underline{\Omega}^{-1}\int_0^1 dS S_2' .$$

Thus

$$T(\hat{B}-B) \Rightarrow \underline{\Omega}_{11\cdot 2}(E'\underline{\Omega}^{-1}\int_0^1 dS S_2')\left[\int_0^1 S_2 S_2'\right]^{-1} = (\int_0^1 dS_{1\cdot 2} S_2')\left[\int_0^1 S_2 S_2'\right]^{-1}$$

since $E'\underline{\Omega}^{-1}S \equiv BM(E'\underline{\Omega}^{-1}E) \equiv BM(\underline{\Omega}_{11\cdot 2}^{-1})$ and $S_{1\cdot 2} = \underline{\Omega}_{11\cdot 2}E'\underline{\Omega}^{-1}S \equiv BM(\Omega_{11\cdot 2})$.

REFERENCES

Basawa, I. V. and D. J. Scott (1983). *Asymptotic Optimal Inference for Non–Ergodic Models.* New York: Spring Verlag.

Billingsley, P. (1968). *Convergence of Probability Measures.* New York: Wiley.

Davies, R. B. (1986). "Asymptotic inference when the amount of information is random," in L. M. LeCam and R. A. Olshen (eds.), *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer,* Vol. II. Wadsworth Inc.

Doan, T., R. B. Litterman and C. Sims (1984). "Forecasting and conditional projections using realistic prior distribution," *Econometric Reviews*, 3, 1–100.

Dunsmuir, W. (1979). "A central limit theorem for parameter estimation in stationary vector time series and its application to models for a signal observed with noise," *Annals of Statistics*, 7, 490–506.

Dunsmuir, W. and E. J. Hannan (1976). "Vector linear time series models," *Advances in Applied Probability*, 8, 339–364.

Engle, R. F., D. F. Hendry and J. F. Richard (1983). "Exogeneity," *Econometrica*, 51, 277–304.

Hall, P. and C. C. Heyde (1980). *Martingale Limit Theory and its Application.* New York: Academic Press.

Hendry, D. F. (1986). "Econometric modeling with cointegrated variables: An overview," *Oxford Bulletin of Economics and Statistics*, 48, 201–212.

Jeganathan, P. (1980). "An extension of a result of L. LeCam concerning asymptotic normality," *Sankhya Series A*, 42, 146–160.

_____ (1982). "On the asymptotic theory of estimation when the limit of the log–likelihood ratios is mixed normal," *Sankhya Series A*, 44, 173–212.

_____ (1988). "Some aspects of asymptotic theory with applications to time series models," U. Michigan (mimeo).

Johansen, S. (1987). "Statistical analysis of cointegration vectors," *Journal of Economic Dynamics and Control*, 12, 231–254.

Lahiri, K. and P. Schmidt (1978). On the estimation of triangular structural systems," *Econometrica*, 45, 1217–1223.

LeCam, L. (1986). *Asymptotic Methods in Statistical Decision Theory.* New York: Springer.

Lütkepohl, H. (1984). "Linear transformations of vector ARMA processes," *Journal of Econometrics*, 26, 283–294.

Metivier, M. (1982). *Semimartingales.* New York: Walter de Gruyter.

Park, J. Y. and P. C. B. Phillips (1988). "Statistical inference in regressions with integrated processes: Part 1," *Econometric Theory*, 4, 468–497.

_____ (1989). "Statistical inference in regressions with integrated processes: Part 2," *Econometric Theory*, 5, 95–131.

Phillips, P. C. B. (1988a). "Multiple regression with integrated processes," in N. U. Prabhu (ed.), *Statistical Inference from Stochastic Processes, Contemporary Mathematics*, 80, 79–106.

_____ (1988b). "Weak convergence of sample covariance matrices to stochastic integrals via martingale approximations," *Econometric Theory*, 4, 528–533.

_____ (1988c). "Spectral regression for cointegrated time series," Cowles Foundation Discussion Paper No. 872, Yale University. Forthcoming in W. Barnett (ed.), *Nonparametric and Semiparametric Methods in Economics and Statistics*, CUP, 1990.

_____ (1988d). "Reflections on econometric methodology," *Economic Record*, 544–559.

_____ (1988e). "Error correction and long run equilibria in continuous time," Cowles Foundation Discussion Paper No. 882, Yale University.

_____ (1989). "Partially identified econometric models," *Econometric Theory*, 5, 181–240.

Phillips, P. C B. and S. N. Durlauf (1986). "Multiple time series with integrated variables," *Review of Economic Studies*, 53, 473–496.

Phillips, P. C. B. and B. E. Hansen (1989). "Statistical inference in instrumental variables regression with I(1) processes," *Review of Economic Studies* (forthcoming).

Prakasa Rao, B. L. S. (1986). *Asymptotic Theory of Statistical Inference.* New York: Wiley.

Sargan, J. D. (1988). *Lectures on Advanced Econometric Theory.* Oxford: Basil Blackwell.

Stock, J. H. (1987). "Asymptotic properties of least squares estimators of cointegrating vectors," *Econometrica*, 55, 1035–1056.

Sweeting, T. J. (1983). "On estimator efficiency in stochastic processes," *Stochastic Processes and their Applications*, 15, 93–98.