Cointegration and Error Correction

James Davidson
University of Exeter

Chapter 7 of *Handbook of Empirical Methods in Macroeconomics*, eds. Michael Thornton and Nigar Hashimzade

Contents

1 The Background

Elementary courses in statistics introduce at an early stage the key assumption of "random sampling". In more technical language, the data set is assumed to be identically and independently distributed (i.i.d.). In this framework a range of simple and elegant results can be derived, for example, that the variance of the mean of n observations is $1/n$ times the variance of the observations themselves. Given a random sample of n pairs (x, y) with sample correlation coefficient r_{xy}, if at least one of the pair has a Gaussian (normal) distribution the existence of a relationship between them is tested by comparing the "t statistic" $r_{xy}/\sqrt{(1-r_{xy}^2)/(n-2)}$ with the Student t distribution with $n-2$ degrees of freedom. All the inference procedures in

classical regression analysis follow the same basic approach. The Gaussianity assumption may be dropped by appeal to a large sample and the central limit theorem, but independent sampling is strictly needed to validate these procedures.

The received theory notwithstanding, often the first data sets that students meet in econometrics class are time series for GDP, aggregate consumption, money stock and the like – samples that are neither independently nor identically distributed. Such disjunctions between theory and practice often sew confusion in the understanding of statistical relationships in economics.

One of the first authors to study the problem of inference in time series was G. Udny Yule (1926), who reflected in his presidential address to the Royal Statistical Society on the high correlation (0.9512) between standardized mortality and the proportion of marriages solemnized by the Church of England, recorded in the years 1866 to 1911. It is interesting with the benefit of hindsight to read of the difficulties that professional statisticians would have – both then and much more recently – with the interpretation of such facts. The two series of Yule's example share a pronounced downward drift over the 46 years of the observations. "Large goes with large and small with small", which is the classic indicator of a positive correlation. In what sense is this correlation to be regarded as spurious? It is true that both variables are subject to systematic variation with the passage of time. However, to be driven by a common factor is a perfectly legitimate way of understanding the phenomenon of correlation between variables. This fact alone does not explain why we regard this particular correlation as spurious.

The true explanation requires us to distinguish between correlation as a description of data, and correlation as a theoretical construct; an expected association as a feature of a fixed joint distribution of random variables. Our problem arises when this fixed joint distribution does not exist. The examples Yule analyses in his paper include integrated processes, formed by a cumulation of independent random shocks. As is well known, such processes – often called *random walks* – can "wander anywhere", having no central tendency. Short realizations often give the appearance of deterministic-seeming time trends. Averages of repeated drawings from such processes do not converge to fixed limits as the sample size increases; in other words, they do not obey the law of large numbers. The sample variances of such processes, and likewise covariances, diverge to infinity. While correlation coefficients are normalized to lie between -1 and $+1$, the correlations of pairs of mutually independent random walk processes do not converge to zero, but remain random variables even asymptotically. As famously demonstrated in a set of computer simulations by Granger and Newbold (1974), independent random walks exhibit "significant" correlations, such that the t statistic defined above diverges to infinity as n increases. Additional data do not serve to resolve a spurious correlation but, rather, to reinforce the false conclusion. It

follows that the conventional equating of sample and theoretical correlations in an estimation exercise has no validity.

These phenomena presented a dilemma for econometricians in the middle years of the 20th century, as they attempted to model macroeconomic and financial data sets that are well-described as the integrals (cumulations) of stationary series. One approach was to model the relationships between the differences (the changes from period to period) but clearly a great deal of information about relationships between series is lost in such transformations. It is easy to construct examples where the correlation between the differences of time series have signs opposite to that between the levels. A second approach is to treat trends as deterministic, and remove them by regression on dummy (straight-line) trend variables. Although the relations between fitted trend components can be discounted as spurious (one straight line always "explains" another) the deviations of economic series from linear trend often exhibit random walk characteristics in practice, so the problem is not resolved.

It was in the context of this unsatisfactory hiatus in the progress of time series econometrics, in the course of the 1970s, that Clive Granger initiated his researches into the modelling of economic trends. The culmination of this research was the key idea that relationships between integrated time series must be understood as a sharing of common trends; not correlation, but cointegration. The story of these discoveries, well told in an article by David Hendry (2004) celebrating Granger's 2003 Nobel Prize, provides a fascinating mix of debates and disagreements, false trails, penetrating intuitions and the insightful re-interpretation of applied studies. Hendry's (1980) inaugural lecture at LSE is often cited as an accessible exposition of the issues, although the term 'cointegration' had yet to be coined at that date.

The complete story of the cointegration concept has to acknowledge the indispensable contributions of two other researchers, Peter C. B. Phillips at Yale, who developed the essential links with mathematical stochastic process theory that were needed for a theory of inference in nonstationary data, and Søren Johansen in Copenhagen, who developed a rigorous theory of vector autoregressions in nonstationary data. The net result of these endeavours is that econometrics can deal effectively with time series data, whether or not the "identically and independently distributed" sampling paradigm has any practical relevance.

2 A Linear Model of Nonstationary Data

To fix ideas, consider first the simplest multiple time series model, the first-order VAR. Let $\boldsymbol{x}_t$ $(m \times 1)$ denote a vector of variables evolving according

to the equation

$$\boldsymbol{x}_t = \boldsymbol{a}_0 + \boldsymbol{A}\boldsymbol{x}_{t-1} + \boldsymbol{\varepsilon}_t \tag{1}$$

where $\boldsymbol{A}$ is an $m \times m$ matrix of coefficients and $\boldsymbol{\varepsilon}_t$ $(m \times 1)$ is i.i.d. with mean vector $\mathbf{0}$ and variance matrix $\boldsymbol{\Sigma}$. Suppose that this process has been running for a large number of periods, that we can treat as effectively infinite. Then the equation has the solution

$$\boldsymbol{x}_t = \sum\nolimits_{j=0}^{\infty} \boldsymbol{A}^j(\boldsymbol{a}_0 + \boldsymbol{\varepsilon}_{t-j})$$

where $\boldsymbol{A}^j = \boldsymbol{A}\boldsymbol{A}\cdots\boldsymbol{A}$ (the j-fold product) and $\boldsymbol{A}^0 = \boldsymbol{I}_m$, the identity matrix of order m.

Write the Jordan canonical form of the matrix as $\boldsymbol{A} = \boldsymbol{P}\boldsymbol{M}\boldsymbol{P}^{-1}$, where if the eigenvalues are all distinct, $\boldsymbol{M}$ is a diagonal matrix with the eigenvalues of $\boldsymbol{A}$ (either real or complex valued) on the diagonal.[1] Provided the eigenvalues all have modulus strictly less than unity, it is easy to see that $\boldsymbol{A}^j = \boldsymbol{P}\boldsymbol{M}^j\boldsymbol{P}^{-1} \to \mathbf{0}$ and $\sum_{j=0}^{\infty} \boldsymbol{A}^j = (\boldsymbol{I}_m - \boldsymbol{A})^{-1} < \infty$. In this case, we note that $\boldsymbol{x}_t$ has a distribution independent of t, with mean $(\boldsymbol{I}_m - \boldsymbol{A})^{-1}\boldsymbol{a}_0$ and variance matrix $\boldsymbol{\Sigma}_x = \sum_{j=0}^{\infty} \boldsymbol{A}^j\boldsymbol{\Sigma}(\boldsymbol{A}^j)'$.[2] We say that the process is stationary.

If $\boldsymbol{A}$ has one or more eigenvalues equal to 1, on the other hand, $\boldsymbol{A}^j$ does not converge to zero and $\boldsymbol{I}_m - \boldsymbol{A}$ is singular, by construction. In this case, the assumption that it has been running for an infinite number of periods is not compatible with a well-defined distribution for $\boldsymbol{x}_t$; such a process has infinite magnitude with probability 1. We must instead postulate a finite initial condition $\boldsymbol{x}_0$ and consider the cases $t = 1, 2, 3, ...$ to see what happens. Clearly, this process is nonstationary, and its variance is increasing with time. A particularly simple case is $\boldsymbol{A} = \boldsymbol{I}_m$, where all m eigenvalues are equal to 1, and

$$\boldsymbol{x}_t = \boldsymbol{x}_0 + t\boldsymbol{a}_0 + \sum\nolimits_{j=0}^{t-1} \boldsymbol{\varepsilon}_{t-j}. \tag{2}$$

This is a vector of so-called random walks, with drifts $\boldsymbol{a}_0$. Note how the equation intercepts no longer measure a unique location, or central tendency of the distribution, but the rate of divergence of the central tendency with time. The variance matrix of the process, treating $\boldsymbol{x}_0$ as fixed, is $t\boldsymbol{\Sigma}$. Even with $\boldsymbol{a}_0 = \mathbf{0}$ the average distance from the starting point, as measured by the standard deviation of the coordinates, increases like $\sqrt{t}$.

[1] With repeated eigenvalues $\boldsymbol{M}$ is generally not diagonal. When $\mu_k = \mu_{k+1}$, a '1' appears in position $\{k, k+1\}$. However, note that $\boldsymbol{A}$ and $\boldsymbol{M}$ have the same rank and $\boldsymbol{M}$ is either diagonal or upper triangular. While only in symmetric matrices is the rank always equal to the number of nonzero eigenvalues, a singular matrix always has one or more zero eigenvalues.

[2] This matrix can be written in closed form only with the use of Vec notation, but it's easy to see that it must satisfy the identity $\boldsymbol{\Sigma}_x - \boldsymbol{A}\boldsymbol{\Sigma}_x\boldsymbol{A}' = \boldsymbol{\Sigma}$.

More generally, we may have some of the eigenvalues of the system equal to unity, and others in the stable range. It is convenient in this case to recast the model in the form in which the singular matrix appears explicitly. Write $\boldsymbol{\Pi} = \boldsymbol{A} - \boldsymbol{I}_m$ and then (1) can be written

$$\Delta \boldsymbol{x}_t = \boldsymbol{a}_0 + \boldsymbol{\Pi} \boldsymbol{x}_{t-1} + \boldsymbol{\varepsilon}_t \tag{3}$$

where $\Delta \boldsymbol{x}_t = \boldsymbol{x}_t - \boldsymbol{x}_{t-1}$.[3] Note that the eigenvalues of $\boldsymbol{\Pi}$ are the diagonal elements of $\boldsymbol{M} - \boldsymbol{I}_m$ and, hence, unit eigenvalues of $\boldsymbol{A}$ are zero eigenvalues of $\boldsymbol{\Pi}$. With one or more zero eigenvalues, $\boldsymbol{\Pi}$ is singular, say with rank $s < m$, and note that the case $s = 0$ implies $\boldsymbol{\Pi} = \boldsymbol{0}$ and hence corresponds to the random walk model (2).[4]

A $m \times m$ matrix with rank s always has a representation $\boldsymbol{\Pi} = \boldsymbol{\alpha}\boldsymbol{\beta}'$ where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are $m \times s$ matrices with full rank s. This decomposition is not of course unique, since we can also write $\boldsymbol{\Pi} = \boldsymbol{\alpha}^*\boldsymbol{\beta}^{*\prime}$ where $\boldsymbol{\alpha}^* = \boldsymbol{\alpha}\boldsymbol{D}^{-1}$ and $\boldsymbol{\beta}^* = \boldsymbol{\beta}\boldsymbol{D}'$ for any $s \times s$ nonsingular matrix $\boldsymbol{D}$. However. the columns of $\boldsymbol{\beta}$ must always span the same space.[5] It is also possible that known restrictions on the model could allow $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ to be identified uniquely, an issue that we discuss further in Section 6.

Consider the relationship between the processes $\boldsymbol{x}_t$ and $\Delta \boldsymbol{x}_t$ appearing in (3). Differencing is the inverse of the operation of integrating (i.e., cumulating) a series. If $\boldsymbol{x}_0 = \boldsymbol{0}$ and $\boldsymbol{x}_t = \boldsymbol{y}_1 + \boldsymbol{y}_2 + \cdots + \boldsymbol{y}_t$, then $\Delta \boldsymbol{x}_t = \boldsymbol{y}_t$ for $t \geq 1$. We define the notion of the "order of integration" of a series, denoted d, such that if $\boldsymbol{x}_t$ has order of integration d then $\Delta \boldsymbol{x}_t$ has order of integration $d - 1$. A convenient shorthand for this is to write $\boldsymbol{x}_t \sim \mathrm{I}(d)$. If we (arbitrarily) assign $d = 0$ to the case where the process is stationary with finite variance, then a random walk of the type shown in (2) must be assigned $d = 1$. Differencing an I(0) process yields the case I(-1), again a stationary process but this one is also stationary after integrating; hence this case, sometimes called an *over-differenced* process, is distinct from I(0).

The interesting feature of (3) is that processes with different orders of integration feature on the two sides of the equation. It is not too difficult to deduce from the definitions that $\mathrm{I}(d) + \mathrm{I}(d-p) \sim \mathrm{I}(d)$ for any $p > 0$, and also that $\boldsymbol{\varepsilon}_t \sim \mathrm{I}(0)$. Writing (3) in the form

$$\Delta \boldsymbol{x}_t = \boldsymbol{a}_0 + \boldsymbol{\alpha}\boldsymbol{\beta}' \boldsymbol{x}_{t-1} + \boldsymbol{\varepsilon}_t \tag{4}$$

we see, given that $\boldsymbol{\alpha}$ is a full-rank matrix, that $\boldsymbol{\beta}'\boldsymbol{x}_t$ must be $\mathrm{I}(d-1)$ when $\boldsymbol{x}_t \sim \mathrm{I}(d)$. Taking a certain linear combination of the variables in the model results in a process of lower integration order than that of the variables themselves. While we have not shown by this argument that $d = 1$ in

[3] The difference operator is $\Delta = 1 - L$ where L is the lag operator.

[4] s cannot be less than the number of nonzero eigenvalues, but could be greater.

[5] The space spanned by $\boldsymbol{\beta}$ is the collection of vectors $\boldsymbol{\beta}\boldsymbol{r}$ for all s-vectors $\boldsymbol{r} \neq \boldsymbol{0}$. Clearly, this is identical with the collection $\boldsymbol{\beta}\boldsymbol{D}'\boldsymbol{r}$, for any $s \times s$ nonsingular $\boldsymbol{D}$.

the "reduced rank VAR" (4), this is intuitively clear from considering the limiting cases $s = m$ and $s = 0$, the stationary and random walk models respectively.

With no loss of generality the intercept may be decomposed as $\boldsymbol{a}_0 = \boldsymbol{\delta} - \boldsymbol{\alpha}\boldsymbol{\mu}$ where $\boldsymbol{\mu}$ is $s \times 1$. Then the model can be further rearranged as

$$\Delta \boldsymbol{x}_t = \boldsymbol{\delta} + \boldsymbol{\alpha}\mathbf{z}_{t-1} + \boldsymbol{\varepsilon}_t \tag{5}$$

where $\boldsymbol{z}_t$ is the s-vector of *cointegrating residuals*, defined as

$$\boldsymbol{z}_t = \boldsymbol{\beta}'\boldsymbol{x}_t - \boldsymbol{\mu}\ (s \times 1). \tag{6}$$

The elements of $\boldsymbol{\alpha}$ are often referred to as the 'loadings coefficients' or 'error correction coefficients'. Premultiplying (5) by $\boldsymbol{\beta}'$ and rearranging yields the VAR(1) representation of the residuals,

$$\boldsymbol{z}_t = \boldsymbol{\beta}'\boldsymbol{\delta} + (\boldsymbol{I}_s + \boldsymbol{\beta}'\boldsymbol{\alpha})\boldsymbol{z}_{t-1} + \boldsymbol{\beta}'\boldsymbol{\varepsilon}_t. \tag{7}$$

This relation defines a modified form of stability condition. If the matrix $\boldsymbol{I}_s + \boldsymbol{\beta}'\boldsymbol{\alpha}$ has all its eigenvalues in the stable region, then the series possess s stationary linear combinations. If $\boldsymbol{\delta} \neq \mathbf{0}$ the system contains a drift, the variables of the system having a persistent tendency to either rise or fall depending on the signs of the elements, although if $\boldsymbol{\beta}'\boldsymbol{\delta} = \mathbf{0}$ the cointegrating relations cancel the drift and $E(\boldsymbol{z}_t) = \mathbf{0}$. On the other hand, if $\boldsymbol{\delta} = \mathbf{0}$ the processes are drift-neutral, their variances increasing with time but as likely to fall as to rise in any period. Such a process is said to exhibit a pure stochastic trend. Take care to note that $\boldsymbol{\mu}$ does not contribute to the drift so that $\boldsymbol{a}_0 = \mathbf{0}$ is not necessary for drift-neutrality.

We have now derived a simple form of the celebrated Granger representation theorem, which says, in essentials, the following. A vector autoregression containing unit roots generates nonstationary processes, but if the number of these roots is smaller than the dimension of the system there must at the same time exist a set of $s < m$ stationary linear combinations of the variables, forming the so-called *cointegrating relations*. s is called the *cointegrating rank* of the system. A necessary feature of the system is that the cointegrating residuals Granger-cause[6] future changes of the process, so that the model can always be cast in the so-called *error-correction* form. The variables of the model are said to exhibit $m - s$ *common trends*. The variables evolve along nonstationary paths, but these paths are tied together by the cointegrating relations. The error correction form has a very natural interpretation, that to maintain the common trends through time requires that changes in the variables must respond to deviations from the cointegrating relations measured by $\boldsymbol{z}_t$. For this to happen requires the elements

[6] A variable x is said to Granger-cause another variable y if knowledge of x_t improves the forecasts of y_{t+j} for $j > 0$. This concept is defined in Granger (1969), Clive Granger's first notable contribution to time series econometrics.

of $\boldsymbol{\alpha}$ to have appropriate signs and magnitudes to ensure stable adjustment, according to (7). This feature is of course implicit in the requirement that the non-unit eigenvalues of $\boldsymbol{A}$ fall in the stable region.

3 The General Linear Case

We next consider the standard generalization of the foregoing simple case. An m-dimensional *linear process* is defined as a process whose nondeterministic component (after subtracting intercepts, trends, etc.) has the representation

$$\boldsymbol{y}_t = \boldsymbol{C}(L)\boldsymbol{\varepsilon}_t \tag{8}$$

where[7] $\boldsymbol{C}(z) = \sum_{j=0}^{\infty} \boldsymbol{C}_j z^j$ $(m \times m)$ and $\{\boldsymbol{\varepsilon}_t,\ -\infty < t < \infty\}$ is an i.i.d. sequence of random m-vectors with mean $\boldsymbol{0}$ and variance $\boldsymbol{\Sigma}$. This is sometimes called the Wold representation of the process (Wold, 1938) although remember that Wold's representation exists for any stationary process if the innovation process is white noise (i.e., stationary and uncorrelated). The definition of a linear process specifies independence of the innovations, a stronger condition than white noise. We assume $\boldsymbol{C}_0 = \boldsymbol{I}_m$, although this entails no loss of generality if $\boldsymbol{\Sigma}$ is arbitrary, and could be replaced by the requirement $\boldsymbol{\Sigma} = \boldsymbol{I}_m$.

If (a) $\sum_{j=0}^{\infty} \|\boldsymbol{C}_j\| < \infty$ [8] and (b) $\sum_{j=0}^{\infty} \boldsymbol{C}_j \neq \boldsymbol{0}$, we call the process I(0). Note that (a) is a stronger condition than is required for stationarity. Define

$$\boldsymbol{\Gamma}_k = E(\boldsymbol{y}_t \boldsymbol{y}'_{t+k}) = \sum\nolimits_{j=0}^{\infty} \boldsymbol{C}_j \boldsymbol{\Sigma} \boldsymbol{C}'_{j+k}$$

for $k > 0$, where $\boldsymbol{\Gamma}_{-k} = \boldsymbol{\Gamma}'_k$. Then, writing $\boldsymbol{C}$ as shorthand for $\boldsymbol{C}(1) = \sum_{j=0}^{\infty} \boldsymbol{C}_j$, note that (a) is sufficient for $\boldsymbol{\Omega} < \infty$ where

$$\boldsymbol{\Omega} = \sum\nolimits_{k=-\infty}^{\infty} \boldsymbol{\Gamma}_k = \boldsymbol{C}\boldsymbol{\Sigma}\boldsymbol{C}'. \tag{9}$$

This matrix is called the 'long-run variance' of the process,[9] and observe that

$$\boldsymbol{\Omega} = \lim_{T\to\infty} \frac{1}{T} E\left(\sum\nolimits_{t=1}^{T} \boldsymbol{y}_t \sum\nolimits_{t=1}^{T} \boldsymbol{y}'_t \right).$$

Thus, the I(0) property embodies the "square root rule", which says that the average variability of the partial sums grows like the square root of the

[7] It is convenient to give the properties of a lag polynomial in the context of a dummy numerical argument z, in general complex-valued.

[8] $\|\boldsymbol{A}\| = \sqrt{\mathrm{tr}(\boldsymbol{A}'\boldsymbol{A})}$ is one of several alternative definitions of the matrix norm. This is a simple way to specify absolute summability, ruling out the possibility of off-setting signs allowing elements to be summable while their squares, for example, are not summable.

[9] In the VAR(1) case (1), $\boldsymbol{\Omega} = (\boldsymbol{I} - \boldsymbol{A})^{-1}\boldsymbol{\Sigma}(\boldsymbol{I} - \boldsymbol{A}')^{-1}$. Be careful to distinguish between this formula and $\boldsymbol{\Sigma}_x$.

sample size. Condition (b) rules out the case of an over-differenced process. It is easy to verify that if $\boldsymbol{y}_t$ is given by (8), then $\Delta\boldsymbol{y}_t$ is a linear process with coefficients $\boldsymbol{C}_0$, $\boldsymbol{C}_1 - \boldsymbol{C}_0$, $\boldsymbol{C}_2 - \boldsymbol{C}_1$,..., and condition (b) is violated in this case if condition (a) holds.

The significance of these properties is that they suffice to validate the standard asymptotic distribution results, such as the central limit theorem for re-scaled sums of the $\boldsymbol{y}_t$. Simple stationarity is not sufficient for this by itself, and over-differencing presents an obvious counter-example, featuring $\boldsymbol{\Omega} = \boldsymbol{0}$. We shall appeal to some stronger assumptions on the sequence of coefficients for our present development, in particular (c) $\sum_{j=0}^{\infty} j \, \|\boldsymbol{C}_j\| < \infty$, which we call 1-summability (the "1" referring to the power of j)[10]. Note that 1-summability is equivalent to the condition $\sum_{j=0}^{\infty} \sum_{k=j+1}^{\infty} \|\boldsymbol{C}_k\| < \infty$. Many operational models in econometrics, in particular stable finite-order vector ARMA models, satisfy the still stronger condition $\|\boldsymbol{C}(z)\| < \infty$ for $|z| \leq 1 + \delta$, for some $\delta > 0$, implying that the coefficients converge to zero at an exponential rate. However, this is not required for present purposes.

The particular case we consider here is the I(1) linear process $\boldsymbol{x}_t$, such that the Wold representation of the differences is

$$\Delta\boldsymbol{x}_t = \boldsymbol{C}(L)\boldsymbol{\varepsilon}_t \tag{10}$$

where conditions (a), (b) and (c) are satisfied in the right-hand side. The key relation in this analysis is commonly known as the *Beveridge-Nelson* (BN) *decomposition* (Beveridge and Nelson 1981). This is nothing but an easily verified identity for polynomials,

$$\boldsymbol{C}(z) = \boldsymbol{C}(1) + (1 - z)\boldsymbol{C}^*(z)$$

where $\boldsymbol{C}^*(z) = \sum_{j=0}^{\infty} \boldsymbol{C}_j^* z^j$ and $\boldsymbol{C}_j^* = -\sum_{k=j+1}^{\infty} \boldsymbol{C}_k$. Thus, we can write,

$$\Delta\boldsymbol{x}_t = \boldsymbol{C}\boldsymbol{\varepsilon}_t + \boldsymbol{\zeta}_t - \boldsymbol{\zeta}_{t-1}.$$

where $\boldsymbol{\zeta}_t = \boldsymbol{C}^*(L)\boldsymbol{\varepsilon}_t$ is a I(0) process, by 1-summability. Integrating[11] this sequence from an initial value $\boldsymbol{x}_0$,[12] which we must assume finite, yields

$$\boldsymbol{x}_t - \boldsymbol{x}_0 = \boldsymbol{C}\boldsymbol{w}_t + \boldsymbol{\zeta}_t \tag{11}$$

[10] Some of the results in this theory can be proved under the weaker condition (c′) $\sum_{j=0}^{\infty} j^{1/2} \, \|\boldsymbol{C}_j\| < \infty$, see for example Phillips and Solo (1992). The conditions stated here are sufficient for the properties we discuss, and are hopefully the most intuitive ones.

[11] Note the conventions governing the use of the difference operator Δ and its inverse, the integration operator $\Delta^{-1} = 1 + L + \cdots + L^t$. Consider a sequence $y_1, \ldots, y_T$. Since $\Delta^{-1} y_1 = y_1$, the operator Δ must be accordingly defined by $\Delta y_1 = y_1$ and $\Delta y_t = y_t - y_{t-1}$ for $t > 1$.

[12] Some care is needed in the treatment of initial conditions. Expressing the observed process as the deviation from an initial value $\boldsymbol{x}_0$ allows assumptions about how $\boldsymbol{x}_0$ is generated to be sidelined. To avoid infinities, this clearly has be by a different mechanism from that generating $\boldsymbol{x}_t$ for $t > 0$.

where $\boldsymbol{w}_t = \sum_{s=1}^{t} \boldsymbol{\varepsilon}_s$ is a random walk process. Thus, we are able to decompose a linear process rather straightforwardly into stationary and nonstationary components. Since the first right-hand side term is $O_p(t^{1/2})$ and the second one is $O_p(1)$,[13] equation (11) can be used to verify directly the result that was previously determined by substitution in (9), that is,

$$\lim_{T\to\infty} \frac{1}{T} E(\boldsymbol{x}_T - \boldsymbol{x}_0)(\boldsymbol{x}_T - \boldsymbol{x}_0)' = \boldsymbol{C}\boldsymbol{\Sigma}\boldsymbol{C}' = \boldsymbol{\Omega}. \tag{12}$$

Now, consider the case where $\boldsymbol{C}$, and hence $\boldsymbol{\Omega}$, is singular with rank $m - s$. There must exist in this case a matrix $\boldsymbol{\beta}$ $(m \times s)$ of rank s such that $\boldsymbol{\beta}'\boldsymbol{C} = \boldsymbol{0}$, and it follows immediately that

$$\boldsymbol{z}_t = \boldsymbol{\beta}'(\boldsymbol{x}_t - \boldsymbol{x}_0) = \boldsymbol{\beta}'\boldsymbol{\zeta}_t$$

is an I(0) process. In other words, deficient rank of the matrix $\boldsymbol{C}$ implies the existence of cointegration in the nonstationary series $\boldsymbol{x}_t$. In the extreme case, $\boldsymbol{C} = \boldsymbol{0}$ implies that $\boldsymbol{x}_t$ is stationary, since the factor Δ cancels in (10).

Next, consider an autoregressive representation of the process. Suppose

$$\boldsymbol{A}(L)(\boldsymbol{x}_t - \boldsymbol{x}_0) = \boldsymbol{\varepsilon}_t.$$

Writing $\boldsymbol{A}(z) = \boldsymbol{A}^+(z)(1 - z)$ shows that the Wold polynomial $\boldsymbol{C}(z)$ must have the representation $\boldsymbol{A}^+(z)^{-1}$.[14] Substituting the BN decomposition $\boldsymbol{A}(z) = (1 - z)\boldsymbol{A}^*(z) + \boldsymbol{A}$ where $\boldsymbol{A} = \boldsymbol{A}(1)$ yields

$$\boldsymbol{\varepsilon}_t = \boldsymbol{A}^*(L)\Delta\boldsymbol{x}_t + \boldsymbol{A}(\boldsymbol{x}_t - \boldsymbol{x}_0). \tag{13}$$

For this equation to balance requires $\boldsymbol{A}(\boldsymbol{x}_t - \boldsymbol{x}_0) \sim \mathrm{I}(0)$, so there must exist a decomposition of the form $\boldsymbol{A} = -\boldsymbol{\alpha}\boldsymbol{\beta}'$ for some $\boldsymbol{\alpha}$ $(m \times s)$ of rank s. Therefore, note from (13) and (10) that

$$\begin{aligned}(1 - z)\boldsymbol{I}_m &= \boldsymbol{C}(z)\boldsymbol{A}(z) \\ &= \boldsymbol{C}(z)\boldsymbol{A}^*(z)(1 - z) - \boldsymbol{C}(z)\boldsymbol{\alpha}\boldsymbol{\beta}' \\ &= \boldsymbol{C}(z)\boldsymbol{B}(z)(1 - z) - z\boldsymbol{C}(z)\boldsymbol{\alpha}\boldsymbol{\beta}' \end{aligned} \tag{14}$$

where $\boldsymbol{B}(z) = \boldsymbol{A}^*(z) - \boldsymbol{\alpha}\boldsymbol{\beta}'$. Evaluating (14) at the point $z = 1$ yields $\boldsymbol{C}\boldsymbol{\alpha} = \boldsymbol{0}$, since $\boldsymbol{\beta}$ has full rank, and hence $\boldsymbol{C}\boldsymbol{A} = \boldsymbol{0}$ and also note that $\boldsymbol{A}\boldsymbol{C} = -\boldsymbol{\alpha}\boldsymbol{\beta}'\boldsymbol{C} = \boldsymbol{0}$. The matrices $\boldsymbol{A}$ and $\boldsymbol{C}$ span orthogonal spaces, respectively

[13] We write $X_T = O_p(T^k)$ to denote that for every $\varepsilon > 0$, there exists $B_\varepsilon < \infty$ such that $P(|X_T| / T^k > B_\varepsilon) < \varepsilon$. In particular, a stationary process is $O_p(1)$.

[14] Be careful to note that $\boldsymbol{A}^+(z)$ is an invertible autoregressive polynomial, of finite or infinite order, driving the stationary differences, whereas $\boldsymbol{A}(z)$ involves the finite-order integration operator Δ^{-1}. Cumulation must be initiated at some finitely remote date. However, considering the sequence $\boldsymbol{x}_t - \boldsymbol{x}_0$ allows us to set this date as $t = 1$ without loss of generality.

the cointegrating space of dimension s and the space of dimension $m - s$ containing the common trends, through (11).

Evaluating (14) at the point $z = 0$, noting $\boldsymbol{C}_0 = \boldsymbol{I}_m$, also yields $\boldsymbol{B}_0 = \boldsymbol{I}_m$. Accordingly, defining $\boldsymbol{\Gamma}(z)$ by $\boldsymbol{B}(z) = \boldsymbol{I}_m - z\boldsymbol{\Gamma}(z)$, the error correction form of the system is obtained from (13), after some rearrangement, as

$$\Delta \boldsymbol{x}_t = \boldsymbol{\Gamma}(L)\Delta \boldsymbol{x}_{t-1} + \boldsymbol{\alpha}\boldsymbol{z}_{t-1} + \boldsymbol{\varepsilon}_t \tag{15}$$

where $\boldsymbol{z}_t = \boldsymbol{\beta}'\boldsymbol{x}_t - \mu$ and $\mu = \boldsymbol{\beta}'\boldsymbol{x}_0$. This is the generalization of (5), although it is also a simplification since the possibility of drift terms has been excluded here. (To re-introduce these would be a useful exercise for the reader.) Note that an intercept appears in the cointegrating relation, in general, unless the data are explicitly initialized at zero.

This system has the feature that $\Delta \boldsymbol{x}_t$ is explained only by lagged variables, whereas the macro-econometrics literature generally allows for the existence of contemporaneous interactions between variables, which might either be truly simultaneous relations, or involve some kind of causal ordering within the period of observation. The extension to cover this case is a simple matter of treating (15) as a solved form. Writing

$$\boldsymbol{B}_0\Delta \boldsymbol{x}_t = \boldsymbol{B}_1(L)\Delta \boldsymbol{x}_{t-1} + \boldsymbol{\rho}\boldsymbol{z}_{t-1} + \boldsymbol{u}_t, \tag{16}$$

where $\boldsymbol{B}_0$ is a square nonsingular matrix, we then recover (15) with the substitutions $\boldsymbol{\Gamma}(L) = \boldsymbol{B}_0^{-1}\boldsymbol{B}_1(L)$, $\boldsymbol{\alpha} = \boldsymbol{B}_0^{-1}\boldsymbol{\rho}$ and $\boldsymbol{\varepsilon}_t = \boldsymbol{B}_0^{-1}\boldsymbol{u}_t$, so that $E(\boldsymbol{u}_t\boldsymbol{u}_t') = \boldsymbol{B}_0\Sigma\boldsymbol{B}_0'$. We call (16) a structural form, where $\boldsymbol{B}_0 = \boldsymbol{I}_m$ is a permissible case but not a requisite.

While (15) is perhaps the commonest representation of a cointegrated system in the applied literature, the Park-Phillips triangular form (see Park and Phillips 1988, 1989, Phillips and Loretan 1991, Phillips 1991 *inter alia*) has considerable virtues of simplicity and ease of manipulation. Partitioning the vector of variables as $\boldsymbol{x}_t = (\boldsymbol{x}_{1t}', \boldsymbol{x}_{2t}')'$ where $\boldsymbol{x}_{1t}$ is $s\times 1$ and $\boldsymbol{x}_{2t}$ $(m-s)\times 1$, write[15]

$$\boldsymbol{x}_{1t} = \boldsymbol{B}\boldsymbol{x}_{2t} + \boldsymbol{v}_{1t} \tag{17a}$$

$$\Delta \boldsymbol{x}_{2t} = \boldsymbol{v}_{2t} \tag{17b}$$

where $\boldsymbol{v}_{1t}$ and $\boldsymbol{v}_{2t}$ are constrained solely to be I(0) stochastic processes. If we form the partition

$$\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix} \begin{matrix} s \times s \\ (m-s) \times s \end{matrix}$$

after re-ordering variables as necessary to ensure $\boldsymbol{\beta}_1$ has full rank, the first equation shows the cointegrating relations expressed as a reduced form with

[15] We follow the cited papers by Phillips and co-authors in using $\boldsymbol{B}$ for the reduced form cointegrating coefficients. Don't confuse this usage with the lag polynomial $\boldsymbol{B}(z)$ appearing earlier.

$\boldsymbol{B} = -\boldsymbol{\beta}_1^{-1}\boldsymbol{\beta}_2$. This matrix is accordingly unique, given this partition of the variables.

The second block of equations is merely the relevant block from the Wold representation (10). Writing the system as $\boldsymbol{A}(L)\boldsymbol{x}_t = \boldsymbol{v}_t$ where

$$\boldsymbol{A}(L) = \begin{bmatrix} \boldsymbol{I}_s & -\boldsymbol{B} \\ \boldsymbol{0} & \Delta\boldsymbol{I}_{m-s} \end{bmatrix},$$

the Wold form is obtained as $\Delta\boldsymbol{x}_t = \Delta\boldsymbol{A}(L)^{-1}\boldsymbol{v}_t$ or, in partitioned form,

$$\begin{bmatrix} \Delta\boldsymbol{x}_{1t} \\ \Delta\boldsymbol{x}_{2t} \end{bmatrix} = \Delta \begin{bmatrix} \boldsymbol{I}_s & \Delta^{-1}\boldsymbol{B} \\ \boldsymbol{0} & \Delta^{-1}\boldsymbol{I}_{m-s} \end{bmatrix} \begin{bmatrix} \boldsymbol{v}_{1t} \\ \boldsymbol{v}_{2t} \end{bmatrix} = \begin{bmatrix} \Delta\boldsymbol{v}_{1t} + \boldsymbol{B}\boldsymbol{v}_{2t} \\ \boldsymbol{v}_{2t} \end{bmatrix}.$$

This simple case gives a little insight into the mechanism of cointegration. The $m - s$ common trends are supplied as the integrals of $\boldsymbol{v}_{2t}$, whereas $\boldsymbol{v}_{1t}$ contributes only the noise component in the cointegrating relations. We discuss below how the triangular form can be the basis for a useful approach to estimation and inference.

Let's summarize the conclusions of this section. We have shown that an arbitrary linear model, that need not have a finite-order VAR representation but has 1-summable coefficients in its Wold representation, satisfies the Granger representation theorem. In other words, if the matrix $\boldsymbol{C}$ has reduced rank $m - s$ in the representation $\Delta\boldsymbol{x}_t = \boldsymbol{C}(L)\boldsymbol{\varepsilon}_t$, then the variables are cointegrated with rank s and the system admits an error-correction representation. Note that the choice of a first-order lag in (15) is completely arbitrary. It can be set to any finite value, p, by a suitable redefinition of the polynomial $\boldsymbol{\Gamma}(L)$. It is customary in the literature to let p match the order of the VAR when this is finite, such that $\boldsymbol{\Gamma}(L)$ is a polynomial of order $p - 1$.

4 Interpreting Cointegration

In his earliest contributions on the topic of cointegration, Granger (1981) was keen to emphasize his debt to the macro-econometric research of the time, in particular Sargan (1964) on wages and prices and Davidson et al. (1978) on consumption and income. These authors had explicitly built dynamic equations for nonstationary series that correlated logarithmic changes with the logarithms of "long-run" ratios, which were now to be recognized as cointegrating relations. In both the cited cases the relations happily involved no unknown parameters so the resulting regressions were easily fitted by ordinary least squares. The technical challenges involved for estimation when $\boldsymbol{z}_t$ in (15) involves unknown parameters (of which more later) did not have to be faced.

However, these models were somewhat casual in their approach to the dynamics of economic behaviour. It was assumed, first, that there existed

identifiable economic relations that described behaviour in a "steady state", abstracting from business cycle fluctuations but possibly allowing for a secular drift; and second, that these relations are not expected to hold period-to-period (nor of course are they observed to) due to unspecified dynamic effects about which economic theory is taken to be mute. There was a simple presumption that in a dynamic setting agents would formulate plans (say, for the consumption/savings balance as income changes) that combined "rule of thumb" responses to changes in driving variables represented by the $\mathbf{\Gamma}(L)$ coefficients in (15) with compensating adjustments, represented by the $\boldsymbol{\alpha}$ coefficients, to achieve a proportion of the required adjustment towards the long-run (steady state) relation in each period. The actual behavioural mechanisms were treated as beyond the reach of economics to explain, and hence this modelling approach is often spoken of as *ad hoc*, with a mildly pejorative tone.

We should not overlook that the error correction form is only nominally dynamic, and subsumes instantaneous adjustment. The static equation $y_t = \beta x_t + \varepsilon_t$, where ε_t is an independent disturbance, can of course be written equivalently as

$$\Delta y_t = \beta \Delta x_t + \alpha(y_{t-1} - \beta x_{t-1}) + \varepsilon_t$$

with $\alpha = -1$. However, empirical work with such equations invariably shows α closer to zero than to -1, and also no match between the 'dynamic' and 'long run' coefficients. These observed adjustment dynamics called for some explanation, and a number of authors have attempted to lay more rigorous economic foundations for the ECM scheme, notably Salmon (1982), Nickell (1985) and Campbell and Shiller (1988). Natural precursors are the partial adjustment model of Lovell (1961) and the habit persistence model of Brown (1952). Assuming that agents face costs associated with speedy adjustment (physical building costs in the case of inventory investment, psychological costs of changing behaviour in the case of decisions by consumers) it is straightforward to formulate a quadratic loss function for a decision variable y_t involving both the costs of change $y_t - y_{t-1}$, and the costs of deviation from equilibrium $y_t - y_t^*$, where y_t^* is the function of forcing variables defining equilibrium. Optimizing with respect to the choice of y_t leads directly to a plan to set y_t to a value intermediate between y_t^* and y_{t-1},

$$y_t = \lambda y_{t-1} + (1 - \lambda) y_t^*, \qquad 0 \leq \lambda \leq 1$$

which after a simple rearrangement, and the addition of a shock representing random deviations from the plan, can be cast in the ECM form

$$\Delta y_t = (1 - \lambda)\Delta y_t^* + (1 - \lambda)(y_{t-1}^* - y_{t-1}) + \varepsilon_t$$

replacing y^* in practice by a linear combination of forcing variables.

The constraints across these dynamic adjustment coefficients are a consequence of the extreme simplicity (or maybe we should say naïveté) of this

particular setup. However, the framework is easily elaborated to allow for forward-looking behaviour and multi-step dynamic optimization. See Nickell (1985) also Davidson (2000, Section 5.5.4) for illustrations. What these examples show is that the solved form of the dynamic adjustment depends not only on the agent's optimization rule but also on the form of the processes generating the forcing variables.

Campbell and Shiller (1988) argue that error-correction behaviour can be observed even without the existence of adjustment costs, and illustrate their case with the class of present value models. Theory has the spread between long and short rates depending mechanically on the difference between the former and rational forecasts of the latter; but if these forecasts use information not available to the modeller, the spread involves a random component that, moreover, must Granger-cause the future changes in the short rate. This gives rise to an error correction structure with the spread representing the cointegrating residual, but note that this structure does not arise through agents reacting to resolve perceived disequilibria, as the classic ECM framework suggests.

Cointegration has been derived in the preceding sections as the attribute of a system of dynamic equations. However, many of the models that appear in the applied literature, the prototypical examples of Sargan (1964), Davidson et al. (1978), Hendry (1979) and many others, are cast as single equations and estimated by least squares. The driving variables are assumed to be *weakly exogenous* within the time frame of observation. Weak exogeneity is a technical concept, defined formally in Engle et al. (1983), but it can be loosely interpreted to describe a variable that is regarded as given and conditionally fixed by agents within the decision period, even though it could be endogenous in the wider sense of depending on past values of the variables it drives. A key implication of weak exogeneity is that the variable is uncorrelated with the shocks in the regression model, and hence ordinary least squares is a consistent estimator for the dynamic equation.

Without loss of generality, assume that the equation of interest is the first equation in the system, and so partition the variables as $\boldsymbol{x}_t = (x_{1t},\, \boldsymbol{x}'_{2t})'$. Further assume, in concert with the cited references, that the cointegrating rank is 1. The structural system (16) is then partitioned as

$$\begin{aligned}
&\begin{bmatrix} 1 & \boldsymbol{b}'_{0,12} \\ \boldsymbol{b}_{0,21} & \boldsymbol{B}_{0,22} \end{bmatrix} \begin{bmatrix} \Delta x_{1t} \\ \Delta \boldsymbol{x}_{2t} \end{bmatrix} \\
&= \begin{bmatrix} b_{1,11}(L) & \boldsymbol{b}'_{1,12}(L) \\ \boldsymbol{b}_{1,21}(L) & \boldsymbol{B}_{1,22}(L) \end{bmatrix} \begin{bmatrix} \Delta x_{1,t-1} \\ \Delta \boldsymbol{x}_{2,t-1} \end{bmatrix} + \begin{bmatrix} \rho_1 \\ \boldsymbol{\rho}_2 \end{bmatrix} z_{t-1} + \begin{bmatrix} u_{1t} \\ \boldsymbol{u}_{2t} \end{bmatrix}
\end{aligned} \tag{18}$$

where $z_t = \boldsymbol{\beta}'\boldsymbol{x}_t - \mu$. The noteworthy feature of this setup is the potential dependence of all the variables on z_{t-1}. If $\boldsymbol{\beta}$ is known then z_t can be treated as a datum and there is no obstacle to estimating the first equation by least squares, subject to the usual weak exogeneity restrictions on the distribution

of $\boldsymbol{x}_{2t}$, specifically that $\boldsymbol{b}_{0,21} = \mathbf{0}$ and $E(u_{1t}\boldsymbol{u}_{2t}) = \mathbf{0}$. On the other hand, if $\boldsymbol{\beta}$ is unknown, then it is potentially involved in all the equations of the system. Weak exogeneity of $\boldsymbol{x}_{2t}$ in the first equation requires the extra condition $\boldsymbol{\rho}_2 = \mathbf{0}$, so that the error correction effect is wholly focused on the evolution of x_{1t}. Under these circumstances, the first equation can be studied in isolation, conditional on $\boldsymbol{x}_{2t}$. Note that $\boldsymbol{\beta}$ could be estimated by nonlinear least squares applied to this equation. We say more about this estimation question below.

5 Estimating Cointegrating Relations

We start the discussion of estimation with the focus of attention on the matrix $\boldsymbol{\beta}$ of cointegrating coefficients. Immediately, we run into the difficulty that this matrix is not in general unique. It is defined merely to span a space of m-vectors having the property that any element of the space cointegrates the variables of the model. One approach to estimation is to impose normalization restrictions, such as having the columns orthogonal and with unit length. The structural modelling approach, on the other hand, supposes that cointegration is to be explained by the existence of some long-run economic relations, and the cointegrating space is relevant because these structural vectors span it, in particular. When the cointegrating rank s is greater than 1, however, any linear combination of the hypothesized structural vectors is also a cointegrating vector. We therefore face a problem of identifying the parameters of interest.

Before approaching that more difficult case, assume initially that $s = 1$. Then $\boldsymbol{\beta}$ $(m \times 1)$ is unique up to a choice of normalization and, normalizing on x_{1t} in the partition $\boldsymbol{x}_t = (x_{1t},\ \boldsymbol{x}_{2t}')'$, with *almost* no loss of generality,[16] we can write the cointegrating relation as a regression model,

$$x_{1t} = \boldsymbol{\gamma}'\boldsymbol{x}_{2t} + \mu + z_t \tag{19}$$

where $\boldsymbol{\beta} = (1, -\boldsymbol{\gamma}')'$, and it is natural to consider the possibility of OLS estimation. If

$$S(\boldsymbol{g}, m) = \sum\nolimits_{t=1}^{T}(x_{1t} - \boldsymbol{g}'\boldsymbol{x}_{2t} - m)^2$$

it can be shown that $S(\boldsymbol{\gamma}, m) = O_p(T)$ for any m, whereas $S(\boldsymbol{g}, m) = O_p(T^2)$ at points where $\boldsymbol{g} \neq \boldsymbol{\gamma}$. The proof of consistency of least squares is therefore very direct, and (letting hats denote the least squares estimators) $\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma} = O_p(T^{-1})$ by comparison with the usual convergence rate of $O_p(T^{-1/2})$ in

[16] Where the choice of normalization has unintended consequences is in the case where the first element of $\boldsymbol{\beta}$ is actually zero, so that x_{1t} is not in the cointegrating set. This is a valid special case of the model and obviously needs be ruled out by assumption. To pre-empt this possibility it's desirable to compare alternative normalizations.

stationary data. This property is known as *superconsistency.*[17] The other features of this regression include $R^2 \to 1$ as $T \to \infty$.

However, notwithstanding these desirable properties, the large-sample distribution of $T(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})$ is non-standard, and depends critically on the structure of the model. Consider the OLS formula in the standard notation $\hat{\boldsymbol{\gamma}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$, the rows of $\boldsymbol{X}$ having the form $(\boldsymbol{x}_{2t} - \bar{\boldsymbol{x}}_2)'$ for $t = 1, \ldots, T$ where $\bar{\boldsymbol{x}}_2$ is the vector of sample means, and $\boldsymbol{y} = (x_{11}, \ldots, x_{1T})'$. The problem with cointegrating regression is that the regressors do not obey the law of large numbers. It can be shown that

$$\begin{aligned} T(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) &= \left(\frac{\boldsymbol{X}'\boldsymbol{X}}{T^2}\right)^{-1} \frac{\boldsymbol{X}'\boldsymbol{u}}{T} \\ &\xrightarrow{d} \boldsymbol{P}^{-1}\boldsymbol{q} \end{aligned} \tag{20}$$

where $\xrightarrow{d}$ denotes convergence in distribution, and $\boldsymbol{u} = (z_1, \ldots, z_T)'$. $\boldsymbol{P}$ and $\boldsymbol{q}$, the limits in distribution of the normalized sums of squares and products matrices, are in general random variables and correlated with each other. Since $\boldsymbol{q}$ typically has a mean different from zero, there can be substantial finite sample biases. Similarly, the usual regression standard errors do not converge to constants, as in the stationary data analysis, but to random elements proportional to the square roots of the diagonal elements of $\boldsymbol{P}^{-1}$. The asymptotic distributions of the regression t-ratios are therefore not merely non-standard, but depend on nuisance parameters and cannot be tabulated. All these facts are bad news for making inferences on cointegrating vectors.

However, there is a favourable special case. Suppose that $\boldsymbol{x}_{2t}$ is strictly exogenous in equation (19), which means that $E(\boldsymbol{x}_{2t-j}z_t) = \boldsymbol{0}$ for $-\infty < j < \infty$. For this condition to be satisfied, note that in (18) the parameters $\boldsymbol{b}_{0,21}$, $\boldsymbol{b}_{1,21}(L)$ and $\boldsymbol{\rho}_2$ will all need to be zero, and in addition, $E(u_{1t}\boldsymbol{u}'_{2t}) = \boldsymbol{0}$. In this case, the distribution of $T(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})$ is *asymptotically mixed normal.* Under strict exogeneity, $\boldsymbol{X}$ in (20) can be treated as conditionally fixed when considering the distribution of $\boldsymbol{u}$. It can be shown that $T(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})$ is asymptotically normally distributed under the *conditional* distribution, holding $\boldsymbol{X}$ fixed, although its variance matrix is a random drawing under the unconditional distribution, hence 'mixed normal'. Further, we can compute t ratios that (on the null hypothesis) are *conditionally* N(0,1) in the limit. However, since this distribution is the same for any set of conditioning variables, the same limit result holds *un*conditionally. This means that standard inference procedures, using tabulations of the standard normal and chi-squared distributions, are asymptotically valid. The only modification of the usual least squares inference procedure that may be necessary, since the residuals are typically autocorrelated, is to use a heteroscedasticity and autocorrelation

[17] Note however that $\hat{\mu} - \mu = O_p(T^{1/2})$ in the usual way.

consistent (HAC) estimator for the residual variance, such as that derived by Newey and West (1987).[18]

Unfortunately, strict exogeneity is a very strong assumption in macroeconomic data, and this favourable case is the exception to the rule. An alternative approach, while maintaining the single-equation framework, is to estimate the dynamic error correction model itself by nonlinear least squares. This method is analysed by Stock (1987). The first equation of (18) may be written

$$\Delta x_{1t} = a_0 - \boldsymbol{b}_{0,12}'\Delta \boldsymbol{x}_{2t} + \boldsymbol{b}_{1,1}(L)'\Delta \boldsymbol{x}_{t-1} + \rho_1(x_{1,t-1} - \boldsymbol{\gamma}'\boldsymbol{x}_{2,t-1}) + u_{1t} \quad (21)$$

This equation can be estimated unrestrictedly by least squares, and $\hat{\boldsymbol{\gamma}}$ recovered by dividing the coefficients of $\boldsymbol{x}_{2,t-1}$ by minus the coefficient of $x_{1,t-1}$. Alternatively, a nonlinear optimization algorithm may be used. This estimator can be shown to be superconsistent, and it is also asymptotically mixed normal (meaning that standard inference applies, as above) subject to the weak exogeneity condition detailed following (18). In particular, in addition to the usual requirements of no simultaneity, the condition $\boldsymbol{\rho}_2 = \mathbf{0}$ is needed to ensure that all the sample information about the cointegrating relation is contained in (21). Without these conditions, there is once again a failure of mixed normality, and a dependence of the limit distributions on nuisance parameters. However, note that these conditions are less severe than those required to obtain the equivalent result for the OLS estimator of the cointegrating relation itself.

To achieve standard asymptotic inference in arbitrary cases of (18), a number of proposals have been made to modify the least squares estimator. Saikkonen (1991) and Stock and Watson (1993) independently proposed similar procedures. Consider the triangular representation in (17), assuming $s = 1$ for present purposes. Saikkonen shows that the $\boldsymbol{x}_{2t}$ variables can be treated as conditionally fixed in the regression of the first block if $E(v_{1t}\boldsymbol{v}_{2,t-j}') = \mathbf{0}$ for $-\infty < j < \infty$ where, in this context, $\boldsymbol{v}_{2t} = \Delta \boldsymbol{x}_{2t}$. However, by augmenting the first equation in (17) with these observed variables, the same condition can be engineered. Substituting from the second block, the ideal set of additional regressors are $\Delta \boldsymbol{x}_{2,t-j}$ for $-\infty < j < \infty$. Whereas this is not a feasible choice, the same asymptotic distribution is obtained by running the finite-order regression

$$x_{1t} = \boldsymbol{\gamma}'\boldsymbol{x}_{2t} + \sum\nolimits_{j=-K_T}^{K_T} \boldsymbol{\pi}_j \Delta \boldsymbol{x}_{2,t-j} + \mu + e_t \quad (22)$$

where K_T increases with T, although at a slower rate. Saikkonen proposes $K_T = o(T^{1/3})$.[19] In this regression, the regressors are "as if" strictly exogenous. The coefficients $\boldsymbol{\pi}_j$ are merely projection parameters and their

[18] Think of this as a method for estimating (an element of) $\boldsymbol{\Omega}$ in (9), rather than the corresponding (element of) $\boldsymbol{\Sigma}$.

[19] The $o()$ notation is a shorthand for the condition $|K_T|/T^{1/3} \to 0$ as $T \to \infty$.

values are generally not of direct interest. The unusual (from an econometric modelling point of view) step of including leads as well as lags in the regression has to be understood as allowing for the possibility that x_{1t} Granger-causes $\boldsymbol{x}_{2t}$ through endogenous feedbacks, hence the disturbance term must be purged of both past and future dependence on $\boldsymbol{x}_{2t}$. Thus, (22) must *not* be confused with a conventional structural equation describing agents' behaviour. Implicitly, we need the full multi-equation system to do this correctly.

The augmented least squares estimator is asymptotically mixed normal when correctly specified. Note that the regression in (22) does not make allowance for autocorrelation in the residual disturbance e_t, which can clearly exist even following the projection onto the $\Delta\boldsymbol{x}_{2,t-j}$ variables. This fact does not invalidate the asymptotic distribution results, provided that the covariance matrix is computed in the correct manner. As already noted for the strictly exogenous case, it is typically necessary to use a HAC estimator for the residual variance. Conventional t and F test statistics then have standard distributions asymptotically and the usual normal and chi-squared tables can be used to get approximate critical values. Saikkonen also shows that the augmented estimator is optimal, in the sense of achieving the maximum concentration of the asymptotic distribution about the true values.

An alternative approach to this type of correction is the fully modified least squares (FMLS) estimator of Phillips and Hansen (1990). The essential idea here is to derive the limiting distribution of $\boldsymbol{P}^{-1}\boldsymbol{q}$ in (20), identify the components of this formula that produce the deviation from the centred mixed normal distribution, and estimate these components using the sample. The ingredients of these modifications include the covariance matrix of the data increments and disturbances, estimated by an HAC formula using the consistent OLS estimator of the parameters computed in a preliminary step. The resulting formulae are somewhat technical, and will not be reproduced here. The main thing to be aware of is that the asymptotic distribution of this estimator matches that of the Saikkonen-Stock-Watson augmented least squares estimator. Both of these methods are suitable for dealing with arbitrary forms of the distribution of the cointegrating VAR, and hence are inherently more robust than the single-equation ECM method of (21).

We have discussed the estimation of the vector $\boldsymbol{\gamma}$, but naturally we shall also be interested in inference on the dynamic parameters of an equation such as (21). In particular, we may be interested in knowing how rapidly the error-correction mechanism moves the variables towards their cointegrating relations. However, given an efficient estimator of $\boldsymbol{\gamma}$, we can now exploit the super-consistency property. Construct $\hat{z}_t = x_{1t} - \hat{\boldsymbol{\gamma}}'\boldsymbol{x}_{2t}$, and insert this constructed sequence into (21) with coefficient ρ_1. These residuals can be treated effectively as data from the standpoint of the asymptotic distribution, and are (by hypothesis) I(0), so the usual asymptotics for stationary data can be used to make inferences about ρ_1 and the other parameters of

the equation.

6 Multiple Cointegrating Vectors

Consider the case when there are two or more linearly independent vectors spanning the cointegrating space. Here is a simple example with $m = 3$. Suppose that $\boldsymbol{x}_t = (x_{1t}, x_{2t}, x_{3t})' \sim \mathrm{I}(1)$ and

$$p_t = x_{1t} - \mu x_{2t} \sim \mathrm{I}(0) \tag{23a}$$

$$q_t = x_{2t} - \nu x_{3t} \sim \mathrm{I}(0). \tag{23b}$$

Then, for any λ,

$$p_t + \lambda q_t = x_{1t} - (\mu - \lambda)x_{2t} - \lambda\nu x_{3t} \sim \mathrm{I}(0).$$

The vectors $\boldsymbol{\beta_\lambda} = (1, -(\mu - \lambda), -\lambda\nu)'$ are cointegrating for all choices of λ. If an attempt is made to estimate this vector, say by OLS regression of x_{1t} onto x_{2t} and x_{3t}, then the estimated coefficients will merely correspond to the case of λ that minimizes the sum of squares, which in turn depends on the relative sample variances of the variables p_t and q_t. It cannot tell us anything about the values of μ or ν, as such. While setting $\lambda = 0$ returns us relation (23a), there is in fact no value of λ that can return (23b) because of the choice of normalization.

Nonetheless, there is a simple way to estimate μ and ν, given that we know the structure. This is to run two regressions[20], the first one excluding x_{3t} and the second one excluding x_{1t} and normalized on x_{2t}. In fact the, regression of x_{1t} onto x_{3t} will estimate a third cointegrating vector of the system, $\boldsymbol{\beta_\mu} = (1, 0, -\mu\nu)'$.

On the other hand, suppose that (23a) holds, but not (23b), and instead there exists a cointegrating relation of the form

$$x_{1t} - \delta_1 x_{2t} - \delta_2 x_{3t} \sim \mathrm{I}(0) \tag{24}$$

It is easy to see that while the same restricted regression procedure will consistently estimate μ, there is no way to estimate the coefficients of (24). Running the regression with all three variables inevitably gives us an arbitrary linear combination of (23a) and (24). We say in this situation that the coefficients δ_1 and δ_2 are *unidentified*.

Generalizing from this example we see that the problem has a strong affinity with the analysis of static simultaneous equations that we now associate with the research agenda of the Cowles Commission at the University of Chicago in the 1940s (see Koopmans 1949, and also any number of econometrics texts, such as Johnston and DiNardo 1997). If $\boldsymbol{\beta}$ $(m \times s)$ is a matrix

[20] Here we use the term "regression" generically, to denote any of the consistent methods described in Section 5

spanning the cointegrating space, any vector of the form $\boldsymbol{\beta r}$ is a cointegrating vector where $\boldsymbol{r}$ $(s \times 1)$ is arbitrary. The only way that one of these vectors can be distinguished from another is by the existence of known restrictions on the coefficients. Assume for the sake of argument that the columns of $\boldsymbol{\beta}$ are "structural" in the sense that the elements have a specific interpretation in terms of economic behaviour. In particular, some of these elements are known to be zero, since structural economic relations do not in general involve all the variables in a system. Such a relation (say, the first column of $\boldsymbol{\beta}$ with no loss of generality) is said to be identified if the only choice of $\boldsymbol{r}$ that preserves the known restrictions is $r = \boldsymbol{e}_1 = (1, 0, \ldots, 0)'$. Assume, without loss of generality, that the variables are ordered so that the first g_1 of the elements of column 1 of $\boldsymbol{\beta}$ are nonzero, with the first element set to 1 as normalization, and the last $m - g_1$ elements are zeros. Accordingly, partition $\boldsymbol{\beta}$ by rows as $\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix}$ where $\boldsymbol{\beta}_2$ $(m - g_1 \times s)$ has first column equal to zero by construction, so that its rank cannot exceed $s - 1$. The following well-known proposition is the *rank condition for identification*:

- Equation 1 is identified if and only if $\boldsymbol{\beta}_2$ has rank $s - 1$.

Clearly, $\boldsymbol{\beta}_2$ having maximum rank means that it is not possible to construct a zero linear combination of its columns except for the specific cases of $a\boldsymbol{e}_1$ for scalar a, where the normalization rules out all of these cases except $a = 1$. An important by-product of this result is the *order condition for identification* (necessary but not sufficient) that requires $g_1 \leq m - s + 1$.

We now have the following result: a structural cointegrating relation that is identified by zero restrictions is consistently estimated by a least squares regression (or efficient counterpart) imposing these zero restrictions.[21] In text-book accounts of the simultaneous equations model, recall that it is necessary to separate the variables of the model into endogenous and exogenous categories, and implement estimation by (for example) two-stage least squares, where the order condition for identification determines whether sufficient instruments are available to estimate the unrestricted coefficients. Here, there is no such separation. All the variables are on the same footing and least squares is consistent, with identification achieved by excluding variables to match the known restrictions. Every identified structural cointegrating relation can be consistently and efficiently estimated by running either the Saikkonen-Stock-Watson or Phillips-Hansen procedures on equations containing only the non-excluded variables. For example, following Saikkonen's notation, equation (22) would become

$$x_{1t} = \boldsymbol{\gamma}' \boldsymbol{x}_{2t} + \sum\nolimits_{j=-K_T}^{K_T} \boldsymbol{\pi}_j \Delta \boldsymbol{x}_{c,t-j} + \mu + e_t \tag{25}$$

[21] Further discussion of this and related results can be found in Davidson (1994, 1998b).

where $\boldsymbol{x}_{ct} = (\boldsymbol{x}'_{2t}, \boldsymbol{x}'_{3t})'$, and the subscript 3 denotes the excluded variables. Each identified relation is estimated with a different partition of the variables into inclusions and exclusions, not overlooking the fact that the identity of the normalizing variable x_{1t} needs to be changed if it is itself to be excluded from the relation.

A further point of interest about identified structural relations is that they are *irreducible*. In other words, no variable can be dropped without the relation ceasing to be cointegrating. The examples in (23) are a good case in point, and this is how in practice we can detect the fact that a relation such as (24) cannot be both structural and identified. To appreciate the role of irreducibility, consider the triangular form (17) once again. We had assumed $s = 1$. Suppose however that, contrary to the implicit assumption, the variables $\boldsymbol{x}_{2t}$ in fact featured a cointegrating relation amongst themselves. Clearly, in this case, the first relation is not irreducible, although to discover this it may be necessary to change the normalization. Likewise if there are two or more cointegrating vectors containing x_{1t}, so that the estimated $\boldsymbol{\gamma}$ is a composite relation, there will necessarily exist a linear combination of these vectors that excludes one of the variables, and is cointegrating. So, again, it cannot be irreducible. Ideally, the irreducibility property should be checked (see Section 8 on cointegration testing) on each postulated structural relation. However, it's important to note that irreducibility is not an exclusive property of identified structures. In the three-variable example, it is of course shared by the solved relation involving x_{1t} and x_{3t}. There is no way except by prior knowledge of the structure that we can be sure of distinguishing structural from irreducible solved forms.

7 Estimating Cointegrated Systems

In a series of papers focusing chiefly on the triangular parameterization (17), Peter Phillips and coauthors (Phillips 1988, Park and Phillips 1988, 1989, Phillips and Hansen 1990, Phillips 1991, Phillips and Loretan 1991) have provided a careful analysis of the issue of valid inference in cointegrated systems. One feature of their approach is that the cointegrated relations are always parameterized in reduced form. In other words, if

$$\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix} \begin{matrix} s \times s \\ (m-s) \times s \end{matrix}$$

then, in (17), $\boldsymbol{B} = -\boldsymbol{\beta}_1^{-1}\boldsymbol{\beta}_2$. While the normalization on $\boldsymbol{x}_{1t}$ is basically arbitrary – any partition of the variables that delivers a $\boldsymbol{\beta}_1$ of full rank will do – there is no reason in principle why the matrix $[\boldsymbol{I} : -\boldsymbol{B}]'$ should not be replaced with a matrix $\boldsymbol{\beta}$ of structural vectors, subject to identifying restrictions. Such an approach is less easy to implement in practice, however.

The primary lesson of this research is that the number of cointegrating vectors in the system is the crucial piece of information for efficient, mixed normal estimation. It's convenient as a pedagogical device to consider the case where $\boldsymbol{v}_t = (\boldsymbol{v}_{1t}', \boldsymbol{v}_{2t}')'$ in (17) is an i.i.d. vector. Then the efficient, asymptotically mixed normal estimator of the system is simply computed by applying least squares to the s augmented equations.

$$\boldsymbol{x}_{1t} = \boldsymbol{B}\boldsymbol{x}_{2t} + \boldsymbol{\Gamma}_{12}\Delta\boldsymbol{x}_{2t} + \boldsymbol{e}_t$$

where we define $\boldsymbol{\Gamma}_{12} = \boldsymbol{\Omega}_{22}^{-1}\boldsymbol{\Omega}_{21}$ with $\boldsymbol{\Omega}_{22} = E(\boldsymbol{v}_{2t}\boldsymbol{v}_{2t}')$, $\boldsymbol{\Omega}_{21} = E(\boldsymbol{v}_{2t}\boldsymbol{v}_{1t}')$ and, with Gaussian disturbances,

$$\boldsymbol{e}_t = \boldsymbol{v}_{1t} - \boldsymbol{\Gamma}_{12}\Delta\boldsymbol{x}_{2t} = \boldsymbol{v}_{1t} - E(\boldsymbol{v}_{1t}|\boldsymbol{v}_{2t}).$$

In the event that $\boldsymbol{v}_t$ is autocorrelated, the further augmentation by leads and lags of $\Delta\boldsymbol{x}_{2t}$ will provide efficiency, as detailed in Section 5 above. Contrast this with the case of the triangular model

$$\begin{aligned} \boldsymbol{x}_{1t} &= \boldsymbol{B}\boldsymbol{x}_{2t} + \boldsymbol{v}_{1t} \\ \boldsymbol{x}_{2t} &= \boldsymbol{\Pi}\boldsymbol{x}_{2t-1} + \boldsymbol{v}_{2t} \end{aligned} \tag{26}$$

where $\boldsymbol{\Pi}$ is unrestricted. The roots of the autoregressive system could be either unity or stable and the identity $\boldsymbol{v}_{2t} = \Delta\boldsymbol{x}_{2t}$ no longer obtains. Phillips (1991) shows that the maximum likelihood estimator of $\boldsymbol{B}$ in this system has an asymptotic distribution contaminated by nuisance parameters such that conventional inference is not possible. The knowledge that $\boldsymbol{\Pi} = \boldsymbol{I}_{m-s}$ is the key to mixed-normal asymptotics.

Thanks largely to the influential contributions of Søren Johansen (1988a,b, 1991, 1995), the most popular approach to system estimation is the *reduced rank regression* estimator. This works with the representation in (15), although specialized by assuming a finite-order vector autoregressive specification. To describe how this method works with the maximum clarity we develop the case of the first-order VECM

$$\Delta\boldsymbol{x}_t = \boldsymbol{a}_0 + \boldsymbol{\alpha}\boldsymbol{\beta}'\boldsymbol{x}_{t-1} + \boldsymbol{\varepsilon}_t \tag{27}$$

as in (4). As before, the key piece of prior information is the cointegrating rank of the system.

The natural estimator for a system of reduced form equations is *least generalized variance* (LGV), which is also the maximum likelihood estimator when the disturbances are Gaussian. This minimizes the determinant of the system covariance matrix,

$$\Lambda_s(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \left| \sum\nolimits_{t=2}^{T} \boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t' \right| = \left| \boldsymbol{S}_{00} - \boldsymbol{\alpha}\boldsymbol{\beta}'\boldsymbol{S}_{10} - \boldsymbol{S}_{01}\boldsymbol{\beta}\boldsymbol{\alpha}' + \boldsymbol{\alpha}\boldsymbol{\beta}'\boldsymbol{S}_{11}\boldsymbol{\beta}\boldsymbol{\alpha}' \right| \tag{28}$$

where

$$\begin{aligned}
\boldsymbol{S}_{00} &= \sum\nolimits_{t=2}^{T} (\Delta \boldsymbol{x}_t - \overline{\Delta \boldsymbol{x}})(\Delta \boldsymbol{x}_t - \overline{\Delta \boldsymbol{x}})' \\
\boldsymbol{S}_{01} &= \sum\nolimits_{t=2}^{T} (\Delta \boldsymbol{x}_t - \overline{\Delta \boldsymbol{x}})(\boldsymbol{x}_{t-1} - \bar{\boldsymbol{x}}_{-1})' \\
\boldsymbol{S}_{11} &= \sum\nolimits_{t=2}^{T} (\boldsymbol{x}_{t-1} - \bar{\boldsymbol{x}}_{-1})(\boldsymbol{x}_{t-1} - \bar{\boldsymbol{x}}_{-1})'
\end{aligned}$$

and $\boldsymbol{S}_{10} = \boldsymbol{S}_{01}'$. Note that the value of s is built into this function through the dimensions of the unknown matrices $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, and so is indicated in the subscript in (28). If additional non-I(1) variables are to be included in (27), such as dummy variables and lagged values of $\Delta \boldsymbol{x}_t$, these are removed by regressing $\Delta \boldsymbol{x}_t$ and $\boldsymbol{x}_{t-1}$ onto them and taking the residuals. The mean-deviations shown here are just the simplest possible case of this "partialling out" operation. It's conventional to replace $\boldsymbol{x}_{t-1}$ by $\boldsymbol{x}_{t-p}$ where p is the maximum lag order, but this is optional. Either choice of lag will yield the same asymptotic properties.

To minimize Λ_s, first fix $\boldsymbol{\beta}$ temporarily and regress $\Delta \boldsymbol{x}_t$ onto $\boldsymbol{\beta}' \boldsymbol{x}_{t-1}$ to get a conditional estimate of $\boldsymbol{\alpha}$; that is,

$$\hat{\boldsymbol{\alpha}} = (\boldsymbol{\beta}' \boldsymbol{S}_{11} \boldsymbol{\beta})^{-1} \boldsymbol{\beta}' \boldsymbol{S}_{10}. \tag{29}$$

Substitution of (29) into (28) yields the concentrated criterion function

$$\Lambda_s^*(\boldsymbol{\beta}) = \left| \boldsymbol{S}_{00} - \boldsymbol{S}_{01} \boldsymbol{\beta} (\boldsymbol{\beta}' \boldsymbol{S}_{11} \boldsymbol{\beta})^{-1} \boldsymbol{\beta}' \boldsymbol{S}_{10} \right| \tag{30}$$

Now, the rule for determinants of partitioned matrices gives the twin identities

$$\begin{aligned}
\begin{vmatrix} \boldsymbol{S}_{00} & \boldsymbol{S}_{01}\boldsymbol{\beta} \\ \boldsymbol{\beta}'\boldsymbol{S}_{10} & \boldsymbol{\beta}'\boldsymbol{S}_{11}\boldsymbol{\beta} \end{vmatrix} &= \left| \boldsymbol{\beta}' \boldsymbol{S}_{11} \boldsymbol{\beta} \right| \left| \boldsymbol{S}_{00} - \boldsymbol{S}_{01} \boldsymbol{\beta} (\boldsymbol{\beta}' \boldsymbol{S}_{11} \boldsymbol{\beta})^{-1} \boldsymbol{\beta}' \boldsymbol{S}_{10} \right| \\
&= \left| \boldsymbol{S}_{00} \right| \left| \boldsymbol{\beta}' \boldsymbol{S}_{11} \boldsymbol{\beta} - \boldsymbol{\beta}' \boldsymbol{S}_{10} \boldsymbol{S}_{00}^{-1} \boldsymbol{S}_{01} \boldsymbol{\beta} \right|
\end{aligned}$$

from which we obtain the alternative form of (30),

$$\Lambda_s^*(\boldsymbol{\beta}) = \left| \boldsymbol{S}_{00} \right| \frac{\left| \boldsymbol{\beta}' (\boldsymbol{S}_{11} - \boldsymbol{S}_{10} \boldsymbol{S}_{00}^{-1} \boldsymbol{S}_{01}) \boldsymbol{\beta} \right|}{\left| \boldsymbol{\beta}' \boldsymbol{S}_{11} \boldsymbol{\beta} \right|} \tag{31}$$

where $|\boldsymbol{S}_{00}|$ does not depend on $\boldsymbol{\beta}$ and so can be omitted from the function. The next step is to appeal to a well-known result from multivariate analysis. The minimum with respect to $\boldsymbol{\beta}$ of the ratio of determinants in (31) is obtained by solving the generalized eigenvalue problem

$$\left| \lambda \boldsymbol{S}_{11} - \boldsymbol{S}_{10} \boldsymbol{S}_{00}^{-1} \boldsymbol{S}_{01} \right| = 0. \tag{32}$$

Specifically, $\Lambda_s^*(\boldsymbol{\beta})$ is minimized uniquely when the columns of $\boldsymbol{\beta}$ are the solutions $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_s$ of the s homogeneous equations

$$(\lambda_j \boldsymbol{S}_{11} - \boldsymbol{S}_{10} \boldsymbol{S}_{00}^{-1} \boldsymbol{S}_{01}) \boldsymbol{q}_j = \boldsymbol{0} \tag{33}$$

where $\lambda_1, \ldots, \lambda_s$ are the s *largest* solutions to (32), subject to the normalization

$$\boldsymbol{q}_i' \boldsymbol{S}_{11} \boldsymbol{q}_j = \begin{cases} 1, & i = j \\ 0, & \text{otherwise.} \end{cases} \tag{34}$$

The eigenvalues must fall in the interval $[0, 1]$. Noting that $\boldsymbol{S}_{11} = O(T^2)$ whereas $\boldsymbol{S}_{10}\boldsymbol{S}_{00}^{-1}\boldsymbol{S}_{01} = O(T)$, observe how necessarily $\lambda_j = O_p(T^{-1})$ unless the solution to (33) is a cointegrating vector. The normalization in (34) is not convenient, but letting $\boldsymbol{L}$ $(m \times m)$ be defined by $\boldsymbol{S}_{11}^{-1} = \boldsymbol{L}\boldsymbol{L}'$, so that $\boldsymbol{L}'\boldsymbol{S}_{11}\boldsymbol{L} = \boldsymbol{I}_m$, the $\lambda_1, \ldots, \lambda_m$ are also the simple eigenvalues of the matrix $\boldsymbol{L}'\boldsymbol{S}_{10}\boldsymbol{S}_{00}^{-1}\boldsymbol{S}_{01}\boldsymbol{L}$, whereas the eigenvectors are $\hat{\boldsymbol{\beta}}_j = \boldsymbol{L}\boldsymbol{q}_j$, which are orthonormal (orthogonal with unit length).

Care is needed in interpreting this result. The orthonormal matrix $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_s)$ asymptotically spans the cointegrating space, but it is not a reduced form nor, of course, a structural form. Given the arbitrary nature of the normalization, it is difficult to give an interpretation to these vectors, but for the fact that any structural cointegrating vector can be found asymptotically as a linear combination of the columns.

While inference on the elements of $\hat{\boldsymbol{\beta}}$ itself is neither possible nor indeed useful, it is possible to impose and test linear restrictions on the cointegrating space. Following Johansen and Juselius (1992), one can write for example

$$\boldsymbol{\beta} = \boldsymbol{H}\boldsymbol{\phi}$$

where $\boldsymbol{H}$ is a $m \times (m-r)$ matrix of known constants (0 and 1 typically) and $\boldsymbol{\phi}$ $(m-r \times s)$ is an unrestricted matrix of parameters. This parameterization allows the cointegrating space to satisfy a certain type of linear restriction, and permits a likelihood ratio test of these restrictions.

Davidson (1998a) shows how to test a set of p restrictions expressed in the form

- There exists a vector $\boldsymbol{a}$ $(s \times 1)$ such that $\boldsymbol{H}\boldsymbol{\beta}\boldsymbol{a} = \boldsymbol{0}$

where here, $\boldsymbol{H}$ is a $p \times m$ matrix of known constants. This approach allows testing of hypotheses such as "a vector subject to p specified zero restrictions lies in the cointegrating space". Given asymptotic mixed normality of the estimators $\hat{\boldsymbol{\beta}}$, which can be demonstrated subject to regularity conditions, these tests can be performed making use of the standard chi-squared tables in large samples.

8 Testing Cointegration

We have shown that the cointegrating rank of a collection of I(1) processes is the key piece of information, without which inference on the system cannot realistically proceed. It is in this context in particular that Søren Johansen's

contributions have proved essential. The discussion of the last section has already provided the clue. Consider the m solutions to (32), ordered as $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$. These are functions of the data set but their computation does not depend on a choice of s. The $m - s$ smallest converge to zero as $T \to \infty$, because the corresponding eigenvectors $\boldsymbol{q}_j$ are not cointegrating vectors, and the terms $\boldsymbol{S}_{11}\boldsymbol{q}_j$ and $\boldsymbol{S}_{10}\boldsymbol{S}_{00}^{-1}\boldsymbol{S}_{01}\boldsymbol{q}_j$ in (33) are therefore diverging at different rates, respectively $O_p(T^2)$ and $O_p(T)$. Balancing the equation imposes $\lambda_j = O_p(T^{-1})$ for $s < j \leq m$, while $\lambda_j = O_p(1)$ for $1 \leq j \leq s$.

This provides the basis for a test based on the statistics $T\lambda_j$, which are computed as a by-product of the reduced rank regression. If $j \leq s$ then $T\lambda_j = O_p(T)$, otherwise $T\lambda_j = O_p(1)$. Suppose that the distributions of the $T\lambda_j$ in the second case can be tabulated. We can then proceed to compare the statistic with this distribution for the cases $j = 1, 2, \ldots, m$, in decreasing order of magnitude, until the chosen critical value is not exceeded. Then, s can be estimated as the last case of j before this non-rejection occurs. For any choice of s, the tests can be formally cast in the form H_0 : *cointegrating rank* $= s$ against the alternative H_1 : *cointegrating rank* $\geq s + 1$.

It is shown by an ingenious argument that, under the null hypothesis, the non-divergent $T\lambda_j$ (the cases $j = s + 1 \ldots, m$) are tending as $T \to \infty$ to the eigenvalues of a certain random matrix of dimension $(m - s) \times (m - s)$ whose distribution is free of nuisance parameters. This limiting distribution is shared with matrices that can be generated on the computer using pseudo-random numbers, so the distribution of its eigenvalues can be tabulated in a Monte Carlo simulation exercise.

There are two ways in which this idea might be implemented as a test. One is to look at $T\lambda_{s+1}$, the largest of the set of the $m - s$ smallest rescaled eigenvalues. This is called the *maximum eigenvalue test*. The second implementation is to look at $\sum_{j=s+1}^{m} T\lambda_j$. This latter statistic converges to the trace of the limit matrix, and so this is known as the *trace test*. Each of these distributions has been tabulated for a range of values of $m - s$, although not depending on m, note, since the cases (m, s) and $(m+1, s+1)$ are equivalent.

It can also be shown that the minimized value of the generalized variance is

$$\Lambda_s^*(\hat{\boldsymbol{\beta}}) = \prod_{j=1}^{s}(1 - \lambda_j)$$

(the product of the s terms) and hence, using the fact that $\log(1 + x) \approx x$ when x is small,

$$T \log \Lambda_s^*(\hat{\boldsymbol{\beta}}) - T \log \Lambda_{s+1}^*(\hat{\boldsymbol{\beta}}) = -T \log(1 - \lambda_{s+1}) \sim T\lambda_{s+1}$$

and

$$T \log \Lambda_s^*(\hat{\boldsymbol{\beta}}) - T \log \Lambda_m^*(\hat{\boldsymbol{\beta}}) = -T \sum_{j=s+1}^{m} \log(1 - \lambda_j) \sim \sum_{j=s+1}^{m} T\lambda_j$$

where '$\sim$' denotes that the ratio of the two sides converges to 1. Hence, asymptotically equivalent tests can be based on the estimation minimands. If the disturbances are Gaussian, the maximized likelihood takes the form $-\frac{1}{2}T \log \Lambda_s^*(\hat{\boldsymbol{\beta}})$ and then these forms have the natural interpretation of likelihood ratio tests. However, be careful to note that these limiting distributions are not chi-squared. It is a general rule of I(1) asymptotics that restrictions affecting the orders of integration of variables – in other words that concern unit roots – give rise to non-standard distributions. Be careful to note that the standard asymptotic tests that we have described in this chapter all share the feature that the cointegrating rank is given and not part of the tested hypothesis.

An interesting special case is the test based on the statistic $T\lambda_m$, for the null hypothesis of a single common trend (cointegrating rank $s = m - 1$) against the alternative that the data are stationary. In this case the trace and maximal eigenvalue statistics coincide and, interestingly enough, the null limiting distribution is none other than the square of the Dickey-Fuller distribution associated with the standard test for a unit root.

An alternative approach to testing cointegration is to estimate a single equation and test whether the resulting residual is I(0). In these tests, non-cointegration is the null hypothesis. This is basically comparable to testing the hypothesis $H_0 : s = 0$ in the cointegrating VECM framework, but avoids modelling the complete system. A well-known paper by Phillips and Ouliaris (1990) compares a range of alternative implementations of this idea. The best known is based on the augmented Dickey-Fuller (ADF) test for a unit root (Dickey and Fuller 1979, 1981, Said and Dickey 1984) applied to the residuals from an ordinary least squares regression. The test statistic takes the form of the t ratio for the parameter estimate $\hat{\phi}$ in the regression

$$\Delta \hat{u}_t = \phi \hat{u}_{t-1} + \sum\nolimits_{j=1}^{K_T} \pi_j \Delta \hat{u}_{t-j} + e_t \tag{35}$$

where $\hat{u}_t = x_{1t} - \hat{\mu} - \hat{\boldsymbol{\gamma}}' \boldsymbol{x}_{2t}$ and $K_T = o(T^{1/3})$.

Although this test closely resembles the augmented Dickey-Fuller test for a unit root, there are a number of important issues to be aware of. When the null hypothesis is true, there is no cointegration and $\hat{\boldsymbol{\gamma}}$ does not converge in probability and is a random vector even in the limit as $T \to \infty$. A linear combination of random walks with random coefficients, where these coefficients are computed specifically to minimize the variance of the combination, is not itself a random walk, in the sense that the regular Dickey-Fuller distribution should be expected to apply. In fact, the asymptotic distribution of this test depends only on the number of elements in $\boldsymbol{x}_{2t}$, and tabulation of the distributions is therefore feasible (see Engle and Yoo 1987). However, while it might be assumed that an efficient single-equation estimator would be a better choice than OLS for the estimator of $\boldsymbol{\gamma}$, in fact the limit distributions have been derived on the assumption of OLS estimation and depend

on this for their validity. The requirement that $K_T \to \infty$ is important because, under H_0, $\Delta\hat{u}_t$ is a random combination of stationary processes. Even if these have finite-order autoregressive structures individually, there is no reason to assume this of the combination. The idea of approximating a finite-order ARMA process by an AR(∞), approximated in finite samples by the AR(K_T), is due to Said and Dickey (1984). In practice it should give an adequate account of the autocorrelation structure of most I(0) processes.

Leading alternatives to the ADF statistic are the statistics denoted $\hat{Z}_\alpha$ and $\hat{Z}_t$ in Phillips (1987), where the coefficient $\hat{\phi}$ is subjected to specifically tailored corrections that play an equivalent role to the lag-difference terms in (35). These corrections are similar in principle to those in the Phillips-Hansen (1990) fully modified least squares estimator of $\boldsymbol{\gamma}$, and make use of HAC estimates of the data long-run covariance matrix.

9 Conclusion

This chapter has aimed to survey the main issues in the specification and estimation of cointegrating relationships in nonstationary data. This is now a very large literature, and inevitably there are many aspects which there has been no space to deal with here. In particular, while a number of conclusions about the large-sample distributions of estimators have been asserted, no attempt has been made to describe the asymptotic analysis on which these conclusions rest. This theory makes a clever use of the calculus of Brownian motion, following from the fundamental idea that nonstationary economic time series, when viewed in the large, move much like pollen grains suspended in water as first observed microscopically by the botanist Robert Brown. The same mathematics can be used to analyse either phenomenon. The key result in this theory is the functional central limit theorem, generalizing the ordinary central limit theorem to show the limiting Gaussianity of every increment of the path of a normalized partial sum process. Interested readers can find many of the details omitted here in Part IV of the present author's text *Econometric Theory* (Davidson 2000).

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 the measurement of the business cycle. *Journal of Monetary Economics* 7, 151–74.

Brown, T. M. 1952: Habit persistence and lags in consumer behaviour. *Econometrica 20*, 355–71.

Campbell, J. Y. and R. J. Shiller 1988: Interpreting cointegrated models. *Journal of Economic Dynamics and Control* 12, 505-522.

Davidson, J. , D. F. Hendry, F. Srba, and S. Yeo 1978: Econometric modelling of the aggregate time-series relationship between consumers expenditure and income in the United Kingdom. *Economic Journal* 88, 661–92.

Davidson, J. 1994: Identifying cointegrating regressions by the rank condition. *Oxford Bulletin of Economics and Statistics*, 56, 103–8.

Davidson, J. 1998a: A Wald test of restrictions on the cointegrating space based on Johansen's estimator. *Economics Letters* 59, 183–7.

Davidson, J. 1998b: Structural relations, cointegration and identification: some simple results and their application. *Journal of Econometrics* 87, 87–113.

Davidson, J. 2000: *Econometric Theory*, Oxford: Blackwell Publishers.

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.

Dickey, D. A. and W. A. Fuller 1981: Likelihood ratio statistics for autoregressive time series with a unit root. *Econometrica* 49, 1057–72.

Engle, R. F., D. F. Hendry and J.-F. Richard 1983: Exogeneity. *Econometrica* 51, 277–304.

Engle, R. F. and B. S. Yoo 1987: Forecasting and testing in cointegrated systems *Journal of Econometrics* 35, 143–59.

Granger, C. W. J. 1969: Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* 37, 424–38.

Granger, C. W. J. 1981: Some properties of time series data and their use in econometric model specification. *Journal of Econometrics* 16, 121–30.

Granger, C. W. J. and P. Newbold 1974: Spurious regressions in econometrics, *Journal of Econometrics* 2, 111–20.

Hendry 1979: Predictive failure and econometric modelling in macroeconomics: the transactions demand for money. Chapter 9 of *Economic Modelling,* ed. P. Ormerod, Heinemann, (reprinted in *Econometrics – Alchemy or Science* by D. F. Hendry, Blackwell Publishers 1993.)

Hendry D. F. 1980: Econometrics – Alchemy or science? *Economica* 47, 387-406.

Hendry D. F. 2004: The Nobel Memorial Prize for Clive W. J. Granger. *Scandinavian Journal of Economics* 106(2), 187-213

Johansen, S. 1988a: The mathematical structure of error correction models. *Contemporary Mathematics* 80, 359–86.

Johansen, S. 1988b: Statistical analysis of cointegration vectors. *Journal of Economic Dynamics and Control* 12, 231–54.

Johansen, S. 1991: Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models. *Econometrica* 59, 1551–80.

Johansen, S. 1995: *Likelihood-Based Inference in Cointegrated Vector Autoregressive Models.* Oxford: Oxford University Press.

Johansen, S. and K. Juselius 1992: Testing structural hypotheses in a multivariate cointegration analysis of the PPP and the UIP for UK. *Journal of Econometrics* 53, 211–44.

Johnston, J., and J. DiNardo 1997: *Econometric Methods,* 4th Edition, McGraw-Hill.

Koopmans 1949: Identification problems in economic model construction, *Econometrica* 17, 125-144

Lovell, M. C. 1961: Manufacturers' inventories, sales expectations and the acceleration principle. *Econometrica* 29, 293–314

Newey, W. K. and K. D. West 1987: A simple, positive semi-definite, heteroscedasticity and autocorrelation consistent covariance matrix. *Econometrica* 55, 703–8.Newey and West (1987)

Nickell, S. J. 1985: Error correction, partial adjustment and all that: An expository note. *Oxford Bulletin of Economics and Statistics* 47, 119–30.

Park, J. Y. and P. C. B. Phillips 1988: Statistical inference in regressions with integrated processes: Part 1. *Econometric Theory* 4, 468-97.

Park, J. Y. and P. C. B. Phillips 1989: Statistical inference in regressions with integrated processes: Part 2, *Econometric Theory* 5, 95–131.

Phillips, P. C. B. 1987: Time series regression with a unit root. *Econometrica* 55, 277–301.

Phillips, P. C. B. 1988: Reflections on econometric methodology. *Economic Record* 64, 544–59.

Phillips, P. C. B. 1991: Optimal inference in cointegrated systems. *Econometrica* 59 283–306.

Phillips, P. C. B. and B. E. Hansen 1990: Statistical inference in instrumental variables regression with I(1) processes, *Review of Economic Studies* 57, 99–125.

Phillips, P. C. B. and M. Loretan 1991: Estimating long-run economic equilibria. *Review of Economic Studies* 58, 407–37.

Phillips, P. C. B. and S. Ouliaris 1990: Asymptotic properties of residual based tests for cointegration *Econometrica* 58, 165–93.

Phillips, P. C. B. and V. Solo 1992: Asymptotics for linear processes. *Annals of Statistics* 2, 971- 1001.

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

Saikkonen, P. 1991: Asymptotically efficient estimation of cointegration regressions, *Econometric Theory* 7, 1–21.

Salmon 1982:Error correction mechanisms, *The Economic Journal* 92, 615-

629
Sargan, J. D. 1964: Wages and prices in the United Kingdom: a study in econometric methodology, in P. E. Hart, G. Mills and J. K. Whitaker (eds.), *Econometric Analysis for National Economic Planning*, Butterworth, pp.25-54, reprinted in *Econometrics and Quantitative Economics*, eds D. F. Hendry and K. F. Wallis, Oxford: Basil Blackwell,.1984..

Stock, J. H. 1987: Asymptotic Properties of least squares estimators of cointegrating vectors, *Econometrica* 55 1035–56.

Stock, J. H. and M. W. Watson 1993: A simple estimator of cointegrating vectors in higher order integrated systems. *Econometrica* 61, 783–820.

Wold, H. 1938: *A Study in the Analysis of Stationary Time Series*. Uppsala: Almqvist and Wiksell.

Yule, G. U. 1926: Why do we sometimes get nonsense correlations between time series? *Journal of the Royal Statistical Society* 89, 1–64.