VECTOR AUTOREGRESSIONS AND COINTEGRATION*

MARK W. WATSON

Northwestern University and Federal Reserve Bank of Chicago

Contents

*The paper has benefited from comments by Edwin Denson, Rob Engle, Neil Ericsson, Michael Horvath, Soren Johansen, Peter Phillips, Greg Reinsel, James Stock and students at Northwestern University and Studienzentrum Gerzensee. Support was provided by the National Science Foundation through grants SES–89–10601 and SES–91–22463.

Handbook of Econometrics, Volume IV, Edited by R.F. Engle and D.L. McFadden

Abstract

This paper surveys three topics: vector autoregressive (VAR) models with integrated regressors, cointegration, and structural VAR modeling. The paper begins by developing methods to study potential "unit root" problems in multivariate models, and then presents a simple set of rules designed to help applied researchers conduct inference in VARs. A large number of examples are studied, including tests for Granger causality, tests for VAR lag length, spurious regressions and OLS estimators of cointegrating vectors. The survey of cointegration begins with four alternative representations of cointegrated systems: the vector error correction model (VECM), and the moving average, common trends and triangular representations. A variety of tests for cointegration and efficient estimators for cointegrating vectors are developed and compared. Finally, structural VAR modeling is surveyed, with an emphasis on interpretation, econometric identification and construction of efficient estimators. Each section of this survey is largely self-contained. Inference in VARs with integrated regressors is covered in Section 2, cointegration is surveyed in Section 3, and structural VAR modeling is the subject of Section 4.

1. Introduction

Multivariate time series methods are widely used by empirical economists, and econometricians have focused a great deal of attention at refining and extending these techniques so that they are well suited for answering economic questions. This paper surveys two of the most important recent developments in this area: vector autoregressions and cointegration.

Vector autoregressions (VARs) were introduced into empirical economics by Sims (1980), who demonstrated that VARs provide a flexible and tractable framework for analyzing economic time series. Cointegration was introduced in a series of papers by Granger (1983), Granger and Weiss (1983) and Engle and Granger (1987). These papers developed a very useful probability structure for analyzing both long-run and short-run economic relations.

Empirical researchers immediately began experimenting with these new models, and econometricians began studying the unique problems that they raise for econometric identification, estimation and statistical inference. Identification problems had to be confronted immediately in VARs. Since these models don't dichotomize variables into "endogenous" and "exogenous," the exclusion restrictions used to identify traditional simultaneous equations models make little sense. Alternative sets of restrictions, typically involving the covariance matrix of the errors, have been used instead. Problems in statistical inference immediately confronted researchers using cointegrated models. At the heart of cointegrated models are "integrated" variables, and statistics constructed from integrated variables often behave in nonstandard ways. "Unit root" problems are present and a large research effort has attempted to understand and deal with these problems.

This paper is a survey of some of the developments in VARs and cointegration that have occurred since the early 1980s. Because of space and time constraints, certain topics have been omitted. For example, there is no discussion of forecasting or data analysis; the paper focuses entirely on structural inference. Empirical

questions are used to motivate econometric issues, but the paper does not include a systematic survey of empirical work. Several other papers have surveyed some of the material covered here. In particular, the reader is referred to the survey on VARs by Canova (1991) and to surveys on statistical issues in integrated and cointegrated systems by Campbell and Perron (1991), Engle and Yoo (1991), Phillips (1988) and Phillips and Loretan (1989). Excellent textbook treatments of many of the issues discussed here can be found in Banerjee et al. (1993) and Hamilton (1994).

Before proceeding, it is useful to digress for a moment and introduce some notation. Throughout this paper, I(d) will denote a variable that is integrated of order d, where d is an integer. For our purposes, an I(d) process can be defined as follows. Suppose that $\phi(L)x_{0,t} = \theta(L)\varepsilon_t$, where the roots of the polynomial $\phi(z)$ and $\theta(z)$ are outside the unit circle and ε_t is a martingale difference sequence with variance σ^2. In other words, $x_{0,t}$ follows a covariance stationary and invertible autoregressive moving average (ARMA) process. Let $x_{d,t}$ be defined recursively by $x_{d,t} = \sum_{s=1}^{t} x_{d-1,s}$, for $d = 1, \ldots$. Then x_t^d is defined as I(d). This definition says that an I(d) process can be interpreted as a d-fold partial sum of stationary and invertible ARMA process.

Many of the statistical techniques surveyed in this chapter were developed to answer questions concerning the dynamic relationship between macroeconomic time series. With this in mind, it is useful to focus the discussion of econometric techniques on a set of concrete economic questions. The questions concern a macroeconomic system composed of eight time series: the logarithms of output (y), consumption (c), investment (i), employment (n), nominal wages (w), money (m), prices (p) and the level of nominal interest rates (r).

Economic hypotheses often restrict the Granger (1969) causal structure of the system. A classic example is Hall's (1978) interpretation of the permanent income/life-cycle model of consumption. In Hall's model, consumption follows a martingale, so that c_{t-1} is an optimal forecast of c_t. Thus, the model predicts that no variables in the system will Granger-cause consumption. When the data are integrated, some important and subtle statistical issues arise when this proposition is tested. For example, Mankiw and Shapiro (1985) demonstrate that unit root problems plague the regression of Δc_t onto y_{t-1}: standard critical values for Granger-causality test statistics lead to rejection of the null hypothesis far too frequently when the null is true. On the other hand, Stock and West (1988) show that these unit root problems disappear when Granger causality is tested using the regression of c_t onto c_{t-1} and y_{t-1}, but then reappear in the regression of c_t onto c_{t-1} and m_{t-1}. The Mankiw–Shapiro/Stock–West results are explained in Section 2 which focuses on the general problem of inference in regression models with integrated regressors.

Economic theories often restrict long-run relationships between economic variables. For example, the proposition that money is neutral in the long run implies that exogenous permanent changes in the level of m_t have no long-run effect on the level of y_t. When the money-output process is stationary, Lucas (1972) and Sargent (1972) show that statistical tests of long-run neutrality require a complete specification of the structural economic model generating the data. However, when money and output are integrated, Fisher and Seater (1993), show that the neutrality

proposition is testable without a complete specification of the structural model. The basic idea is that when money and output are integrated, the historical data contain permanent shocks. Long-run neutrality can be investigated by examining the relationship between the permanent changes in money and output. This raises two important econometric questions. First, how can the permanent changes in the variables be extracted from the historical time series? Second, the neutrality proposition involves "exogenous" components of changes in money; can these components be econometrically identified? The first question is addressed in Section 3, where, among other topics, trend extraction in integrated processes is discussed. The second question concerns structural identification and is discussed in Section 4.

One important restriction of economic theory is that certain "Great Ratios" are stable. In the eight-variable system, five of these restrictions are noteworthy. The first four are suggested by the standard neoclassical growth model. In response to exogenous growth in productivity and population, the neoclassical growth model predicts that output, consumption and investment will grow in a balanced way. That is, even though y_t, c_t, and i_t increase permanently in response to increases in productivity and population, there are no permanent shifts in $c_t - y_t$ and $i_t - y_t$. The model also predicts that the marginal product of capital will be stable in the long run, suggesting that similar long-run stability will be present in ex-post real interest rates, $r - \Delta p$. Absent long-run frictions in competitive labor markets, real wages equal the marginal product of labor. Thus, when the production function is Cobb–Douglas (so that marginal and average products are proportional), $(w - p) - (y - n)$ is stable in the long run. Finally, many macroeconomic models of money [e.g. Lucas (1988)] imply a stable long-run relation between real balances $(m - p)$, output (y) and nominal interest rates (r), such as $m - p = \beta_y y + \beta_r r$; that is, these models imply a stable long-run "money demand" equation.

Kosobud and Klein (1961) contains one of the first systematic investigations of these stability propositions. They tested whether the deterministic growth rates in the series were consistent with the propositions. However, in models with stochastic growth, the stability propositions also restrict the stochastic trends in the variables. These restrictions can be described succinctly. Let x_t denote the 8×1 vector $(y_t, c_t, i_t, n_t, w_t, m_t, p_t, r_t)$. Assume that the forcing processes of the system (productivity, population, outside money, etc.) are such that the elements of x_t are potentially I(1). The five stability propositions imply that $z_t = \alpha' x_t$ is I(0), where

$$\alpha = \begin{bmatrix} 1 & 1 & -1 & -\beta_y & 0 \\ -1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & -1 & -1 & 0 \\ 0 & 0 & 0 & -\beta_r & 1 \end{bmatrix}.$$

The first two columns of α are the balanced growth restrictions, the third column is the real wage – average labor productivity restriction, the fourth column is stable long-run money demand restriction, and the last column restricts nominal interest rates to be I(0). If money and prices are I(1), Δp is I(0) so that stationary real rates imply stationary nominal rates.[1]

These restrictions raise two econometric questions. First, how should the stability hypotheses be tested? This is answered in Section 3.3 which discusses tests for cointegration. Second, how should the coefficients β_y and β_r be estimated from the data, and how should inference about their values be carried out?[2] This is the subject of Section 3.4 which considers the problem of estimating cointegrating vectors.

In addition to these narrow questions, there are two broad and arguably more important questions about the business cycle behavior of the system. First, how do the variables respond dynamically to exogenous shocks? Do prices respond sluggishly to exogenous changes in money? Does output respond at all? And if so, for how long? Second, what are the important sources of fluctuations in the variables. Are business cycles largely the result of supply shocks, like shocks to productivity? Or do aggregate demand shocks, associated with monetary and fiscal policy, play the dominant role in the business cycle?

If the exogenous shocks of econometric interest – supply shocks, monetary shocks, etc. – can be related to one-step-ahead forecast errors, then VAR models can be used to answer these questions. The VAR, together with a function relating the one-step-ahead forecast errors to exogenous structural shocks is called a "structural" VAR. The first question – what is the dynamic response of the variables to exogenous shocks? – is answered by the moving average representation of the structural VAR model and its associated impulse response functions. The second question – what are the important sources of economic fluctuations? – is answered by the structural VAR's variance decompositions. Section 4 shows how the impulse responses and variance decompositions can be computed from the VAR. Their calculation and interpretation are straightforward. The more interesting econometric questions involve issues of identification and efficient estimation in structural VAR models. The bulk of Section 4 is devoted to these topics.

Before proceeding to the body of the survey, three organizational comments are useful. First, the sections of this survey are largely self contained. This means that the reader interested in structural VARs can skip Sections 2 and 3 and proceed directly to Section 4. The only exception to this is that certain results on inference in cointegrated systems, discussed in Section 3, rely on asymptotic results from Section 2. If the reader is willing to take these results on faith, Section 3 can be read without the benefit of Section 2. The second comment is that Sections 2 and

[1] Since nominal rates are I(0) from the last column of α, the long run interest semielasticity of money demand, β_r, need not appear in the fourth column of α.

[2] The values of β_y and β_r are important to macroeconomists because they determine (i) the relationship between the average growth rate of money, output and prices and (ii) the steady-state amount of seignorage associated with any given level of money growth.

3 are written at a somewhat higher level than Section 4. Sections 2 and 3 are based on lecture notes developed for a second year graduate econometrics course and assumes that students have completed a traditional first year econometrics sequence. Section 4, on structural VARs, is based on lecture notes from a first year graduate course in macroeconomics and assumes only that students have a basic understanding of econometrics at the level of simultaneous equations. Finally, this survey focuses only on the classical statistical analysis of I(1) and I(0) systems. Many of the results presented here have been extended to higher order integrated systems, and these extensions will be mentioned where appropriate.

2. Inference in VARs with integrated regressors

2.1. *Introductory comments*

Time series regressions that include integrated variables can behave very differently than standard regression models. The simplest example of this is the AR(1) regression: $y_t = \rho y_{t-1} + \varepsilon_t$, where $\rho = 1$ and ε_t is independent and identically distributed with mean zero and variance σ^2, i.i.d.$(0, \sigma^2)$. As Stock shows in his chapter of the Handbook, $\hat{\rho}$, the ordinary least squares (OLS) estimator of ρ, has a non-normal asymptotic distribution, is asymptotically biased, and yet is "super consistent," converging to its true value at rate T.

Estimated coefficients in VARs with integrated components, can also behave differently than estimators in covariance stationary VARs. In particular, some of the estimated coefficients behave like $\hat{\rho}$, with non-normal asymptotic distributions, while other estimated coefficients behave in the standard way, with asymptotic normal large sample distributions. This has profound consequences for carrying out statistical inference, since in some instances, the usual test statistics will not have asymptotic χ^2 distributions, while in other circumstances they will. For example, Granger causality test statistics will often have nonstandard asymptotic distributions, so that conducting inference using critical values from the χ^2 table is incorrect. On the other hand, test statistics for lag length in the VAR will usually be distributed χ^2 in large samples. This section investigates these subtleties, with the objective of developing a set of simple guidelines that can be used for conducting inference in VARs with integrated components. We do this by studying a model composed of I(0) and I(1) variables. Although results are available for higher order integrated systems [see Park and Phillips (1988, 1989), Sims et al. (1990) and Tsay and Tiao (1990)], limiting attention to I(1) processes greatly simplifies the notation with little loss of insight.

2.2. *An example*

Many of the complications in statistical inference that arise in VARs with unit

roots can be analyzed in a simple univariate AR(2) model[3]

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \eta_t. \tag{2.1}$$

Assume that $\phi_1 + \phi_2 = 1$ and $|\phi_2| < 1$, so that process contains one unit root. To keep things simple, assume that η_t is i.i.d.(0, 1) and normally distributed [n.i.i.d.(0, 1)]. Let $x_t = (y_{t-1}\ y_{t-2})'$ and $\phi = (\phi_1\ \phi_2)'$, so that the OLS estimator is $\hat{\phi} = (\sum x_t x_t')^{-1} \times (\sum x_t y_t)$ and $(\hat{\phi} - \phi) = (\sum x_t x_t')^{-1}(\sum x_t \eta_t)$. (Unless noted otherwise, $\sum$ will denote $\sum_{t=1}^{T}$ throughout this paper.)

In the covariance stationary model, the large sample distribution of $\hat{\phi}$ is deduced by writing $T^{1/2}(\hat{\phi} - \phi) = (T^{-1}\sum x_t x_t')^{-1}(T^{-1/2}\sum x_t \eta_t)$, and then using a law of large numbers to show that $T^{-1}\sum x_t x_t' \xrightarrow{p} E(x_t x_t') \equiv V$, and a central limit theorem to show that $T^{-1/2}\sum x_t \eta_t \xrightarrow{\mathscr{L}} N(0, V)$. These results, together with Slutsky's theorem, imply that $T^{1/2}(\hat{\phi} - \phi) \xrightarrow{\mathscr{L}} N(0, V^{-1})$.

When the process contains a unit root, this argument fails. The most obvious reason is that, when $\rho = 1$, $E(x_t x_t')$ is not constant, but rather grows with t. Because of this, $T^{-1}\sum x_t x_t'$ and $T^{-1/2}\sum x_t \eta_t$ no longer converge: convergence requires that $\sum x_t x_t'$ be divided by T^2 instead of T, and that $\sum x_t \eta_t$ be divided by T instead of $T^{1/2}$. Moreover, even with these new scale factors, $T^{-2}\sum x_t x_t'$ converges to a random matrix rather than a constant, and $T^{-1}\sum x_t \eta_t$ converges to a non-normal random vector.

However, even this argument is too simple, since the standard approach can be applied to a specific linear combination of the regressors. To see this, rearrange the regressors in (2.1) so that

$$y_t = \gamma_1 \Delta y_{t-1} + \gamma_2 y_{t-1} + \eta_t, \tag{2.2}$$

where $\gamma_1 = -\phi_2$ and $\gamma_2 = \phi_1 + \phi_2$. Regression (2.2) is equivalent to regression (2.1) in the sense that the OLS estimates of ϕ_1 and ϕ_2 are linear transformations of the OLS estimators of γ_1 and γ_2. In terms of the transformed regressors

$$\begin{bmatrix} \hat{\gamma}_1 - \gamma_1 \\ \hat{\gamma}_2 - \gamma_2 \end{bmatrix} = \begin{bmatrix} \sum \Delta y_{t-1}^2 & \sum \Delta y_{t-1} y_{t-1} \\ \sum y_{t-1} \Delta y_{t-1} & \sum y_{t-1}^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum \Delta y_{t-1} \eta_t \\ \sum y_{t-1} \eta_t \end{bmatrix} \tag{2.3}$$

and the asymptotic behavior of $\hat{\gamma}_1$ and $\hat{\gamma}_2$ (and hence $\hat{\phi}$) can be analyzed by studying the large sample behavior of the cross products $\sum \Delta y_{t-1}^2, \sum \Delta y_{t-1} y_{t-1}, \sum y_{t-1}^2$, $\sum \Delta y_{t-1} \eta_t$ and $\sum y_{t-1} \eta_t$.

To begin, consider the terms $\sum \Delta y_{t-1}^2$ and $\sum \Delta y_{t-1} \eta_t$. Since $\phi_1 + \phi_2 = \gamma_2 = 1$,

$$\Delta y_t = -\phi_2 \Delta y_{t-1} + \eta_t. \tag{2.4}$$

[3] Many of the insights developed by analyzing this example are discussed in Fuller (1976) and Sims (1978).

Since $|\phi_2| < 1, \Delta y_t$ (and hence Δy_{t-1}) is covariance stationary with mean zero. Thus, standard asymptotic arguments imply that $T^{-1}\sum \Delta y_{t-1}^2 \xrightarrow{p} \sigma_{\Delta y}^2$ and $T^{-1/2}\sum \Delta y_{t-1}\eta_t \xrightarrow{\mathscr{L}} N(0, \sigma_{\Delta y}^2)$. This means that the first regressor in (2.2) behaves in the usual way. "Unit root" complications arise only because of the second regressor, y_{t-1}. To analyze the behavior of this regressor, solve (2.4) backwards for the level of y_t:

$$y_t = (1+\phi_2)^{-1}\xi_t + y_0 + s_t, \tag{2.5}$$

where $\xi_t = \sum_{s=1}^{t}\eta_s$ and $s_t = -(1+\phi_2)^{-1}\sum_{i=0}^{t-1}(-\phi_2)^{i+1}\eta_{t-i}$, and $\eta_i = 0$ for $i \leqslant 0$ has been assumed for simplicity. Equation (2.5) is the Beveridge–Nelson (1981) decomposition of y_t. It decomposes y_t into the sum of a martingale or "stochastic trend" $[(1+\phi_2)^{-1}\xi_t]$, a constant (y_0) and an I(0) component (s_t). The martingale component has a variance that grows with t, and (as is shown below) it is this component that leads to the nonstandard behavior of the cross products $\sum y_{t-1}^2$, $\sum y_{t-1}\Delta y_{t-1}$ and $\sum y_{t-1}\eta_t$.

Other types of trending regressors also arise naturally in time series models and their presence affects the sampling distribution of coefficient estimators. For example, suppose that the AR(2) model includes a constant, so that

$$y_t = \alpha + \gamma_1 \Delta y_{t-1} + \gamma_2 y_{t-1} + \eta_t. \tag{2.6}$$

This constant introduces two additional complications. First, a column of 1's must be added to the list of regressors. Second, solving for the level of y_t as above:

$$y_t = (1+\phi_2)^{-1}\alpha t + (1+\phi_2)^{-1}\xi_t + y_0 + s_t. \tag{2.7}$$

The key difference between (2.5) and (2.7) is that now y_t contains the linear trend $(1+\phi_2)^{-1}\alpha t$. This means that terms involving y_{t-1} now contain cross products that involve linear time trends. Estimators of the coefficients in equation (2.6) can be studied systematically by investigating the behavior of cross products of (i) zero mean stationary components (like η_t and Δy_{t-1}), (ii) constant terms, (iii) martingales and (iv) time trends. We digress to present a useful lemma that shows the limiting behavior of these cross products. This lemma is the key to deriving the asymptotic distribution for coefficient estimators and test statistics for linear regressions involving I(0) and I(1) variables, for tests for cointegration and for estimators of cointegrating vectors. While the AR(2) example involves a scalar process, most of the models considered in this survey are multivariate, and so the lemma is stated for vector processes.

2.3. *A useful lemma*

Three key results are used in the lemma. The first is the functional central limit theorem. Letting η_t denote an $n \times 1$ martingale difference sequence, this theorem

expresses the limiting behavior of the sequence of partial sums $\xi_t = \sum_{s=1}^{t} \eta_s$, $t = 1, \ldots, T$, in terms of the behavior of an $n \times 1$ standardized Wiener or Brownian motion process $B(s)$ for $0 \leqslant s \leqslant 1$.[4] That is, the limiting behavior of the discrete time random walk ξ_t is expressed in terms of the continuous time random walk $B(s)$. The result implies, for example, that $T^{-1/2}\xi_{[sT]} \Rightarrow B(s) \sim N(0, s)$, for $0 \leqslant s \leqslant 1$, where $[sT]$ denotes the first integer less than or equal to sT. The second result used in the lemma is the continuous mapping theorem. Loosely, this theorem says that the limit of a continuous function is equal to the function evaluated at the limit of its arguments. The nonstochastic version of this theorem implies that $T^{-2}\sum_{t=1}^{T} t = T^{-1}\sum_{t=1}^{T}(t/T) \to \int_0^1 s\,ds = \frac{1}{2}$. The stochastic version implies that $T^{-3/2}\sum_{t=1}^{T}\xi_t = T^{-1}\sum_{t=1}^{T}(T^{-1/2}\xi_t) \Rightarrow \int_0^1 B(s)\,ds$. The final result is the convergence of $T^{-1}\sum y_{t-1}\eta_t'$ to the stochastic integral $\int_0^1 B(s)\,dB(s)'$, which is one of the moments directly under study. These key results are discussed in Wooldridge's chapter of the Handbook. For our purposes they are important because they lead to the following lemma.

Lemma 2.3

Let η_t be an $n \times 1$ vector of random variables with $E(\eta_t | \eta_{t-1}, \ldots, \eta_1) = 0$, $E(\eta_t \eta_t' | \eta_{t-1}, \ldots, \eta_1) = I_n$, and bounded fourth moments. Let $F(L) = \sum_{i=0}^{\infty} F_i L^i$ and $G(L) = \sum_{i=0}^{\infty} G_i L^i$ denote two matrix polynomials in the lag operator with $\sum_{i=0}^{\infty} i|F_i| < \infty$ and $\sum_{i=0}^{\infty} i|G_i| < \infty$. Let $\xi_t = \sum_{s=1}^{t}\eta_s$, and let $B(s)$ denote an $n \times 1$ dimensional Brownian motion process. Then the following converge jointly:

(a) $T^{-1/2}\sum F(L)\eta_t \Rightarrow F(1)\int B(s)\,ds$,
(b) $T^{-1}\sum \xi_t \eta_{t+1}' \Rightarrow \int B(s)\,dB(s)'$,
(c) $T^{-1}\sum \xi_t [F(L)\eta_t]' \Rightarrow F(1)' + \int B(s)\,dB(s)'F(1)'$,
(d) $T^{-1}\sum [F(L)\eta_t][G(L)\eta_t]' \xrightarrow{p} \sum_{i=1}^{\infty} F_i G_i'$,
(e) $T^{-3/2}\sum t[F(L)\eta_{t+1}]' \Rightarrow \int s\,dB(s)'F(1)'$,
(f) $T^{-3/2}\sum \xi_t \Rightarrow \int B(s)\,ds$,
(g) $T^{-2}\sum \xi_t \xi_t' \Rightarrow \int B(s)B(s)'\,ds$,
(h) $T^{-5/2}\sum t\xi_t \Rightarrow \int sB(s)\,ds$,

where, to simplify notation $\int_0^1$ is denoted by $\int$. The lemma follows from results in Chan and Wei (1988) together with standard versions of the law of large numbers and the central limit theorem for martingale difference sequences [see White (1984)]. Many versions of this lemma (often under assumptions slightly different from those stated here) have appeared in the literature. For example, univariate versions can be found in Phillips (1986, 1987a), Phillips and Perron (1988) and Solo (1984), while multivariate versions (in most cases covering higher order integrated processes) can be found in Park and Phillips (1988, 1989), Phillips and

[4] Throughout this paper $B(s)$ will denote a multivariate standard Brownian motion process, i.e., an $n \times 1$ process with independent increments $B(r) - B(s)$ that are distributed $N(0, (r-s)I_n)$ for $r > s$.

Durlauf (1986), Phillips and Solo (1992), Sims et al. (1990) and Tsay and Tiao (1990).

The specific regressions that are studied below fall into two categories: (i) regressions that include a constant and a martingale as regressors or, (ii) regressions that include a constant, a time trend and a martingale as regressors. In either case, the coefficient on the martingale is the parameter of interest. The estimated value of this coefficient can be calculated by including a constant or a constant and time trend in the regression, or, alternatively, by first demeaning or detrending the data. It is convenient to introduce some notation for the demeaned and detrended martingales and their limiting Brownian motion representations. Thus, let $\xi_t^{\mu} = \xi_t - T^{-1}\sum_{s=1}^{T}\xi_s$ denote the demeaned martingale, and let $\xi_t^{\tau} = \xi_t - \hat{\beta}_1 - \hat{\beta}_2 t$ denote the detrended martingale, where $\hat{\beta}_1$ and $\hat{\beta}_2$ are the OLS estimators obtained from the regression of ξ_t onto $(1\ t)$. Then, from the lemma, a straightforward calculation yields

$$T^{-1/2}\xi^{\mu}_{[sT]} \Rightarrow B(s) - \int_0^1 B(r)\,dr \equiv B^{\mu}(s)$$

and

$$T^{-1/2}\xi^{\tau}_{[sT]} \Rightarrow B(s) - \int_0^1 a_1(r)B(r)\,dr - s\int_0^1 a_2(r)B(r)\,dr \equiv \beta^{\tau}(s),$$

where $a_1(r) = 4 - 6r$ and $a_2(r) = -6 + 12r$.

2.4. *Continuing with the example*

We are now in a position to complete the analysis of the AR(2) example. Consider a scaled version of (2.3),

$$\begin{bmatrix} T^{1/2}(\hat{\gamma}_1 - \gamma_1) \\ T(\hat{\gamma}_2 - \gamma_2) \end{bmatrix} = \begin{bmatrix} T^{-1}\sum \Delta y_{t-1}^2 & T^{-3/2}\sum \Delta y_{t-1} y_{t-1} \\ T^{-3/2}\sum y_{t-1}\Delta y_{t-1} & T^{-2}\sum y_{t-1}^2 \end{bmatrix}^{-1} \times \begin{bmatrix} T^{-1/2}\sum \Delta y_{t-1}\eta_t \\ T^{-1}\sum y_{t-1}\eta_t \end{bmatrix}.$$

From (2.5) and result (g) of the lemma, $T^{-2}\sum y_{t-1}^2 \Rightarrow (1+\phi_2)^{-2}\int B(s)^2\,ds$ and from (b) $T^{-1}\sum y_{t-1}\eta_t \Rightarrow (1+\phi_2)^{-1}\int B(s)\,dB(s)$. Finally, noting from (2.4) that $\Delta y_t = (1+\phi_2 L)^{-1}\eta_t$, (c) implies that $T^{-3/2}\sum \Delta y_{t-1} y_{t-1} \xrightarrow{p} 0$. This result is particularly important because it implies that the limiting scaled "$X'X$" matrix for the regression is block diagonal. Thus,

$$T^{1/2}(\hat{\gamma}_1 - \gamma_1) = (T^{-1}\sum \Delta y_{t-1}^2)^{-1} T^{-1/2}\sum \Delta y_{t-1}\eta_t + o_p(1) \xrightarrow{\mathcal{L}} N(0, \sigma_{\Delta y}^{-2}),$$

and

$$T(\hat{\gamma}_2 - \gamma_2) = (T^{-2}\sum y_{t-1}^2)^{-1}(T^{-1}\sum y_{t-1}\eta_t) + o_p(1)$$
$$\Rightarrow (1+\phi_2)\left[\int B(s)^2\,ds\right]^{-1}\left[\int B(s)\,dB(s)\right].$$

Two features of these results are important. First, $\hat{\gamma}_1$ and $\hat{\gamma}_2$ converge at different rates. These rates are determined by the variability of their respective regressors: γ_1 is the coefficient on a regressor with bounded variance, while γ_2 is the coefficient on a regressor with a variance that increases at rate t. The second important feature is that $\hat{\gamma}_1$ has an asymptotic normal distribution, while the asymptotic distribution of $\hat{\gamma}_2$ is non-normal. Unit root complications will affect statistical inference about γ_2 but not γ_1.

Now consider the estimated regression coefficients $\hat{\phi}_1$ and $\hat{\phi}_2$ in the untransformed regression. Since $\hat{\phi}_2 = -\hat{\gamma}_1$, $T^{1/2}(\hat{\phi}_2 - \phi_2) \xrightarrow{p} N(0, \sigma_{\Delta y}^{-2})$. Furthermore, since $\hat{\phi}_1 = \hat{\gamma}_1 + \hat{\gamma}_2$, $T^{1/2}(\hat{\phi}_1 - \phi_1) = T^{1/2}(\hat{\gamma}_1 - \gamma_1) + T^{1/2}(\hat{\gamma}_2 - \gamma_2) = T^{1/2}(\hat{\gamma}_1 - \gamma_1) + o_p(1)$. That is, even though $\hat{\phi}_1$ depends on both $\hat{\gamma}_1$ and $\hat{\gamma}_2$, the "super consistency" of $\hat{\gamma}_2$ implies that its sampling error can be ignored in large samples. Thus, $T^{1/2}(\hat{\phi}_1 - \phi_1) \xrightarrow{\mathcal{L}} N(0, \sigma_{\Delta y}^{-2})$, so that both $\hat{\phi}_1$ and $\hat{\phi}_2$ converge at rate $T^{1/2}$ and have asymptotic normal distributions. Their joint distribution is more complicated. Since $\phi_1 + \phi_2 = \gamma_2$, $T^{1/2}(\hat{\phi}_1 - \phi_1) + T^{1/2}(\hat{\phi}_2 - \phi_2) = T^{1/2}(\hat{\gamma}_2 - \gamma_2) \xrightarrow{p} 0$ and the joint asymptotic distribution of $T^{1/2}(\hat{\phi}_1 - \phi_1)$ and $T^{1/2}(\hat{\phi}_2 - \phi_2)$ is singular. The linear combination $\hat{\phi}_1 + \hat{\phi}_2$ converges at rate T to a non-normal distribution: $T[(\hat{\phi}_1 + \hat{\phi}_2) - (\phi_1 + \phi_2)] = T(\hat{\gamma}_2 - \gamma_2) \Rightarrow (1+\phi_2)[\int B(s)^2\,ds]^{-1}[\int B(s)\,dB(s)]$.

There are two important practical consequences of these results. First, inference about ϕ_1 or about ϕ_2 can be conducted in the usual way. Second, inference about the sum of coefficients $\phi_1 + \phi_2$ must be carried out using nonstandard asymptotic distributions. Under the null hypothesis, the t-statistic for testing the null H_o: $\phi_1 = c$ converges to a standard normal random variable, while the t-statistic for testing the null hypothesis H_o: $\phi_1 + \phi_2 = 1$ converges to $[\int B(s)^2\,ds]^{-1/2}[\int B(s)\,dB(s)]$, which is the distribution of the Dickey–Fuller τ statistic (see Stock's chapter of the Handbook).

As we will see, many of the results developed for the AR(2) carry over to more general settings. First, estimates of linear combinations of regression coefficients converge at different rates. Estimators that correspond to coefficients on stationary regressors, or that can be written as coefficients on stationary regressors in a transformed regression (γ_1 in this example), converge at rate $T^{1/2}$ and have the usual asymptotic normal distribution. Estimators that correspond to coefficients on I(1) regressors, and that cannot be written as coefficients on I(0) regressors in a transformed regression (γ_2 in this example), converge at rate T and have a nonstandard asymptotic distribution. The asymptotic distribution of test statistics is also affected by these results. Wald statistics for restrictions on coefficients corresponding to I(0) regressors have the usual asymptotic normal or χ^2 distributions. In

general, Wald statistics for restrictions on coefficients that cannot be written as coefficients on I(0) regressors have nonstandard limiting distributions. We now demonstrate these results for the general VAR model with I(1) variables.

2.5. *A general framework*

Consider the VAR model

$$Y_t = \alpha + \sum_{i=1}^{p} \Phi_i Y_{t-i} + \varepsilon_t, \tag{2.8}$$

where Y_t is an $n \times 1$ vector and ε_t is a martingale difference sequence with constant conditional variance Σ_ε (abbreviated mds(Σ_ε)) with finite fourth moments. Assume that the determinant of the autoregressive polynomial $|I - \Phi_1 z - \Phi_2 z^2 - \cdots - \Phi_p z^p|$ has all of its roots outside the unit circle or at $z = 1$, and continue to maintain the simplifying assumption that all elements of Y_t are individually I(0) or I(1).[5] For simplicity, assume that there are no cross equation restrictions, so that the efficient linear estimators correspond to the equation-by-equation OLS estimators. We now study the distribution of these estimators and commonly used test statistics.[6]

2.5.1. *Distribution of estimated regression coefficients*

To begin, write the ith equation of the model as

$$y_{i,t} = X_t'\beta + \varepsilon_{i,t}, \tag{2.9}$$

where $y_{i,t}$ is the ith element of Y_t, $X_t = (1\ Y_{t-1}'\ Y_{t-2}' \cdots Y_{t-p}')'$ is the $(np+1)$ vector of regressors, β is the corresponding vector of regression coefficients, and $\varepsilon_{i,t}$ is the ith element of ε_t. (For notational convenience the dependence of β on i has been suppressed.) The OLS estimator of β is $\hat{\beta} = (\sum X_t X_t')^{-1}(\sum X_t y_{i,t})$, so that $\hat{\beta} - \beta = (\sum X_t X_t')^{-1}(\sum X_t \varepsilon_{i,t})$.

As in the univariate AR(2) model, the asymptotic behavior of $\hat{\beta}$ is facilitated by

[5] Higher order integrated processes can also be studied using the techniques discussed here, see Park and Phillips (1988) and Sims et al. (1990). Seasonal unit roots (corresponding to zeroes elsewhere on the unit circle) can also be studied using a modification of these procedures. See Tsay and Tiao (1990) for a careful analysis of this case.

[6] The analysis in this section is based on a large body of work on estimation and inference in multivariate time series models with unit roots. A partial list of relevant references includes Chan and Wei (1988), Park and Phillips (1988, 1989), Phillips (1988), Phillips and Durlauf (1986), Sims et al. (1990), Stock (1987), Tsay and Tiao (1990), and West (1988). Additional references are provided in the body of the text.

transforming the regressors in a way that isolates the various stochastic and deterministic trends. In particular, the regressors are transformed as $Z_t = DX_t$, where D is nonsingular and $Z_t = (z_{1,t} z_{2,t} \cdots z_{4,t})'$, where the $z_{i,t}$ will be referred to as "canonical" regressors. These regressors are related to the deterministic and stochastic trends given in Lemma 2.3 by the transformation

$$\begin{bmatrix} z_{1,t} \\ z_{2,t} \\ z_{3,t} \\ z_{4,t} \end{bmatrix} = \begin{bmatrix} F_{11}(L) & 0 & 0 & 0 \\ 0 & F_{22} & 0 & 0 \\ F_{31}(L) & F_{32} & F_{33} & 0 \\ F_{41}(L) & F_{42} & F_{43} & F_{44} \end{bmatrix} \begin{bmatrix} \eta_{t-1} \\ 1 \\ \xi_{t-1} \\ t \end{bmatrix}$$

or

$$Z_t = F(L)v_{t-1},$$

where $v_t = (\eta_t' \, 1 \, \xi_t' \, t)'$. The advantage of this transformation is that it isolates the terms of different orders of probability. For example, $z_{1,t}$ is a zero mean I(0) regressor, $z_{2,t}$ is a constant, the asymptotic behavior of the regressor $z_{3,t}$ is dominated by the martingale component $F_{33}\xi_{t-1}$, and $z_{4,t}$ is dominated by the time trend $F_{44}t$. The canonical regressors $z_{2,t}$ and $z_{4,t}$ are scalars, while $z_{1,t}$ and $z_{3,t}$ are vectors. In the AR(2) example analyzed above, $z_{1,t} = \Delta y_{t-1} = (1 + \phi_2 L)^{-1}\eta_{t-1}$, so that $F_{11}(L) = (1 + \phi_2 L)^{-1}$; $z_{2,t}$ is absent, since the model did not contain a constant; $z_{3,t} = y_{t-1} = (1 + \phi_2)^{-1}\xi_{t-1} + y_0 + s_{t-1}$, so that $F_{33} = (1 + \phi_2)^{-1}$, $F_{32} = y_0$ and $F_{31}(L) = \phi_2(1 + \phi_2)^{-1}(1 + \phi_2 L)^{-1}$; and $z_{4,t}$ is absent since y_t contains no deterministic drift.

Sims et al. (1990) provide a general procedure for transforming regressors from an integrated VAR into canonical form. They show that Z_t can always be formed so that the diagonal blocks, $F_{ii}, i > 2$ have full row rank, although some blocks may be absent. They also show that $F_{12} = 0$, as shown above, whenever the VAR includes a constant. The details of their construction need not concern us since, in practice, there is no need to construct the canonical regressors. The transformation from the X_t to the Z_t regressors is merely an analytic device. It is useful for two reasons. First, $X_t'D'(D')^{-1}\beta = Z_t'\gamma$, with $\gamma = (D')^{-1}\beta$. Thus the OLS estimators of the original and transformed models are related by $D'\hat{\gamma} = \hat{\beta}$. Second, the asymptotic properties of $\hat{\gamma}$ are easy to analyze because of the special structure of the regressors. Together these imply that we can study the asymptotic properties of $\hat{\beta}$ by first studying the asymptotic properties of $\hat{\gamma}$ and then transforming these coefficients into the $\hat{\beta}$'s.

The transformation from X_t to Z_t is not unique. All that is required is *some* transformation that yields a lower triangular $F(L)$ matrix. Thus, in the AR(2) example we set $z_{1,t} = \Delta y_{t-1}$ and $z_{3,t} = y_{t-1}$, but an alternative transformation would have set $z_{1,t} = \Delta y_{t-1}$ and $z_{3,t} = y_{t-2}$. Since we always transform results for

the canonical regressors Z_t back into results for the "natural" regressors X_t, this non-uniqueness is of no consequence.

We now derive the asymptotic properties of $\hat{\gamma}$ constructed from the regression $y_{i,t} = Z_t'\gamma + \varepsilon_{i,t}$. Writing $\varepsilon_t = \Sigma_\varepsilon^{1/2}\eta_t$, where η_t is the standardized $n \times 1$ martingale difference sequence from Lemma 2.3, then $\varepsilon_{i,t} = \omega'\eta_t = \eta_t'\omega$, where ω' is the ith row of $\Sigma_\varepsilon^{1/2}$, and $\hat{\gamma} - \gamma = (\sum Z_t Z_t')^{-1}(\sum Z_t \eta_t'\omega)$. Lemma 2.3 can be used to deduce the asymptotic behavior of $\sum Z_t Z_t'$ and $\sum Z_t \eta_t'\omega$. Some care must be taken, however, since all of the $z_{j,t}$ elements of Z_t are growing at different rates. Assume that $z_{1,t}$ contains k_1 elements, $z_{3,t}$ contains k_3 elements, and partition γ conformably with Z_t as $\gamma = (\gamma_1\ \gamma_2\ \gamma_3\ \gamma_4)'$, where γ_j are the regression coefficients corresponding to $z_{j,t}$. Let

$$\Psi_T = \begin{bmatrix} T^{1/2}I_{k_1} & 0 & 0 & 0 \\ 0 & T^{1/2} & 0 & 0 \\ 0 & 0 & TI_{k_3} & 0 \\ 0 & 0 & 0 & T^{3/2} \end{bmatrix}$$

and consider $\Psi_T(\hat{\gamma} - \gamma) = (\Psi_T^{-1}\sum Z_t Z_t' \Psi_T^{-1})^{-1}(\Psi_T^{-1}\sum Z_t \eta_t'\omega)$. The matrix Ψ_T multiplies the various blocks of $(\hat{\gamma}_i - \gamma_i)$, $\sum Z_t Z_t'$, and $\sum Z_t \eta_t$ by the scaling factors appropriate from the lemma. The first block of coefficients, γ_1, are coefficients on zero mean stationary components and are scaled up by the usual factor of $T^{1/2}$; the same scaling factor is appropriate for γ_2, the constant term; the parameters making up γ_3 are coefficients on regressors that are dominated by martingales, and these need to be scaled by T; finally, γ_4 is a coefficient on a regressor that is dominated by a time trend and is scaled by $T^{3/2}$.

Applying the lemma, we have $\Psi_T^{-1}\sum Z_t Z_t' \Psi_T^{-1} \Rightarrow V$, where, partitioning V conformably with Z_t:

$$\begin{aligned}
T^{-1}\textstyle\sum z_{1,t}z_{1,t}' &\xrightarrow{p} \sum_j F_{11,j}F_{11,j}' && = V_{11}, \\
T^{-1}\textstyle\sum (z_{2,t})^2 &\to F_{22}^2 && = V_{22}, \\
T^{-2}\textstyle\sum z_{3,t}z_{3,t}' &\Rightarrow F_{33}\left[\int B(s)B(s)'\,ds\right]F_{33}' && = V_{33}, \\
T^{-3}\textstyle\sum (z_{4,t})^2 &\xrightarrow{p} \frac{F_{44}^2}{3} && = V_{44}, \\
T^{-j/2}\textstyle\sum z_{1,t}z_{j,t}' &\xrightarrow{p} 0 && = V_{1j} = V_{j1}' \quad \text{for} \quad j = 2,3,4, \\
T^{-3/2}\textstyle\sum z_{2,t}z_{3,t}' &\Rightarrow F_{22}\int B(s)'\,ds F_{33}' && = V_{23} = V_{32}', \\
T^{-2}\textstyle\sum z_{2,t}z_{4,t} &\xrightarrow{p} \frac{F_{22}F_{44}}{2} && = V_{24} = V_{42}, \\
T^{-5/2}\textstyle\sum z_{3,t}z_{4,t} &\Rightarrow F_{33}\int sB(s)\,ds F_{44} && = V_{34} = V_{43}',
\end{aligned}$$

where the notation reflects the fact that F_{22} and F_{44} are scalars. The limiting value of this scaled moment matrix shares two important characteristics with its analogue in the univariate AR(2) model. First, V is block diagonal with $V_{1j}=0$ for $j\neq 1$. (Recall that in the AR(2) model $T^{-3/2}\sum\Delta y_{t-1}y_{t-1}\xrightarrow{p}0$.) Second, many of the blocks of V contain random variables. (In the AR(2) model $T^{-2}\sum y_{t-1}^2$ converged to a random variable.)

Now, applying the lemma to $\Psi_T^{-1}\sum Z_t\eta_t'\omega$ yields $\Psi_T^{-1}\sum Z_t\eta_t'\omega\Rightarrow A$, where, partitioning A conformably with Z_t:

$$
\begin{aligned}
T^{-1/2}\textstyle\sum z_{1,t}\eta_t'\omega &\xrightarrow{\mathcal{L}} N[0,(\omega'\omega)V_{11}] &&= A_1,\\
T^{-1/2}\textstyle\sum z_{2,t}\eta_t'\omega &\Rightarrow F_{22}\int \mathrm{d}B(s)'\omega &&= A_2,\\
T^{-1}\textstyle\sum z_{3,t}\eta_t'\omega &\Rightarrow F_{33}\int B(s)\mathrm{d}B(s)'\omega &&= A_3,\\
T^{-3/2}\textstyle\sum z_{4,t}\eta_t'\omega &\Rightarrow F_{44}\int s\,\mathrm{d}B(s)'\omega &&= A_4.
\end{aligned}
$$

Putting the results together, $\Psi_T(\hat{\gamma}-\gamma)\Rightarrow V^{-1}A$, and three important results follow. First, the individual coefficients converge to their values at different rates: $\hat{\gamma}_1$ and $\hat{\gamma}_2$ converge to their values at rate $T^{1/2}$, while all of the other coefficients converge more quickly. Second, the block diagonality of V implies that $T^{1/2}(\hat{\gamma}_1-\gamma_1)\xrightarrow{\mathcal{L}} N(0,\sigma_i^2V_{11}^{-1})$, where $\sigma_i^2=\omega'\omega=\mathrm{var}(\varepsilon_t^i)$. Moreover, A_1 is independent of A_j for $j>1$ [Chan and Wei (1988, Theorem 2.2)], so that $T^{1/2}(\hat{\gamma}_1-\gamma_1)$ is asymptotically independent of the other estimated coefficients. Third, all of the other coefficients will have non-normal limiting distributions, in general. This follows because $V_{j3}\neq 0$ for $j>1$, and A_3 is non-normal. A notable exception to this general result is when the canonical regressors do not contain any stochastic trends, so that $z_{3,t}$ is absent from the model. In this case V is a constant and A is normally distributed, so that the estimated coefficients have a joint asymptotic normal distribution.[7] The leading example of this is polynomial regression, when the set of regressors contains covariance stationary regressors and polynomials in time. Another important example is given by West (1988), who considers the scalar unit root AR(1) model with drift.

The asymptotic distribution of the coefficients β that correspond to the "natural" regressors X_t can now be deduced. It is useful to begin with a special case of the general model,

$$y_{i,t}=\beta_1+x_{2,t}'\beta_2+x_{3,t}'\beta_3+\varepsilon_{i,t}, \tag{2.10}$$

[7] A_1, A_2, and A_4 are jointly normally distributed since $\int s^k\mathrm{d}B(s)'\omega$ is a normally distributed random variable with mean 0 and variance $(\omega'\omega)\int s^{2k}\mathrm{d}s$.

where $x_{1,t} = 1$ for all $t, x_{2,t}$ is an $h \times 1$ vector of zero mean I(0) variables and $x_{3,t}$ contains the other regressors. It is particularly easy to transform this model into canonical form. First, since $x_{1,t} = 1$, we can set $z_{2,t} = x_{1,t}$; thus, in terms of the transformed regression, $\beta_1 = \gamma_2$. Second, since the elements of $x_{2,t}$ are zero mean I(0) variables, we can set the first h elements of $z_{1,t}$ equal to $x_{2,t}$; thus β_2 is equal to the first h elements of γ_1. The remaining elements of z_t are linear combination of the regressors that need not concern us here. In this example, since β_2 is a subset of the elements of γ_1, $T^{1/2}(\hat{\beta}_2 - \beta_2)$ is asymptotically normal and independent of the coefficients corresponding to trend and unit root regressors. This result is very useful because it provides a constructive sufficient condition for estimated coefficients to have an asymptotic normal limiting distribution: whenever the block of coefficients can be written as coefficients on zero mean I(0) regressors in a model that includes a constant term they will have a joint asymptotic normal distribution.

Now consider the general model. Recall that $\hat{\beta} = D'\hat{\gamma}$. Let d_j denote the jth column of D, and partition this conformably with γ, so that $d_j = (d'_{1j} d'_{2j} d'_{3j} d'_{4j})'$, where d_{ij} and $\hat{\gamma}_i$ are the same dimension. Then the jth element of $\hat{\beta}$ is $\hat{\beta}_j = \sum_i d'_{ij}\hat{\gamma}_i$. Since the components of $\hat{\gamma}$ converge at different rates, $\hat{\beta}_j$ will converge at the slowest rate of the $\hat{\gamma}_i$ included in the sum. Thus, when $d_{1j} \neq 0, \hat{\beta}_j$ will converge at rate $T^{1/2}$, the rate of convergence of $\hat{\gamma}_1$.

2.5.2. *Distribution of Wald test statistics*

Consider Wald test statistics for linear hypotheses of the form $R\beta = r$, where R is a $q \times k$ matrix with full row rank,

$$W = \frac{(R\hat{\beta} - r)'[R(\sum X_t X_t')^{-1} R']^{-1}(R\hat{\beta} - r)}{\hat{\sigma}_i^2}.$$

(Recall that β corresponds to the coefficients in the ith equation, so that W tests within-equation restrictions.) Letting $Q = R(D')$, an equivalent way of writing the Wald statistic is in terms of the canonical regressors Z_t and their estimated coefficients $\hat{\gamma}$,

$$W = \frac{(Q\hat{\gamma} - r)'[Q(\sum Z_t Z_t')^{-1} Q']^{-1}(Q\hat{\gamma} - r)}{\hat{\sigma}_i^2}.$$

Care must be taken when analyzing the large sample behavior of W because the individual coefficients in $\hat{\gamma}$ converge at different rates. To isolate the different components, it is useful to assume (without loss of generality) that Q is upper triangular.[8]

[8] This assumption is made without loss of generality since the constraint $Q\gamma = r$ (and the resulting Wald statistic) is equivalent to $CQ\gamma = Cr$, for nonsingular C. For any matrix Q, C can chosen so that CQ is upper triangular.

Now, partition Q, conformably with $\hat{\gamma}$ and the canonical regressors making up Z_t, so that $Q = [q_{ij}]$ where q_{ij} is a $q_i \times k_j$ matrix representing q_i constraints on the k_j elements in γ_j. These blocks are chosen so that q_{ii} has full row rank and $q_{ij} = 0$ for $i < j$. Since the set of constraints $Q\gamma = r$ may not involve γ_i, the blocks q_{ij} might be absent for some i. Thus, for example, when the hypothesis concerns only γ_3, then Q is written as $Q = [q_{31} q_{32} q_{33} q_{34}]$, where $q_{31} = 0$, $q_{32} = 0$ and q_{33} has full row rank. Partition $r = (r_1' r_2' r_3' r_4')'$ conformably with Q, where again some of the r_i may be absent.

Now consider the first q_1 elements of $Q\hat{\gamma}$: $q_{11}\hat{\gamma}_1 + q_{12}\hat{\gamma}_2 + q_{13}\hat{\gamma}_3 + q_{14}\hat{\gamma}_4$. Since $\hat{\gamma}_j$, for $j > 2$, converges more quickly than $\hat{\gamma}_1$ and $\hat{\gamma}_2$, the sampling error in this vector will be dominated asymptotically by the sampling error in $q_{11}\hat{\gamma}_1 + q_{12}\hat{\gamma}_2$. Similarly, the sampling error in the next group of q_2 elements of $Q\hat{\gamma}$ is dominated by $q_{22}\hat{\gamma}_2$, in the next q_3 by $q_{33}\hat{\gamma}_3$, etc. Thus, the appropriate scaling matrix for $Q\hat{\gamma} - r$ is

$$\tilde{\Psi}_T = \begin{bmatrix} T^{1/2} I_{q_1} & 0 & 0 & 0 \\ 0 & T^{1/2} I_{q_2} & 0 & 0 \\ 0 & 0 & T I_{q_3} & 0 \\ 0 & 0 & 0 & T^{3/2} I_{q_4} \end{bmatrix}.$$

Now, write the Wald statistic as

$$W = \frac{(Q\hat{\gamma} - r)' \tilde{\Psi}_T' [\tilde{\Psi}_T Q (\sum Z_t Z_t')^{-1} Q' \tilde{\Psi}_T]^{-1} \tilde{\Psi}_T (Q\hat{\gamma} - r)}{\hat{\sigma}_i^2}.$$

But, under the null,

$$\begin{aligned} &T^{1/2}(q_{11}\hat{\gamma}_1 + q_{12}\hat{\gamma}_2 + q_{13}\hat{\gamma}_3 + q_{14}\hat{\gamma}_4 - r_1) \\ &\quad = T^{1/2}(q_{11}\hat{\gamma}_1 + q_{12}\hat{\gamma}_2 - r_1) + o_p(1), \quad \text{and} \end{aligned}$$

$$T^{(j-1)/2}(q_{jj}\hat{\gamma}_j + \cdots + q_{j4}\hat{\gamma}_4 - r_j) = T^{(j-1)/2}(q_{jj}\hat{\gamma}_j - r_j) + o_p(1), \quad \text{for} \quad j > 1.$$

Thus, if we let

$$\tilde{Q} = \begin{bmatrix} q_{11} & q_{12} & 0 & 0 \\ 0 & q_{22} & 0 & 0 \\ 0 & 0 & q_{33} & 0 \\ 0 & 0 & 0 & q_{44} \end{bmatrix},$$

then

$$\tilde{\Psi}_T (Q\hat{\gamma} - r) = \tilde{Q} \Psi_T (\hat{\gamma} - \gamma) + o_p(1)$$

under the null.[9] Similarly, it is straightforward to show that

$$\tilde{\Psi}_T Q(\sum Z_t Z_t')^{-1} Q' \tilde{\Psi}_T = \tilde{Q}(\Psi_T^{-1} \sum Z_t Z_t' \Psi_T^{-1})^{-1} \tilde{Q}' + o_p(1).$$

Finally, since $\Psi_T(\hat{\gamma} - \gamma) \Rightarrow V^{-1}A$ and $\Psi_T^{-1} \sum Z_t Z_t' \Psi_T^{-1} \Rightarrow V$, then $W \Rightarrow (\tilde{Q}V^{-1}A)' \times (\tilde{Q}V^{-1}\tilde{Q})^{-1}(\tilde{Q}V^{-1}A)$.

The limiting distribution of W is particularly simple when $q_{ii} = 0$ for $i \geq 2$. In this case, all of the hypotheses of interest concern linear combinations of zero mean I(0) regressors, together with the other regression coefficients. When $q_{12} = 0$, so that the constant term is unrestricted, we have

$$\sigma_i^2 W = [q_{11}(\hat{\gamma}_1 - \gamma_1)]'[q_{11}(\sum z_{1,t} z_{1,t}')^{-1} q_{11}']^{-1}[q_{11}(\hat{\gamma}_1 - \gamma_1)] + o_p(1),$$

so that $W \xrightarrow{\mathscr{L}} \chi^2_{q_1}$. When the constraints involve other linear combinations of the regression coefficients, the asymptotic χ^2 distribution of the regression coefficients will not generally obtain.

This analysis has only considered tests of restrictions on coefficients from the same equation. Results for cross equation restrictions are contained in Sims et al. (1990). The same general results carry over to cross equation restrictions. Namely, restrictions that involve subsets of coefficients, that can be written as coefficients on zero mean stationary regressors in regressions that include constant terms, can be tested using standard asymptotic distribution theory. Otherwise, in general, the statistics will have nonstandard limiting distributions.

2.6. *Applications*

2.6.1. *Testing lag length restrictions*

Consider the VAR$(p+s)$ model,

$$Y_t = \alpha + \sum_{i=1}^{p+s} \Phi_i Y_{t-i} + \varepsilon_t$$

and the null hypothesis H_0: $\Phi_{p+1} = \Phi_{p+2} = \cdots = \Phi_{p+s} = 0$, which says that the true model is a VAR(p). When $p \geq 1$, the usual Wald (and LR and LM) test statistic for H_0 has an asymptotic χ^2 distribution under the null. This can be demonstrated by rewriting the regression so that the restrictions in H_o concern coefficients on zero mean stationary regressors. Assume that ΔY_t is I(0) with mean μ, and then

[9] q_{12} is the only off-diagonal element appearing in $\tilde{Q}$. It appears because $\hat{\gamma}_1$ and $\hat{\gamma}_2$ both converge at rate $T^{1/2}$.

rewrite the model as

$$Y_t = \tilde{\alpha} + \Lambda Y_{t-1} + \sum_{i=1}^{p+s-1} \Theta_i(\Delta Y_{t-i} - \mu) + \varepsilon_t,$$

where $\Lambda = \sum_{i=1}^{p+s} \Phi_i$, $\Theta_i = -\sum_{j=i+1}^{p+s} \Phi_j$ and $\tilde{\alpha} = \alpha + \sum_{i=1}^{p+s-1} \Theta_i \mu$. The restrictions $\Phi_{p+1} = \Phi_{p+2} = \cdots = \Phi_{p+s} = 0$, in the original model are equivalent to $\Theta_p = \Theta_{p+1} = \cdots = \Theta_{p+s-1}$ in the transformed model. Since these coefficients are zero mean I(0) regressors in regression equations that contain a constant term, the test statistics will have the usual large sample χ^2 distribution.

2.6.2. *Testing for Granger causality*

Consider the bivariate VAR model

$$y_{1,t} = \alpha_1 + \sum_{i=1}^{p} \phi_{11,i} y_{1,t-i} + \sum_{i=1}^{p} \phi_{12,i} y_{2,t-i} + \varepsilon_{1,t},$$

$$y_{2,t} = \alpha_2 + \sum_{i=1}^{p} \phi_{21,i} y_{1,t-i} + \sum_{i=1}^{p} \phi_{22,i} y_{2,t-i} + \varepsilon_{2,t}.$$

The restriction that $y_{2,t}$ does not Granger-cause $y_{1,t}$ corresponds to the null hypothesis $H_o: \phi_{12,1} = \phi_{12,2} = \cdots = \phi_{12,p} = 0$. When $(y_{1,t}\, y_{2,t})$ are covariance stationary, the resulting Wald, LR or LM test statistic for this hypothesis will have a large sample χ_p^2 distribution. When $(y_{1,t}\, y_{2,t})$ are integrated, the distribution of the test statistic depends on the location of unit roots in the system. For example, suppose that $y_{1,t}$ is I(1), but that $y_{2,t}$ is I(0). Then, by writing the model in terms of deviations of $y_{2,t}$ from its mean, the restrictions involve only coefficients on zero mean I(0) regressors. Consequently, the test statistic has a limiting χ_p^2 distribution.

When $y_{2,t}$ is I(1), then the distribution of the statistic will be asymptotically χ^2 when $y_{1,t}$ and $y_{2,t}$ are cointegrated. When $y_{1,t}$ and $y_{2,t}$ are not cointegrated, the Granger-causality test statistic will not be asymptotically χ^2, in general. Again, the first result is easily demonstrated by writing the model so the coefficients of interest appear as coefficients on zero mean stationary regressors. In particular, when $y_{1,t}$ and $y_{2,t}$ are cointegrated, there is an I(0) linear combination of the variables, say $w_t = y_{2,t} - \lambda y_{1,t}$, and the model can be rewritten as

$$y_{1,t} = \tilde{\alpha}_1 + \sum_{i=1}^{p} \tilde{\phi}_{11,i} y_{1,t-i} + \sum_{i=1}^{p} \phi_{12,i}(w_{t-i} - \mu_w) + \varepsilon_{1,t},$$

where μ_w is the mean of w_t, $\tilde{\alpha}_1 = \alpha + \sum_{i=1}^{p} \phi_{12,i}\mu_w$ and $\tilde{\phi}_{11,i} = \phi_{11,i} + \phi_{12,i}\lambda$, $i = 1, \ldots, p$. In the transformed regression, the Granger-causality restriction corresponds to the restriction that the terms $w_{t-i} - \mu_w$ do not enter the regression. But

these are zero mean I(0) regressors in a regression that includes a constant, so that the resulting test statistics will have a limiting χ_p^2 distribution. When $y_{1,t}$ and $y_{2,t}$ are not cointegrated, the regression cannot be transformed in this way, and the resulting test statistic will not, in general, have a limiting χ^2 distribution.[10]

The Mankiw–Shapiro (1985)/Stock–West (1988) results concerning Hall's test of the life-cycle/permanent income model can now be explained quite simply. Mankiw and Shapiro considered tests of Hall's model based on the regression of Δc_t (the logarithm of consumption) onto y_{t-1} (the lagged value of the logarithm of income). Since y_{t-1} is (arguably) integrated, its regression coefficient and t-statistic will have a nonstandard limiting distribution. Stock and West, following Hall's (1978) original regressions, considered regressions of c_t onto c_{t-1} and y_{t-1}. Since, according to the life-cycle/permanent income model, c_{t-1} and y_{t-1} are cointegrated, the coefficient on y_{t-1} will be asymptotically normal and its t-statistic will have a limiting standard normal distribution. However, when y_{t-1} is replaced in the regression with m_{t-1} (the lagged value of the logarithm of money), the statistic will not be asymptotically normal, since c_{t-1} and m_{t-1} are not cointegrated. A more detailed discussion of this example is contained in Stock and West (1988).

2.6.3. *Spurious regressions*

In a very influential paper in the 1970's, Granger and Newbold (1974) presented Monte Carlo evidence reminding economists of Yule's (1926) spurious correlation results. Specifically, Granger and Newbold showed that a large R^2 and a large t-statistic were not unusual when one random walk was regressed on another, statistically independent, random walk. Their results warned researchers that standard measures of fit can be very misleading in "spurious" regressions. Phillips (1986) showed how these results could be interpreted quite simply using the framework outlined above, and his analysis is summarized here.

Let $y_{1,t}$ and $y_{2,t}$ be two independent random walks

$$y_{1,t} = y_{1,t-1} + \varepsilon_{1,t},$$
$$y_{2,t} = y_{2,t-1} + \varepsilon_{2,t},$$

where $\varepsilon_t = (\varepsilon_{1,t}\, \varepsilon_{2,t})'$ is an mds(Σ_ε) with finite fourth moments, and $\{\varepsilon_{1,t}\}_{t=1}^T$ and $\{\varepsilon_{2,t}\}_{t=1}^T$ are mutually independent. For simplicity, set $y_{1,0} = y_{2,0} = 0$. Consider the linear regression of $y_{2,t}$ onto $y_{1,t}$,

$$y_{2,t} = \beta y_{1,t} + u_t, \tag{2.11}$$

where u_t is the regression error. Since $y_{1,t}$ and $y_{2,t}$ are statistically independent $\beta = 0$ and $u_t = y_{2,t}$.

[10] A detailed discussion of Granger-causality tests in integrated systems in contained in Sims et al. (1990) and Toda and Phillips (1993a, b).

Now consider three statistics, the OLS regression estimator, the regression R^2 and the usual t-statistic for testing the null that $\beta = 0$;

$$\hat{\beta} = \frac{[\sum y_{2,t} y_{1,t}]}{[\sum (y_{1,t})^2]},$$

$$R^2 = \frac{[\sum y_{2,t} y_{1,t}]^2}{\sum (y_{1,t})^2 \sum (y_{2,t})^2},$$

$$\tau = \frac{\hat{\beta}}{S_{\hat{\beta}}},$$

where $(S_{\hat{\beta}})^2 = T^{-1}[\sum (y_{2,t})^2 - \hat{\beta} \sum y_{2,t} y_{1,t}]/[\sum (y_{1,t})^2]$ is the usual formula for the variance of $\hat{\beta}$. When $y_{1,t}$ and $y_{2,t}$ are mutually independent and i.i.d., standard results imply that $\hat{\beta} \xrightarrow{p} 0$, $R^2 \xrightarrow{p} 0$ and $\tau \xrightarrow{\mathcal{L}} N(0,1)$. When $y_{1,t}$ and $y_{2,t}$ are mutually independent random walks, things are quite different. Let $B(s)$ denote a 2×1 Brownian motion process, $V = \int B(s) B(s)' ds$ and v_{ij} denote the ijth element of V. Then, utilizing Lemma 2.3,

$$\hat{\beta} \Rightarrow \left(\frac{\sigma_2}{\sigma_1}\right)\left(\frac{v_{21}}{v_{11}}\right), \tag{2.12}$$

$$R^2 \Rightarrow \frac{(v_{21}^2)}{(v_{11} v_{22})}, \tag{2.13}$$

$$T^{-1/2}\tau \Rightarrow \frac{v_{21}}{(v_{11} v_{22} - v_{21}^2)^{1/2}}, \tag{2.14}$$

where $\sigma_i^2 = \mathrm{var}(\varepsilon_{i,t}), i = 1, 2$. Thus, both $\hat{\beta}$ and R^2 converge to non-degenerate random variables, while τ diverges. This shows that large absolute values of the t-statistic should be expected in "spurious" regressions.

When the estimated regression (2.11) contains a constant or a constant and time trend, similar results obtain with the demeaned and detrended Brownian motion processes $B^{\mu}(s)$ and $B^{\tau}(s)$ replacing $B(s)$. When the regression contains a constant, the results are invariant to the initial conditions for $y_{1,t}$ and $y_{2,t}$; when the regression contains a constant and a time trend the results are invariant to initial conditions and drift terms in $y_{1,t}$ and $y_{2,t}$. See Phillips (1986) for a more detailed discussion.

2.6.4. *Estimating cointegrating vectors by ordinary least squares*

Now suppose that $y_{1,t}$ and $y_{2,t}$ are generated by

$$\Delta y_{1,t} = u_{1,t}, \tag{2.15}$$

$$y_{2,t} = \beta y_{1,t} + u_{2,t}, \tag{2.16}$$

where $u_t = (u_{1,t}\, u_{2,t})' = D\varepsilon_t$, where ε_t is an mds(I_2) with finite fourth moments. Like the spurious regression model, both $y_{1,t}$ and $y_{2,t}$ are individually I(1): $y_{1,t}$ is a random walk, while $\Delta y_{2,t}$ follows a univariate ARMA(1, 1) process. Unlike the spurious regression model, one linear combination of the variables $y_{2,t} - \beta y_{1,t} = u_{2,t}$ is I(0), and so the variables are cointegrated.

Stock (1987) derives the asymptotic distribution of the OLS estimator of cointegrating vectors. In this example, the limiting distribution is quite simple. Write

$$\hat{\beta} - \beta = \frac{\sum y_{1,t} u_{2,t}}{\sum (y_{1,t})^2} \tag{2.17}$$

and let d_{ij} denote the ijth element of D, and $D_i = (d_{i1}\, d_{i2})$ denote the ith row of D. Then the limiting behavior, or the denominator of $\hat{\beta} - \beta$, follows directly from the lemma:

$$T^{-2}\sum (y_{1,t})^2 = D_1[T^{-2}\sum \xi_t \xi_t']D_1' \Rightarrow D_1\left[\int B(s)B(s)'\,\mathrm{d}s\right]D_1', \tag{2.18}$$

where ξ_t is the bivariate random walk, with $\Delta\xi_t = \varepsilon_t$ and $B(s)$ is a 2×1 Brownian motion process. The numerator is only slightly more difficult:

$$\begin{aligned} T^{-1}\sum y_{1,t} u_{2,t} &= T^{-1}\sum y_{1,t-1} u_{2,t} + T^{-1}\sum \Delta y_{1,t} u_{2,t} \\ &= D_1[T^{-1}\sum \xi_{t-1}\varepsilon_t']D_2' + D_1[T^{-1}\sum \varepsilon_t \varepsilon_t']D_2' \\ &\Rightarrow D_1\left[\int B(s)\,\mathrm{d}B(s)'\right]D_2' + D_1 D_2'. \end{aligned} \tag{2.19}$$

Putting these two results together,

$$T(\hat{\beta} - \beta) \Rightarrow \left[D_1 \int B(s)\,\mathrm{d}B(s)' D_2' + D_1 D_2'\right]\left[D_1 \int B(s)B(s)'\,\mathrm{d}s\, D_1'\right]^{-1}. \tag{2.20}$$

There are three interesting features of the limiting representation (2.20). First, $\hat{\beta}$ is "super consistent," converging to its true value at rate T. Second, while super consistent, $\hat{\beta}$ is asymptotically biased, in the sense that the mean of the asymptotic distribution is not centered at zero. The constant term $D_1 D_2' = d_{12}d_{22} + d_{11}d_{21}$ that appears in the numerator of (2.20) is primarily responsible for this bias. To see the source of this bias, notice that the regressor $y_{1,t}$ is correlated with the error term $u_{2,t}$. In standard situations, this "simultaneous equation bias" is reflected in

large samples as an inconsistency in $\hat{\beta}$. With cointegration, the regressor is I(1) and the error term is I(0), so no inconsistency results; the "simultaneous equations bias" shows up as bias in the asymptotic distribution of $\hat{\beta}$. In realistic examples this bias can be quite large. For example, Stock (1988) calculates the asymptotic bias that would obtain in the OLS estimator of the marginal propensity to consume, obtained from a regression of consumption onto income using annual observations with a process for u_t similar to that found in U.S. data. He finds that the bias is still -0.10 even when 53 years of data are used.[11] Thus, even though the OLS estimators are "super" consistent, they can be quite poor.

The third feature of the asymptotic distribution in (2.20) involves the special case in which $d_{12} = d_{21} = 0$ so that $u_{1,t}$ and $u_{2,t}$ are statistically independent. In this case the OLS estimator corresponds to the Gaussian maximum likelihood estimation (MLE). When $d_{12} = d_{21} = 0$, (2.20) simplifies to

$$T(\hat{\beta} - \beta) \Rightarrow \left(\frac{d_{22}}{d_{11}}\right) \frac{\int B_1(s)\,\mathrm{d}B_2(s)}{\int B_1(s)^2\,\mathrm{d}s}, \tag{2.21}$$

where $B(s)$ is partitioned as $B(s) = [B_1(s)\, B_2(s)]'$. This result is derived in Phillips and Park (1988) where the distribution is given a particularly simple and useful interpretation. To develop the interpretation, suppose for the moment that $u_{2,t} = d_{22}\varepsilon_{2,t}$ was n.i.i.d. (In large samples the normality assumption is not important; it is made here to derive simple and exact small sample results.) Now, consider the distribution of $\hat{\beta}$ conditional on the regressors $\{y_{1,t}\}_{t=1}^{T}$. Since $u_{2,t}$ is n.i.i.d., the restriction $d_{12} = d_{21} = 0$ implies that $u_{2,t}$ is independent of $\{y_{1,t}\}_{t=1}^{T}$. This means that $\hat{\beta} - \beta | \{y_{1,t}\}_{t=1}^{T} \sim N\{0, d_{22}^2[\sum(y_{1,t})^2]^{-1}\}$, so that the unconditional distribution $\hat{\beta} - \beta$ is normal with mean zero and random covariance matrix, $d_{22}^2[\sum(y_{1,t})^2]^{-1}$. In large samples, $T^{-2}\sum(y_{1,t})^2 \Rightarrow d_{11}^2 \int B_1(s)^2\,\mathrm{d}s$, so that $T(\hat{\beta} - \beta)$ converges to a normal random variable with a mean of zero and random covariance matrix, $(d_{22}/d_{11})^2[\int B_1(s)^2\,\mathrm{d}s]^{-1}$. Thus, $T(\hat{\beta} - \beta)$ has an asymptotic distribution that is a random mixture of normals. Since the normal distributions in the mixture have a mean of zero, the asymptotic distribution is distributed symmetrically about zero, and thus $\hat{\beta}$ is asymptotically median unbiased.

The distribution is useful, not so much for what it implies about the distribution of $\hat{\beta}$, but for what it implies about the t-statistic for $\hat{\beta}$. When d_{12} or d_{21} are not equal to zero, the t-statistic for testing the null $\beta = \beta_0$ has a nonstandard limiting distribution, analogous to the distribution of the Dickey–Fuller t-statistic for testing the null of a unit AR coefficient in a univariate regression. However, when $d_{12} = d_{21} = 0$, the t-statistic has a limiting standard normal distribution. To see

[11] Stock (1988, Table 4). These results are for durable plus nondurable consumption. When nondurable consumption is used, Stock estimates the bias to be -0.15.

why this is true, again consider the situation in which $u_{2,t}$ is n.i.i.d. When $d_{12}=d_{21}=0$, the distribution of the t-statistic for testing $\beta=\beta_0$ conditional on $\{y_{1,t}\}_{t=1}^{T}$ has an exact Student's t distribution with $T-1$ degrees of freedom. Since this distribution does not depend on $\{y_{1,t}\}_{t=1}^{T}$, this is the unconditional distribution as well. This means that in large samples, the t-statistic has a standard normal distribution. As we will see in this next section, the Phillips and Park (1988) result carries over to a much more general setting.

In the example developed here, $u_t=D\varepsilon_t$ is serially uncorrelated. This simplifies the analysis, but all of the results hold more generally. For example, Stock (1987) assumes that $u_t=D(L)\varepsilon_t$, where $D(L)=\sum_{i=0}^{\infty}D_iL^i$, $|D(1)|\neq 0$ and $\sum_{i=1}^{\infty}i|D_i|<\infty$. In this case,

$$T(\hat{\beta}-\beta)\Rightarrow\left\{D_1(1)\left[\int B(s)\mathrm{d}B(s)'\right]D_2(1)'+\sum_{i=0}^{\infty}D_{1,i}D_{2,i}'\right\}\times\left\{D_1(1)\left[\int B(s)B(s)'\mathrm{d}s\right]D_1(1)\right\}^{-1}, \tag{2.22}$$

where $D_j(1)$ is the jth row of $D(1)$ and $D_{j,i}$ is the jth row of D_i. Under the additional assumption that $d_{12}(1)=d_{21}(1)=0$ and $\sum_{i=0}^{\infty}D_{1,i}D_{2,i}'=0$, $T(\hat{\beta}-\beta)$ is distributed as a mixed normal (asymptotically) and the t-statistic for testing $\beta=\beta_0$ has an asymptotic normal distribution when $d_{12}(1)=d_{21}(1)=0$ [see Phillips and Park (1988) and Phillips (1991a)].

2.7. *Implications for econometric practice*

The asymptotic results presented above are important because they determine the appropriate critical values for tests of coefficient restrictions in VAR models. The results lead to three lessons that are useful for applied practice.

(1) Coefficients that can be written as coefficients on zero mean I(0) regressors in regressions that include a constant term are asymptotically normal. Test statistics for restrictions on these coefficients have the usual asymptotic χ^2 distributions. For example, in the model

$$y_t=\gamma_1 z_{1,t}+\gamma_2+\gamma_3 z_{3,t}+\gamma_4 t+\varepsilon_t, \tag{2.23}$$

where $z_{1,t}$ is a mean zero I(0) scalar regressor and $z_{3,t}$ is a scalar martingale regressor, this result implies that Wald statistics for testing $H_o\colon\gamma_1=c$ is asymptotically χ^2.

(2) Linear combinations of coefficients that include coefficients on zero mean I(0) regressors together with coefficients on stochastic or deterministic trends will have asymptotic normal distributions. Wald statistics for testing restrictions on these

linear combinations will have large sample χ^2 distributions. Thus in (2.23), Wald statistics for testing $H_o: R_1\gamma_1 + R_3\gamma_3 + R_4\gamma_4 = r$, will have an asymptotic χ^2 distribution if $R_1 \neq 0$.

(3) Coefficients that cannot be written as coefficients on zero mean I(0) regressors (e.g. constants, time trends, and martingales) will, in general, have nonstandard asymptotic distributions. Test statistics that involve restrictions on these coefficients that are not a function of coefficients on zero mean I(0) regressors will, in general, have nonstandard asymptotic distributions. Thus in (2.23), the Wald statistic for testing: $H_o: R(\gamma_2\ \gamma_3\ \gamma_4)' = r$ has a non-χ^2 asymptotic distribution, as do test statistics for composite hypotheses of the form $H_o: R(\gamma_2\ \gamma_3\ \gamma_4)' = r$ and $\gamma_1 = c$.

When test statistics have a nonstandard distribution, critical values can be determined by Monte Carlo methods by simulating approximations to the various functionals of $B(s)$ appearing in Lemma 2.3. As an example, consider using Monte Carlo methods to calculate the asymptotic distribution of sum of coefficients $\phi_1 + \phi_2 = \gamma_2$ in the univariate AR(2) regression model (2.1). Section 2.4 showed that $T(\hat{\gamma}_2 - \gamma_2) \Rightarrow (1 + \phi_2)[\int B(s)^2 ds]^{-1}[\int B(s) dB(s)]$, where $B(s)$ is a scalar Brownian motion process. If x_t is generated as a univariate Gaussian random walk, then one draw of the random variable $[\int B(s)^2 ds]^{-1}[\int B(s) dB(s)]$ is well approximated by $(T^{-2}\sum x_t^2)^{-1}(T^{-1}\sum x_t \Delta x_{t+1})$ with T large. (A value of $T = 500$ provides an adequate approximation for most purposes.) The distribution of $T(\hat{\gamma}_2 - \gamma_2)$ can then be approximated by taking repeated draws of $(T^{-2}\sum x_t^2)^{-1}(T^{-1}\sum x_t \Delta x_{t+1})$ multiplied by $(1 + \hat{\phi}_2)$. An example of this approach in a more complicated multivariate model is provided in Stock and Watson (1988).

Application of these rules in practice requires that the researcher know about the presence and location of unit roots in the VAR. For example, in determining the asymptotic distribution of Granger-causality test statistics, the researcher has to know whether the candidate causal variable is integrated and, if it is integrated, whether it is cointegrated with any other variable in the regression. If it is cointegrated with the other regressors, then the test statistic has a χ^2 asymptotic distribution. Otherwise the test statistic is asymptotically non-χ^2, in general. In practice such prior information is often unavailable, and an important question is what is to be done in this case?[12]

The general problem can be described as follows. Let W denote the Wald test statistic for a hypothesis of interest. Then the asymptotic distribution of the Wald statistic when a unit root is present, say $F(W|U)$, is not equal to the distribution of the statistic when no unit root is present, say $F(W|N)$. Let c_U and c_N denote

[12]Toda and Phillips (1993a, b) discuss testing for Granger causality in a situation in which the researcher knows the number of unit roots in the model but doesn't know the cointegrating vectors. They develop a sequence of asymptotic χ^2 tests for the problem. When the number of unit roots in the system in unknown, they suggest pretesting for the number of unit roots. While this will lead to sensible results in many empirical problems, examples such as the one presented at the end of this section show that large pretest biases are possible.

the "unit root" and "no unit root" critical values for a test with size α. That is, c_U and c_N satisfy: $P(W > c_U|U) = P(W > c_N|N) = \alpha$ under the null. The problem is that $c_U \neq c_N$, and the researcher does not know whether U or N is the correct specification.

In one sense, this is not an ususual situation. Usually, the distribution of statistics depends on characteristics of the probability distribution of the data that are unknown to the researcher, even under the null hypothesis. Typically, there is uncertainty over certain "nuisance parameters," that affect the distribution of the statistic of interest. Yet, typically the distribution depends on the nuisance parameters in a continuous fashion, in the sense that critical values are continuous functions of the nuisance parameters. This means that asymptotically valid inference can be carried out by replacing the unknown parameters with consistent estimates.

This is not possible in the present situation. While it is possible to represent the uncertainty in the distribution of test statistics as a function of nuisance parameters that can be consistently estimated, the critical values are not continuous functions of these prameters. Small changes in the nuisance parameters – associated with sampling error in estimates – may lead to large changes in critical values. Thus, inference cannot be carried out by replacing unknown nuisance parameters with consistent estimates. Alternative procedures are required.[13]

Development of these alternative procedures is currently an active area of research, and it is too early to speculate on which procedures will prove to be the most useful. It is possible to mention a few possibilities and highlight the key issues.

The simplest procedure is to carry out conservative inference. That is, to use the largest of the "unit root" and "no unit root" critical values, rejecting the null when $W > \max(c_U, c_N)$. By construction, the size of the test is less than or equal to α. Whenever $W > \max(c_U, c_N)$, so that the null is rejected using either distribution, or $W < \min(c_U, c_N)$, so that the null is not rejected using either distribution, one need not proceed further. However a problem remains when $\min(c_U, c_N) < W < \max(c_U, c_N)$. In this case, an intuitively appealing procedure is to look at the data to see which hypothesis – unit root or no unit root – seems more plausible.

This approach is widely used in applications. Formally, it can be described as follows. Let γ denote a statistic helpful in classifying the stochastic process as a unit root or no unit root process. (For example, γ might denote a Dickey–Fuller "t-statistic" or one of the test statistics for cointegration discussed in the next section.) The procedure is then to define a region for γ, say R_U, and when $\gamma \in R_U$, the critical value c_U is used; otherwise the critical value c_N is used. (For example, the unit root critical value might be used if the Dickey–Fuller "t-statistic" was greater than -2, and the no unit root critical value used when the DF statistic

[13] Alternatively, using "local-to-unity" asymptotics, the critical values can be represented as continuous functions of the local-to-unity parameter, but this parameter cannot be consistently estimated from the data. See Bobkoski (1983), Cavanagh (1985), Chan and Wei (1987), Chan (1988), Phillips (1987b) and Stock (1991).

was less than -2.) In this case, the probability of type 1 error is

$$P(\text{Type 1 error}) = P(W > c_U | \gamma \in R_U)P(\gamma \in R_U) + P(W > c_N | \gamma \notin R_U)P(\gamma \notin R_U).$$

The procedure will work well, in the sense of having the correct size and a power close to the power that would obtain when the correct unit root or no unit root specification were known, if two conditions are met. First, $P(\gamma \in R_U)$ should be near 1 when the unit root specification is true, and $P(\gamma \notin R_U)$ should be near 1 when the unit root specification is false, respectively. Second, $P(W > c_U | \gamma \in R_U)$ and $P(W > c_N | \gamma \notin R_U)$ should be near $P(W > c_U | U)$ and $P(W > c_N | N)$, respectively. Unfortunately, in practice neither of these conditions may be true. The first requires statistics that perfectly discriminate between the unit root and non-unit root hypotheses. While significant progress has been made in developing powerful inference procedures [e.g. Dickey and Fuller (1979), Elliot et al. (1992), Phillips and Ploberger (1991), Stock (1992)], a high probability of classification errors is unavoidable in moderate sample sizes.

In addition, the second condition may not be satisfied. An example presented in Elliot and Stock (1992) makes this point quite forcefully. [Also see Cavanagh and Stock (1985).] They consider the problem of testing whether the price-divided ratio helps to predict future changes in stock prices.[14] A stylized version of the model is

$$p_t - d_t = \phi(p_{t-1} - d_{t-1}) + \varepsilon_{1,t}, \tag{2.24}$$

$$\Delta p_t = \beta(p_{t-1} - d_{t-1}) + \varepsilon_{2,t}, \tag{2.25}$$

where p_t and d_t are the logs of prices and dividends, respectively, and $(\varepsilon_{1,t}\, \varepsilon_{2,t})'$ is an mds(Σ_ε). The hypothesis of interest is H_o: $\beta = 0$. Under the null, and when $|\phi| < 1$, the t-statistic for this null will have an asymptotic standard normal distribution; when the hypothesis $\phi = 1$, the t-statistic will have a unit root distribution. (The particular form of the distribution could be deduced using Lemma 2.3, and critical values could be constructed using numerical methods.) The pretest procedure involves carrying out a test of $\phi = 1$ in (2.24), and using the unit root critical value for the t-statistic for $\beta = 0$ in (2.25) when $\phi = 1$ is not rejected. If $\phi = 1$ is rejected, the critical value from the standard normal distribution is used.

Elliot and Stock show that the properties of this procedure depends critically on the correlation between $\varepsilon_{1,t}$ and $\varepsilon_{2,t}$. To see why, consider an extreme example. In the data, dividends are much smoother than prices, so that most of the variance in the price-dividend ratio comes from movements in prices and not from dividends. Thus, $\varepsilon_{1,t}$ and $\varepsilon_{2,t}$ are likely to be highly correlated. In the extreme case, when

[14] Hodrick (1992) contains an overview of the empirical literature on the predictability of stock prices using variables like the price-dividend ratio. Also see, Fama and French (1988) and Campbell (1990).

they are perfectly correlated, $(\hat{\beta} - \beta)$ is proportional to $(\hat{\phi} - \phi)$, and the "t-statistic" for testing $\beta = 0$ is exactly equal to the "t-statistic" for testing $\phi = 1$. In this case $F(W|\hat{\gamma})$ is degenerate and does not depend on the null hypothesis. All of the information in the data about the hypothesis $\beta = 0$ is contained in the pretest. While this example is extreme, it does point out the potential danger of relying on unit root pretests to choose critical values for subsequent tests.

3. Cointegrated systems

3.1. *Introductory comments*

An important special case of the model analyzed in Section 4 is the cointegrated VAR. This model provides a framework for studying the long-run economic relations discussed in the introduction. There are three important econometric questions that arise in the analysis of cointegrated systems. First, how can the common stochastic trends present in cointegrated systems be extracted from the data? Second, how can the hypothesis of cointegration be tested? And finally, how should unknown parameters in cointegrating vectors be estimated, and how should inference about their values be conducted? These questions are answered in this section.

We begin, in Section 3.2, by studying different representations for cointegrated systems. In addition to highlighting important characteristics of cointegrated systems, this section provides an answer to the first question by presenting a general trend extraction procedure for cointegrated systems. Section 3.3 discusses the problem of testing for the order of cointegration, and Section 3.4 discusses the problem of estimation and inference for unknown parameters in cointegrating vectors. To keep the notation simple, the analysis in Sections 3.2–3.4 abstracts from deterministic components (constants and trends) in the data. The complications in estimation and testing that arise when the model contains constants and trends is the subject of Section 3.5. Only I(1) systems are considered here. Using Engle and Granger's (1987) terminology, the section discusses only CI(1,1) systems; that is, systems in which linear combinations of I(1) and I(0) variables are I(0). Extensions for CI(d, b) systems with d and b different from 1 are presented in Johansen (1988b, 1992c), Granger and Lee (1990) and Stock and Watson (1993).

3.2. *Representations for the I(1) cointegrated model*

Consider the VAR

$$x_t = \sum_{i=1}^{p} \Pi_i x_{t-i} + \varepsilon_t, \tag{3.1}$$

where x_t is an $n \times 1$ vector composed of I(0) and I(1) variables, and ε_t is an mds(Σ_ε). Since each of the variables in the system are I(0) or I(1), the determinantal polynomial $|\Pi(z)|$ contains at most n unit roots, with $\Pi(z) = I - \sum_{i=1}^{p} \Pi_i z^i$. When there are fewer than n unit roots, then the variables are cointegrated, in the sense that certain linear combinations of the x_t's are I(0). In this subsection we derive four useful representations for cointegrated VARs: (1) the vector error correction VAR model, (2) the moving average representation of the first differences of the data, (3) the common trends representation of the levels of the data, and (4) the triangular representation of the cointegrated model.

All of these representations are readily derived using a particular Smith–McMillan factorization of the autoregressive polynomial $\Pi(L)$. The specific factorization used here was originally developed by Yoo (1987) and was subsequently used to derive alternative representations of cointegrated systems by Engle and Yoo (1991). Some of the discussion presented here parallels the discussion in this latter reference. Yoo's factorization of $\Pi(z)$ isolates the unit roots in the system in a particularly convenient fashion. Suppose that the polynomial $\Pi(z)$ has all of its roots on or outside the unit circle, then the polynomial can be factored as $\Pi(z) = U(z)M(z)V(z)$, where $U(z)$ and $V(z)$ are $n \times n$ matrix polynomials with all of their roots outside the unit circle, and $M(z)$ is an $n \times n$ diagonal matrix polynomial with roots on or outside the unit circle. In the case of the I(1) cointegrated VAR, $M(L)$ can be written as

$$M(L) = \begin{bmatrix} \Delta_k & 0 \\ 0 & I_r \end{bmatrix},$$

where $\Delta_k = (1 - L)I_k$ and $k + r = n$. This factorization is useful because it isolates all of the VAR's nonstationarities in the upper block of $M(L)$.

We now derive alternative representations for the cointegrated system.

3.2.1. *The vector error correction VAR model (VECM)*

To derive the VECM, subtract x_{t-1} from both sides of (3.1) and rearrange the equation as

$$\Delta x_t = \Pi x_{t-1} + \sum_{i=1}^{p-1} \Phi_i \Delta x_{t-i} + \varepsilon_t, \tag{3.2}$$

where $\Pi = -I_n + \sum_{i=1}^{p} \Pi_i = -\Pi(1)$, and $\Phi_i = -\sum_{j=i+1}^{p} \Pi_j$, $i = 1, \ldots, p-1$. Since $\Pi(1) = U(1)M(1)V(1)$, and $M(1)$ has rank r, $\Pi = -\Pi(1)$ also has rank r. Let α denote an $n \times r$ matrix whose columns form a basis for the row space of Π, so that every row of Π can be written as a linear combination of the rows of α'. Thus, we can write $\Pi = \delta\alpha'$, where δ is an $n \times r$ matrix with full column rank.

Equation (3.2) then becomes

$$\Delta x_t = \delta\alpha' x_{t-1} + \sum_{i=1}^{p-1} \Phi_i \Delta x_{t-i} + \varepsilon_t \tag{3.3}$$

or

$$\Delta x_t = \delta w_{t-1} + \sum_{i=1}^{p-1} \Phi_i \Delta x_{t-i} + \varepsilon_t, \tag{3.4}$$

where $w_t = \alpha' x_t$. Solving (3.4) for w_{t-1} shows that $w_{t-1} = (\delta'\delta)^{-1}\delta'[\Delta x_t - \sum_{i=1}^{p-1} \Phi_i \Delta x_{t-i} - \varepsilon_t]$, so that w_t is I(0). Thus, the linear combinations of the potentially I(1) elements of x_t formed by the columns of α are I(0), and the columns of α are cointegrating vectors.

The VECM imposes $k < n$ unit roots in the VAR by including first differences of all of the variables and $r = n - k$ linear combinations of levels of the variables. The levels of x_t are introduced in a special way – as $w_t = \alpha' x_t$ – so that all of the variables in the regression are I(0). Equations of this form appeared in Sargan (1964) and the term "error correction model" was introduced in Davidson et al. (1978).[15] As explained there and in Hendry and von Ungern-Sternberg (1981), $\alpha' x_t = 0$ can be interpreted as the "equilibrium" of the dynamical system, w_t as the vector of "equilibrium errors" and equation (3.4) describes the self correcting mechanism of the system.

3.2.2. *The moving average representation*

To derive the moving average representation for Δx_t, let

$$\bar{M}(L) = \begin{bmatrix} I_k & 0 \\ 0 & \Delta_r \end{bmatrix},$$

so that $\bar{M}(L)M(L) = (1 - L)I_n$. Then,

$$\bar{M}(L)M(L)V(L)x_t = \bar{M}(L)U(L)^{-1}\varepsilon_t,$$

so that

$$V(L)\Delta x_t = \bar{M}(L)U(L)^{-1}\varepsilon_t,$$

[15] As Phillips and Loretan (1991) point out in their survey, continuous time formulations of error correction models were used extensively by A.W. Phillips in the 1950's. I thank Peter Phillips for drawing this work to any attention.

and

$$\Delta x_t = C(L)\varepsilon_t, \tag{3.5}$$

where $C(L) = V(L)^{-1}\bar{M}(L)U(L)^{-1}$.

There are two special characteristics of the moving average representation. First, $C(1) = V(1)^{-1}\bar{M}(1)U(1)^{-1}$ has rank k and is singular when $k < n$. This implies that the spectral density matrix of Δx_t evaluated at frequency zero, $(2\pi)^{-1}C(1)\Sigma_\varepsilon C(1)'$, is singular in a cointegrated system. Second, there is a close relationship between $C(1)$ and the matrix of cointegrating vectors α. In particular, $\alpha'C(1) = 0$.[16] Since $w_t = \alpha' x_t$ is I(0), $\Delta w_t = \alpha'\Delta x_t$ is I(−1) so that its spectrum at frequency zero, $(2\pi)^{-1}\alpha'C(1)\Sigma_\varepsilon C(1)'\alpha$, vanishes.

The equivalence of vector error correction models and cointegrated variables with moving average representations of the form (3.5) is provided in Granger (1983) and forms the basis of the Granger Representation Theorem [see Engle and Granger (1987)].

3.2.3. *The common trends representation*

The common trends representation follows directly from (3.5). Adding and subtracting $C(1)\varepsilon_t$ from the right hand side of (3.5) yields

$$\Delta x_t = C(1)\varepsilon_t + [C(L) - C(1)]\varepsilon_t. \tag{3.6}$$

Solving backwards for the level of x_t,

$$x_t = C(1)\xi_t + C^*(L)\varepsilon_t + x_0, \tag{3.7}$$

where $\xi_t = \sum_{s=1}^{t}\varepsilon_s$ and $C^*(L) = (1-L)^{-1}[C(L) - C(1)] = \sum_{i=0}^{\infty} C_i^* L^i$, where $C_i^* = -\sum_{j=i+1}^{\infty} C_j$ and $\varepsilon_i = 0$ for $i \leqslant 0$ is assumed. Equation (3.7) is the multivariate Beveridge–Nelson (1981) decomposition of x_t; it decomposes x_t into its "permanent component," $C(1)\xi_t + x_0$, and its "transitory component," $C^*(L)\varepsilon_t$.[17] Since $C(1)$ has rank k, we can find a nonsingular matrix G, such that $C(1)G = [A\,0_{n\times r}]$, where A is an $n \times k$ matrix with full column rank.[18] Thus $C(1)\xi_t = C(1)GG^{-1}\xi_t$,

[16] To derive this result, note from (3.2) and (3.3) that $\Pi = -\Pi(1) = -U(1)M(1)V(1) = \delta\alpha'$. Since $M(1)$ has zeroes everywhere, except the lower diagonal block which is I_r, α' must be a nonsingular transformation of the last r rows of $V(1)$. This implies that the first k columns of $\alpha'V(1)^{-1}$ contain only zeroes, so that $\alpha'V(1)^{-1}\bar{M}(1)U(1) = \alpha'C(1) = 0$.

[17] The last component can be viewed as transitory because it has a finite spectrum at frequency zero. Since $U(z)$ and $V(z)$ are finite order with roots outside the unit circle, the C_i coefficients decline exponentially for large i, and thus $\sum_i i|C_i|$ is finite. Thus the C_i^* matrices are absolutely summable, and $C^*(1)\Sigma_\varepsilon C^*(1)'$ is finite.

[18] The matrix G is not unique. One way to construct G is from the eigenvectors of A. The first k columns of G are the eigenvectors corresponding to the nonzero eigenvalues of A and the remaining eigenvectors are the last $n-k$ columns of G.

so that

$$x_t = A\tau_t + C^*(L)\varepsilon_t + x_0, \tag{3.8}$$

where τ_t denotes the first k components of $G^{-1}\xi_t$.

Equation (3.8) is the common trends representation of the cointegrated system. It decomposes the $n \times 1$ vector x_t into k "permanent components" τ_t and n "transitory components" $C^*(L)\varepsilon_t$. These permanent components often have natural interpretations. For example, in the eight variable (y, c, i, n, w, m, p, r) system introduced in Section 1, five cointegrating vectors were suggested. In an eight variable system with five cointegrating vectors there are three common trends. In the (y, c, i, n, m, p, r) systems these trends can be interpreted as population growth, technological progress and trend growth in money.

The common trends representation (3.8) is used in King et al. (1991) as a device to "extract" the single common trend in a three variable system consisting of y, c and i. The derivation of (3.8) shows exactly how to do this: (i) estimate the VECM (3.3) imposing the cointegration restrictions; (ii) invert the VECM to find the moving average representation (3.5); (iii) find the matrix G introduced below equation (3.7); and, finally, (iv) construct τ_t recursively from $\tau_t = \tau_{t-1} + e_t$, where e_t is the first element of $G^{-1}\varepsilon_t$, and where ε_t denotes the vector of residuals from the VECM. Other interesting applications of trend extraction in cointegrated systems are contained in Cochrane and Sbordone (1988) and Cochrane (1994).

3.2.4. *The triangular representation*

The triangular representation also represents x_t in terms of a set of k non-cointegrated I(1) variables. Rather than construct these stochastic trends as the latent variables τ_t in the common trends representation, a subset of the x_t variables are used. In particular, the triangular representation takes the form:

$$\Delta x_{1,t} = u_{1,t}, \tag{3.9}$$

$$x_{2,t} - \beta x_{1,t} = u_{2,t}, \tag{3.10}$$

where $x_t = (x'_{1,t}\ x'_{2,t})'$, $x_{1,t}$ is $k \times 1$ and $x_{2,t}$ is $r \times 1$. The transitory components are $u_t = (u'_{1,t}\ u'_{2,t})' = D(L)\varepsilon_t$, where (as we show below) $D(1)$ has full rank. In this representation, the first k elements of x_t are the common trends and $x_{2,t} - \beta x_{1,t}$ are the I(0) linear combinations of the data.

To derive this representation from the VAR (3.2), use $\Pi(L) = U(L)M(L)V(L)$ to write

$$U(L)M(L)V(L)x_t = \varepsilon_t, \tag{3.11}$$

so that

$$M(L)V(L)x_t = U(L)^{-1}\varepsilon_t. \tag{3.12}$$

Now, partition $V(L)$ as

$$V(L) = \begin{bmatrix} v_{11}(L) & v_{12}(L) \\ v_{21}(L) & v_{22}(L) \end{bmatrix},$$

where $v_{11}(L)$ is $k \times k$, $v_{12}(L)$ is $k \times r$, $v_{21}(L)$ is $r \times k$ and $v_{22}(L)$ is $r \times r$. Assume that the data have been ordered so that $v_{22}(L)$ has all of its roots outside the unit circle. (Since $V(L)$ has all of its roots outside the unit circle, this assumption is made with no loss of generality.) Now, let

$$C(L) = \begin{bmatrix} I_k & 0 \\ \beta(L) & I_r \end{bmatrix},$$

where $\beta(L) = -v_{22}(L)^{-1}v_{21}(L)$. Then

$$M(L)V(L)C(L)C(L)^{-1}x_t = U(L)^{-1}\varepsilon_t \tag{3.13}$$

or, rearranging and simplifying,

$$\begin{bmatrix} v_{11}(L) + v_{12}(L)\beta & (1-L)v_{12}(L) \\ -v_{22}(L)\beta^*(L) & v_{22}(L) \end{bmatrix} \begin{bmatrix} \Delta x_{1,t} \\ x_{2,t} - \beta x_{1,t} \end{bmatrix} = U(L)^{-1}\varepsilon_t, \tag{3.14}$$

where $\beta^*(L) = (1-L)^{-1}[\beta(L) - \beta(1)]$ and $\beta = \beta(1)$. Letting $G(L)$ denote the matrix polynomial on the left hand side of (3.14), the triangular representation is obtained by multiplying equation (3.14) by $G(L)^{-1}$. Thus, in equations (3.9) and (3.10), $u_t = D(L)\varepsilon_t$, with $D(L) = G(L)^{-1}U(L)^{-1}$.

When derived from the VAR (3.2), $D(L)$ is seen to have a special structure that was inherited from the assumption that the data were generated by a finite order VAR. But of course, there is nothing inherently special or natural about the finite order VAR; it is just one flexible parameterization for the x_t process. When the triangular representation is used, an alternative approach is to parameterize the matrix polynomial $D(L)$ directly.

An early empirical study using this formulation is contained in Campbell and Shiller (1987). They estimate a bivariate model of the term structure that includes long term and short term interest rates. Both interest rates are assumed to be I(1), but the "spread" or difference between the variables is assumed to be I(0). Thus, in terms of (3.9)–(3.10), $x_{1,t}$ is the short term interest rate, $x_{2,t}$ is the long rate and $\beta = 1$. In their empirical work, Campbell and Shiller modeled the process u_t in (3.10) as a finite order VAR.

In empirical work, the triangular representation is no more nor less convenient than the VECM. However, in theoretical econometric work concerned with estimating cointegrating vectors, the triangular representation is very convenient. The reason is that coefficients making up the cointegrating vectors appear only in (3.10), and the system (3.9)–(3.10) "looks like" a standard triangular simultaneous equation system; estimators developed for that model, suitably modified, can be used to estimate the cointegrating vectors. Phillips (1991a) gave the triangular representation its name and demonstrated its usefulness for developing and analyzing the properties of estimators of cointegrating vectors.[19] The representation has subsequently been used by many other researchers who have developed a large number of asymptotically efficient estimators.

Regardless of the representation used, model building for cointegrated systems involves two steps. In the first step, the degree of cointegration (or equivalently the number of unit roots in the model) is determined. In the second step, the unknown parameters of the model are estimated. Statistical procedures for carrying out these steps are the subject of the next two sections.

3.3. *Testing for cointegration in I(1) systems*

It is convenient to cast our discussion in terms of the VAR in equation (3.2). We are interested in tests concerning $r = \text{rank}(\Pi)$ for this equation. The null and alternative hypotheses are

$$H_o: \text{rank}(\Pi) = r = r_o,$$
$$H_a: \text{rank}(\Pi) = r = r_o + r_a,$$

where $r_a > 0$. The hypotheses are written so that r_a denotes the additional cointegrating vectors that are present under the alternative. For example, when $r_o = 0$, the null specifies that there are no cointegrating vectors, while the alternative implies that there are $r_a > 0$ cointegrating vectors. Specifying the null as "no cointegration" and the alternative as "cointegration" is natural, since when $r = 0$, then $\Pi = 0$ in equation (3.2), while when $r \neq 0$, then $\Pi \neq 0$; the null and alternative are then $H_o: \Pi = 0$ and $H_a: \Pi \neq 0$ (but restricted to have rank r_a).

As might be expected, the distribution of test statistics for cointegration are complicated by the presence of unit roots. Using the results developed in Section 2, these complications are easily overcome. To keep things as simple as possible, this section ignores constant terms and deterministic growth in the model. In terms of the analysis in Section 2, this eliminates the canonical regressors corresponding

[19]This representation was also used in important antecedents of Phillips (1991a), notably Phillips and Durlauf (1986), Phillips (1988) and Park and Phillips (1988, 1989).

to the constant ($z_{2,t}$) and the deterministic time trends ($z_{4,t}$). Hypothesis testing when deterministic components are present is discussed in Section 3.5.

There are a many tests for cointegration: some are based on likelihood methods, using a Gaussian likelihood and the VECM representation for the model, while others are based on more ad hoc methods. Section 3.3.1 presents likelihood based (Wald and Likelihood Ratio) tests for cointegration constructed from the VECM. The non-likelihood-based methods of Engle and Granger (1987) and Stock and Watson (1988) are the subject of Section 3.3.2, and the various tests are compared in Section 3.3.3.

3.3.1. *Likelihood based tests for cointegration*[20]

In Section 3.2.1 the general VECM was written as

$$\Delta x_t = \delta\alpha' x_{t-1} + \sum_{i=1}^{p-1} \Phi_i \Delta x_{t-i} + \varepsilon_t. \tag{3.3}$$

To develop the restrictions on the parameters in (3.3) implicit in the null hypothesis, first partition the matrix of cointegrating vectors as $\alpha = [\alpha_o\ \alpha_a]$ where α_o is an $n \times r_o$ matrix whose columns are the cointegrating vectors present under the null and α_a is the $n \times r_a$ matrix of additional cointegrating vectors present under the alternative. Partition δ conformably as $\delta = [\delta_o\ \delta_a]$, let $\Gamma = (\Phi_1\ \Phi_2 \cdots \Phi_{p-1})$ and let $z_t = (\Delta x'_{t-1}\ \Delta x'_{t-2} \cdots \Delta x'_{t-p+1})'$. The VECM can then be written as

$$\Delta x_t = \delta_o\alpha'_o x_{t-1} + \delta_a\alpha'_a x_{t-1} + \Gamma z_t + \varepsilon_t, \tag{3.15}$$

where, under the null hypothesis, the term $\delta_a\alpha'_a x_{t-1}$ is absent. This suggests writing the null and alternative hypotheses as $H_o\colon \delta_a = 0$ vs. $H_a\colon \delta_a \neq 0$.[21] Written in this way, the null is seen as a linear restriction on the regression coefficients in (3.15). An important complication is that the regressor $\alpha'_a x_{t-1}$ depends on parameters in α_a that are potentially unknown. Moreover, when $\delta_a = 0$, $\alpha'_a x_{t-1}$ does not enter the regression, and so the data provide no information about any unknown parameters in α_a. This means that these parameters are econometrically identified only under the alternative hypothesis, and this complicates the testing problem in ways discussed by Davies (1977, 1987), and (in the cointegration context) by Engle and Granger (1987).

In many applications, this may not be a problem of practical consequence, since the coefficients in α are determined by the economic theory under consideration. For example, in the (y, c, i, w, n, r, m, p) system, candidate error correction terms

[20] Much of the discussion in this section is based on material in Horvath and Watson (1993).

[21] Formally, the restriction $\text{rank}(\delta_a\alpha'_a) = r_a$ should be added as a qualifier to H_a. Since this constraint is satisfied almost surely by unconstrained estimators of (3.15) it can safely be ignored when constructing likelihood ratio test statistics.

with no unknown parameters are $y-c, y-i, (w-p)-(y-n)$ and r. Only one error correction term, $m-p-\beta_y y-\beta_r r$, contains potentially unknown parameters. Yet, when testing for cointegration, a researcher may not want to impose specific values of potential cointegrating vectors, particularly during the preliminary data analytic stages of the empirical investigation. For example, in their investigation of long-run purchasing power parity, Johansen and Juselius (1992) suggest a two-step testing procedure. In the first step cointegration is tested without imposing any information about the cointegrating vector. If the null hypothesis of no cointegration is rejected, a second stage test is conducted to see if the cointegrating vector takes on the value predicted by economic theory. The advantage of this two-step approach is that it can uncover cointegrating relations not predicted by the specific economic theory under study. The disadvantage is that the first stage test for cointegration will have low power relative to a test that imposes the correct cointegrating vector.

It is useful to have testing procedures that can be used when cointegrating vectors are known and when they are unknown. With these two possibilities in mind, we write $r=r_k+r_u$, where r_k denotes the number of cointegrating vectors with known coefficients, and r_u denotes the number of cointegrating vectors with unknown coefficients. Similarly, write $r_o=r_{ok}+r_{ou}$ and $r_a=r_{ak}+r_{au}$, where the subscripts "k" and "u" denote known and unknown respectively. Of course, the r_{ak} subset of "known cointegrating vectors" are present only under the alternative, and $\alpha_a' x_t$ is I(1) under the null.

Likelihood ratio tests for cointegration with unknown cointegrating vectors (i.e. $H_o: r=r_{ou}$ vs. $H_a: r=r_{ou}+r_{au}$) are developed in Johansen (1988a), and these tests are modified to incorporate known cointegrating vectors (nonzero values of r_{ok} and r_{ak}) in Horvath and Watson (1993). The test statistics and their asymptotic null distributions are developed below.

For expositional purposes it is convenient to consider three special cases. In the first, $r_a=r_{ak}$, so that all of the additional cointegrating vectors present under the alternative are assumed to be known. In the second, $r_a=r_{au}$, so that they are all unknown. The third case allows nonzero values of both r_{ak} and r_{au}. To keep the notation simple, the tests are derived for the $r_o=0$ null. In one sense, this is without loss of generality, since the LR statistic for $H_o: r=r_o$ vs. $H_a: r=r_o+r_a$ can always be calculated as the difference between the LR statistics for $[H_o: r=0$ vs. $H_a: r=r_o+r_a]$ and $[H_o: r=0$ vs. $H_a: r=r_o]$. However, the asymptotic null distribution of the test statistic does depend on r_{ok} and r_{ou}, and this will be discussed at the end of this section.

Testing $H_o: r=0$ vs. $H_a: r=r_{ak}$ When $r_o=0$, equation (3.15) simplifies to

$$\Delta x_t=\delta_a(\alpha_a' x_{t-1})+\Gamma z_t+\varepsilon_t. \tag{3.16}$$

Since $\alpha_a' x_{t-1}$ is known, (3.16) is a multivariate linear regression, so that the LR, Wald

and LM statistics have their standard regression form. Letting $X = [x_1\, x_2 \cdots x_T]'$, $X_{-1} = [x_0\, x_1 \cdots x_{T-1}]'$, $\Delta X = X - X_{-1}$, $Z = [z_1\, z_2 \cdots z_T]'$, $\varepsilon = [\varepsilon_1\, \varepsilon_2 \cdots \varepsilon_T]'$ and $M_Z = [I - Z(Z'Z)^{-1}Z']$, the OLS estimator of δ_a is $\hat{\delta}_a = (\Delta X' M_Z X_{-1}\alpha_a)(\alpha_a' X'_{-1} M_Z X_{-1}\alpha_a)^{-1}$, which is the Gaussian MLE. The corresponding Wald test statistic for H_o vs. H_a is

$$\begin{aligned} W &= [\mathrm{vec}(\hat{\delta}_a)]'[(\alpha_a' X'_{-1} M_Z X_{-1}\alpha_a)^{-1} \otimes \hat{\Sigma}_\varepsilon]^{-1}[\mathrm{vec}(\hat{\delta}_a)] \\ &= [\mathrm{vec}(\Delta X' M_Z X_{-1}\alpha_a)]'[(\alpha_a' X'_{-1} M_Z X_{-1}\alpha_a)^{-1} \otimes \hat{\Sigma}_\varepsilon^{-1}] \\ &\quad \times [\mathrm{vec}(\Delta X' M_Z X_{-1}\alpha_a)], \end{aligned} \tag{3.17}$$

where $\hat{\Sigma}_\varepsilon$ is the usual estimator value of Σ_ε ($\hat{\Sigma}_\varepsilon = T^{-1}\hat{\varepsilon}'\hat{\varepsilon}$, where $\hat{\varepsilon}$ is the matrix of OLS residuals from (3.16)), "vec" is the operator that stacks the column of a matrix, and the second line uses the result that $\mathrm{vec}(ABC) = (C' \times A)\,\mathrm{vec}(B)$ for conformable matrices A, B and C. The corresponding LR and LM statistics are asymptotically equivalent to W under the null and local alternatives.

The asymptotic null distribution of W is derived in Horvath and Watson (1993), where it is shown that

$$W \Rightarrow \mathrm{Trace}\left\{\left[\int B_1(s)\,\mathrm{d}B(s)'\right]'\left[\int B_1(s)B_1(s)'\,\mathrm{d}s\right]^{-1}\left[\int B_1(s)\,\mathrm{d}B(s)'\right]\right\}, \tag{3.18}$$

where $B(s)$ is an $n \times 1$ Wiener process partitioned into r_a and $n - r_a$ components $B_1(s)$ and $B_2(s)$, respectively. A proof of this result will not be offered here, but the form of the limiting distribution can be understood by considering a special case with $\Gamma = 0$ (so that there are no lags of Δx_t in the regression), $\Sigma_\varepsilon = I_n$ and $\alpha_a' = [I_{r_a} 0]$. In this case, x_t is a random walk with n.i.i.d.$(0, I_n)$ innovations, and (3.16) is the regression of Δx_t onto the first r_a elements of x_{t-1}, say $x_{1,t-1}$. Using the true value of Σ_ε, the Wald statistic in (3.17) simplifies to

$$\begin{aligned} W &= [\mathrm{vec}(\textstyle\sum \Delta x_t x'_{1,t-1})]'[(\sum x_{1,t-1}x'_{1,t-1})^{-1} \otimes I_n][\mathrm{vec}(\sum \Delta x_t x'_{1,t-1})]. \\ &= \mathrm{Trace}[(\textstyle\sum \Delta x_t x'_{1,t-1})(\sum x_{1,t-1}x'_{1,t-1})^{-1}(\sum x_{1,t-1}\Delta x_t')] \\ &= \mathrm{Trace}[(T^{-1}\textstyle\sum \Delta x_t x'_{1,t-1})(T^{-2}\sum x_{1,t-1}x'_{1,t-1})^{-1}(T^{-1}\sum x_{1,t-1}\Delta x_t')] \\ &\Rightarrow \mathrm{Trace}\left[\left(\int B_1(s)\,\mathrm{d}B(s)'\right)'\left(\int B_1(s)B_1(s)'\right)^{-1}\left(\int B_1(s)\,\mathrm{d}B(s)'\right)\right], \end{aligned}$$

where the second line uses the result that for square matrices, $\mathrm{Trace}(AB) = \mathrm{Trace}(BA)$, and for conformable matrices, $\mathrm{Trace}(ABCD) = [\mathrm{vec}(D)]'(A \times C')\,\mathrm{vec}(B')$ [Magnus and Neudecker (1988, page 30)], and the last line follows from Lemma 2.3. This verifies (3.18) for the example.

Testing $H_o: r = 0$ vs. $H_a: r = r_{a_u}$ When α_a is unknown, the Wald test in (3.17) cannot be calculated because the regressor $\alpha_a' X_{t-1}$ depends on unknown parameters. However, the LR statistic can be calculated, and useful formulae for the LR statistic are developed in Anderson (1951) (for the reduced rank regression model) and Johansen (1988a) (for the VECM). In the context of the VECM (3.3), Johansen (1988a) shows that the LR statistics can be written as

$$\text{LR} = -T\sum_{i=1}^{r_{a_u}} \ln(1-\gamma_i), \tag{3.19}$$

where γ_i are the ordered squared canonical correlations between Δx_t and x_{t-1}, after controlling for $\Delta x_{t-1}, \ldots, \Delta x_{t-p+1}$. These canonical correlations can be calculated as the eigenvalues of $T^{-1}S$, where $S = \tilde{\Sigma}_\varepsilon^{-1/2}(\Delta X' M_Z X_{-1})(X'_{-1} M_Z X_{-1})^{-1} \times (X'_{-1} M_Z \Delta X)(\tilde{\Sigma}_\varepsilon^{-1/2})'$, and where $\tilde{\Sigma}_\varepsilon = T^{-1}(\Delta X' M_Z \Delta X)$ is the estimated covariance matrix of ε_t, computed under the null [see Anderson (1984, Chapter 12) or Brillinger (1980, Chapter 10)]. Letting $\lambda_i(S)$ denote the eigenvalues of S ordered as $\lambda_1(S) \geq \lambda_2(S) \geq \cdots \geq \lambda_n(S)$, then γ_i from (3.19) is $\gamma_i = T^{-1}\lambda_i(S)$. Since the elements of S are $O_p(1)$ from Lemma 2.3, a Taylor series expansion of $\ln(1-\gamma_i)$ shows that the LR statistic can be written as

$$\text{LR} = \sum_{i=1}^{r_{a_u}} \lambda_i(S) + o_p(1). \tag{3.20}$$

Equation (3.20) shows why the LR statistic is sometimes called the "Maximal eigenvalue statistic" when $r_{a_u} = 1$ and the "Trace-statistic" when $r_{a_u} = n$ [Johansen and Juselius (1990)].[22]

One way to motivate the formula for the LR statistic given in (3.20), is by manipulating the Wald statistic in (3.17).[23] To see the relationship between LR and W in this case, let $L(\delta_a, \alpha_a)$ denote the log likelihood written as a function of δ_a and α_a, and let $\hat{\delta}_a(\alpha_a)$ denote the MLE of δ_a for fixed α_a. When Σ_ε is known, then the well known relation between the Wald and LR statistics in the linear regression model [Engle (1984)] implies that the Wald statistic can be written as

$$\begin{aligned} W(\alpha_a) &= 2[L(\hat{\delta}_a(\alpha_a), \alpha_a) - L(0, \alpha_a)] \\ &= 2[L(\hat{\delta}_a(\alpha_a), \alpha_a) - L(0, 0)], \end{aligned} \tag{3.21}$$

where the last line follows since α_a does not enter the likelihood when $\delta_a = 0$, and where $W(\alpha_a)$ is written to show the dependence of W on α_a. From (3.21), with Σ_ε

[22] In standard jargon, when $r_{o_u} \neq 0$, the trace statistic corresponds to the test for the alternative $r_{a_u} = n - r_{o_u}$.

[23] See Hansen (1990b) for a general discussion of the relationship between Wald, LR and LM tests in the presence of unidentified parameters.

known,

$$\begin{aligned}\operatorname{Sup}_{\alpha_a} W(\alpha_a) &= \operatorname{Sup}_{\alpha_a} 2[L(\hat{\delta}_a(\alpha_a), \alpha_a) - L(0,0)] \\ &= 2[L(\hat{\delta}_a, \hat{\alpha}_a) - L(0,0)] \\ &= \mathrm{LR} \end{aligned} \tag{3.22}$$

where the Sup is taken over all $n \times r_a$ matrices α_a. When Σ_ε is unknown, this equivalence is asymptotic, i.e. $\operatorname{Sup}_{\alpha_a} W(\alpha_a) = \mathrm{LR} + o_p(1)$.

To calculate $\operatorname{Sup}_{\alpha_a} W(\alpha_a)$, rewrite (3.17) as

$$\begin{aligned} W(\alpha_a) &= [\mathrm{vec}(\Delta X' M_Z X_{-1}\alpha_a)]'[(\alpha_a' X'_{-1} M_Z X_{-1}\alpha_a)^{-1} \otimes \hat{\Sigma}_\varepsilon^{-1}] \\ &\quad \times [\mathrm{vec}(\Delta X' M_Z X_{-1}\alpha_a)] \\ &= \mathrm{TR}[\hat{\Sigma}_\varepsilon^{-1/2}(\Delta X' M_Z X_{-1}\alpha_a)(\alpha_a' X'_{-1} M_Z X_{-1}\alpha_a)^{-1} \\ &\quad \times (\alpha_a' X'_{-1} M_Z \Delta X)(\hat{\Sigma}_\varepsilon^{-1/2})'] \\ &= \mathrm{TR}[\hat{\Sigma}_\varepsilon^{-1/2}(\Delta X' M_Z X_{-1}) DD' (X'_{-1} M_Z \Delta X)(\hat{\Sigma}_\varepsilon^{-1/2})'], \quad \text{where} \\ &\qquad\qquad D = \alpha_a(\alpha_a' X'_{-1} M_Z X_{-1}\alpha_a)^{-1/2} \\ &= \mathrm{TR}[D'(X'_{-1} M_Z \Delta X)\hat{\Sigma}_\varepsilon^{-1}(\Delta X' M_Z X_{-1}) D] \\ &= \mathrm{TR}[F'CC'F], \end{aligned} \tag{3.23}$$

where $F = (X'_{-1} M_Z X_{-1})^{1/2}\alpha_a(\alpha_a' X'_{-1} M_Z X_{-1}\alpha_a)^{-1/2}$, and $C = (X'_{-1} M_Z X_{-1})^{-1/2} \times (X'_{-1} M_Z \Delta X)\hat{\Sigma}_\varepsilon^{-1/2\prime}$. Since $F'F = I_{r_{a_u}}$,

$$\begin{aligned}\operatorname{Sup}_{\alpha_a} W(\alpha_a) &= \operatorname{Sup}_{F'F = I} \mathrm{TR}[F'(CC')F] = \sum_{i=1}^{r_{a_u}} \lambda_i(CC') = \sum_{i=1}^{r_{a_u}} \lambda_i(C'C) \\ &= \mathrm{LR} + o_p(1), \end{aligned} \tag{3.24}$$

where $\lambda_i(CC')$ denote the ordered eigenvalues of (CC'), and the final two equalities follow from the standard principal components argument [for example, see Theil (1971, page 46)] and $\lambda_i(CC') = \lambda_i(C'C)$. Equation (3.24) shows that the likelihood ratio statistic can then be calculated (up to an $o_p(1)$ term) as the largest r_a eigenvalues of

$$C'C = \hat{\Sigma}_\varepsilon^{-1/2}(\Delta X' M_Z X_{-1})(X'_{-1} M_Z X_{-1})^{-1}(X'_{-1} M_Z \Delta X)(\hat{\Sigma}_\varepsilon^{-1/2})'.$$

To see the relationship between the formulae for the LR statistics in (3.24) and (3.20), notice that $C'C$ in (3.24) and S in (3.20) differ only in the estimator of Σ_ε; $C'C$ uses an estimator constructed from residuals calculated under the alternative, while S uses an estimator constructed from residuals calculated under the null.

In general settings, it is not possible to derive a simple representation for the asymptotic distribution of the Likelihood Ratio statistic when some parameters are present only under the alternative. However, the special structure of the VECM makes such a simple representation possible. Johansen (1988a) shows that the LR statistic has the limiting asymptotic null distribution given by

$$\mathrm{LR} \Rightarrow \sum_{i=1}^{r_{a_u}} \lambda_i(H) \tag{3.25}$$

where $H = [\int B(s)\mathrm{d}B(s)']'[\int B(s)B(s)'\mathrm{d}s]^{-1}[\int B(s)\mathrm{d}B(s)']$, and $B(s)$ is an $n \times 1$ Wiener process. To understand Johansen's result, again consider the special case with $\Gamma = 0$ and $\Sigma_\varepsilon = I_n$. In this case, $C'C$ becomes

$$\begin{aligned} C'C &= (\Delta X' X_{-1})(X'_{-1}X_{-1})^{-1}(X'_{-1}\Delta X) \\ &= [\textstyle\sum \Delta x_t x'_{t-1}][\sum x_{t-1}x'_{t-1}]^{-1}[\sum x_{t-1}\Delta x'_t] \\ &= [T^{-1}\textstyle\sum \Delta x_t x'_{t-1}]'[T^{-2}\sum x_{t-1}x'_{t-1}]^{-1}[T^{-1}\sum x_{t-1}\Delta x'_t] \\ &\Rightarrow \left[\int B(s)\mathrm{d}B(s)'\right]'\left[\int B(s)B(s)'\mathrm{d}s\right]^{-1}\left[\int B(s)\mathrm{d}B(s)'\right] \end{aligned} \tag{3.26}$$

from Lemma 2.3. This verifies (3.25) for the example.

Testing $H_o: r = 0$ *vs.* $H_a: r_a = r_{a_k} + r_{a_u}$ The model is now

$$\Delta X_t = \delta_{a_k}(\alpha'_{a_k}X_{t-1}) + \delta_{a_u}(\alpha'_{a_u}X_{t-1}) + \beta Z_t + \varepsilon_t, \tag{3.27}$$

where α_a has been partitioned so that α_{a_k} contains the r_{a_k} known cointegrating vectors, α_{a_u} contains the r_{a_u} unknown cointegrating vectors and δ_a has been partitioned conformably as $\delta = (\delta_{a_k}\ \delta_{a_u})$. As above, the LR statistic can be approximated up to an $o_p(1)$ term by maximizing the Wald statistic over the unknown parameters in α_{a_u}. Let $M_{zk} = M_z - M_z X_{-1}\alpha_{a_k}(\alpha'_{a_k}X'_{-1}M_z X_{-1}\alpha_{a_k})^{-1}\alpha'_{a_k}X'_{-1}M_z$ denote the matrix that partials both Z and $X_{-1}\alpha_{a_k}$ out of the regression (3.27), The Wald statistic (as a function of α_{a_k} and α_{a_u}) can then be written as[24]

$$\begin{aligned} W(\alpha_{a_k}, \alpha_{a_u}) = {} & [\mathrm{vec}(\Delta X' M_z X_{-1}\alpha_{a_k})]'[(\alpha'_{a_k}X'_{-1}M_z X_{-1}\alpha_{a_k})^{-1} \otimes \hat{\Sigma}_\varepsilon^{-1}] \\ & \times [\mathrm{vec}(\Delta X' M_z X_{-1}\alpha_{a_k})] \\ & \times [\mathrm{vec}(\Delta X' M_{zk} X_{-1}\alpha_{a_u})]'[(\alpha'_{a_u}X'_{-1}M_{zk}X_{-1}\alpha_{a_u})^{-1} \otimes \hat{\Sigma}_\varepsilon^{-1}] \\ & \times [\mathrm{vec}(\Delta X' M_{zk} X_{-1}\alpha_{a_u})]. \end{aligned} \tag{3.28}$$

[24]The first term in (3.28) is the Wald statistic for testing $\delta_{a_k} = 0$ imposing the constraint that $\delta_{a_u} = 0$. The second term is the Wald statistic for testing $\delta_{a_u} = 0$ with $\alpha'_{a_k}X_{t-1}$ and Z_t partialled out of the regression. This form of the Wald statistic can be deduced from the partitioned inverse formula.

The first term is identical to equation (3.17) above, and the second term has the same form except that both Z and $X_{-1}\alpha_{a_k}$ have been partialled out of the regression. We can derive the LR statistic as above with one modification: when maximizing $W(\alpha_{a_k}, \alpha_{a_u})$ over the unknown cointegrating vectors in α_{a_u}, attention can be restricted to cointegrating vectors that are linearly independent of α_{a_k}. Thus, the LR statistic is obtained by maximizing (3.28) over all $n \times r_{a_u}$ matrices α_{a_u} satisfying $\alpha'_{a_u}\alpha_{a_k} = 0$. Let G denote an (arbitrary) $n \times (n - r_{a_k})$ matrix whose columns span the null space of the columns of α_{a_k}. Then α_{a_u} can be written as a linear combination of the columns of G, so that $\alpha_{a_u} = G\tilde{\alpha}_{a_u}$, where $\tilde{\alpha}_{a_u}$ is an $(n - r_{a_k}) \times r_{a_u}$ matrix, so that $\alpha'_{a_u}\alpha_{a_k} = \tilde{\alpha}'_{a_u}G'\alpha_{a_k} = 0$ for all $\tilde{\alpha}_{a_u}$. Substituting $G\tilde{\alpha}_{a_u}$ into (3.28) and carrying out the maximization yields

$$\begin{aligned}\operatorname*{Sup}_{\alpha_{a_u}} W(\alpha_{a_k}, \alpha_{a_u}) &= [\operatorname{vec}(\Delta X' M_z X_{-1}\alpha_{a_k})]'[(\alpha'_{a_k}X'_{-1}M_z X_{-1}\alpha_{a_k})^{-1} \otimes \hat{\Sigma}_\varepsilon^{-1}] \\ &\quad \times [\operatorname{vec}(\Delta X' M_z X_{-1}\alpha_{a_k})] + \sum_{i=1}^{r_{a_u}} \lambda_i(H'H) \\ &= \mathrm{LR} + o_p(1), \end{aligned} \tag{3.29}$$

where $H'H = \hat{\Sigma}_\varepsilon^{-1/2}(\Delta X' M_{zk} X_{-1} G)(G'X'_{-1}M_{zk}X_{-1}G)^{-1}(G'X'_{-1}M_{zk}\Delta X)(\hat{\Sigma}_\varepsilon^{-1/2})'$.

The statistic is calculated as follows. Regress ΔX onto $\alpha'_{a_k}X_{-1}$ and Z and form the usual Wald statistic. This is the first term on the right hand side of (3.29). Let G be an arbitrary matrix whose columns span the null space of the columns of α_{a_k}. (The columns of G can be formed in a number of ways, for example using the Gram–Schmidt orthogonalization procedure.) The second term on the right hand side of (3.29) is the sum of the r_{a_u} largest eigenvalues of $\hat{\Sigma}_\varepsilon^{-1/2}(\Delta X' M_{zk} X_{-1} G) \times (G'X'_{-1}M_{zk}X_{-1}G)^{-1}(G'X_{-1}M_{zk}\Delta X)(\hat{\Sigma}_\varepsilon^{-1/2})'$. Σ_ε can be replaced by any consistent estimator of Σ_ε without affecting the large sample behavior of the statistic. Two particularly simple estimators are $\tilde{\Sigma}_\varepsilon = T^{-1}\Delta X' M_z \Delta X$ and the residual covariance matrix from the regression of X_t onto p lagged levels of X_t.

The asymptotic null distribution of the LR statistic in (3.29) is derived in Horvath and Watson (1993). They show

$$\begin{aligned}\mathrm{LR} \Rightarrow \operatorname{Trace}\left\{\left[\int B_1(s)\,\mathrm{d}B(s)'\right]'\left[\int B_1(s)B_1(s)'\,\mathrm{d}s\right]^{-1}\left[\int B_1(s)\,\mathrm{d}B(s)'\right]\right\} \\ + \sum_{i=1}^{r_{a_u}} \lambda_i(\tilde{C}'\tilde{C}), \end{aligned} \tag{3.30}$$

where $B(s)$ is an $(n \times 1)$ Wiener process partitioned into r_{a_k} and $n - r_{a_k}$ components $B_1(s)$ and $B_2(s)$, respectively, $\tilde{C}'\tilde{C} = [\int \tilde{B}_2(s)\mathrm{d}B(s)']'[\int \tilde{B}_2(s)\tilde{B}_2(s)']^{-1}[\int \tilde{B}_2(s)\mathrm{d}B(s)']$, and $\tilde{B}_2(s)$ is the residual from the regression of $B_2(s)$ onto $B_1(s)$, i.e. $\tilde{B}_2(s) = B_2(s) - \gamma B_1(s)$, where $\gamma = \int B_2(s)B_1(s)'[\int B_1(s)B_1(s)']^{-1}$.[25]

[25] This compact way of writing the limiting distributions, using projections of Wiener processes, is taken from Park and Phillips (1988, 1989).

As pointed out above, when the null hypothesis is $H_o: r = r_{o_k} + r_{o_u}$, the LR test statistic can be calculated as the difference between the LR statistics for $[H_o: r = 0$ vs. $H_a: r = r_o + r_a]$ and $[H_o: r = 0$ vs. $H_a: r = r_o]$. So, for example, when testing $H_o: r = r_{o_u}$ vs. $H_a: r = r_{o_u} + r_{a_u}$, the LR statistic is

$$\text{LR} = -T \sum_{i=r_{o_u}+1}^{r_{a_u}} \ln(1-\gamma_i) = \sum_{i=r_{o_u}+1}^{r_{a_u}} \lambda_i(S) + o_p(1), \tag{3.31}$$

where γ_i are the canonical correlations defined below equation (3.19) [see Anderson (1951) and Johansen (1988a)]. Critical values for the case $r_{o_k} = r_{a_k} = 0$ and $n - r_{o_u} \leqslant 5$ are given in Johansen (1988a) for the trace-statistic (so that the alternative is $r_{a_u} = n - r_{o_u}$); these are extended for $n - r_{o_u} \leqslant 11$ in Osterwald-Lenum (1992), who also tabulates asymptotic critical values for the maximal eigenvalue statistic (so that $r_{o_k} = r_{a_k} = 0$ and $r_{a_u} = 1$). Finally, asymptotic critical values for all combinations of $r_{o_k}, r_{o_u}, r_{a_k}$ and r_{a_u} with $n - r_{o_u} \leqslant 9$ are tabulated in Horvath and Watson (1992).

3.3.2. *Non-likelihood-based approaches*

In addition to the likelihood based tests discussed in the last section, standard univariate unit root tests and their multivariate generalizations have also been used as tests for cointegration. To see why these tests are useful, consider the hypotheses $H_o: r = 0$ vs. $H_a: r = 1$, and suppose that α is known under the alternative. Since the data are not cointegrated under the null, $w_t = \alpha' x_t$ is I(1), while under the alternative it is I(0). Thus, cointegration can be tested by applying a standard unit root test to the univariate series w_t. To be useful in more general cointegrated models, standard unit root tests have been modified in two ways. First, modifications have been proposed so that the tests can be applied when α is unknown. Second, multivariate unit root tests have been developed for the general testing problem $H_o: r = r_o$ vs. $H_a: r = r_o + r_a$. We discuss these two modifications in turn.

Engle and Granger (1987) develop a test for the hypotheses $H_o: r = 0$ vs. $H_a: r = 1$ when α is unknown. They suggest using OLS to estimate the single cointegrating vector and applying a standard unit root test (they suggest an augmented Dickey–Fuller t-test) to the OLS residuals, $\hat{w}_t = \hat{\alpha} x_t$. Under the alternative, $\hat{\alpha}$ is a consistent estimator of α, so that $\hat{w}_t$ will behave like w_t. However, under the null, $\hat{\alpha}$ is obtained from a "spurious" regression (see Section 2.6.3) and the residuals from a spurious regression ($\hat{w}_t$) behave differently than non-stochastic linear combinations of I(1) variables (w_t). This affects the null distribution of unit root statistics calculated using $\hat{w}_t$. For example, the Dickey–Fuller t-statistic constructed using $\hat{w}_t$ has a different null distribution than the statistic calculated using w_t, so that the usual critical values given in Fuller (1976) cannot be used for the Engle–Granger test. The correct asymptotic null distribution of the statistic is derived in Phillips and Ouliaris (1990), and is tabulated in Engle and Yoo (1987) and MacKinnon (1991). Hansen

(1990a) proposes a modification of the Engle–Granger test that is based on an iterated Cochrane–Orcutt estimator which eliminates the "spurious regression" problem and results in test statistics with standard Dickey–Fuller asymptotic distributions under the null.

Stock and Watson (1988), building on work by Fountis and Dickey (1986), propose a multivariate unit root test. Their procedure is most easily described by considering the VAR(1) model, $x_t = \Phi x_{t-1} + \varepsilon_t$, together with the hypotheses H_o: $r = 0$ vs. H_a: $r = r_a$. Under the null the data are not cointegrated, so that $\Phi = I_n$. Under the alternative there are r_a covariance stationary linear combinations of the data, so that Φ has r_a eigenvalues that are less than one in modulus. The Stock–Watson test is based on the ordered eigenvalues of $\hat{\Phi}$, the OLS estimator of Φ. Writing these eigenvalues as $|\hat{\lambda}_1| \geq |\hat{\lambda}_2| \geq \cdots$, the test is based on $\hat{\lambda}_{n-r_a+1}$, the r_ath smallest eigenvalue. Under the null, $\lambda_{n-r_a+1} = 1$, while under the alternative, $|\lambda_{n-r_a+1}| < 1$. The asymptotic null distribution of $T(\hat{\Phi} - I)$ and $T(|\hat{\lambda}_i| - 1)$ are derived in Stock and Watson (1988), and critical values for $T(|\hat{\lambda}_{n-r_a+1}| - 1)$ are tabulated. This paper also develops the required modifications for testing in a general VAR(p) model with $r_o \neq 0$.

3.3.3. *Comparison of the tests*

The tests discussed above differ from one another in two important respects. First, some of the tests are constructed using the true value of the cointegrating vectors under the alternative, while others estimate the cointegrating vectors. Second, the likelihood based tests focus their attention on δ in (3.3), while the non-likelihood-based tests focus on the serial correlation properties of certain linear combinations of the data. Of course, knowledge of the cointegrating vectors, if available, will increase the power of the tests. The relative power of tests that focus on δ and tests that focus on the serial correlation properties of $w_t = \alpha' x_t$ is less clear.

Some insight can be obtained by considering a special case of the VECM (3.3)

$$\Delta x_t = \delta_a(\alpha_a' x_{t-1}) + \varepsilon_t. \tag{3.32}$$

Suppose that α_a is known and that the competing hypotheses are H_o: $r = 0$ vs. H_a: $r = 1$. Multiplying both sides of (3.31) by α_a' yields

$$\Delta w_t = \theta w_{t-1} + e_t, \tag{3.33}$$

where $w_t = \alpha_a' x_t$, $\theta = \alpha_a' \delta_a$ and $e_t = \alpha_a' \varepsilon_t$. Unit root tests constructed from w_t test the hypotheses H_o: $\theta = (\alpha_a' \delta_a) = 0$ vs. H_a: $\theta = (\alpha_a' \delta_a) < 0$, while the VECM-based LR and Wald statistics test H_o: $\delta_a = 0$ vs. H_a: $\delta_a \neq 0$. Thus, unit root tests constructed from w_t focus on departures from the $\delta_a = 0$ null in the direction of the cointegrating vector α_a. In contrast, the VECM likelihood based tests are invariant to transformations of the form $P\alpha_a' x_{t-1}$ when α_a is known and Px_{t-1} when α_a is unknown,

where P is an arbitrary nonsingular matrix. Thus, the likelihood based tests aren't focused in a single direction like the univariate unit root test. This suggests that tests based on w_t should perform relatively well for departures in the direction of α_a, but relatively poorly in other directions. As an extreme case, when $\alpha_a'\delta_a = 0$, the elements of x_t are I(2) and w_t is I(1). [The system is $CI(2,1)$ in Engle and Granger's (1987) notation.] The elements are still cointegrated, at least in the sense that a particular linear combination of the variables is less persistent than the individual elements of x_t, and this form of cointegration can be detected by a nonzero value of δ_a in equation (3.32) even though $\theta = 0$ in (3.33).[26]

A systematic comparison of the power properties of the various tests will not be carried out here, but one simple Monte Carlo experiment, taken from a set of experiments in Horvath and Watson (1993), highlights the power tradeoffs. Consider a bivariate model of the form given in (3.32) with $\varepsilon_t \sim$ n.i.i.d.$(0, I_2)$, $\alpha_a = (1 \; -1)'$ and $\delta_a = (\delta_{a_1} \; \delta_{a_2})'$. This design implies that $\theta = \delta_{a_1} - \delta_{a_2}$ in (3.33), so that the unit root tests should perform reasonably well when $|\delta_{a_1} - \delta_{a_2}|$ is large and reasonably poorly when $|\delta_{a_1} - \delta_{a_2}|$ is small. Changes in δ_a have two effects on the power of the VECM likelihood based tests. In the classical multivariate regression model, the power of the likelihood based tests increase with $\zeta = \delta_{a_1}^2 + \delta_{a_2}^2$. However, in the VECM, changes in δ_{a_1} and δ_{a_2} also affect the serial correlation properties of the regressor, $w_{t-1} = \alpha' x_{t-1}$, as well as ζ. Indeed, for this design, $w_t \sim$ AR(1) with

Table 1
Comparing power of tests for cointegration.[a]
Size for 5 percent asymptotic critical values and power for tests carried out at 5 percent level[b]

	$(\delta_{a_1}, \delta_{a_2})$			
Test	(0,0)	(0.05, 0.055)	(−0.05, 0.055)	(−0.0105, 0)
DF (α known)	5.0	6.5	81.5	81.9
EG–DF (α unknown)	4.7	2.9	31.9	32.5
Wald (α known)	4.7	95.0	54.2	91.5
LR (α unknown)	4.4	86.1	20.8	60.7

[a]Design is

$$\begin{bmatrix} x_t^1 \\ x_t^2 \end{bmatrix} = \begin{bmatrix} \delta_{a_1} \\ \delta_{a_2} \end{bmatrix} [x_t^1 - x_t^2] + \begin{bmatrix} \varepsilon_t^1 \\ \varepsilon_t^2 \end{bmatrix},$$

where $\varepsilon_t = (\varepsilon_t^1 \varepsilon_t^2)' \sim$ n.i.i.d.$(0, I_2)$ and $t = 1, \ldots, 100$.

[b]These results are based on 10,000 replications. The first column shows rejection frequencies using asymptotic critical values. The other columns show rejection frequencies using 5 percent critical values calculated from the experiment in column 1.

[26]This example was pointed out to me by T. Rothenberg.

AR coefficient $\theta = \delta_{a_1} - \delta_{a_2}$ [see equation (3.33)]. Increases in θ lead to increases in the variability of the regressor and increases in the power of the test.

Table 1 shows size and power for four different values of δ_a when $T = 100$ in this bivariate system. Four tests are considered: (1) the Dickey–Fuller (DF) t-test using the true value of α; (2) the Engle–Granger test (EG–DF, the Dickey–Fuller t-test using a value of α estimated by OLS); (3) the Wald statistic for $H_o: \delta_a = 0$ using the true value of α; and (4) the LR statistic for $H_o: \delta_a = 0$ for unknown α.

The table contains several noteworthy results. First, for this simple design, the size of the tests is close to the size predicted by asymptotic theory. Second, as expected, the DF and EG–DF tests perform quite poorly when $|\delta_{a_1} - \delta_{a_2}|$ is small. Third, increasing the serial correlation in $w_t = \alpha_a' x_t$, while holding $\delta_{a_1}^2 + \delta_{a_2}^2$ constant, increases the power of the likelihood based tests. [This can be seen by comparing the $\delta_a = (0.05, 0.055)$ and $\delta_a = (-0.05, 0.055)$ columns.] Fourth, increasing $\delta_{a_1}^2 + \delta_{a_2}^2$, while holding the serial correlation in w_t constant, increases the power of the likelihood based tests. [This can be seen by comparing the $\delta_a = (-0.05, 0.055)$ and $\delta_a = (-0.105, 0.00)$ columns.] Fifth, when the DF and EG–DF are focused on the correct direction, their power exceeds the likelihood based tests. [This can be seen from the $\delta_a = (-0.05, 0.055)$ column.] Finally, there is a gain in power from incorporating the true value of the cointegrating vector. (This can be seen by comparing the DF test to the EG–DF test and the Wald test to the LR test.) A more thorough comparison of the tests is contained in Horvath and Watson (1993).

3.4. *Estimating cointegrating vectors*

3.4.1. *Gaussian maximum likelihood estimation (MLE) based on the triangular representation*

In Section 3.2.4 the triangular representation of the cointegrated system was written as

$$\Delta x_{1,t} = u_{1,t}, \tag{3.9}$$

$$x_{2,t} - \beta x_{1,t} = u_{2,t}, \tag{3.10}$$

where $u_t = D(L)\varepsilon_t$. In this section we discuss the MLE estimator of β under the assumption that $\varepsilon_t \sim$ n.i.i.d.$(0, I)$. The n.i.i.d. assumption is used only to motivate the Gaussian MLE. The asymptotic distribution of estimators that are derived below follow under the weaker distributional assumptions listed in Lemma 2.3. In Section 2.6.4 we considered the OLS estimator of β in a bivariate model, and paid particular attention to the distribution of the estimator when $D(L) = D$ with $d_{12} = d_{21} = 0$. In this case, $x_{1,t}$ is weakly exogenous for β and the MLE estimator corresponds

to the OLS estimator. Recall (see Section 2.6.4) that when $d_{12} = d_{21} = 0$, the OLS estimator of β has an asymptotic distribution that can be represented as a variance mixture of normals and that the t-statistic for $\hat{\beta}$ has an asymptotic null distribution that is standard normal. This means that tests concerning the valuc of β and confidence intervals for β can be constructed in the usual way; complications from the unit roots in the system can be ignored. These results carry over immediately to the vector case where $x_{1,t}$ is $k \times 1$ and $x_{2,t}$ is $r \times 1$ when u_t is serially uncorrelated and $u_{1,t}$ is independent of $u_{2,t}$. Somewhat surprisingly, they also carry over to the MLE of β in the general model with $u_t = D(L)\varepsilon_t$, so that the errors are both serially and cross correlated.

Intuition for this result can be developed by considering the static model with $u_t = D\varepsilon_t$ and D is not necessarily block diagonal. Since $u_{1,t}$ and $u_{2,t}$ are correlated, the MLE of β corresponds to the seemingly unrelated regression (SUR) estimator from (3.9)–(3.10). But, since there are no unknown regression coefficients in (3.9), the SUR estimator can be calculated by OLS in the regression

$$x_{2,t} = \beta x_{1,t} + \gamma \Delta x_{1,t} + e_{2,t}, \tag{3.34}$$

where γ is the coefficient from the regression of $u_{2,t}$ onto $u_{1,t}$, and $e_{2,t} = u_{2,t} - E[u_{2,t}|u_{1,t}]$ is the residual from this regression. By construction, $e_{2,t}$ is independent of $\{x_{1,\tau}\}_{\tau=1}^{T}$ for all t. Moreover, since γ is a coefficient on a zero mean stationary regressor and β is a coefficient on a martingale, the limiting scaled "$X'X$" matrix from the regression is block diagonal (Section 2.5.1). Thus from Lemma 2.3,

$$\begin{aligned} T(\hat{\beta} - \beta) &= (T^{-1}\textstyle\sum e_{2,t}x'_{1,t})(T^{-2}\sum x_{1,t}x'_{1,t})^{-1} + o_p(1) \\ &\Rightarrow \left(\Sigma_{u_1}^{1/2}\int B_1(s)\mathrm{d}B_2(s)'\Sigma_{e_2}^{1/2}\right)\left(\Sigma_{u_1}^{1/2}\int B_1(s)B_1(s)'\mathrm{d}s(\Sigma_{u_1}^{1/2})'\right)^{-1}, \end{aligned} \tag{3.35}$$

where $\Sigma_{u_1} = \mathrm{var}(u_{1,t})$, $\Sigma_{e_2} = \mathrm{var}(e_{2,t})$ and $B(s)$ is an $n \times 1$ Brownian motion process, partitioned as $B(s) = [B_1(s)'\ B_2(s)']'$, where $B_1(s)$ is $k \times 1$ and $B_2(s)$ is $r \times 1$. Except for the change in scale factors and dimensions, equation (3.35) has the same form as (2.21), the asymptotic distribution of $\hat{\beta}$ in the case $d_{12} = d_{21} = 0$. Thus, the asymptotic distribution of $\hat{\beta}$ can be represented as a variance mixture of normals. Moreover, the same conditioning argument used when $d_{12} = d_{21} = 0$ implies that the asymptotic distribution of Wald test statistics concerning β have their usual large sample χ^2 distribution. Thus, inference about β can be carried out using standard procedures and standard distributions.

Now suppose that $u_t = D(L)\varepsilon_t$. The dynamic analogue of (3.34) is

$$x_{2,t} = \beta x_{1,t} + \gamma(L)\Delta x_{1,t} + e_{2,t}, \tag{3.36}$$

where $\gamma(L)\Delta x_{1,t} = E[u_{2,t}|\{\Delta x_{1,\tau}\}_{\tau=1}^{T}] = E[u_{2,t}|\{u_{1,\tau}\}_{\tau=1}^{T}]$, and $e_{2,t} = u_{2,t} - E[u_{2,t}|\{u_{1,\tau}\}_{\tau=1}^{T}]$. Letting $D_1(L)$ denote the first k rows of $D(L)$ and $D_2(L)$ denote

the remaining r rows, then from classical projection formulae [e.g. Whittle (1983, Chapter 5)], $\gamma(L) = D_2(L)D_1(L)'[D_1(L)D_1(L^{-1})']^{-1}$.[27] Equation (3.36) differs from (3.34) in two ways. First, there is potential serial correlation in the error term of (3.36), and second, $\gamma(L)$ in (3.36) is a two-sided polynomial. These differences complicate the estimation process.

To focus on the first complication, assume that $\gamma(L) = 0$. In this case, (3.36) is a regression model with a serially correlated error, so that (asymptotically) the MLE of β is just the feasible GLS estimator in (3.36). But, as shown in Phillips and Park (1988), the GLS correction has no effect on the asymptotic distribution of the estimator: the OLS estimator and GLS estimators of β in (3.17) are asymptotically equivalent.[28] Since the regression error $e_{2,t}$ and the regressors $\{x_{1,\tau}\}_{\tau=1}^{T}$ are independent, by analogy with the serially uncorrelated case, $T(\hat{\beta} - \beta)$ will have an asymptotic distribution that can be represented as a variance mixture of normals. Indeed, the distribution will be exactly of the form (3.35), where now Σ_{u_1} and Σ_{e_2} represent "long-run" covariance matrices.[29]

Using conditioning arguments like those used in Section 2.6.4, it is straightforward to show that the Wald test statistics constructed from the GLS estimators of β have large sample χ^2 distributions. However, since the errors in (3.36) are serially correlated, the usual estimator of the covariance matrix for the OLS estimators of β is inappropriate, and a serial correlation robust covariance matrix is required.[30] Wald test statistics constructed from OLS estimators of β together with serial correlation robust estimators of covariance matrices will be asymptotically χ^2 and

[27] This is the formula for the projection onto the infinite sample, i.e. $\gamma(L)\Delta x_t^1 = E[u_t^2 | \{\Delta x_\tau^1\}_{\tau=-\infty}^{\infty}]$. In general, $\gamma(L)$ is two-sided and of infinite order, so that this is an approximation to $E[u_t^2 | \{\Delta x_t^1\}_{\tau=1}^{T}]$. The effect of this approximation error on estimators of β is discussed below.

[28] This can be demonstrated as follows. When $\gamma(L) = 0$, $e_{2,t} = D_{22}(L)\varepsilon_{2,t}$ and $x_{1,t} = D_{11}(L)\varepsilon_{1,t}$ where $\varepsilon_{1,t}$ and $\varepsilon_{2,t}$ are the first k and last r elements of ε_t, and $D_{11}(L)$ and $D_{22}(L)$ are the appropriate diagonal blocks of $D(L)$. Let $C(L) = [D_{22}(L)]^{-1}$ and assume that the matrix coefficients in $C(L)$, $D_{11}(L)$ and $D_{22}(L)$ are 1-summable. Letting $\delta = \text{vec}(\beta)$, the GLS estimator and OLS estimators satisfy

$$T(\hat{\delta}_{\text{OLS}} - \delta) = (T^{-2}\sum q_t q_t')^{-1}(T^{-1}\sum q_t e_{2,t}),$$
$$T(\hat{\delta}_{\text{GLS}} - \delta) = (T^{-2}\sum \tilde{q}_t \tilde{q}_t')^{-1}(T^{-1}\sum \tilde{q}_t \varepsilon_{2,t}),$$

where $q_t = [x_{1,t} \otimes I_r]$, and defining the operator L so that $z_t L = L z_t = z_{t-1}$, $\tilde{q}_t = [x_{1,t} \otimes C(L)']$. Using the Lemma 2.3

$$\begin{aligned} T(\hat{\delta}_{\text{OLS}} - \delta) &= [T^{-2}\textstyle\sum x_{1,t}x_{1,t}' \otimes I_r]^{-1}[T^{-1}\sum (x_{1,t}' \otimes I_r)D_{22}(L)\varepsilon_{2,t}] \\ &= [T^{-2}\textstyle\sum x_{1,t}x_{1,t}' \otimes I_r]^{-1}[T^{-1}\sum (x_{1,t}' \otimes D_{22}(1)\varepsilon_{2,t})] + o_p(1), \\ T(\hat{\delta}_{\text{GLS}} - \delta) &= [T^{-2}\textstyle\sum \{C(L)x_{1,t}\}\{x_{1,t}'C(L)'\} \otimes I_r]^{-1}[T^{-1}\sum (x_{1,t} \otimes C(L)')\varepsilon_{2,t}] \\ &= [T^{-2}\textstyle\sum x_{1,t}x_{1,t}' \otimes C(1)'C(1)]^{-1}[T^{-1}\sum (x_{1,t} \otimes C(1)'\varepsilon_{2,t})] + o_p(1). \end{aligned}$$

Since $C(1)^{-1} = D_{22}(1)$, $T(\hat{\delta}_{\text{OLS}} - \delta) = T(\hat{\delta}_{\text{GLS}} - \delta) + o_p(1)$, so that $T(\hat{\delta}_{\text{OLS}} - \hat{\delta}_{\text{GLS}}) \xrightarrow{p} 0$.

[29] The long-run covariance matrix for an $n \times 1$ covariance stationary vector y_t with absolutely summable autocovariances is $\sum_{i=-\infty}^{\infty} \text{Cov}(y_t, y_{t-i})$, which is 2π times the spectral density matrix for y_t at frequency zero.

[30] See Wooldridge's chapter of the Handbook for a thorough discussion of robust covariance matrix estimators.

will be asymptotically equivalent to the statistics calculated using the GLS estimators of β [Phillips and Park (1988)]. In summary, the serial correlation in (3.36) poses no major obstacles.

The two-sided polynomial $\gamma(L)$ poses more of a problem, and three different solutions have developed. In the first approach, $\gamma(L)$ is approximated by a finite order (two-sided) polynomial.[31] In this case, equation (3.36) can be estimated by GLS, yielding what Stock and Watson (1993) call the "Dynamic GLS" estimator of β. Alternatively, utilizing the Phillips and Park (1988) result, an asymptotically equivalent "Dynamic OLS" estimator can be constructed by applying OLS to (3.36).

To motivate the second approach, assume for a moment that $\gamma(L)$ were known. The OLS estimator of β would then be formed by regressing $x_{2,t} - \gamma(L)\Delta x_{1,t}$ onto $x_{1,t}$. But $T^{-1}\sum[\gamma(L)\Delta x_{1,t}]x_{1,t} = T^{-1}\sum[\gamma(1)\Delta x_{1,t}]x_{1,t} + B + o_p(1)$, where $B = \lim_{t\to\infty} E(y_t x_{1,t})$, where $y_t = [\gamma(L) - \gamma(1)]\Delta x_{1,t}$. (This can be verified using (c) and (d) of Lemma 2.3.) Thus, an asymptotically equivalent estimator can be constructed by regressing $x_{2,t} - \gamma(1)\Delta x_{1,t}$ onto $x_{1,t}$ and correcting for the "bias" term B. Park's (1992) "Canonical Cointegrating Regression" estimator and Phillips and Hansen's (1990) "Fully Modified" estimator use this approach, where in both cases $\gamma(1)$ and B are replaced by consistent estimators.

The final approach is motivated by the observation that the low frequency movements in the data asymptotically dominate the estimator of β. Phillips (1991b) demonstrates that an efficient band spectrum regression, concentrating on frequency zero, can be used to calculate an estimator asymptotically equivalent to the MLE estimator in (3.36).[32]

All of these suggestions lead to asymptotically equivalent estimators. The estimators have asymptotic representations of the form (3.35) (where Σ_{u_1} and Σ_{e_2} represent long-run covariance matrices), and thus their distributions can be represented as variance mixtures of normals. Wald test statistics computed using the estimators (and serial correlation robust matrices) have the usual large sample χ^2 distributions under the null.

3.4.2. *Gaussian maximum likelihood estimation based on the VECM*

Most of the empirical work with cointegrated systems has utilized parameterizations based on the finite order VECM representation shown in equation (3.3). Exact MLEs calculated from the finite order VECM representation of the model are different from the exact MLE's calculated from the triangular representations that were developed in the last section. The reason is that the VECM imposes constraints on the coefficients in $\gamma(L)$ and the serial correlation properties of $e_{2,t}$ in (3.36).

[31] This suggestion can be found in papers by Hansen (1988), Phillips (1991a), Phillips and Loretan (1991), Saikkonen (1991) and Stock and Watson (1993). Saikkonen (1991) contains a careful discussion of the approximation error that arises when $\gamma(L)$ is approximated by a finite order polynomial. Using results of Berk (1974) and Said and Dickey (1984) he shows that a consistent estimator of $\gamma(1)$ (which, as we show below is required for an asymptotically efficient estimator of β) obtains if the order of the estimated polynomial $\gamma(L)$ increases at rate T^{δ} for $0 < \delta < \frac{1}{3}$.

[32] See Hannan (1970) and Engle (1976) for a general discussion of band spectrum regression.

These restrictions were not exploited in the estimators discussed in Section 3.4.1. While these restrictions are asymptotically uninformative about β, they impact the estimator in small samples.

Gaussian MLEs of β constructed from the finite order VECM (3.2) are analyzed in Johansen (1988a, 1991) and Ahn and Reinsel (1990) using the reduced rank regression framework originally studied by Anderson (1951). Both papers discuss computational approaches for computing the MLEs, and more importantly, derive the asymptotic distribution of the Gaussian MLE. There are two minor differences between the Johansen (1988a, 1991) and Ahn and Reinsel (1990) approaches. First, different normalizations are employed. Since $\Pi = \delta\alpha' = \delta FF^{-1}\alpha$ for any nonsingular $r \times r$ matrix F, the parameters in δ and α are not econometrically identified without further restrictions. Ahn and Reinsel (1990) use the same identifying restriction imposed in the triangular model, i.e., $\alpha' = [-\beta I_r]$; Johansen (1991) uses the normalization $\hat{\alpha}' R \hat{\alpha} = I_r$, where R is the sample moment matrix of residuals from the regression of x_{t-1} onto $\Delta x_{t-i}, i = 1, \ldots, p-1$. Both sets of restrictions are normalizations in the sense that they "just" identify the model, and lead to identical values of the maximized likelihood. Partitioning Johansen's MLE as $\hat{\alpha} = (\hat{\alpha}_1' \; \hat{\alpha}_2')'$, where $\hat{\alpha}_1$ is $k \times r$ and $\hat{\alpha}_2$ is $r \times r$, implies that the MLE of β using Ahn and Reinsel's normalization is $\hat{\beta} = -(\hat{\alpha}_1 \hat{\alpha}_2^{-1})'$.

The approaches also differ in the computational algorithm used to maximize the likelihood function. Johansen (1988a), following Anderson (1951), suggests an algorithm based on partial canonical correlation analysis between Δx_t and x_{t-1} given $\Delta x_{t-i}, i = 1, \ldots, p-1$. This framework is useful because likelihood ratio tests for cointegration are computed as a byproduct (see Equation 3.19). Ahn and Reinsel (1990) suggests an algorithm based on iterative least squares calculations. Modern computers quickly find the MLEs for typical economic systems using either algorithm.

Some key results derived in both Johansen (1988a) and Ahn and Reinsel (1990) are transparent from the latter's regression formulae. As in Section 3.3, write the VECM as

$$\begin{aligned} \Delta x_t &= \delta\alpha' x_{t-1} + \Gamma z_t + \varepsilon_t \\ &= \delta[x_{2,t-1} - \beta x_{1,t-1}] + \Gamma z_t + \varepsilon_t, \end{aligned} \tag{3.37}$$

where z_t includes the relevant lags of Δx_t and the second line imposes the Ahn–Reinsel normalization of α. Let $w_{t-1} = x_{2,t-1} - \beta x_{1,t-1}$ denote the error correction term, and let $\theta = [\text{vec}(\delta)' \; \text{vec}(\Gamma)' \; \text{vec}(\beta)']'$ denote the vector of unknown parameters. Using the well known relations between the vec operator and Kronecker products, $\text{vec}(\Gamma z_t) = (z_t' \otimes I_n)\,\text{vec}(\Gamma)$, $\text{vec}(\delta w_{t-1}) = (w_{t-1}' \otimes I_n)\,\text{vec}(\Gamma)$ and $\text{vec}(\delta\beta x_{1,t-1}) = (x_{1,t-1}' \otimes \delta)\,\text{vec}(\beta)$. Using these expressions, and defining $Q_t = [(z_t \otimes I_n)(W_{t-1} \otimes I_n)(x_{1,t-1} \otimes \delta)]'$, then the Gauss–Newton iterations for estimating θ are

$$\hat{\theta}^{i+1} = \hat{\theta}^i + [\textstyle\sum Q_t' \hat{\Sigma}_\varepsilon^{-1} Q_t]^{-1} [\sum Q_t' \hat{\Sigma}_\varepsilon^{-1} \varepsilon_t] \tag{3.38}$$

where $\hat{\theta}^i$ denotes the estimator of θ at the ith iteration, $\hat{\Sigma}_\varepsilon = T^{-1}\sum \varepsilon_t \varepsilon_t'$, and Q_t and ε_t are evaluated at $\hat{\theta}^i$.[33] Thus, the Gauss–Newton regression corresponds to the GLS regression of ε_t onto $(z_t' \otimes I_n)$, $(w_{t-1}' \otimes I_n)$ and $(x_{1,t-1}' \otimes \delta)$. Since the z_t and w_t are I(0) with zero mean and $x_{1,t}$ is I(1), the analysis in Section 2 suggests that the limiting regression "$X'X$" matrix will be block diagonal, and the MLEs of δ and Γ will be asymptotically independent of the MLE of β. Johansen (1988a) and Ahn and Reinsel (1990) show that this is indeed the case. In addition they demonstrate that the MLE of β has a limiting distribution of the same form as shown in equation (3.35) above, so that $T(\hat{\beta} - \beta)$ can be represented as a variance mixture of normals. Finally, paralleling the result for MLEs from triangular representation, Johansen (1988a) and Ahn and Reinsel (1990) demonstrate that

$$[\sum Q_t' \hat{\Sigma}_\varepsilon^{-1} Q_t]^{-1/2}(\hat{\theta} - \theta) \xrightarrow{\mathscr{L}} N(0, I),$$

so that hypothesis tests and confidence intervals for all of the parameters in the VECM can be constructed using the normal and χ^2 distributions.

3.4.3. *Comparison and efficiency of the estimators*

The estimated cointegrating vectors constructed from the VECM (3.3) or the triangular representation (3.9)–(3.10) differ only in the way that the I(0) dynamics of the system are parameterized. The VECM models these dynamics using a VAR involving the first differences $\Delta x_{1,t}, \Delta x_{2,t}$ and the error correction terms, $x_{2,t} - \beta x_{1,t}$; the triangular representation uses only $\Delta x_{1,t}$ and the error correction terms. Section 3.4.1 showed that the exact parameterization of the I(0) dynamics – $\gamma(L)$ and the serial correlation of the error term in (3.36) – mattered little for the asymptotic behavior of the estimator from the triangular representation. In particular, estimators of β that ignore residual serial correlation and replace $\gamma(L)$ with $\gamma(1)$ and adjust for bias are asymptotically equivalent to the exact MLE in (3.36). Saikkonen (1991) shows that this asymptotic equivalence extends to Gaussian MLEs constructed from the VECM. Estimators of β constructed from (3.36) with appropriate nonparametric estimators of $\gamma(1)$ are asymptotically equivalent to Gaussian MLEs constructed from the VECM (3.3). Similarly, test statistics for H_o: $R[\text{vec}(\beta)] = r$ constructed from estimators based on the triangular representation and the VECM are also asymptotically equivalent.

[33] Consistent initial conditions for the iterations are easily constructed from the OLS estimators of the parameters in the VAR (3.2). Let $\hat{\Pi}$ denote the OLS estimator of Π, partitioned as $\hat{\Pi} = [\hat{\Pi}_1 \hat{\Pi}_2]$, where $\hat{\Pi}_1$ is $n \times (n-r)$ and $\hat{\Pi}_2$ is $n \times r$; further partition $\hat{\Pi}_1 = [\hat{\Pi}_{11}' \hat{\Pi}_{21}']'$ and $\hat{\Pi}_2 = [\hat{\Pi}_{12}' \hat{\Pi}_{22}']'$, where $\hat{\Pi}_{11}$ is $(n-r) \times (n-r)$, $\hat{\Pi}_{21}$ is $r \times (n-r)$, $\hat{\Pi}_{12}$ is $(n-r) \times r$ and $\hat{\Pi}_{22}$ is $r \times r$. Then $\hat{\Pi}_2$ serves as an initial consistent estimator of δ and $-(\Pi_{22})^{-1}\hat{\Pi}_{21}$ serves as an estimator of β. Ahn and Reinsel (1990) and Saikkonen (1992) develop efficient two-step estimators of β constructed from $\hat{\Pi}$, and Engle and Yoo (1991) develop an efficient three-step estimator of all the parameters in the model using iterations similar to those in (3.38).

Since estimators of cointegrating vectors do not have asymptotic normal distributions, the standard analysis of asymptotic efficiency – based on comparing estimator's asymptotic covariance matrices – cannot be used. Phillips (1991a) and Saikkonen (1991) discuss efficiency of cointegrating vectors using generalizations of the standard efficiency definitions.[34] Loosely speaking, these generalizations compare two estimators in terms of the relative probability that the estimators are contained in certain convex regions that are symmetric about the true value of the parameter vector. Phillips (1991a) shows that when u_t in the triangular representation (3.9)–(3.10) is generated by a Gaussian ARMA process, then the MLE is asymptotically efficient. Saikkonen (1991) considers estimators whose asymptotic distributions can be represented by a certain class of functionals of Brownian motion. This class contains the OLS and nonlinear least squares estimators analyzed in Stock (1987), the instrumental variable estimators analyzed in Phillips and Hansen (1990), all of the estimators discussed in Sections 3.4.1 and 3.4.2, and virtually every other estimator that has been suggested. Saikkonen shows that the Gaussian MLE or (any of the estimators that are asymptotically equivalent to the Gaussian MLE) are asymptotically efficient members of this class.

Several studies have used Monte Carlo methods to examine the small sample behavior of the various estimators of cointegrating vectors. A partial list of the Monte Carlo studies is Ahn and Reinsel (1990), Banerjee et al. (1986), Gonzalo (1989), Park and Ogaki (1991), Phillips and Hansen (1990), Phillips and Loretan (1991) and Stock and Watson (1993). A survey of these studies suggests three general conclusions. First, the static OLS estimator can be very badly biased even when the sample size is reasonably large. This finding is consistent with the bias in the asymptotic distribution of the OLS estimator (see equation (2.22)) that was noted by Stock (1987).

The second general conclusion concerns the small sample behavior of the Gaussian MLE based on the finite order VECM. The Monte Carlo studies discovered that, when the sample size is small, the estimator has a very large mean squared error, caused by a few extreme outliers. Gaussian MLEs based on the triangular representation do not share this characteristic. Some insight into this phenomenon is provided in Phillips (1991c) which derives the exact (small sample) distribution of the estimators in a model in which the variables follow independent Gaussian random walks. The MLE constructed from the VECM is shown to have a Cauchy distribution and so has no integer moments, while the estimator based on the triangular representation has integer moments up to order $T - n + r$. While Phillips' results concern a model in which the variables are not cointegrated, it is useful because it suggests that when the data are "weakly" cointegrated – as might be the case in small samples – the estimated cointegrating vector will (approximately) have these characteristics.

The third general conclusion concerns the approximate Gaussian MLEs based

[34] See Basawa and Scott (1983) and Sweeting (1983).

on the triangular representation. The small sample properties of these estimators and test statistics depend in an important way on the estimator used for the long-run covariance matrix of the data (spectrum at frequency zero), which is used to construct an estimator of $\gamma(1)$ and the long-run residual variance in (3.36). Experiments in Park and Ogaki (1991), Stock and Watson (1993) and (in a different context) Andrews and Moynihan (1990), suggest that autoregressive estimators or estimators that rely on autoregressive pre-whitening outperform estimators based on simple weighted averages of sample covariances.

3.5. *The role of constants and trends*

3.5.1. *The model of deterministic components*

Thus far, deterministic components in the time series (constants and trends) have been ignored. These components are important for three reasons. First, they represent the average growth or nonzero level present in virtually all economic time series; second, they affect the efficiency of estimated cointegrating vectors and the power of tests for cointegration; finally, they affect the distribution of estimated cointegrating vectors and cointegration test statistics. Accordingly, suppose that the observed $n \times 1$ time series y_t can be represented as

$$y_t = \mu_0 + \mu_1 t + x_t, \tag{3.39}$$

where x_t is generated by the VAR (3.1). In (3.39), $\mu_0 + \mu_1 t$ represents the deterministic component of y_t, and x_t represents the stochastic component. In this section we discuss how the deterministic components affect the estimation and testing procedures that we have already surveyed.[35]

There is a simple way to modify the procedures so that they can be applied to y_t. The deterministic components can be eliminated by regressing y_t onto a constant and time trend. Letting y_t^τ denote the detrended series, the estimation and testing procedures developed above can then be used by replacing x_t with y_t^τ. This changes the asymptotic distribution of the statistics in a simple way: since the detrended values of y_t and x_t are identical, all statistics have the same limiting representation with the Brownian motion process $B(s)$ replaced by $B^\tau(s)$, the detrended Brownian motion introduced in Section 2.3.

While this approach is simple, it is often statistically inefficient because it discards all of the deterministic trend information in the data, and the relationship between these trends is often the most useful information about cointegration. To see this,

[35] We limit discussion to linear trends in y_t for reasons of brevity and because this is the most important model for empirical applications. The results are readily extended to higher order trend polynomials and other smooth trend functions.

let α denote a cointegrating vector and consider the "stable" linear combination

$$\alpha' y_t = \lambda_0 + \lambda_1 t + w_t, \tag{3.40}$$

where $\lambda_0 = \alpha'\mu_0, \lambda_1 = \alpha'\mu_1$, and $w_t = \alpha' x_t$. In most (if not all) applications, the cointegrating vector will annihilate both the stochastic trend and deterministic trend in $\alpha' y_t$. That is, w_t will be I(0) and $\lambda_1 = 0$.[36] As shown below, this means that one linear combination of the coefficients in the cointegrating vector can be consistently estimated at rate $T^{3/2}$. In contrast, when detrended data are used, the cointegrating vectors are consistently estimated at rate T. Thus, the data's deterministic trends are the dominant source of information about the cointegrating vector and detrending the data throws this information away.

The remainder of this section discusses estimation and testing procedures that utilize the data's deterministic trends. Most of these procedures are simple modifications of the procedures that were developed above.

3.5.2. *Estimating cointegrating vectors*

We begin with a discussion of the MLE of cointegrating vectors based on the triangular representation. Partitioning y_t into $(n-r)\times 1$ and $r\times 1$ components, $y_{1,t}$ and $y_{2,t}$, the triangular representation for y_t is

$$\Delta y_{1,t} = \gamma + u_{1,t}, \tag{3.41}$$

$$y_{2,t} - \beta y_{1,t} = \lambda_0 + \lambda_1 t + u_{2,t}. \tag{3.42}$$

This is identical to the triangular representation for x_t given in (3.9)–(3.10) except for the constant and trend terms. The constant term in (3.41) represents the average growth in $y_{1,t}$. In most situations $\lambda_1 = 0$ in (3.42) since the cointegrating vector annihilates the deterministic trend in the variables. In this case, λ_0 denotes the mean of the error correction terms, which is unrestricted in most economic applications.

Assuming that $\lambda_1 = 0$ and λ_0 and γ are unrestricted, efficient estimation of β in (3.42) proceeds as in Section 3.3.1. The only difference is that the equations now include a constant term. As in Section 3.3.1, Wald, LR or LM test statistics for testing $H_o: R[\text{vec}(\beta)] = r$ will have limiting χ^2 distributions, and confidence intervals for the elements of β can be constructed in the usual way. The only result from Section 3.3.1 that needs to be modified is the asymptotic distribution of $\hat{\beta}$. This estimator is calculated from the regression of $y_{2,t}$ onto $y_{1,t}$, leads and lags of $\Delta y_{1,t}$ and a constant term. When the $y_{1,t}$ data contain a trend [$\gamma \neq 0$ in (3.41)],

[36] Ogaki and Park (1990) define these two restrictions as "stochastic" and "deterministic" cointegration. Stochastic cointegration means that w_t is I(0), while deterministic cointegration means that $\lambda_1 = 0$.

one of the canonical regressors is a time trend ($z_{4,t}$ from Section 2.5.1), and the estimated coefficient on the time trend converges at rate $T^{3/2}$. This means that one linear combination of the estimated coefficients in the cointegrating vector converges to its true value very quickly; when the model did not contain a linear trend the estimator converged at rate T.

The results for MLEs based on the finite order VECM representation are analogous to those from the triangular representation. The VECM representation for y_t is derived directly from (3.2) and (3.39),

$$\begin{aligned}\Delta y_t &= \mu_1 + \delta(\alpha' x_{t-1}) + \sum_{i=1}^{p-1} \Phi_i \Delta x_{t-i} + \varepsilon_t \\ &= \tilde{\mu}_1 + \delta(\alpha' y_{t-1} - \lambda_0 - \lambda_1 t) + \sum_{i=1}^{p-1} \Phi_i \Delta y_{t-i} + \varepsilon_t, \end{aligned} \tag{3.43}$$

where $\tilde{\mu}_1 = (I - \sum_{i=1}^{p-1} \Phi_i)\mu_1$, $\lambda_0 = \alpha'\mu_0$ and $\lambda_1 = \alpha'\mu_1$. Again, in most applications $\lambda_1 = 0$, and the VECM is

$$\Delta y_t = \theta + \delta(\alpha' y_{t-1}) + \sum_{i=1}^{p-1} \Phi_i \Delta y_{t-i} + \varepsilon_t, \tag{3.44}$$

where $\theta = \tilde{\mu}_1 - \delta\lambda_0$. When the only restriction on μ_1 is $\alpha'\mu_1 = 0$, the constant term θ is unconstrained, and (3.44) has the same form as (3.2) except that a constant term has been added. Thus, the Gaussian MLE from (3.44) is constructed exactly as in Section 3.4.2 with the addition of a constant term in all equations. The distribution of test statistics is unaffected, but for the reasons discussed above, the asymptotic distribution of the cointegrating vector changes because of the presence of the deterministic trend.

In some situations the data are not trending in a deterministic fashion, so that $\mu_1 = 0$. (For example, this is arguably the case when y_t is a vector of U.S. interest rates.) When $\mu_1 = 0$, then $\tilde{\mu}_1 = 0$ in (3.43), and this imposes a constraint on θ in (3.44). To impose this constraint, the model can be written as

$$\Delta y_t = \delta(\alpha' y_{t-1} - \lambda_0) + \sum_{i=1}^{p-1} \Phi_i \Delta y_{t-i} + \varepsilon_t \tag{3.45}$$

and estimated using a modification of the Gauss–Newton iterations in (3.38), or a modification of Johansen's canonical correlation approach [see Johansen and Juselius (1990)].

3.5.3. Testing for cointegration

Deterministic trends have important effects on tests for cointegration. As discussed in Johansen and Juselius (1990) and Johansen (1991, 1992a), it is useful to consider two separate effects. First, as in (3.43)–(3.44) nonzero values of μ_0 and μ_1 affect the form of the VECM, and this, in turn, affects the form of the cointegration test statistic. Second, the deterministic components affect the properties of the regressors, and this, in turn, affects the distribution of cointegration test statistics. In the most general form of the test considered in Section 3.3.1, α was partitioned into known and unknown cointegrating vectors under both the null and alternative; that is, α was written as $\alpha = (\alpha_{o_K} \alpha_{o_u} \alpha_{a_K} \alpha_{o_u})$. When nonzero values of μ_0 and μ_1 are allowed, the precise form of the statistic and resulting asymptotic null distribution depends on which of these cointegrating vectors annihilate the trend or constant [see Horvath and Watson (1993)]. Rather than catalogue all of the possible cases, the major statistical issues will be discussed in the context of two examples. The reader is referred to Johansen and Juselius (1990), Johansen (1992a) and Horvath and Watson (1993) for a more systematic treatment.

In the first example, suppose that $r = 0$ under the null, that α is known under the alternative, that μ_0 and μ_1 are nonzero, but that $\alpha'\mu_1 = 0$ is known. To be concrete, suppose that the data are aggregate time series on the logarithms of income, consumption and investment for the United States. The balanced growth hypothesis suggests two possible cointegrating relations with cointegrating vectors $(1, -1, 0)$ and $(1, 0, -1)$. The series exhibit deterministic growth, so that $\mu_1 \neq 0$, and the sample growth rates are approximately equal, so that $\alpha'\mu_1 = 0$ is reasonable. In this example, (3.44) is the correct specification of the VECM with θ unrestricted under both the null and alternative and $\delta = 0$ under the null. Comparing (3.44) and the specification with no deterministic components given in (3.3), the only difference is that x_t in (3.3) becomes y_t in (3.44) and the constant term θ is added. Thus, the Wald test for H_o: $\delta = 0$ is constructed as in (3.17) with y_t replacing x_t and Z augmented by a column of 1's. Since $\alpha'\mu_1 = 0$, the regressor is $\alpha' y_{t-1} = \alpha' x_{t-1} + \alpha'\mu_0$, but since a constant is included in the regression, all of the variables are deviated from their sample means. Since the demeaned values of $\alpha' y_{t-1}$ and $\alpha' x_{t-1}$ are the same, the asymptotic null distribution of the Wald statistic for testing H_o: $\delta = 0$ in (3.44) is given by (3.18) with $\beta^{\mu}(s)$, the demeaned Wiener process defined below Lemma 2.3, replacing $B(s)$.

The second example is the same as the first, except that now α is unknown. Equation (3.44) is still the correct VECM with θ unrestricted under the null and alternative. The LR test statistic is calculated as in (3.19), again with y_t replacing x_t and Z augmented by a vector of 1's. Now, however, the distribution of the test statistic changes in an important way. Since the regressor y_{t-1} contains a nonzero trend, it behaves like a combination of time trend and martingale components. When the $n \times 1$ vector y_{t-1} is transformed into the canonical regressors of Section 2, this yields one regressor dominated by a time trend and $n - 1$ regressors dominated

by martingales. As shown in Johansen and Juselius (1990), the distribution of the resulting LR statistic has a null distribution given by (3.25) where now

$$H = \left[\int F(s)\mathrm{d}B(s)'\right]'\left[\int F(s)F(s)'\mathrm{d}s\right]^{-1}\left[\int F(s)\mathrm{d}B(s)'\right],$$

where $F(s)$ is an $n \times 1$ vector, with first $n-1$ elements given by the first $n-1$ elements of $\beta^{\mu}(s)$ and the last element given by the demeaned time trend, $s - \frac{1}{2}$. (The components are demeaned because of the constant term in the regression.)

Johansen and Juselius (1990) also derive the asymptotic null distribution of the LR test for cointegration with unknown cointegrating vectors when $\mu_1 = 0$, so that (3.45) is the appropriate specification of the VECM. Tables of critical values are presented in Johansen and Juselius (1990) for $n - r_{o_u} \leqslant 5$ for the various deterministic trend models, and these are extended in Osterwald-Lenum for $n - r_{o_u} \leqslant 11$. Horvath and Watson (1992) extend the tables to include nonzero values of r_{o_k} and r_{a_k}.

The appropriate treatment of deterministic components in cointegration and unit root tests is still unsettled, and remains an active area of research. For example, Elliot et al. (1992) report that large gains in power for univariate unit root tests can be achieved by modifying standard Dickey–Fuller tests by an alternative method of detrending the data. They propose detrending the data using GLS estimators or μ_0 and μ_1 from (3.39) together with specific assumptions about initial conditions for the x_t process. Analogous procedures for likelihood based tests for cointegration can also be constructed. Johansen (1992b) develops a sequential testing procedure for cointegration in which the trend properties of the data and potential error corrections terms are unknown.

4. Structural vector autoregressions

4.1. *Introductory comments*

Following the work of Sims (1980), vector autoregressions have been extensively used by economists for data description, forecasting and structural inference. Canova (1991) surveys VARs as a tool for data description and forecasting; this survey focuses on structural inference. We begin the discussion in Section 4.2 by introducing the structural moving average model, and show that this model provides answers to the "impulse" and "propagation" questions often asked by macroeconomists. The relationship between the structural moving average model and structural VAR is the subject of Section 4.3. That section discusses the conditions under which the structural moving average polynomial can be inverted, so that the structural shocks can be recovered from a VAR. When this is possible, a structural VAR obtains. Section 4.4 shows that the structural VAR can be interpreted

as a dynamic simultaneous equations model, and discusses econometric identification of the model's parameters. Finally, Section 4.5 discusses issues of estimation and statistical inference.

4.2. *The structural moving average model, impulse response functions and variance decompositions*

In this section we study the model

$$y_t = C(L)\varepsilon_t, \tag{4.1}$$

where y_t is an $n_y \times 1$ vector of economic variables and ε_t is an $n_\varepsilon \times 1$ vector of shocks. For now we allow $n_y \neq n_\varepsilon$. Equation (4.1) is called the structural moving average model, since the elements of ε_t are given a structural economic interpretation. For example, one element of ε_t might be interpreted as an exogenous shock to labor productivity, another as an exogenous shock to labor supply, another as an exogenous change in the quantity of money, and so forth. In the jargon developed for the analysis of dynamic simultaneous equations models, (4.4) is the final form of an economic model, in which the endogenous variables y_t are expressed as a distributed lag of the exogenous variables, given here by the elements of ε_t. It will be assumed that the endogenous variables y_t are observed, but that the exogenous variables ε_t are not directly observed. Rather, the elements of ε_t are indirectly observed through their effect on the elements of y_t. This assumption is made without loss of generality, since any observed exogenous variables can always be added to the y_t vector.

In Section 1, a typical macroeconomic system was introduced and two broad questions were posed. The first question asked how the system's endogenous variables responded dynamically to exogenous shocks. The second question asked which shocks were the primary causes of variability in the endogenous variables. Both of these questions are readily answered using the structural moving average model.

First, the dynamic effects of the elements of ε_t on the elements of y_t are determined by the elements of the matrix lag polynomial $C(L)$. Letting $C(L) = C_0 + C_1 L + C_2 L^2 + \cdots$, where C_k is an $n_y \times n_\varepsilon$ matrix with typical element $[c_{ij,k}]$, then

$$c_{ij,k} = \frac{\partial y_{i,t}}{\partial \varepsilon_{j,t-k}} = \frac{\partial y_{i,t+k}}{\partial \varepsilon_{j,t}}, \tag{4.2}$$

where $y_{i,t}$ is the ith element of y_t, $\varepsilon_{j,t}$ is the jth element of ε_t, and the last equality follows from the time invariance of (4.1). Viewed as a function of k, $c_{ij,k}$ is called the impulse response function of $\varepsilon_{j,t}$ for $y_{i,t}$. It shows how $y_{i,t+k}$ changes in response to a one unit "impulse" in $\varepsilon_{j,t}$. In the classic econometric literature on distributed lag models, the impulse responses are called dynamic multipliers.

To answer the second question concerning the relative importance of the shocks, the probability structure of the model must be specified and the question must be refined. In most applications the probability structure is specified by assuming that the shocks are i.i.d.$(0, \Sigma_\varepsilon)$, so that any serial correlation in the exogenous variables is captured in the lag polynomial $C(L)$. The assumption of zero mean is inconsequential, since deterministic components such as constants and trends can always be added to (4.1). Viewed in this way, ε_t represents innovations or unanticipated shifts in the exogenous variables. The question concerning the relative importance of the shocks can be made more precise by casting it in terms of the h-step-ahead forecast errors of y_t. Let $y_{t/t-h} = E(y_t | \{\varepsilon_s\}_{s=-\infty}^{t-h})$ denote the h-step-ahead forecast of y_t made at time $t-h$, and let $a_{t/t-h} = y_t - y_{t/t-h} = \sum_{k=0}^{h-1} C_k \varepsilon_{t-k}$ denote the resulting forecast error. For small values of h, $a_{t/t-h}$ can be interpreted as "short-run" movements in y_t, while for large values of h, $a_{t/t-h}$ can be interpreted as "long-run" movements. In the limit as $h \to \infty$, $a_{t/t-h} = y_t$. The importance of a specific shock can then be represented as the fraction of the variance in $a_{t/t-h}$ that is explained by that shock; it can be calculated for short-run and long-run movements in y_t by varying h. When the shocks are mutually correlated there is no unique way to do this, since their covariance must somehow be distributed. However, when the shocks are uncorrelated the calculation is straightforward. Assume Σ_ε is diagonal with diagonal elements σ_j^2, then the variance of the ith element of $a_{t/t-h}$ is

$$\sum_{j=1}^{n_\varepsilon} \left[\sigma_j^2 \sum_{k=0}^{h-1} c_{ij,k}^2 \right],$$

so that

$$R_{ij,h}^2 = \frac{\left[\sigma_j^2 \sum_{k=0}^{h-1} c_{ij,k}^2 \right]}{\sum_{m=1}^{n_\varepsilon} \left[\sigma_m^2 \sum_{k=0}^{h-1} c_{im,k}^2 \right]} \tag{4.3}$$

shows the fraction of the h-step-ahead forecast error variance in $y_{i,t}$ attributed to $\varepsilon_{j,t}$. The set of n_ε values of $R_{ij,h}^2$ are called the variance decomposition of $y_{i,t}$ at horizon h.

4.3. *The structural VAR representation*

The structural VAR representation of (4.1) is obtained by inverting $C(L)$ to yield

$$A(L) y_t = \varepsilon_t, \tag{4.4}$$

where $A(L) = A_0 - \sum_{k=1}^{\infty} A_k L^k$ is a one-sided matrix lag polynomial. In (4.4), the exogenous shocks ε_t are written as a distributed lag of current and lagged values of y_t. The structural VAR representation is useful for two reasons. First, when the model parameters are known, it can be used to construct the unobserved exogenous shocks as a function of current and lagged values of the observed variables y_t. Second, it provides a convenient framework for estimating the model parameters: with $A(L)$ approximated by a finite order polynomial, Equation (4.4) is a dynamic simultaneous equations model, and standard simultaneous methods can be used to estimate the parameters.

It is not always possible to invert $C(L)$ and move from the structural moving average representation (4.1) to the VAR representation (4.4). One useful way to discuss the invertibility problem [see Granger and Anderson (1978)] is in terms of estimates of ε_t constructed from (4.4) using truncated versions of $A(L)$. Since a semi-infinite realization of the y_t process, $\{y_s\}_{s=-\infty}^{T}$, is never available, estimates of ε_t must be constructed from (4.4) using $\{y_s\}_{s=1}^{T}$. Consider the estimator $\tilde{\varepsilon}_t = \sum_{i=0}^{t-1} A_i y_{t-i}$ constructed from the truncated realization. If $\tilde{\varepsilon}_t$ converges to ε_t in mean square as $t \to \infty$, then the structural moving average process (4.1) is said to be invertible. Thus, when the process is invertible, the structural errors can be recovered as a one-sided moving average of the observed data, at least in large samples.

This definition makes it clear that the structural moving average process cannot be inverted if $n_y \leqslant n_\varepsilon$. Even in the static model $y_t = C\varepsilon_t$, a necessary condition for obtaining a unique solution for ε_t in terms of y_t is that $n_y \geqslant n_\varepsilon$. This requirement has a very important implication for structural analysis using VAR models: in general, small scale VARs can only be used for structural analysis when the endogenous variables can be explained by a small number of structural shocks. Thus, a bivariate VAR of macroeconomic variables is not useful for structural analysis if there are more than two important macroeconomic shocks affecting the variables.[37] In what follows we assume that $n_y = n_\varepsilon$. This rules out the simple cause of noninvertibility just discussed; it also assumes that any dynamic identities relating the elements of y_t when $n_y > n_\varepsilon$ have been solved out of the model.

With $n_y = n_\varepsilon \equiv n$, $C(L)$ is square and the general requirement for invertibility is that the determinantal polynomial $|C(z)|$ has all of its roots outside the unit circle. Roots on the unit circle pose no special problems; they are evidence of over-differencing and can be handled by appropriately transforming the variables (e.g. accumulating the necessary linear combinations of the elements of y_t). In any event, unit roots can be detected, at least in large samples, by appropriate statistical tests. Roots of $|C(z)|$ that are inside the unit circle pose a much more difficult problem, since models with roots inside the unit circle have the same second moment properties as models with roots outside the unit circle. The simplest example of this

[37] Blanchard and Quah (1989) and Faust and Leeper (1993) discuss special circumstances when some structural analysis is possible when $n_y < n_\varepsilon$. For example, suppose that y_t is a scalar and the n_ε elements of ε_t affect y_t only through the scalar "index" $e_t = D'\varepsilon_t$, where D is an $n_\varepsilon \times 1$ vector. In this case the impulse response functions can be recovered up to scale.

is the univariate MA(1) model $y_t = (1 - cL)\varepsilon_t$, where ε_t is i.i.d.$(0, \sigma_\varepsilon^2)$. The same first and second moments of y_t obtain for the model $y_t = (1 - \tilde{c}L)\tilde{\varepsilon}_t$, where $\tilde{c} = c^{-1}$ and $\tilde{\varepsilon}_t$ is IID$(0, \sigma_{\tilde{\varepsilon}}^2)$ with $\sigma_{\tilde{\varepsilon}}^2 = c^2\sigma_\varepsilon^2$. Thus, the first two moments of y_t cannot be used to discriminate between these two different models. This is important because it can lead to large specification errors in structural VAR models that cannot be detected from the data. For example, suppose that the true structural model is $y_t = (1 - cL)\varepsilon_t$ with $|c| > 1$ so that the model is not invertible. A researcher using the invertible model would not recover the true structural shocks, but rather $\tilde{\varepsilon}_t = (1 - \tilde{c}L)^{-1}y_t = (1 - \tilde{c}L)^{-1}(1 - cL)\varepsilon_t = \varepsilon_t - (\tilde{c} - c)\sum_{i=1}^{\infty}\tilde{c}^i\varepsilon_{t-1-i}$. A general discussion of this subject is contained in Hannan (1970) and Rozanov (1967). Implications of these results for the interpretation of structural VARs are discussed in Hansen and Sargent (1991) and Lippi and Reichlin (1993). For related discussion see Quah (1986).

Hansen and Sargent (1991) provides a specific economic model in which non-invertible structural moving average processes arise. In the model, one set of economic variables, say x_t, are generated by an invertible moving average process. Another set of economic variables, say y_t, are expectational variables, formed as discounted sums of expected future x_t's. Hansen and Sargent then consider what would happen if only the y_t data were available to the econometrician. They show that the implied moving average process of y_t, written in terms of the structural shocks driving x_t, is not invertible.[38] The Hansen–Sargent example provides an important and constructive lesson for researchers using structural VARs: it is important to include variables that are directly related to the exogenous shocks under consideration (x_t in the example above). If the only variables used in the model are indirect indicators with important expectational elements (y_t in the example above), severe misspecification may result.

4.4. *Identification of the structural VAR*

Assuming that the lag polynomial of $A(L)$ in (4.4) is of order p, then structural VAR can be written as

$$A_0 y_t = A_1 y_{t-1} + A_2 y_{t-2} + \cdots + A_p y_{t-p} + \varepsilon_t. \tag{4.5}$$

[38] A simple version of their example is as follows: suppose that y_1 and x_t are two scalar time series, with x_t generated by the MA(1) process $x_t = \varepsilon_t - \theta\varepsilon_{t-1}$. Suppose that y_t is related to x_t by the expectational equation

$$\begin{aligned} y_t &= E_t \sum_{i=0}^{\infty} \beta^i x_{t+i} \\ &= x_t + \gamma E_t x_{t+1} \\ &= (1 - \beta\theta)\varepsilon_t - \theta\varepsilon_{t-1} \equiv C(L)\varepsilon_t, \end{aligned}$$

where the second and third lines follow from the MA(1) process for x_t. It is readily verified that the root of $C(z)$ is $(1 - \beta\theta)/\theta$, which may be less than 1 even when the root of $(1 - \theta z)$ is greater than 1. (For example, if $\theta = \beta = 0.8$, the root of $(1 - \theta z)$ is 1.25 and the root of $C(z)$ is 0.45.)

Since A_0 is not restricted to be diagonal, equation (4.5) is a dynamic simultaneous equations model. It differs from standard representations of the simultaneous equations model [see Hausman (1983)] because observable exogenous variables are not included in the equations. However, since exogenous and predetermined variables – lagged values of y_{t-1} – are treated identically for purposes of identification and estimation, equation (4.5) can be analyzed using techniques developed for simultaneous equations.

The reduced form of (4.5) is

$$y_t = \Phi_1 y_{t-1} + \Phi_2 y_{t-2} + \cdots + \Phi_p y_{t-p} + e_t, \tag{4.6}$$

where $\Phi_i = A_0^{-1} A_i$, for $i = 1, \ldots, p$, and $e_t = A_0^{-1} \varepsilon_t$. A natural first question concerns the identifiability of the structural parameters in (4.5), and this is the subject taken up in this section.

The well known "order" condition for identification is readily deduced. Since y_t is $n \times 1$, there are pn^2 elements in $(\Phi_1, \Phi_2, \ldots, \Phi_p)$ and $n(n+1)/2$ elements in $\Sigma_e = A_0^{-1} \Sigma_\varepsilon (A_0^{-1})'$, the covariance matrix of the reduced form disturbances. When the structural shocks are n.i.i.d.$(0, \Sigma_\varepsilon)$, these $[n^2 p + n(n+1)/2]$ parameters completely characterize the probability distribution of the data. In the structural model (4.5) there are $(p+1)n^2$ elements in $(A_0, A_1, \ldots, A_p)$ and $n(n+1)/2$ elements in Σ_ε. Thus, there are n^2 more parameters in the structural model than are needed to characterize the likelihood function, so that n^2 restrictions are required for identification. As usual, setting the diagonal elements of A_0 equal to 1, gives the first n restrictions. This leaves $n(n-1)$ restrictions that must be deduced from economic considerations.

The identifying restrictions must be dictated by the economic model under consideration. It makes little sense to discuss the restrictions without reference to a specific economic system. Here, some general remarks on identification are made in the context of a simple bivariate model explaining output and money; a more detailed discussion of identification in structural VAR models is presented in Giannini (1991). Let the first element of y_t, say $y_{1,t}$, denote the rate of growth of real output, and the second element of y_t, say $y_{2,t}$ denote the rate of growth of money.[39] Writing the typical element of A_k as $a_{ij,k}$, equation (4.5) becomes

$$y_{1,t} = -a_{12,0} y_{2,t} + \sum_{i=1}^{p} a_{11,i} y_{1,t-i} + \sum_{i=1}^{p} a_{12,i} y_{2,t-i} + \varepsilon_{1,t}, \tag{4.7a}$$

$$y_{2,t} = -a_{21,0} y_{1,t} + \sum_{i=1}^{p} a_{21,i} y_{1,t-i} + \sum_{i=1}^{p} a_{22,i} y_{2,t-i} + \varepsilon_{2,t}. \tag{4.7b}$$

Equation (4.7a) is interpreted as an output or "aggregate supply" equation, with

[39] Much of this discussion concerning this example draws from King and Watson (1993).

$\varepsilon_{1,t}$ interpreted as an aggregate supply or productivity shock. Equation (4.7b) is interpreted as a money supply "reaction function" showing how the money supply responds to contemporaneous output, lagged variables, and a money supply shock $\varepsilon_{2,t}$. Identification requires $n(n-1)=2$ restrictions on the parameters of (4.7).

In the standard analysis of simultaneous equation models, identification is achieved by imposing zero restrictions on the coefficients for the predetermined variables. For example, the order condition is satisfied if $y_{1,t-1}$ enters (4.7a) but not (4.7b), and $y_{1,t-2}$ enters (4.7b) but not (4.7a); this imposes the two constraints $a_{21,1}=a_{11,2}=0$. In this case, $y_{1,t-1}$ shifts the output equation but not the money equation, while $y_{1,t-2}$ shifts the money equation but not the output equation. Of course, this is a very odd restriction in the context of the output–money model, since the lags in the equations capture expectational effects, technological and institutional inertia arising production lags and sticky prices, information lags, etc.. There is little basis for identifying the model with the restriction $a_{21,1}=a_{11,2}=0$. Indeed there is little basis for identifying the model with any zero restrictions on lag coefficients. Sims (1980) persuasively makes this argument in a more general context, and this has led structural VAR modelers to avoid imposing zero restrictions on lag coefficients. Instead, structural VARs have been identified using restrictions on the covariance matrix of structural shocks Σ_ε, the matrix of contemporaneous coefficients A_0 and the matrix of long-run multipliers $A(1)^{-1}$.

Restrictions on Σ_ε have generally taken the form that Σ_ε is diagonal, so that the structural shocks are assumed to be uncorrelated. In the example above, this means that the underlying productivity shocks and money supply shocks are uncorrelated, so that any contemporaneous cross equation impacts arise through nonzero values of $a_{12,0}$ and $a_{21,0}$. Some researchers have found this a natural assumption to make, since it follows from a modeling strategy in which unobserved structural shocks are viewed as distinct phenomena which give rise to comovement in observed variables only through the specific economic interactions studied in the model. The restriction that Σ_ε is diagonal imposes $n(n-1)$ restrictions on the model, leaving only $n(n-1)/2$ additional necessary restrictions.[40]

These additional restrictions can come from a priori knowledge about the A_0 matrix in (4.5). In the bivariate output–money model in (4.7), if Σ_ε is diagonal, then only $n(n-1)/2=1$ restriction on A_0 is required for identification. Thus, a priori knowledge of $a_{12,0}$ or $a_{21,0}$ will serve to identify the model. For example, if it was assumed that the money shocks affect output only with a lag, so that $\partial y_{1,t}/\partial \varepsilon_{2,t}=0$, then $a_{12,0}=0$, and this restriction identifies the model. The generalization of this restriction in the n-variable model produces the Wold causal chain [see Wold (1954) and Malinvaud (1980, pp. 605–608)], in which $\partial y_{i,t}/\partial \varepsilon_{j,t}=0$ for $i<j$. This restriction leads to a recursive model with A_0 lower triangular, yielding the required $n(n-1)/2$ identifying restrictions. This restriction was used in Sims (1980), and has

[40]Other restrictions on the covariance matrix are possible, but will not be discussed here. A more general discussion of identification with covariance restrictions can be found in Hausman and Taylor (1983), Fisher (1966), Rothenberg (1971) and the references cited there.

become the "default" identifying restriction implemented automatically in commercial econometric software. Like any identifying restriction, it should never be used automatically. In the context of the output–money example, it is appropriate under the maintained assumption that exogenous money supply shocks, and the resulting change in interest rates, have no contemporaneous effect on output. This may be a reasonable assumption for data sampled at high frequencies, but loses its appeal as the sampling interval increases.[41,42]

Other restrictions on A_0 can also be used to identify the model. Blanchard and Watson (1986), Bernanke (1986) and Sims (1986) present empirical models that are identified by zero restrictions on A_0 that don't yield a lower triangular matrix. Keating (1990) uses a related set of restrictions. Of course, nonzero equality restrictions can also be used; see Blanchard and Watson (1986) and King and Watson (1993) for examples.

An alternative set of identifying restrictions relies on long-run relationships. In the context of structural VARs these restrictions were used in papers by Blanchard and Quah (1989) and King et al. (1991).[43] These papers relied on restrictions on $A(1) = A_0 - \sum_{i=1}^{p} A_i$ for identification. Since $C(1) = A(1)^{-1}$, these can alternatively be viewed as restrictions on the sum of impulse responses. To motivate these restrictions, consider the output–money example.[44] Let $x_{1,t}$ denote the logarithm of the level of output and $x_{2,t}$ denote the logarithm of the level of money, so that $y_{1,t} = \Delta x_{1,t}$ and $y_{2,t} = \Delta x_{2,t}$. Then from (4.1),

$$\frac{\partial x_{i,t+k}}{\partial \varepsilon_{j,t}} = \sum_{m=0}^{k} \frac{\partial y_{i,t+m}}{\partial \varepsilon_{j,t}} = \sum_{m=0}^{k} c_{ij,m}, \tag{4.8}$$

for $i, j = 1, 2$, so that

$$\lim_{k \to \infty} \frac{\partial x_{i,t+k}}{\partial \varepsilon_{j,t}} = \sum_{m=0}^{\infty} c_{ij,m}, \tag{4.9}$$

which is the ijth element of $C(1)$. Now, suppose that money is neutral in the long run, in the sense that shocks to money have no permanent effect on the level of output. This means that $\lim_{k \to \infty} \partial x_{1,t+k} / \partial \varepsilon_{2,t} = 0$, so that $C(1)$ is a lower triangular

[41] The appropriateness of the Wold causal chain was vigorously debated in the formative years of simultaneous equations. See Malinvaud (1980, pp. 55–58) and the references cited there.

[42] Applied researchers sometimes estimate a variety of recursive models in the belief (or hope) that the set of recursive models somehow "brackets" the truth. There is no basis for this. Statements like "the ordering of the Wold causal chain didn't matter for the results" say little about the robustness of the results to different identifying restrictions.

[43] For other early applications of this approach, see Shapiro and Watson (1988) and Gali (1992).

[44] The empirical model analyzed in Blanchard and Quah (1989) has the same structure as the output–money example with the unemployment rate used in place of money growth.

matrix. Since $A(1) = C(1)^{-1}$, this means that $A(1)$ is also lower triangular, and this yields the single extra identifying restriction that is required to identify the bivariate model. The analogous restriction in the general n-variable VAR, is the long-run Wold causal chain in which $\varepsilon_{i,t}$ has no long-run effect on $y_{j,t}$ for $j < i$. This restriction implies that $A(1)$ is lower triangular yielding the necessary $n(n-1)/2$ identifying restrictions.[45]

4.5. *Estimating structural VAR models*

This section discusses methods for estimating the parameters of the structural VAR (4.5). The discussion is centered around generalized method of moment (GMM) estimators. The relationship between these estimators and FIML estimators constructed from a Gaussian likelihood is discussed below. The simplest version of the GMM estimator is indirect least squares, which follows from the relationship between the reduced form parameters in (4.6) and the structural parameters in (4.5):

$$A_0^{-1} A_i = \Phi_i, \qquad i = 1, \ldots, p, \tag{4.10}$$

$$A_0 \Sigma_\varepsilon A_0' = \Sigma_e. \tag{4.11}$$

Indirect least squares estimators are formed by replacing the reduced form parameters in (4.10) and (4.11) with their OLS estimators and solving the resulting equations for the structural parameters. Assuming that the model is exactly identified, a solution will necessarily exist. Given estimators $\hat{\Phi}_i$ and $\hat{A}_0$, equation (4.10) yields $\hat{A}_i = \hat{A}_0 \hat{\Phi}_i$. The quadratic equation (4.11) is more difficult to solve. In general, iterative techniques are required, but simpler methods are presented below for specific models.

To derive the large sample distribution of the estimators and to "solve" the indirect least squares equations when there are overidentifying restrictions, it is convenient to cast the problem in the standard GMM framework [see Hansen (1982)]. Hausman et al. (1987) show how this framework can be used to construct efficient estimators for the simultaneous equations model with covariance restrictions on the error terms, thus providing a general procedure for forming efficient estimators in the structural VAR model.

Some additional notation is useful. Let $z_t = (y_{t-1}', y_{t-2}', \ldots, y_{t-p}')'$ denote the vector of predetermined variables in the model, and let θ denote the vector of unknown parameters in $A_0, A_1, \ldots, A_p$ and Σ_ε. The population moment conditions that implicitly define the structural parameters are

$$E(\varepsilon_t z_t') = 0, \tag{4.12}$$

[45] Of course, restrictions on A_0 and $A(1)$ can be used in concert to identify the model. See Gali (1992) for an empirical example.

$$E(\varepsilon_t \varepsilon_t') = \Sigma_\varepsilon, \tag{4.13}$$

where ε_t and Σ_ε are functions of the unknown θ. GMM estimators are formed by choosing $\hat{\theta}$ so that (4.12) and (4.13) are satisfied, or satisfied as closely as possible, with sample moments used in place of the population moments.

The key ideas underlying the GMM estimator in the structural VAR model can be developed using the bivariate output–money example in (4.7). This avoids the cumbersome notation associated with the n-equation model and arbitrary covariance restrictions. [See Hausman et al. (1987) for discussion of the general case.] Assume that the model is identified by linear restrictions on the coefficients of $A_0, A_1, \ldots, A_p$ and the restriction that $E(\varepsilon_{1,t}\varepsilon_{2,t}) = 0$. Let $w_{1,t}$ denote the variables appearing on the right hand side of (4.7a) after the restrictions on the structural coefficients have been solved out, and let δ_1 denote the corresponding coefficients. Thus, if $a_{12,0} = 0$ is the only coefficient restriction in (4.7a), then only lags of y_t appear in the equation and $w_{1,t} = (y_{t-1}', y_{t-2}', \ldots, y_{t-p}')'$. If the long-run neutrality assumption $\sum_{i=0}^{p} a_{12,i} = 0$ is imposed in (4.7a), then $w_{1,t} = (y_{1,t-1}, y_{1,t-2}, \ldots, y_{1,t-p}, \Delta y_{2,t}, \Delta y_{2,t-1}, \ldots, \Delta y_{2,t-p+1})'$.[46] Defining $w_{2,t}$ and δ_2 analogously for equation (4.7b), the model can be written as

$$y_{1,t} = w_{1,t}'\delta_1 + \varepsilon_{1,t}, \tag{4.14a}$$

$$y_{2,t} = w_{2,t}'\delta_2 + \varepsilon_{2,t}, \tag{4.14b}$$

and the GMM moment equations are:

$$E(z_t \varepsilon_{1,t}) = 0, \tag{4.15a}$$

$$E(z_t \varepsilon_{2,t}) = 0, \tag{4.15b}$$

$$E(\varepsilon_{1,t}\varepsilon_{2,t}) = 0, \tag{4.15c}$$

$$E(\varepsilon_{i,t}^2 - \sigma_{\varepsilon_i}^2) = 0, \quad i = 1, 2. \tag{4.15d}$$

The sample analogues of (4.15a)–(4.15c) determine the estimators $\hat{\delta}_1$ and $\hat{\delta}_2$, while (4.15d) determines $\hat{\sigma}_{\varepsilon_1}^2$ and $\hat{\sigma}_{\varepsilon_2}^2$ as sample averages of sums of squared residuals.

Since the estimators of $\hat{\sigma}_{\varepsilon_1}^2$ and $\hat{\sigma}_{\varepsilon_2}^2$ are standard, we focus on (4.15a)–(4.15c) and the resulting estimators of δ_1 and δ_2. Let $u_t = (z_t'\varepsilon_{1,t}, z_t'\varepsilon_{2,t}, \varepsilon_{1,t}\varepsilon_{2,t})'$ and $\bar{u} = T^{-1}\sum u_t$ denote the sample values of the second moments in (4.15a)–(4.15c). Then the GMM estimators, $\hat{\delta}_1$ and $\hat{\delta}_2$, are values of δ_1 and δ_2 that minimize

[46] If $a_{12}(L) = \sum_{i=0}^{p} a_{12,i} L^i$ and $a_{12}(1) = 0$, then $a_{12}(L)y_{2,t} = a_{12}^*(L)(1-L)y_{2,t} = a_{12}^*(L)\Delta y_{2,t}$, where $a_{12}^*(L) = \sum_{i=0}^{p-1} a_{12,i}^* L^i$, where $a_{12,i}^* = -\sum_{j=i+1}^{p} a_{12,j}$. The discussion that follows assumes the linear restrictions on the structural coefficients are homogeneous (or zero). As usual, the only change required for nonhomogeneous (or nonzero) linear restrictions is a redefinition of the dependent variable.

$$J = \bar{u}' \hat{\Sigma}_u^{-1} \bar{u}, \tag{4.16}$$

where $\hat{\Sigma}_u$ is a consistent estimator of $E(u_t u_t')$.[47] These estimators have a simple GLS or instrumental variable interpretation. To see this, let $Z = (z_1\, z_2 \cdots z_T)'$ denote the $T \times 2p$ matrix of instruments; let $W_1 = (w_{1,1}\, w_{1,2} \cdots w_{1,T})'$ and $W_2 = (w_{2,1}\, w_{2,2} \cdots w_{2,T})'$ denote the $T \times k_1$ and $T \times k_2$ matrices of right hand side variables; finally, let Y_1, Y_2, ε_1 and ε_2 denote the $T \times 1$ vectors composed of $y_{1,t}, y_{2,t}, \varepsilon_{1,t}$ and $\varepsilon_{2,t}$, respectively. Multiplying equations (4.14a) and (4.14b) by z_t and summing yields

$$Z'Y_1 = (Z'W_1)\delta_1 + Z'\varepsilon_1, \tag{4.17a}$$

$$Z'Y_2 = (Z'W_2)\delta_2 + Z'\varepsilon_2. \tag{4.17b}$$

Now, letting $\bar{\varepsilon}_i = Y_i - W_i\bar{\delta}_i$, for some $\bar{\delta}_i$

$$\bar{\varepsilon}_1'\bar{\varepsilon}_2 + \bar{\varepsilon}_1' W_2\bar{\delta}_2 + \bar{\varepsilon}_2' W_1\bar{\delta}_1 = (\bar{\varepsilon}_2' W_1)\delta_1 + (\bar{\varepsilon}_1' W_2)\delta_2 + \varepsilon_1'\varepsilon_2 + \text{quadratic terms}. \tag{4.17c}$$

Stacking equations (4.17a)–(4.17c) and ignoring the quadratic terms in (4.17c) yields

$$Q = P_1\delta_1 + P_2\delta_2 + V, \tag{4.18}$$

where $Q = [(Z'Y_1)|(Z'Y_2)|(\bar{\varepsilon}_1'\bar{\varepsilon}_2 + \bar{\varepsilon}_1' W_2\bar{\delta}_2 + \bar{\varepsilon}_2' W_1\bar{\delta}_1)]$, $P_1 = [(Z'W_1)|0_{2p\times k_1}|(\bar{\varepsilon}_2' W_1)]$, $P_2 = [0_{2p\times k_2}|(Z'W_2)|(\bar{\varepsilon}_1' W_2)]$, and $V = [(Z'\varepsilon_1)|(Z'\varepsilon_2)|(\varepsilon_1'\varepsilon_2)]$, and where "|" denotes vertical concatenation ("stacking"). By inspection $V = T\bar{u}$ from (4.16). Thus when Q, P_1 and P_2 are evaluated at $\bar{\delta}_1 = \hat{\delta}_1$ and $\bar{\delta}_2 = \hat{\delta}_2$, the GMM estimators coincide with the GLS estimators from (4.18). This means that the GMM estimators can be formed by iterative GLS estimation of equations (4.18), updating $\bar{\delta}_1$ and $\bar{\delta}_2$ at each iteration and using $T^{-1}\sum \hat{u}_t\hat{u}_t'$ as the GLS covariance matrix.

Hausman et al. (1987) show that the resulting GMM estimators of $\delta_1, \delta_2, \sigma_{\varepsilon_1}^2$ and $\sigma_{\varepsilon_2}^2$ are jointly asymptotically normally distributed when the vectors $(z_t'\varepsilon_t')'$ are independently distributed and standard regularity conditions hold. These results are readily extended to the structural VAR when the roots of $\Phi(z)$ are outside the unit circle, so that the data are covariance stationary. Expressions for the asymptotic variance of the GMM estimators are given in their paper. When some of the variables in the model are integrated, the asymptotic distribution of the estimators changes in a way like that discussed in Section 2. This issue does not seem to have been studied explicitly in the structural VAR model, although such an analysis would seem to be reasonably straightforward.[48]

[47] When elements of u_t and u_τ are correlated for $t \neq \tau$, $\hat{\Sigma}_u$ is replaced by a consistent estimator of the limiting value of the variance of $T^{1/2}\bar{u}$.

[48] Instrumental variable estimators constructed from possibly integrated regressors and instruments are discussed in Phillips and Hansen (1990).

The paper by Hausman et al. (1987) also discuss the relationship between efficient GMM estimators and the FIML estimator constructed under the assumption that the errors are normally distributed. It shows that the FIML estimator can be written as the solution to (4.16), using a specific estimator of Σ_u appropriate under the normality assumption. In particular, FIML uses a block diagonal estimator of Σ_u, since $E[(\varepsilon_{1,t}\varepsilon_{2,t})(\varepsilon_{1,t}z_t)] = E([\varepsilon_{1,t}\varepsilon_{2,t})(\varepsilon_{2,t}z_t)] = 0$ when the errors are normally distributed. When the errors are not normally distributed, this estimator of Σ_u may be inconsistent, leading to a loss of efficiency in the FIML estimator relative to the efficient GMM estimator.

Estimation is simplified when there are no overidentifying restrictions. In this case, iteration is not required, and the GMM estimators can be constructed as instrumental variable (IV) estimators. When the model is just identified, only one restriction is imposed on the coefficient in equation (4.7). This implies that one of the vectors δ_1 or δ_2 is $2p \times 1$, while the other is $(2p+1) \times 1$, and (4.17) is a set of $4p+1$ linear equation in $4p+1$ unknowns. Suppose, without loss of generality, that δ_1 is $2p \times 1$. Then $\hat{\delta}_1$ is determined from (4.17a) as $\hat{\delta}_1 = (Z'W_1)^{-1}(Z'Y_1)$, which is the usual IV estimator of equation (4.14a) using z_t as instruments. Using this value for $\bar{\delta}_1$ in (4.17c) and noting that $Y_2 = W_2\bar{\delta}_2 + \bar{\varepsilon}_2$, equation (4.17c) becomes

$$\hat{\varepsilon}_1' Y_2 = (\hat{\varepsilon}_1' W_2)\delta_2 + \varepsilon_1'\varepsilon_2, \tag{4.19}$$

where $\hat{\varepsilon}_1 = Y_1 - W_1\hat{\delta}_1$ is the residual from the first equation. The GMM estimator of δ_2 is formed by solving (4.17b) and (4.19) for $\hat{\delta}_2$. This can be recognized as the IV estimator of equation (4.14b) using z_t and the residual from (4.14a) as an instrument. The residual is a valid instrument because of the covariance restriction (4.15c).[49]

In many structural VAR exercises, the impulse response functions and variance decompositions defined in Section 4.2 are of more interest than the parameters of the structural VAR. Since $C(L) = A(L)^{-1}$, the moving average parameters/impulse responses and the variance decompositions are differentiable functions of the structural VAR parameters. The continuous mapping theorem directly yields the asymptotic distribution of these parameters from the distribution of the structural VAR parameters. Formulae for the resulting covariance matrix can be determined by delta method calculations. Convenient formulae for these covariance matrices can be found in Lutkepohl (1990), Mittnik and Zadrozny (1993) and Hamilton (1994).

[49] While this instrumental variables scheme provides a simple way to compute the GMM estimator using standard computer software, the covariance matrix of the estimators constructed using the usual formula will not be correct. Using $\hat{\varepsilon}_{1,t}$ as an instrument introduces "generated regressor" complications familiar from Pagan (1984). Corrections for the standard formula are provided in King and Watson (1993). An alternative approach is to carry out one GMM iteration using the IV estimators as starting values. The point estimates will remain unchanged, but standard GMM software will compute a consistent estimator of the correct covariance matrix. The usefulness of residuals as instruments is discussed in more detail in Hausman (1983), Hausman and Taylor (1983) and Hausman et al. (1987).

Many applied researchers have instead relied on Monte Carlo methods for estimating standard errors of estimated impulse responses and variance decompositions. Runkle (1987) reports on experiments comparing the small sample accuracy of the estimators. He concludes that the delta method provides reasonably accurate estimates of the standard errors for the impulse responses, and the resulting confidence intervals have approximately the correct coverage. On the other hand, delta method confidence intervals for the variance decompositions are often unsatisfactory. This undoubtedly reflects the [0, 1] bounded support of the variance decompositions and the unbounded support of the delta method normal approximation.

References

Ahn, S.K. and G.C. Reinsel (1990) "Estimation for Partially Nonstationary Autoregressive Models", *Journal of the American Statistical Association*, 85, 813–823.

Anderson, T.W. (1951) "Estimating Linear Restrictions on Regression Coefficients for Multivariate Normal Distributions", *Annals of Mathematical Statistics*, 22, 327–51.

Anderson, T.W. (1984) *An Introduction to Multivariate Statistical Analysis, 2nd Edition.* Wiley: New York.

Andrews, D.W.K. and J.C. Moynihan (1990) An Improved Heteroskedastic and Autocorrelation Consistent Covariance Matrix Estimator, Cowles Foundation Discussion Paper No. 942, Yale University.

Banerjee, A., J.J. Dolado, D.F. Hendry and G.W. Smith (1986) "Exploring Equilibrium Relationships in Econometrics through Static Models: Some Monte Carlo Evidence", *Oxford Bulletin of Economics and Statistics*, 48(3), 253–70.

Banerjee, A., J. Dolado, J.W. Galbraith and D.F. Hendry (1993) *Co-Integration, Error-Correction, and the Econometric Analysis of Non-Stationary Data.* Oxford University Press: Oxford.

Basawa, I.V. and D.J. Scott (1983) *Asymptotic Optimal Inference for Nonergodic Models.* Springer Verlag: New York.

Berk, K.N. (1974) "Consistent Autoregressive Spectral Estimates", *Annals of Statistics*, 2, 489–502.

Bernanke, B. (1986) "Alternative Explanations of the Money–Income Correlation", *Carnegie-Rochester Conference Series on Public Policy.* Amsterdam: North-Holland Publishing Company.

Beveridge, S. and C.R. Nelson (1981) "A New Approach to Decomposition of Time Series in Permanent and Transitory Components with Particular Attention to Measurement of the 'Business Cycle'", *Journal of Monetary Economics*, 7, 151–74.

Blanchard, O.J. and D. Quah (1989) "The Dynamic Effects of Aggregate Demand and Supply Disturbances", *American Economic Review*, 79, 655–73.

Blanchard, O.J. and M.W. Watson (1986) "Are Business Cycles All Alike?", in: R. Gordon, ed., *The American Business Cycle: Continuity and Change.* Chicago: University of Chicago Press, 123–179.

Bobkosky, M.J. (1983) Hypothesis Testing in Nonstationary Time Series, Ph.D. thesis, Department of Statistics, University of Wisconsin.

Brillinger, D.R. (1980) *Time Series, Data Analysis and Theory.* Expanded Edition, Holden–Day: San Francisco.

Campbell, J.Y. (1990) "Measuring the Persistence of Expected Returns", *American Economic Review*, 80(2), 43–47.

Campbell, J.Y. and P. Perron (1991) "Pitfalls and Opportunities: What Macroeconomists Should Know about Unit Roots", *NBER Macroeconomics Annual.* MIT Press: Cambridge, Mass.

Campbell, J.Y. and R.J. Shiller (1987) "Cointegration and Tests of Present Value Models", *Journal of Political Economy*, 95, 1062–1088. Reprinted in: R.F. Engle and C.W.J. Granger, eds., *Long-Run Economic Relations: Readings in Cointegration.* Oxford University Press: New York, 1991.

Canova, Fabio (1991) Vector Autoregressive Models: Specification Estimation and Testing, manuscript, Brown University.

Cavanagh, C.L. (1985) Roots Local to Unity, manuscript, Department of Economics, Harvard University.
Cavanagh, C.L. and J.H. Stock (1985) Inference in Econometric Models with Nearly Nonstationary Regressors, Manuscript, Kennedy School of Government, Harvard University.
Chan, N.H. (1988) "On Parameter Inference for Nearly Nonstationary Time Series", *Journal of the American Statistical Association*, 83, 857–62.
Chan, N.H. and C.Z. Wei (1987) "Asymptotic Inference for Nearly Nonstationary AR(1) Processes", *The Annals of Statistics*, 15, 1050–63.
Chan, N.H. and C.Z. Wei (1988) "Limiting Distributions of Least Squares Estimates of Unstable Autoregressive Processes", *The Annals of Statistics*, 16(1), 367–401.
Cochrane, J.H. (1994) "Permanent and Transitory Components of GNP and Stock Prices", *Quarterly Journal of Economics*, 109, 241–266.
Cochrane, J.H. and A.M. Sbordone (1988) "Multivariate Estimates of the Permanent Components of GNP and Stock Prices", *Journal of Economic Dynamics and Control*, 12, 255–296.
Davidson, J.E., D.F. Hendry, F. Srba and S. Yeo (1978) "Econometric Modelling of the Aggregate Time-Series Relationship Between Consumer's Expenditures and Income in the United Kingdom", *Economic Journal*, 88, 661–692.
Davies, R.B. (1977) "Hypothesis Testing When a Parameter is Present Only Under the Alternative", *Biometrika*, 64, 247–54.
Davies, R.B. (1987) "Hypothesis Testing When a Parameter is Present Only Under the Alternative", *Biometrika*, 74, 33–43.
Dickey, D.A. and W.A. Fuller (1979) "Distribution of the Estimators for Autoregressive Time Series with a Unit Root", *Journal of the American Statistical Association*, 74, 427–31.
Elliot, G. and J.H. Stock (1992) Inference in Time Series Regressions when there is Uncertainty about Whether a Regressor Contains a Unit Root, manuscript, Harvard University.
Elliot, G., T.J. Rothenberg and J.H. Stock (1992) Efficient Tests of an Autoregressive Unit Root, NBER Technical Working Paper 130.
Engle, R.F. (1976) "Band Spectrum Regression", *International Economic Review*, 15, 1–11.
Engle, R.F. (1984) "Wald, Likelihood Ratio, and Lagrange Multiplier Tests in Econometrics", in: Z. Griliches and M. Intriligator, eds., *Handbook of Econometrics*. North-Holland: New York, Vol. 2, 775–826.
Engle, R.F. and C.W.J. Granger (1987) "Cointegration and Error Correction: Representation, Estimation, and Testing", *Econometrica*, 55, 251–276. Reprinted in: R.F. Engle and C.W.J. Granger, eds., *Long-Run Economic Relations: Readings in Cointegration*. Oxford University Press: New York, 1991.
Engle, R.F. and B.S. Yoo (1987) "Forecasting and Testing in Cointegrated Systems", *Journal of Econometrics*, 35, 143–59. Reprinted in: R.F. Engle and C.W.J. Granger, eds., *Long-Run Economic Relations: Readings in Cointegration*. Oxford University Press: New York, 1991.
Engle, R.F. and B.S. Yoo (1991) "Cointegrated Economic Time Series: An Overview with New Results", in R.F. Engle and C.W.J. Granger, eds., *Long-Run Economic Relations: Readings in Cointegration*. Oxford University Press: New York.
Engle, R.F., D.F. Hendry, and J.F. Richard (1983) "Exogeneity", *Econometrica*, 51(2), 277–304.
Fama, E.F. and K.R. French (1988) "Permanent and Transitory Components of Stock Prices", *Journal of Political Economy*, 96, 246–73.
Faust, J. and E.M. Leeper (1993) Do Long-Run Identifying Restrictions Identify Anything?, manuscript, Board of Governors of the Federal Reserve System.
Fisher, F. (1966) *The Identification Problem in Econometrics*. New York: McGraw–Hill.
Fisher, M.E. and J.J. Seater (1993) "Long-Run Neutrality and Superneutrality in an ARIMA Framework", *American Economic Review*, 83(3), 402–415.
Fountis, N.G. and D.A. Dickey (1986) Testing for a Unit Root Nonstationarity in Multivariate Time Series, manuscript, North Carolina State University.
Fuller, W.A. (1976) *Introduction to Statistical Time Series*. New York: Wiley.
Gali, J. (1992) "How Well does the IS-LM Model Fit Postwar U.S. Data", *Quarterly Journal of Economics*, 107, 709–738.
Geweke, J. (1986) "The Superneutrality of Money in the United States: An Interpretation of the Evidence", *Econometrica*, 54, 1–21.
Gianini, C. (1991) Topics in Structural VAR Econometrics, manuscript, Department of Economics, Universita Degli Studi Di Ancona.

Gonzalo, J. (1989) Comparison of Five Alternative Methods of Estimating Long Run Equilibrium Relationships, manuscript, UCSD.
Granger, C.W.J. (1969) "Investigating Causal Relations by Econometric Methods and Cross Spectral Methods", *Econometrica*, 34, 150–61.
Granger, C.W.J. (1983) Co-Integrated Variables and Error-Correcting Models, UCSD Discussion Paper 83–13.
Granger, C.W.J. and A.P. Andersen (1978) *An Introduction to Bilinear Time Series Models*. Vandenhoeck & Ruprecht: Gottingen.
Granger, C.W.J. and T.-H. Lee (1990) "Multicointegration", *Advances in Econometrics*, 8, 71–84. Reprinted in: R.F. Engle and C.W.J. Granger, eds., *Long-Run Economic Relations: Readings in Cointegration*. Oxford University Press: New York, 1991.
Granger, C.W.J. and P. Newbold (1974) "Spurious Regressions in Econometrics", *Journal of Econometrics*, 2, 111–20.
Granger, C.W.J. and P. Newbold (1976) *Forecasting Economic Time Series*. Academic Press: New York.
Granger, C.W.J. and A.A. Weiss (1983) "Time Series Analysis of Error-Correcting Models", in: *Studies in Econometrics, Time Series and Multivariate Statistics*. Academic Press: New York, 255–78.
Hall, R.E. (1978) "Stochastic Implications of the Life Cycle – Permanent Income Hypothesis: Theory and Evidence", *Journal of Political Economy*, 86(6), 971–87.
Hamilton, J.D. (1994) *Time Series Analysis*. Princeton University Press: Princeton, NJ.
Hannan, E.J. (1970) *Multiple Time Series* Wiley: New York.
Hansen, B.E. (1988) Robust Inference in General Models of Cointegration, manuscript, Yale University.
Hansen, B.E. (1990a) A Powerful, Simple Test for Cointegration Using Cochrane–Orcutt, Working Paper No. 230, Rochester Center for Economic Research.
Hansen, B.E. (1990b) Inference When a Nuisance Parameter is Not Identified Under the Null Hypothesis, manuscript, University of Rochester.
Hansen, B.E. and P.C.B. Phillips (1990) "Estimation and Inference in Models of Cointegration: A Simulation Study", *Advances in Econometrics*, 8, 225–248.
Hansen, L.P. (1982) "Large Sample Properties of Generalized Method of Moments Estimators", *Econometrica*, 50, 1029–54.
Hansen, L.P. and T.J. Sargent (1991) "Two Problems in Interpreting Vector Autoregressions", in: L. Hansen and T. Sargent, eds., *Rational Expectations Econometrics*. Westview: Boulder.
Hausman, J.A. (1983) "Specification and Estimation of Simultaneous Equation Models", in: Z. Griliches and M. Inteligator, eds., *Handbook of Econometrics*. North Holland: New York, Vol. 1, 391–448.
Hausman, J.A. and W.E. Taylor (1983) "Identification in Linear Simultaneous Equations Models with Covariance Restrictions: An Instrumental Variables Interpretation", *Econometrica*, 51(5), 1527–50.
Hausman, J.A., W.K. Newey and W.E. Taylor (1987) "Efficient Estimation and Identification of Simultaneous Equation Models with Covariance Restrictions", *Econometrica*, 55(4), 849–874.
Hendry, D.F. and T. von Ungern-Sternberg (1981) "Liquidity and Inflation Effects on Consumer's Expenditure", in: A.S. Deaton, ed., *Essays in the Theory and Measurement of Consumer's Behavior*. Cambridge University Press: Cambridge.
Hodrick, R.J. (1992) "Dividend Yields and Expected Stock Returns: Alternative Procedures for Inference and Measurement", *The Review of Financial Studies*, 5(3), 357–86.
Horvath, M. and M. Watson (1992) Critical Values for Likelihood Based Tests for Cointegration When Some Cointegrating May be Known, manuscript, Northwestern University.
Horvath, M. and M.W. Watson (1993) Testing for Cointegration When Some of the Cointegrating Vectors are Known, manuscript, Northwestern University.
Johansen, S. (1988a) "Statistical Analysis of Cointegrating Vectors", *Journal of Economic Dynamics and Control*, 12, 231–54. Reprinted in: R.F. Engle and C.W.J. Granger, eds., *Long-Run Economic Relations: Readings in Cointegration*. Oxford University Press: New York, 1991.
Johansen, S. (1988b) "The Mathematical Structure of Error Correction Models", in: N.U. Prabhu, ed., *Contemporary Mathematics, vol. 80: Structural Inference from Stochastic Processes*. American Mathematical Society: Providence, RI.
Johansen, S. (1991) "Estimation and Hypothesis Testing of Cointegrating Vectors in Gaussian Vector Autoregression Models", *Econometrica*, 59, 1551–1580.
Johansen, S. (1992a) The Role of the Constant Term in Cointegration Analysis of Nonstationary Variables, Preprint No. 1, Institute of Mathematical Statistics, University of Copenhagen.
Johansen, S. (1992b) "Determination of Cointegration Rank in the Presence of a Linear Trend", *Oxford Bulletin of Economics and Statistics*, 54, 383–397.

Johansen, S. (1992c) "A Representation of Vector Autoregressive Processes Integrated of Order 2", *Econometric Theory*, 8(2), 188–202.
Johansen, S. and K. Juselius (1990) "Maximum Likelihood Estimation and Inference on Cointegration – with Applications to the Demand for Money", *Oxford Bulletin of Economics and Statistics*, 52(2), 169–210.
Johansen, S. and K. Juselius (1992) "Testing Structural Hypotheses in a Multivariate Cointegration Analysis of the PPP and UIP of UK", *Journal of Econometrics*, 53, 211–44.
Keating, J. (1990) "Identifying VAR Models Under Rational Expectations", *Journal of Monetary Economics*, 25(3), 453–76.
King, R.G. and M.W. Watson (1993) Testing for Neutrality, manuscript, Northwestern University.
King, R.G., C.I. Plosser, J.H. Stock and M.W. Watson (1991) "Stochastic Trends and Economic Fluctuations", *American Economic Review*, 81, 819–840.
Kosobud, R. and L. Klein (1961) "Some Econometrics of Growth: Great Ratios of Economics", *Quarterly Journal of Economics*, 25, 173–98.
Lippi, M. and L. Reichlin (1993) "The Dynamic Effects of Aggregate Demand and Supply Disturbances: Comments", *American Economic Review*, 83(3), 644–652.
Lucas, R.E. (1972) "Econometric Testing of the Natural Rate Hypothesis", in: O. Eckstein, ed., *The Econometrics of Price Determination*. Washington, D.C.: Board of Governors of the Federal Reserve System.
Lucas, R.E. (1988) "Money Demand in the United States: A Quantitative Review", *Carnegie–Rochester Conference Series on Public Policy*, 29, 137–68.
Lutkepohl, H. (1990) "Asymptotic Distributions of Impulse Response Functions and Forecast Error Variance Decompositions of Vector Autoregressive Models", *Review of Economics and Statistics*, 72, 116–25.
MacKinnon, J.G. (1991) "Critical Values for Cointegration Tests", in: R.F. Engle and C.W.J. Granger, eds., *Long-Run Economic Relations: Readings in Cointegration*. Oxford University Press: New York.
Magnus, J.R. and H. Neudecker (1988) *Matrix Differential Calculus*. Wiley: New York.
Malinvaud, E. (1980) *Statistical Methods of Econometrics*. Amsterdam: North-Holland.
Mankiw, N.G. and M.D. Shapiro (1985) "Trends, Random Walks and the Permanent Income Hypothesis", *Journal of Monetary Economics*, 16, 165–74.
Mittnik, S. and P.A. Zadrozny (1993) "Asymptotic Distributions of Impulse Responses, Step Responses and Variance Decompositions of Estimated Linear Dynamic Models", *Econometrica*, 61, 857–70.
Ogaki, M. and J.Y. Park (1990) A Cointegration Approach to Estimating Preference Parameters, manuscript, University of Rochester.
Osterwald-Lenum, M. (1992) "A Note with Quantiles of the Asymptotic Distribution of the Maximum Likelihood Cointegration Rank Test Statistics", *Oxford Bulletin of Economics and Statistics*, 54, 461–71.
Pagan, A. (1984) "Econometric Issues in the Analysis of Regressions with Generated Regressors", *International Economic Review*, 25, 221–48.
Park, J.Y. (1992) "Canonical Cointegrating Regression", *Econometrica*, 60(1), 119–144.
Park, J.Y. and M. Ogaki (1991) Inference in Cointegrated Models Using VAR Prewhitening to Estimate Shortrun Dynamics, Rochester Center for Economic Research Working Paper No. 281.
Park, J.Y. and P.C.B. Phillips (1988) "Statistical Inference in Regressions with Integrated Regressors I", *Econometric Theory*, 4, 468–97.
Park, J.Y. and P.C.B. Phillips (1989) "Statistical Inference in Regressions with Integrated Regressors: Part 2", *Econometric Theory*, 5, 95–131.
Phillips, P.C.B. (1986) "Understanding Spurious Regression in Econometrics", *Journal of Econometrics*, 33, 311–40.
Phillips, P.C.B. (1987a) "Time Series Regression with a Unit Root", *Econometrica*, 55, 277–301.
Phillips, P.C.B. (1987b) "Toward a Unified Asymptotic Theory for Autoregression", *Biometrika*, 74, 535–47.
Phillips, P.C.B. (1988) "Multiple Regression with Integrated Regressors", *Contemporary Mathematics*, 80, 79–105.
Phillips, P.C.B. (1991a) "Optimal Inference in Cointegrated Systems", *Econometrica*, 59(2), 283–306.
Phillips, P.C.B. (1991b) "Spectral Regression for Cointegrated Time Series", in: W. Barnett, ed., *Nonparametric and Semiparametric Methods in Economics and Statistics*. Cambridge University Press: Cambridge, 413–436.

Phillips, P.C.B. (1991c) The Tail Behavior of Maximum Likelihood Estimators of Cointegrating Coefficients in Error Correction Models, manuscript, Yale University.
Phillips, P.C.B. (1991d) "To Criticize the Critics: An Objective Bayesian Analysis of Stochastic Trends", *Journal of Applied Econometrics*, 6, 333–364.
Phillips, P.C.B. and S.N. Durlauf (1986) "Multiple Time Series Regression with Integrated Processes", *Review of Economic Studies*, 53, 473–96.
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, 407–436.
Phillips, P.C.B. and S. Ouliaris (1990) "Asymptotic Properties of Residual Based Tests for Cointegration", *Econometrica*, 58, 165–94.
Phillips, P.C.B. and J.Y. Park (1988) "Asymptotic Equivalence of OLS and GLS in Regression with Integrated Regressors", *Journal of the American Statistical Association*, 83, 111–115.
Phillips, P.C.B. and P. Perron (1988) "Testing for Unit Root in Time Series Regression", *Biometrika*, 75, 335–46.
Phillips, P.C.B. and W. Ploberger (1991) Time Series Modeling with a Bayesian Frame of Reference: I. Concepts and Illustrations, manuscript, Yale University.
Phillips, P.C.B. and V. Solo (1992) "Asymptotics for Linear Processes", *Annals of Statistics*, 20, 971–1001.
Quah, D. (1986) Estimation and Hypothesis Testing with Restricted Spectral Density Matrices: An Application to Uncovered Interest Parity, Chapter 4 of Essays in Dynamic Macroeconometrics, Ph.D. Dissertation, Harvard University.
Rothenberg, T.J. (1971) "Identification in Parametric Models", *Econometrica*, 39, 577–92.
Rozanov, Y.A. (1967) *Stationary Random Processes*. San Francisco: Holden Day.
Runkle, D. (1987) "Vector Autoregressions and Reality", *Journal of Business and Economic Statistics*, 5(4), 437–442.
Said, S.E. and D.A. Dickey (1984) "Testing for Unit Roots in Autoregressive-Moving Average Models of Unknown Order", *Biometrika*, 71, 599–608.
Saikkonen, P. (1991) "Asymptotically Efficient Estimation of Cointegrating Regressions", *Econometric Theory*, 7(1), 1–21.
Saikkonen, P. (1992) "Estimation and Testing of Cointegrated Systems by an Autoregressive Approximation", *Econometric Theory*, 8(1), 1–27.
Sargan, J.D. (1964) "Wages and Prices in the United Kingdom: A Study in Econometric Methodology", in: P.E. Hart, G. Mills and J.N. Whittaker, eds., *Econometric Analysis for National Economic Planning*. London: Butterworths.
Sargent, T.J. (1971) "A Note on the Accelerationist Controversy", *Journal of Money, Banking and Credit*, 3, 50–60.
Shapiro, M. and M.W. Watson (1988) "Sources of Business Cycle Fluctuations", *NBER Macroeconomics Annual*, 3, 111–156.
Sims, C.A. (1972) "Money, Income and Causality", *American Economic Review*, 62, 540–552.
Sims, C.A. (1978) Least Squares Estimation of Autoregressions with Some Unit Roots, University of Minnesota, Discussion Paper No. 78–95.
Sims, C.A. (1980) "Macroeconomics and Reality", *Econometrica*, 48, 1–48.
Sims, C.A. (1986) "Are Forecasting Models Usable for Policy Analysis?", *Quarterly Review*, Federal Reserve Bank of Minneapolis, Winter.
Sims, C.A. (1989) "Models and Their Uses", *American Journal of Agricultural Economics*, 71, 489–494.
Sims, C.A., J.H. Stock and M.W. Watson (1990) "Inference in Linear Time Series Models with Some Unit Roots", *Econometrica*, 58(1), 113–44.
Solo, V. (1984) "The Order of Differencing in ARIMA Models", *Journal of the American Statistical Association*, 79, 916–21.
Stock, J.H. (1987) "Asymptotic Properties of Least Squares Estimators of Cointegrating Vectors", *Econometrica*, 55, 1035–56.
Stock, J.H. (1988) "A Reexamination of Friedman's Consumption Puzzle", *Journal of Business and Economic Statistics*, 6(4), 401–14.
Stock, J.H. (1991) "Confidence Intervals of the Largest Autoregressive Root in U.S. Macroeconomic Time Series", *Journal of Monetary Economics*, 28, 435–60.

Stock, J.H. (1992) Deciding Between I(0) and I(1), manuscript, Harvard University.
Stock, J.H. (1993) Forthcoming in: R.F. Engle and D. McFadden, eds., *Handbook of Econometrics*. Vol. 4, North Holland: New York.
Stock, J.H. and M.W. Watson (1988a) "Interpreting the Evidence on Money-Income Causality", *Journal of Econometrics*, 40(1), 161–82.
Stock, J.H. and M.W. Watson (1988b) "Testing for Common Trends", *Journal of the American Statistical Association*, 83, 1097–1107. Reprinted in: R.F. Engle and C.W.J. Granger, eds., *Long-Run Economic Relations: Readings in Cointegration*. Oxford University Press: New York, 1991.
Stock, J.H. and M.W. Watson (1993) "A Simple Estimator of Cointegrating Vectors in Higher Order Integrated Systems", *Econometrica*, 61, 783–820.
Stock, H.H. and K.D. West (1988) "Integrated Regressors and Tests of the Permanent Income Hypothesis", *Journal of Monetary Economics*, 21(1), 85–95.
Sweeting, T. (1983) "On Estimator Efficiency in Stochastic Processes", *Stochastic Processes and their Applications*, 15, 93–98.
Theil, H. (1971) *Principles of Econometrics*. Wiley: New York.
Toda, H.Y. and P.C.B. Phillips (1993a) "Vector Autoregressions and Causality", *Econometrica*, 62(1), 1367–1394.
Toda, H.Y. and P.C.B. Phillips (1993b) "Vector Autoregressions and Causality. A Theoretical Overview and Simulation Study", *Econometric Reviews*, 12, 321–364.
Tsay, R.S. and G.C. Tiao (1990) "Asymptotic Properties of Multivariate Nonstationary Processes with Applications to Autoregressions", *Annals of Statistics*, 18, 220–50.
West, K.D. (1988) "Asymptotic Normality when Regressors Have a Unit Root", *Econometrica*, 56, 1397–1418.
White, H. (1984) *Asymptotic Theory for Econometricians*. New York: Academic Press.
Whittle, P. (1983) *Prediction and Regulation by Linear Least-Square Methods*. Second Edition, Revised. University of Minnesota Press: Minneapolis.
Wold, H. (1954) "Causality and Econometrics", *Econometrica*, 22, 162–177.
Wooldridge, J. (1993) Forthcoming in: R.F. Engle and D. McFadden, eds., *Handbook of Econometrics*. Vol. 4, North-Holland: New York.
Yoo, B.S. (1987) Co-Integrated Time Series Structure, Forecasting and Testing, Ph.D. Dissertation, UCSD.
Yule, G.C. (1926) "Why Do We Sometimes Get Nonsense-Correlations Between Time-Series", *Journal of the Royal Statistical Society B*, 89, 1–64.