FORECASTING AND TESTING IN CO-INTEGRATED SYSTEMS*

Robert F. ENGLE and Byung Sam YOO

University of California, San Diego, CA 92093, USA

This paper examines the behavior of forecasts made from a co-integrated system as introduced by Granger (1981), Granger and Weiss (1983) and Engle and Granger (1987). It is established that a multi-step forecast will satisfy the co-integrating relation exactly and that this particular linear combination of forecasts will have a finite limiting forecast error variance. A simulation study compares the multi-step forecast accuracy of unrestricted vector autoregression with the two-step estimation of the vector autoregression imposing the co-integration restriction.

To test whether a system exhibits co-integration, the procedures introduced in Engle and Granger (1987) are extended to allow different sample sizes and numbers of variables.

1. Introduction

Vector autoregressions provide a convenient representation for both estimation and forecasting of systems of economic time series. Sims (1980) and Litterman (1986) are two of the influential papers which use VAR's for economic analysis and there are many other examples. The forecasting performance of unrestricted VAR's has not been particularly good and the question of whether to preprocess the data by transformations such as differencing so that the analysis will be conducted on stationary series, has perplexed investigators. In this paper it is proposed that the problems are related and that a useful and simple solution is available.

It was first pointed out in Granger (1981) that a vector of time series, all of which are stationary only after differencing, may have linear combinations which are stationary without differencing. In such a case, those variables are said to be co-integrated [see Granger (1986) for a survey]. For such a system there is obviously a dilemma deciding how much differencing to do as the differencing required depends upon the linear combination under consideration. It is easily shown that if all variables are differenced as would appear appropriate from their univariate properties, then the system no longer has a

multi-variate linear time series representation with an invertible moving average. Essentially, the system has been over-differenced.

Engle and Granger (1987) establish that a co-integrated system can be represented in an error correction structure which incorporates both changes and levels of variables such that all the elements are stationary. This error correction structure provides the framework for estimation, forecasting and testing of co-integrated systems. They developed a representation theorem which is reproduced in part here, an estimation procedure and testing procedures. This paper will explore the multi-step forecasting behavior of co-integrated systems and examine the performance of the two-step estimator proposed by Engle and Granger. It will finally extend the tables of test critical values to the multi-variate case for a variety of sample sizes.

2. Some properties of co-integrated systems

The properties of multi-step forecasts of co-integrated systems differ markedly from either stationary VAR's or VAR's in differences. This section will develop the relationships.

Assumptions

An $N \times 1$ vector process $\{x_t;\ t = 1, 2, 3, \dots\}$ has been generated by

$$(1 - B)x_t = C(B)(\varepsilon_t + m)$$
$$= \mu + C(B)\varepsilon_t, \tag{1}$$

where m is an arbitrary $N \times 1$ vector of constants and $\mu = C(1)m$.

(i) $\varepsilon_j = 0,\ \forall j \leq 0$ and $x_0 = 0$.

(ii) $\{\varepsilon_t\}_1^\infty$ is a vector white noise process with

$$\mathrm{E}(\varepsilon_t) = 0, \qquad \forall t \geq 1,$$
$$\mathrm{E}(\varepsilon_i \varepsilon_j') = \Omega\delta(i - j), \qquad \forall i, j \geq 1,$$

where Ω is an $N \times N$ positive definite matrix and δ is the delta function.

(iii) $C(B)$ is an $N \times N$ matrix function in B, the lag operator, given by

$C(B) = C_0 + C_1 B + C_2 B^2 + \dots$, such that

(a) no row of $C(1)$ is identically zero,
(b) $C(e^{i\omega})\Omega C'(e^{-i\omega}) < \infty, \qquad \forall \omega \in [0, \pi]$,
(c) $C(0) = I_N$.

(iv) The rank of $C(1) = N - r$.

Assumption (iii) implies that all the elements of x_t need to be differenced once to become individually stationary. Such time series are called integrated of order one and denoted $I(1)$. Assumption (iv) implies that the variables are co-integrated. Any $N \times r$ matrix α which spans the null space of $C(1)$ so that $\alpha'C(1) = 0$, will produce an $r \times 1$ time series $z_t = \alpha'x_t$ which is stationary. To see this, write $C(B) = C(1) + C^*(B)(1 - B)$. Then

$$(1 - B)\alpha'x_t = \alpha'C(1)(\varepsilon_t + m) + (1 - B)\alpha'C^*(B)\varepsilon_t,$$

so that

$$z_t = \alpha'C^*(B)\varepsilon_t \quad \text{with} \quad \alpha'C^*(1) < \infty.$$

In Engle and Granger (1987) it is assumed that all deterministic components are extracted before analysis so that $m = 0$. Here we slightly generalize the notion of co-integration. In particular, even though each element in x_t is $I(1)$ with drift so that it has a deterministic trend and a variance which goes to infinity with t [i.e., $\text{var}(x_t) = O(t)$], the linear combination $\alpha'x_t$ will be stationary. Thus μ must have property that $\alpha'\mu = 0$ so that the same linear combination which eliminates the unit root will also eliminate the trend.

The time series represented by (1) can be integrated as follows:

$$x_1 = \mu + \varepsilon_1,$$

$$x_2 = 2\mu + \varepsilon_2 + (C_1 + I)\varepsilon_1,$$

$$x_3 = 3\mu + \varepsilon_3 + (C_1 + I)\varepsilon_2 + (C_2 + C_1 + I)\varepsilon_1,$$

so that

$$x_t = t\mu + \sum_{i=1}^{t} \sum_{j=0}^{t-i} C_j \varepsilon_i, \tag{2}$$

and finally for $h \geq 1$,

$$x_{t+h} = (t+h)\mu + \sum_{i=1}^{t} \sum_{j=0}^{t+h-i} C_j \varepsilon_i + \sum_{i=1}^{h} \sum_{j=0}^{h-i} C_j \varepsilon_{t+i}. \tag{3}$$

This form is the basis for multi-step forecasts.

An alternative and closely related representation is introduced by Stock and Watson (1986) which they call the Common Trends representation. This form is convenient for discussing long-term forecasts. A set of time series can be decomposed into deterministic trends, stochastic trends or unit root processes, and stationary components. To see this, rewrite (2) as

$$x_t = \mu t + (1 + B + B^2 + \cdots + B^{t-1})C(B)\varepsilon_t.$$

Using $C(B) = C(1) + (1 - B)C^*(B)$ and denoting

$$y_t = \sum_{i=1}^{t} \varepsilon_i,\tag{4}$$

we get

$$x_t = \mu t + C(1)y_t + C^*(B)\varepsilon_t,\tag{5}$$

since $B'C(B)\varepsilon_t = 0$ by initial conditions.

The deterministic trend is now reflected in μt, while the stochastic trend is given by $C(1)y_t$. The remainder is stationary for large t, if the C_j's are exponentially decaying for large j. This is called a Common Trends representation because the reduced rank of $C(1)$ implies that there are fewer stochastic trends (or unit roots) than variables in the system.

3. Forecasting

Let $\mathscr{F}_t = \sigma(\varepsilon_j; \ j = 1, \ldots, t)$ be the information set generated by the random variables ε_j ($1 \le j \le t$). We shall consider the orthogonal projection of x_{t+h} on $\mathscr{F}_t$, denoted by $x_{t+h|t}$.

From (3) it follows that

$$x_{t+h|t} = \mu(t+h) + \sum_{i=1}^{t} \sum_{j=0}^{t+h-i} C_j \varepsilon_i.\tag{6}$$

Notice that

$$\lim_{h \to \infty} \sum_{j=0}^{t+h-i} C_j = C(1),$$

so that

$$\lim_{h \to \infty} a' x_{t+h|t} = 0,\tag{7}$$

under the co-integration assumption (iv). Thus the long-run forecasts of co-integrated systems are tied together regardless of the fact that individual forecasts diverge to infinity. The co-integrating relationship will hold exactly in the long-run forecast. This is also implied by the Common Trends representation, where the trends in the individual series are such that the linear combination with weights α has no trend.

Note that $\{C_j; \ j = 0, 1, \ldots\}$ will typically be an exponentially decaying sequence for large j and therefore the convergence

$$\sum_{j=0}^{t+h-i} C_j \to C(1)$$

will be rather rapid as $h \to \infty$. Hence (7) may be observed in a moderate forecasting horizon. Whenever h is sufficiently large that this limit is satisfied, (6) can be rewritten as

$$x_{t+h|t} = \mu(t+h) + C(1)y_t, \tag{8}$$

using the notation from (4). Thus the trend component at time t can be given by

$$\mu t + C(1)y_t = x_{t+h|t} - \mu h, \quad \text{for } h \text{ sufficiently large.}$$

This is the approach to trend extraction suggested for univariate series by Beveridge and Nelson (1981) and Watson (1986) and extended to the multivariate common trend set-up by Stock and Watson (1986).

We next consider the forecast error variance. Denoting the h-step forecast error by $e_{t+h|t}$, we have

$$e_{t+h|t} \equiv x_{t+h} - x_{t+h|t} = \sum_{i=1}^{h} \sum_{j=0}^{h-i} C_j \varepsilon_{t+i}. \tag{9}$$

It follows that

$$\text{var}[e_{t+h|t}] = \sum_{i=1}^{h} \left[\left(\sum_{j=0}^{h-i} C_j \right) \Omega \left(\sum_{j=0}^{h-i} C_j' \right) \right]. \tag{10}$$

From (9), it also follows that

$$e_{t+h|t} - e_{t+h-1|t} = \sum_{i=1}^{h} C_{h-i} \varepsilon_{t+i}.$$

The right-hand-side quantity is $C(B)\varepsilon_{t+h}$ taking $\varepsilon_j = 0$ for all $j \leq t$. Therefore the forecast error process has the same stochastic structure as the original process x_t and is co-integrated. Writing the forecast error equation similarly to (5), without the deterministic component, immediately reveals that

$$\text{var}[e_{t+h|t}] = \text{O}(h), \tag{11}$$

and

$$\text{var}[\alpha' e_{t+h|t}] = \alpha' K \alpha < \infty,$$ (12)

for large h, where the constant matrix K comes from the stationary component of the forecast error.

4. Methods for forecasting co-integrated variables

Multi-step forecasts from co-integrated systems have a property not shared by general integrated systems: linear combinations of forecasts are identically zero for large horizons regardless of the forecast origin, and the forecast error variance for this linear combination remains finite while that for all other linear combinations goes to infinity as the horizon goes to infinity. The practical implications of these observations are best seen when it is assumed that the data follow a finite vector autoregression (VAR). The question of whether the data are co-integrated then becomes a question of how many unit roots are in the population autoregressive polynomial, and the question of whether forecasts from the estimated VAR exhibit the desired properties depends upon the roots of the estimated VAR.

Engle and Granger (1987) present and prove a representation theorem due to Granger (1983) which establishes the connection between the moving average representation in (1) and the VAR when $\mu = 0$. This is briefly presented below.

Granger representation theorem

If the $N \times 1$ vector x_t given in (1) satisfies the assumptions (i)–(iv) and is therefore co-integrated with co-integrating rank r, and if this vector can be represented by a finite vector autoregression, then:

(a) $A(1)$ has rank r and $A(0) = I_N$ in

$$A(B)x_t = \varepsilon_t + m.$$ (13)

(b) There exist $N \times r$ matrices, α, γ, of rank r such that

$$\alpha' C(1) = 0, \qquad C(1)\gamma = 0, \qquad A(1) = \gamma\alpha'.$$

(c) There exists an error correction representation with $z_t = \alpha' x_t$, an $r \times 1$ vector of stationary random variables:

$$A^*(B)(1 - B)x_t = -\gamma z_{t-1} + \varepsilon_t + m \quad \text{with} \quad A^*(0) = I_N.$$ (14)

Eq. (13) describes the unrestricted VAR which can be used to estimate and forecast the vector x_t. This representation, however, has restrictions implied by the $N - r$ unit roots which would not be imposed by such an unrestricted estimation. On the other hand, (14) shows that a VAR in differences is inappropriate as the levels should appear through z. In fact the vector $(1 - B)x_t$ does not have a vector ARMA representation with an invertible moving average. Such a series is 'over-differenced'.

Three alternative approaches to the estimation of this system may be identified:

(i) Unrestricted estimation of (13) or equivalently of (14) without imposing the restriction that the level variables in z must be the same in each equation. If the lags are the same in each equation this is simply OLS equation by equation. Call this an unrestricted vector autoregression (UVAR).

(ii) Maximum likelihood of (14) under normality which imposes the cross-equation restrictions and uses SURE estimation.

(iii) Two-step estimation where α is estimated by a static least squares regression called the 'co-integrating regression' and then this value of α is used in estimating (14) thereby imposing the cross-equation constraints. The estimation procedure is again OLS equation by equation including the estimated value of z_{t-1} in each equation. As this was proposed by Engle and Granger we call this the EG 2-step.

It is clear that the second approach is theoretically attractive as it imposes all the restrictions known to be true and estimates the system fully efficiently. The first and third are, however, serious contenders. Engle and Granger introduce the two-step estimator and prove that (in the case $\mu = 0$) it is asymptotically just as efficient an estimate of $A^*(B)$ as the estimator using the true value of α which is of course not a feasible estimator. Phillips and Durlauf (1986) and Said and Dickey (1984) in a more restricted context show that the least squares estimate of (14) without restrictions gives estimates of the parameters of the $A^*(B)$ polynomial which are asymptotically the same as those assuming knowledge of α. Consequently, all three of these estimators have the same limiting distribution for the $A^*(B)$ parameters.

For finite samples however, the three differ in an important fashion. The rank of $A(1)$ will be exactly r for estimates (ii) and (iii), but for (i) it will generally be N except for an event with probability measure zero. In particular, the well-known downward bias of the autoregressive parameter in a univariate model may suggest that $A(B)$ will too often look like it has no unit roots for finite samples.

Differences in the forecasting performance of VAR's of co-integrated systems estimated by (i) and (iii) should therefore show up particularly for small samples and for multi-step forecasts. In the next section a small simulation is carried out to examine this contention.

5. Simulation

To examine the relative forecasting accuracy of models estimated by these methods, a small simulation experiment was designed. Only the least squares methods (i), UVAR, and (iii), EG 2-step, were examined and the problem was set up so that both correctly specified the model. The data generation process is given in three different representations:

Moving Average

$$(1 - B)\begin{pmatrix} x_t \\ y_t \end{pmatrix} = (1 - 0.4B)^{-1}\begin{pmatrix} 1 - 0.8B & 0.8B \\ 0.1B & 1 - 0.6B \end{pmatrix}\varepsilon_t, \tag{15}$$

Autoregressive

$$\begin{pmatrix} 1 - 0.6B & -0.8B \\ -0.1B & 1 - 0.8B \end{pmatrix}\begin{pmatrix} x_t \\ y_t \end{pmatrix} = \varepsilon_t, \tag{16}$$

Error Correction

$$(1 - B)\begin{pmatrix} x_t \\ y_t \end{pmatrix} = \begin{pmatrix} -0.4 \\ 0.1 \end{pmatrix}(1 \quad -2)\begin{pmatrix} x_{t-1} \\ y_{t-1} \end{pmatrix} + \varepsilon_t, \tag{17}$$

where

$$\mathrm{var}(\varepsilon_t) = \begin{pmatrix} 100 & 0 \\ 0 & 100 \end{pmatrix}, \qquad x_0 = y_0 = 0.$$

As can best be seen from the error correction representation in (17), the long-run relation between x and y is $x = 2y$.

One hundred replications of this design were computed in each case, with 100 observations used for fitting the model and 20 observations used for post sample forecasting. The two-step error correction estimation proposed by Engle and Granger (EG) was compared with the unrestricted vector autoregression in terms of mean square error for horizons from 1 to 20 periods. Here the mean square error is the trace of the sample covariance matrix of the forecast errors. The results are presented in table 1.

Table 1

Mean square forecast errors for multi-step forecasts from co-integrated systems.

Horizon	EG 2-step	UVAR	Ratio (%)
1	253	230	110
2	445	392	113
3	608	526	116
4	888	798	111
5	1008	995	101
6	1282	1300	99
7	1350	1445	93
8	1647	1855	89
9	1960	2247	87
10	2164	2574	84
11	2363	2833	83
12	2550	3081	83
13	2712	3371	80
14	3223	4095	79
15	3661	4687	78
16	4022	5158	78
17	4452	5831	76
18	4763	6455	74
19	5012	6870	73
20	4978	7021	71

From these results it is clear that for the shortest-period forecasts there are gains to using the unrestricted vector autoregression. The maximum gain is 16% which occurs at the three-step forecast. Through step 4 there is an apparent advantage to the VAR, however for longer-run forecasts the two-step estimator is more accurate. The increase in forecast accuracy rises to 40% ($1/0.71 = 1.40$) after 20 periods and probably would go to infinity for sufficiently long forecast horizons. The fact that the restricted model performed better than the unrestricted model is expected a priori and hence may not be surprising. However, what is being advocated here is the importance of imposing long-run constraint rather than the restriction per se. One might not expect such a difference in the forecasting performances of two correctly specified models in standard cases. In fact, the two-step procedure does not necessarily lead to better parameter estimates as a whole as revealed in the short-term forecasts. One could imagine that the non-linear least squares estimator which imposes the cross-equation restrictions implicit in the error correction representation would dominate both of these estimators. Much however would depend upon the convergence of this estimator, the starting values chosen and the treatment of forecasts from non-convergent samples.

One could also compare these results with estimates which are obviously misspecified such as least squares on differences or Litterman's (1986) Bayesian Vector Autoregression which shrinks the parameter vector toward

the first difference model which is itself misspecified for this system. The finding that such methods provided inferior forecasts would hardly be surprising.

6. Testing for co-integration

Engle and Granger (1987) investigated tests for the null hypothesis that a pair of the time series which were each $I(1)$ were 'non-co-integrated' against the alternative that they were co-integrated. That is, the null hypothesis is that the system has two unit roots, while the alternative is one unit root. For a sample size of 100 several test statistics were considered. Monte Carlo methods were used to obtain the finite sample critical values and then to examine the power properties of the tests. In a first-order system, two procedures were found to be the best: a Durbin–Watson (DW) test and a Dickey–Fuller (DF) test. In higher-order systems, it is simple to generalize the Dickey–Fuller test to the Augmented Dickey–Fuller test (ADF) which was the recommendation of Engle and Granger.

The tests are computed by performing two regressions. The first, called the co-integrating regression, fits the static bivariate model

$$y_t = \hat{\varphi} + x_t \hat{\pi} + z_t, \tag{18}$$

where z_t is the residual term which is also interpreted as the co-integrating linear relation. The Durbin–Watson test simply examines the DW of this regression to see if it is significantly greater than zero, which would be its probability limit if z_t contains a unit root as required by the null hypothesis. At the second stage, the DF and ADF tests are obtained respectively as the t-statistics of $\hat{\rho}$ in the following regressions $[\Delta \equiv (1 - B)]$:

$$\Delta \hat{z}_t = \hat{\rho} z_{t-1}, \tag{19}$$

$$\Delta \hat{z}_t = \hat{\rho} z_{t-1} + \sum_{i=1}^{p} \hat{\delta}_i \Delta z_{t-i}. \tag{20}$$

In the next section, similar critical values will be developed for other sample sizes and with more than two variables using a simulation approach. In the remainder of this section we discuss the theoretical results which are available for testing unit roots. If φ and π are known, the problem becomes simply a test for a unit root in a univariate process. The more complex case where φ is unknown has been treated by Dickey and Fuller (1979) and the co-integration problem is therefore seen as a further extension to unknown π.

We assume that (y_t, x_t) is generated by

$$y_t = y_{t-1} + v_t,$$

$$x_t = x_{t-1} + w_t,$$

$$\begin{pmatrix} v_t \\ w_t \end{pmatrix} \sim \text{i.i.d.} \left[0, \begin{pmatrix} \sigma_v^2 & 0 \\ 0 & \sigma_w^2 \end{pmatrix} \right], \qquad y_0 = x_0 = 0.$$

For known φ and π, z_t follows a random walk. In particular, this model implies that z_t is equal to y_t since the true values of φ and π are zero. Note that allowing the initial values of the covariance term to be non-zero results in non-zero values of φ and π. However, provided that the estimators of these parameters are understood to be centered around the true values, relaxing these assumptions will not alter the discussion that follows but will only complicate the notation. Hence we maintain this simple set-up. The weak convergence results for the OLS estimator $\hat{\rho}$ in (19) [White (1958)] and associated t-statistics [Dickey and Fuller (1979) and Phillips (1985a)] are known to be

$$T\hat{\rho} \xrightarrow{d} \frac{\left[V(1)^2 - 1 \right]/2}{\int_0^1 V(t)^2 \, dt}, \tag{21}$$

$$t_{\hat{\rho}} \xrightarrow{d} \frac{\left[V(1)^2 - 1 \right]/2}{\left[\int_0^1 V(t)^2 \, dt \right]^{1/2}}, \tag{22}$$

where T is the sample size and $V(t)$ is a standard Brownian motion on the unit interval $[0, 1]$. Therefore, the limiting distribution of $T\hat{\rho}$, for example, is the same as the distribution of the RHS variable in (21). Note that we have a faster convergence rate than the standard rate $\sqrt{T}$. This is because the sample covariance between an $I(1)$ variable and its innovation term is $O_p(1)$, while the sample variance of an $I(1)$ variable is $O_p(T)$. Needless to say, $\hat{\rho}$ is the ratio of these two quantities calculated using the true mean values, zero. Despite the fact that the numerator, which is a simple linear transformation of the χ_1^2 random variable $V^2(1)$, has a distribution which is skewed to the right, the limiting distribution of $T\hat{\rho}$ is highly skewed to the left as shown in Fuller (1976, p. 371). This is also true for the t-statistic in (22), so that the critical values are larger in absolute value than the usual critical values. These critical values are available in Fuller (1976, p. 373) under the name $\hat{\tau}$.

When φ is unknown but π is known to be zero, z_t is equal to the mean-corrected y_t [or equivalently an intercept is allowed in (19)]. The sample

mean of an $I(1)$ variable is $O_p(T^{1/2})$ so that it diverges as $T \to \infty$. Hence using the sample mean in calculating the variance and the covariance does make a difference in the limiting distributions from the previous case, although the order properties remain the same. Dickey and Fuller (1979) also consider this case and the result in our notation becomes

$$T\hat{\rho} \xrightarrow{d} \frac{\left[V(1)^2 - 1\right]/2 - V(1)\int_0^1 V(t)\,dt}{\int_0^1 V(t)^2\,dt - \left[\int_0^1 V(t)\,dt\right]^2}, \tag{23}$$

$$t_{\hat{\rho}} \xrightarrow{d} \frac{\left[V(1)^2 - 1\right]/2 - V(1)\int_0^1 V(t)\,dt}{\left\{\int_0^1 V(t)^2\,dt - \left[\int_0^1 V(t)\,dt\right]^2\right\}^{1/2}}, \tag{24}$$

$$T^{-1/2}\bar{y} \xrightarrow{d} \sigma_v \int_0^1 V(t)\,dt. \tag{25}$$

The second expressions in the numerators and denominators of (23) and (24) are purely due to the sample mean adjustment in z_t. The empirical distribution of $t_{\hat{\rho}}$ of the current case, which is skewed more to the left than the previous one, is also available in Fuller (1976) under the name $\hat{\tau}_\mu$. For comparison, part of the critical values of $\hat{\tau}$ and $\hat{\tau}_\mu$ are reproduced in table 2.

Although our discussion is confined to the case of i.i.d. noise terms, Phillips (1985a) shows that the limiting distributions are also valid for heterogeneous error terms under a mixing condition. However, he shows that, if the noise term is intertemporally dependent, then the limiting distribution depends upon nuisance parameters, namely the variance of v_t and the spectral density of v_t at the zero frequency. When this dependence is autoregressive, Fuller (1976) corrects the problem by including lagged dependent variables in the regression as in (20). See also Said and Dickey (1984) for a more general treatment.

We consider now the case when both φ and π are unknown. The limiting distributions of the OLS estimators of the co-integrating regression (18) has been studied by Stock (1984) and Phillips (1985b):

$$\hat{\pi} \xrightarrow{d} \frac{\sigma_v\left[\int_0^1 V(t)W(t)\,dt - \int_0^1 V(t)\,dt\int_0^1 W(t)\,dt\right]}{\sigma_w\left[\int_0^1 W(t)^2\,dt - \left(\int_0^1 W(t)\,dt\right)^2\right]} \equiv \frac{\sigma_v}{\sigma_w}\xi, \tag{26}$$

$$T^{-1/2}\hat{\varphi} \xrightarrow{d} \sigma_v\left[\int_0^1 V(t)\,dt - \xi\int_0^1 W(t)\,dt\right] \equiv \sigma_v\eta, \tag{27}$$

where $V(t)$ and $W(t)$ are mutually independent standard Brownian motions on the unit interval $[0, 1]$.

Hence $\hat{\varphi}$ diverges at the same speed as the sample mean $\bar{y}$. On the other hand, $\hat{\pi}$ converges to a random variable. If x_t and y_t were co-integrated, $\hat{\pi}$ would converge to a constant with the convergence rate T which is faster than $T^{1/2}$ of the standard cases [Stock (1984)]. Following the work of Phillips (1985b) it is a bit tedious but straightforward to show that

$$T\hat{\rho} \overset{\Delta}{\to} \theta/\psi, \tag{28}$$

$$t_{\hat{\rho}} \overset{\Delta}{\to} \theta/\left[\psi(1+\xi^2)\right]^{1/2}, \tag{29}$$

where

$$\theta = \left\{ \left[V(1) - \xi W(1)\right]^2 - 1 - \xi^2 \right\}/2 - \eta\left[V(1) - \xi W(1)\right],$$

$$\psi = \int_0^1 V(t)^2 \, dt - \left(\int_0^1 V(t) \, dt\right)^2$$

$$- \xi\left[\int_0^1 V(t)W(t) \, dt - \int_0^1 V(t) \, dt \int_0^1 W(t) \, dt\right].$$

Had the $\hat{\pi}$ converged to zero (i.e., $\xi = 0$), the limiting behavior of $\hat{\rho}$ and $t_{\hat{\rho}}$ would be the same as those from the mean adjusted Dickey and Fuller regression. In the Dickey–Fuller case, the standard error of the regression converges to a constant almost surely. In the present case, however, this statistic converges to a random variable:

$$s^2 \overset{\Delta}{\to} \sigma_v^2(1 + \xi^2), \tag{30}$$

where s is the standard error of the regression, which adds a further complication to the limiting distribution.

The large sample behavior of the t-statistic, $t_{\hat{\rho}}$, depends on the number of variables in the co-integrating regression. This can be seen from the expressions (26) through (30): the right-hand-side random variables are functions of independent Brownian motions, the number of which is equal to the number of the variables in the co-integrating regression. Intuition also tells us that the critical values will get larger in absolute value because the residual series from the co-integrating regression would look more close to a stationary series as the number of variables increases. In the next section, we examine these critical values.

7. Critical values for the co-integration test

Critical values are first constructed for various values of N, the number of variables, and T, the sample size. We assume that the data are generated by

$$x_t = x_{t-1} + \varepsilon_t, \quad x_0 = 0, \quad x_t' = (x_{1t}, x_{2t}, \dots, x_{Nt}), \tag{31}$$

with

$$\varepsilon_t \sim IN(0, I_N),$$

and that the co-integrating regression takes the form

$$x_{1t} = \alpha + \beta_2 x_{2t} + \beta_3 x_{3t} + \cdots + \beta_N x_{Nt} + z_t. \tag{32}$$

It is clear that there is a normalization imposed in the co-integrating regression (32) which would mean that different test statistics could be found from the same data set. Stock and Watson, in testing a related hypothesis on the co-integrating rank r, use principle components estimation rather than OLS. Such a convention eliminates the ambiguity. Because the test statistics have the same distribution for all normalizations and in our experience differ little across such choices, and because it is extremely convenient to use OLS with possibly a natural normalization, we continue in the vein of Engle and Granger to pick a particular normalization for the test. The power property of the test, when $N = 2$, has been examined in their paper.

The covariance matrix of the innovations is taken to be identity without any loss of generality. The test statistics will be the same for any value of the covariance matrix as the tests have the property of 'similarity' as discussed for example by Cox and Hinkley (1974, p. 134–136).

Table 2 reports the critical values of the t-statistics in the regression (19) omitting minus signs for simplicity. These values as well as those in other tables in this section have been obtained through ten thousand replications. For comparison, we also reproduce the critical values of $\hat{\tau}$ and $\hat{\tau}_\mu$ discussed in the previous section from Fuller (1976, p. 373).

The 95% confidence intervals of these critical values [see Rohatgi (1984, pp. 496–500)] vary little bit. They are mostly less than ± 0.08 about the 1% values, ± 0.05 about the 5% values, and ± 0.03 about the 10% values.

To examine the movement of critical values in higher-order systems, we have generated data according to the following model in place of (31):

$$x_t = x_{t-1} + U_t,$$

$$u_{it} = 0.8 u_{it-4} + \varepsilon_{it}, \qquad i = 1, \ldots, N, \tag{33}$$

maintaining the assumptions on x_0 and ε_t as before.

The critical values of the t-statistic of the augmented regression (20) with $p = 4$ are reported in table 3. Note that this regression includes more variables (i.e., Δz_{t-i}, $i = 1, 2, 3$) than necessary. Hence the values are not fully efficient ones but reflect the ignorance about the lag length encountered in practice. Of course, such an inefficiency will disappear in a large sample. In theory, the

Table 2

Critical values for the co-integration test.

Number of var's N	Sample size T	Significance level		
		1%	5%	10%
1[a]	50	2.62	1.95	1.61
	100	2.60	1.95	1.61
	250	2.58	1.95	1.62
	500	2.58	1.95	1.62
	∞	2.58	1.95	1.62
1[b]	50	3.58	2.93	2.60
	100	3.51	2.89	2.58
	250	3.46	2.88	2.57
	500	3.44	2.87	2.57
	∞	3.43	2.86	2.57
2	50	4.32	3.67	3.28
	100	4.07	3.37	3.03
	200	4.00	3.37	3.02
3	50	4.84	4.11	3.73
	100	4.45	3.93	3.59
	200	4.35	3.78	3.47
4	50	4.94	4.35	4.02
	100	4.75	4.22	3.89
	200	4.70	4.18	3.89
5	50	5.41	4.76	4.42
	100	5.18	4.58	4.26
	200	5.02	4.48	4.18

[a] Critical values of $\hat{\tau}$.
[b] Critical values of $\hat{\tau}_{\mu}$. Both cited from Fuller (1976, p. 373).

unknown lag structure in z_t might be handled by allowing p to be a slowly increasing function of the sample size as in Said and Dickey (1984). However, it does not provide practical guidance in the choice of p. A readily available way seems to be to use a standard model selection procedure based upon some information criterion (e.g., AIC).

As expected, these critical values differ from those in table 2 when the sample size is small. It might suggest that the values in table 2 are likely to be conservative in small samples. For a sample of size about two hundred, however, the difference becomes qualitatively negligible so that the values in table 2 seem to be fairly accurate approximations for higher-order systems as well.

We have also examined the behavior of the Durbin–Watson statistic from the co-integrating regression. Unfortunately, the discrepancy between the critical values for different systems remains significant even for the sample of size two hundred. This is not surprising since the statistic is not asymptotically

Table 3

Critical values for a higher-order system (33).

Number of var's N	Sample size T	Significance level		
		1%	5%	10%
2	50	4.12	3.29	2.90
	100	3.73	3.17	2.91
	200	3.78	3.25	2.98
3	50	4.45	3.75	3.36
	100	4.22	3.62	3.32
	200	4.34	3.78	3.51
4	50	4.61	3.98	3.67
	100	4.61	4.02	3.71
	200	4.72	4.13	3.83
5	50	4.80	4.15	3.85
	100	4.98	4.36	4.06
	200	4.97	4.43	4.14

Table 4

Critical values for the Durbin–Watson statistic (number of variables $N = 2$).

Sample size	Canonical system (31)			Higher-order system (33)		
	1%	5%	10%	1%	5%	10%
50	1.00	0.78	0.69	1.49	1.03	0.83
100	0.51	0.39	0.32	0.46	0.28	0.21
200	0.29	0.20	0.16	0.13	0.08	0.06

similar as are the preceding tests. Hence this statistic does not appear to be too useful for testing co-integration. For the sake of illustration, we report in table 4 the critical values obtained from the systems (31) and (33) with $N = 2$.

8. Conclusion

It is commonly believed that many economic time series are tied together even though they are all trending. When a forecasting model is needed for such time series, a vector autoregressive model in differences is inappropriate. This is because, even though the residuals may appear to be white, such a model suffers misspecification and the forecasts will diverge from each other. The VAR formulation (in levels) does not suffer such misspecification. However, conventional estimation techniques appear to underestimate the parameters near the unit circle. Since forecasts are made conditional on estimated parameters, the forecasts from a VAR model are likely to be suboptimal

especially in multi-step horizons. Our simulation result in section 5 supports this argument even though more practical experience is needed.

Hence it appears important to build models designed for co-integrated time series. The error correction models provide a simple class. The critical values reported in this paper should be a useful guide to decide when to impose the co-integration constraint. The two-step estimator proposed in Engle and Granger (1987) can then be used to model the error correction structure and achieve the multi-step forecast gains described above.

References

Beveridge, S. and C.R. Nelson, 1981, A new approach to decomposition of economic time series into permanent and transitory components with particular attention to measurement of the business cycle, Journal of Monetary Economics 7, 151–174.

Cox, D.R. and C.V. Hinkley, 1974, Theoretical statistics (Chapman and Hall, London).

Dickey, David A. and W.A. Fuller, 1979, Distribution of estimates for autoregressive time series with unit root, Journal of the American Statistical Association 74, 427–431.

Engle, Robert F. and C.W.J. Granger, 1987, Cointegration and error correction: Representation, estimation and testing, Econometrica 55, 251–276.

Fuller, W.A. 1976, Introduction to statistical time series (Wiley, New York).

Granger, C.W.J., 1981, Some properties of time series data and their use in econometric model specification, Journal of Econometrics 16, 121–130.

Granger, C.W.J., 1983, Cointegrated variables and error-correcting models, Discussion paper (University of California, San Diego, CA).

Granger, C.W.J., 1986, Developments in the study of cointegrated economic variables, Oxford Bulletin of Economics and Statistics 48, 213–228.

Granger, C.W.J. and A.A. Weiss, 1983, Time series analysis of error-correcting models, in: S. Karlin et al., eds., Studies in econometrics, time series, and multivariate analysis (Academic Press, New York) 255–278.

Litterman, R.B., 1986, Forecasting with Bayesian vector autoregressions: Five years of experience, Journal of Business and Economic Statistics 4, 25–38.

Phillips, P.C.B., 1985a, Time series regression with unit roots, Cowles Foundation discussion paper no. 740 (Yale University, New Haven, CT).

Phillips, P.C.B., 1985b, Understanding spurious regressions in econometrics, Cowles Foundation discussion paper no. 757 (Yale University, New Haven, CT).

Phillips, P.C.B. and S.N. Durlauf, 1986, Multiple time series regression with integrated processes, Review of Economic Studies 53, 473–495.

Rohatgi, V.K., 1984, Statistical inference (Wiley, New York).

Said, S.E. and D.A. Dickey, 1984, Testing for unit roots in autoregressive–moving average models of unknown order, Biometrica 71, 599–607.

Sims, C.A., 1980, Macroeconomics and reality, Econometrica 48, 1–48.

Stock, J., 1984, Asymptotic properties of least squares estimators of cointegrating vectors, Mimeo. (Harvard University, Cambridge, MA).

Stock, J.H. and M.W. Watson, 1986, Testing for common trends, Mimeo. (Harvard University, Cambridge, MA).

Watson, M.W., 1986, Univariate detrending methods with stochastic trends, Journal of Monetary Economics 18, 49–75.

White, J.S., 1958, The limiting distribution of the serial correlation coefficient in the explosive case, Annals of Mathematical Statistics 29, 1188–1197.