Chapter 16

Non-Stationary Time Series

16.1 Introduction

At the beginning of Chapter 14 we displayed a set of economic time series. Several (real GDP, exchange rate, interest rate, crude oil price) did not appear to be stationary. In Section 14.23 we introduced the non-stationary unit root process which is an autoregressive process with an autoregressive root at unity. Plots of two simulated examples (Figure 14.5) displayed time-paths with wandering behavior similar to the economic time series. This suggests that perhaps a unit root autoregression is a reasonable model for these series. In this chapter we explore econometric estimation and inference for non-stationary unit root time series.

16.2 Partial Sum Process and Functional Convergence

Take the multivariate random walk

$$Y_t = Y_{t-1} + e_t$$

where $(e_t, \mathscr{F}_t)$ is a vector MDS with finite covariance matrix Σ. By back-substitution we find $Y_t = Y_0 + S_t$ where

$$S_t = \sum_{i=1}^{t} e_i$$

is the cumulative sum of the errors up to time t. We call S_t a **partial sum process**.

The time index t ranges from 0 to n. Write[1] $t = \lfloor nr \rfloor$ as a fraction r of the sample size n. This allows us to write $S_{\lfloor nr \rfloor}$ as a function of the fraction r. Divide by $\sqrt{n}$ so that the variance is stabilized. With these modifications we define the standardized partial sum process.

$$S_n(r) = \frac{1}{\sqrt{n}} S_{\lfloor nr \rfloor} = \frac{1}{\sqrt{n}} \sum_{t=1}^{\lfloor nr \rfloor} e_t.$$

The random process $S_n(r)$ is a scaled version of the time-series Y_t and is a function of the sample fraction $r \in [0,1]$. It is a stochastic process meaning that it is a random function. For any finite n, $S_n(r)$ is a step function with n jumps.

Let's consider the behavior of $S_n(r)$ as n increases. It's largest discrete jump equals $n^{-1/2} \max_{1 \le t \le n} \|e_t\|$. Theorem 6.15 shows that this is $o_p(1)$. This suggests that the jumps in $S_n(r)$ asymptotically vanish. We would like to find its asymptotic distribution. We expect the limit distribution to be a stochastic process as well.

[1] The notation $\lfloor x \rfloor$ means "round down to the nearest integer".

To do so we need to define the asymptotic distribution of a random function. The primary tool is the functional central limit theorem (FCLT) which is a component of empirical process theory (Chapter 18 of *Probability and Statistics for Economists*). It turns out that the FCLT depends on how we measure the difference between two functions. The most commonly used measure is the **uniform metric**. On the space of functions from $[0,1]$ to $\mathbb{R}^m$ it is

$$\rho\left(v_1, v_2\right) = \sup_{0\le r\le 1} \left\| v_1(r) - v_2(r)\right\|.$$

Convergence in distribution for random processes (e.g. Definition 18.6 of *Probability and Statistics for Economists*) is defined with respect to a specific metric. While we don't repeat the details here the important consequence is that continuity is defined with respect to this metric and this impacts applications such as the continuous mapping theorem.

The **Functional Central Limit Theorem** (Theorem 18.9 of *Probability and Statistics for Economists*) states that $S_n(r) \underset{d}{\longrightarrow} S(r)$ as a function over $r \in [0,1]$ if two conditions hold:

1. The limit distributions of $S_n(r)$ coincide with those of $S(r)$.
2. $S_n(r)$ is asymptotically equicontinuous.

The first condition means that for any fixed $r_1,...,r_m$, $(S_n(r_1),...,S_n(r_m)) \underset{d}{\longrightarrow} (S(r_1),...,S(r_m))$. The second condition is technical but essentially requires that $S_n(r)$ is approximately continuous with respect to the uniform metric in large samples.

We now characterize the limit distributions of $S_n(r)$. There are three important properties.

1. $S_n(0) = 0$.
2. For any r, $S_n(r) \underset{d}{\longrightarrow} \mathrm{N}(0, r\Sigma)$.
3. For $r_1 < r_2$, $S_n(r_1)$ and $S_n(r_2) - S_n(r_1)$ are asymptotically independent.

The first property follows from the definition of $S_n(r)$. For the second, set $N = \lfloor nr \rfloor$. For $r > 0$, $N \to \infty$ as $n \to \infty$. The MDS CLT (Theorem 14.11) implies that

$$S_n(r) = \sqrt{\frac{\lfloor nr \rfloor}{n}} \frac{1}{\sqrt{N}} \sum_{t=1}^{N} e_t \underset{d}{\longrightarrow} \sqrt{r}\mathrm{N}(0,\Sigma) = \mathrm{N}(0, r\Sigma)$$

as claimed. For the third property the assumption that e_t is a MDS implies that $S_n(r_1)$ and $S_n(r_2) - S_n(r_1)$ are uncorrelated. An extension of the above previous asymptotic argument shows that they are jointly asymptotically normal with a zero covariance and hence are asymptotically independent.

The above three limit properties of $S_n(r)$ are asymptotic versions of the definition of Brownian motion.

Definition 16.1 A vector **Brownian motion** $B(r)$ for $r \geq 0$ is defined by the properties:

1. $B(0) = 0$.
2. For any r, $B(r) \sim \mathrm{N}(0, r\Sigma)$.
3. For any $r_1 \leq r_2$, $B(r_1)$ and $B(r_2) - B(r_1)$ are independent.

We call Σ the covariance matrix of $B(r)$. If $\Sigma = \boldsymbol{I}_m$ we say that $B(r)$ is a **standard Brownian motion** and denote it as $W(r)$. It satisfies $B(r) = \Sigma^{1/2} W(r)$.

A Brownian motion $B(r)$ is continuous with probability one but is nowhere differentiable. In physics, Brownian motion is used to describe the movement of particles. The wandering properties of particles suspended in liquid was described as far back as the Roman poet Lucretius (*On the Nature of the Universe*, 55 BCE). The name Brownian motion credits the pioneering observational studies of botanist Robert Brown. The mathematical process is often called a **Wiener process** crediting the work of Norbert Wiener.

The above discussion has shown that the limit distributions of the partial sum process $S_n(r)$ coincide with those of Brownian motion $B(r)$. In Section 16.22 we demonstrate that $S_n(r)$ is asymptotically equicontinuous. Together with the FCLT this establishes that $S_n(r)$ converges in distribution to $B(r)$.

Theorem 16.1 Weak Convergence of Partial Sum Process If $(e_t, \mathscr{F}_t)$ is a strictly stationary and ergodic MDS and $\Sigma = \mathbb{E}\left[e_t e_t'\right] < \infty$ then as a function over $r \in [0,1]$, $S_n(r) \underset{d}{\longrightarrow} B(r)$, a Brownian motion with covariance matrix Σ.

We extend Theorem 16.1 to serially correlated processes in Section 16.4.

Let's connect our analysis of $S_n(r)$ with the random walk series Y_t. Since $Y_t = Y_0 + S_t$, we find

$$\frac{1}{\sqrt{n}} Y_{\lfloor nr \rfloor} = S_n(r) + \frac{1}{\sqrt{n}} Y_0.$$

The second term is $o_p(1)$ when Y_0 is finite with probability one. Thus under this latter assumption $n^{-1/2} Y_{\lfloor nr \rfloor} = S_n(r) + o_p(1) \underset{d}{\longrightarrow} B$. For simplicity we will frequently implicitly assume $Y_0 = 0$ to simplify the notation, as the case with $Y_0 \neq 0$ does not fundamentally change the analysis.

16.3 Beveridge-Nelson Decomposition

The previous section focused on random walk processes. A unit root process more broadly is an autoregression with a single root at unity, which means that the differenced process ΔY_t is serially correlated but stationary.

Beveridge and Nelson (1981) introduced a clever way to decompose a unit root process into a permanent (random walk) component and a transitory (stationary) component. This allows a straightforward generalization of Theorem 16.1 to incorporate serial correlation.

Recall that a stationary process has a Wold representation $\Delta Y_t = \Theta(\mathrm{L}) e_t$ where $\Theta(z) = \sum_{j=0}^{\infty} \Theta_j z^j$.

Assumption 16.1 ΔY_t is strictly stationary with no deterministic component, mean zero, and finite covariance matrix Σ. The coefficients of its Wold representation $\Delta Y_t = \Theta(\mathrm{L}) e_t$ satisfy

$$\sum_{j=0}^{\infty} \left\| \sum_{\ell=j+1}^{\infty} \Theta_\ell \right\| < \infty. \qquad (16.1)$$

The condition (16.1) on the coefficients is stronger than absolute summability but holds (for example) if ΔY_t is generated by a stationary AR process. It is similar to the condition used for the autoregressive Wold representation (Theorem 14.19).

Consider the following factorization of the lag polynomial

$$\Theta(z) = \Theta(1) + (1-z)\Theta^*(z) \qquad (16.2)$$

where $\Theta(1) = \sum_{\ell=0}^{\infty} \Theta_\ell$ and $\Theta^*(z)$ is the lag polynomial

$$\Theta^*(z) = \sum_{j=0}^{\infty} \Theta_j^* z^j \qquad (16.3)$$

$$\Theta_j^* = -\sum_{\ell=j+1}^{\infty} \Theta_\ell. \qquad (16.4)$$

At the end of this section we demonstrate (16.2)-(16.4). Assumption (16.1) is the same as $\sum_{j=0}^{\infty} \left\| \Theta_j^* \right\| < \infty$, which implies that $U_t = \Theta^*(\mathrm{L}) e_t$ is convergent, strictly stationary, and ergodic (by Theorem 15.4).

The factorization (16.2) means that we can write

$$\Delta Y_t = \xi_t + U_t - U_{t-1}.$$

where $\xi_t = \Theta(1) e_t$. This decomposes ΔY_t into the innovation e_t plus the first-difference of the stochastic process U_t. Summing the differences we find

$$Y_t = S_t + U_t + V_0$$

where $S_t = \sum_{i=1}^{t} \xi_t$ and $V_0 = Y_0 - U_0$. This decomposes the unit root process Y_t into the random walk S_t, the stationary process U_t, and an initial condition V_0.

We have established the following.

Theorem 16.2 Under Assumption 16.1 then (16.2)-(16.4) holds with $\sum_{j=0}^{\infty} \left\| \Theta_j^* \right\| < \infty$. The process ΔY_t satisfies

$$\Delta Y_t = \xi_t + U_t - U_{t-1}$$

and

$$Y_t = S_t + U_t + V_0$$

where $S_t = \sum_{i=1}^{t} \xi_t$ is a random walk, ξ_t is white noise with variance $\Theta(1)\Sigma\Theta(1)'$, U_t is strictly stationary, and V_0 is an initial condition.

Beveridge and Nelson (1981) called S_t the **permanent** (trend) component of Y_t and U_t the **transitory** component. They called S_t the permanent component as it determines the long-run behavior of Y_t.

As an example, take the MA(1) case $\Delta Y_t = e_t + \Theta_1 e_{t-1}$. This has decomposition $\Delta Y_t = (\boldsymbol{I}_m + \Theta_1)e_t - \Theta_1(e_t - e_{t-1})$. In this case $U_t = -\Theta_1 e_t$.

The Beveridge-Nelson decomposition of a series is unique but it is not the only way to construct a permanent/transitory decomposition. The Beveridge-Nelson decomposition has the characteristic that the innovations driving the permanent and transitory components S_t and U_t are perfectly correlated. Other decompositions do not use this restriction.

We close this section by verifying (16.2)-(16.4). Observe that the right-side of (16.2) is

$$\begin{aligned}\sum_{j=0}^{\infty}\Theta_j - \sum_{j=0}^{\infty}\sum_{\ell=j+1}^{\infty}\Theta_\ell z^j(1-z) &= \sum_{j=0}^{\infty}\Theta_j - \sum_{j=0}^{\infty}\sum_{\ell=j+1}^{\infty}\Theta_\ell z^j + \sum_{j=0}^{\infty}\sum_{\ell=j+1}^{\infty}\Theta_\ell z^{j+1}\\ &= \Theta_0 - \sum_{j=1}^{\infty}\sum_{\ell=j+1}^{\infty}\Theta_\ell z^j + \sum_{j=1}^{\infty}\sum_{\ell=j}^{\infty}\Theta_\ell z^j\\ &= \Theta_0 + \sum_{j=1}^{\infty}\Theta_j z^j\end{aligned}$$

which is $\Theta(z)$ as claimed.

16.4 Functional CLT

Theorem 16.1 showed that a random walk process converges in distribution to a Brownian motion. We now extend this result to the case of a unit root process with correlated differences.

Under Assumption 16.1 a unit root process can be written as $Y_t = S_t + U_t + V_0$ where $S_t = \sum_{i=1}^{t}\xi_t$. Define the scaled processes $Z_n(r) = n^{-1/2}Y_{\lfloor nr\rfloor}$ and $S_n(r) = n^{-1/2}S_{\lfloor nr\rfloor}$. We find

$$Z_n(r) = S_n(r) + \frac{1}{\sqrt{n}}V_0 + \frac{1}{\sqrt{n}}U_{\lfloor nr\rfloor}.$$

If the errors e_t are a MDS with covariance matrix Σ then by Theorem 16.1, $S_n(r) \underset{d}{\longrightarrow} B(r)$, a vector Brownian motion with covariance matrix $\Omega = \Theta(1)\Sigma\Theta(1)'$. The initial condition $n^{-1/2}V_0$ is $o_p(1)$. The third term $n^{-1/2}U_{\lfloor nr\rfloor}$ is $o_p(1)$ if $\sup_{1\le t\le n}\left|\frac{1}{\sqrt{n}}U_t\right| = o_p(1)$, which holds under Theorem 6.15 if U_t has a finite variance. We now show that this holds under Assumption 16.1. The latter implies that $\sum_{j=0}^{\infty}\left\|\Theta_j^*\right\| < \infty$, as discussed before Theorem 16.2. This implies

$$\|\mathrm{var}[U_t]\| = \left\|\sum_{j=0}^{\infty}\Theta_j^*\Sigma\Theta_j^{*\prime}\right\| \le \|\Sigma\|\sum_{j=0}^{\infty}\left\|\Theta_j^*\right\|^2 \le \|\Sigma\|\max_j\left\|\Theta_j^*\right\|\sum_{j=0}^{\infty}\left\|\Theta_j^*\right\| < \infty$$

as needed.

Together we find that

$$Z_n(r) = S_n(r) + o_p(1) \underset{d}{\longrightarrow} B(r).$$

The variance of the limiting process is $\Omega = \Theta(1)\Sigma\Theta(1)'$. This is the "long-run variance" of ΔY_t.

Theorem 16.3 Under Assumption 16.1 and in addition $(e_t, \mathscr{F}_t)$ is a MDS with covariance matrix Σ, then as a function over $r \in [0,1]$, $Z_n(r) \xrightarrow[d]{} B(r)$ a vector Brownian motion with covariance matrix Ω.

Our derivation used the assumption that the linear projection errors are a MDS. This is not essential for the basic result; the FCLT holds under a variety of dependence conditions. A flexible version can be stated using mixing conditions.

Theorem 16.4 If ΔY_t is strictly stationary, $\mathbb{E}[\Delta Y_t] = 0$, with mixing coefficients $\alpha(\ell)$, and for some $r > 2$, $\mathbb{E}\|\Delta Y_t\|^r < \infty$ and $\sum_{\ell=1}^{\infty} \alpha(\ell)^{1-2/r} < \infty$, then as a function over $r \in [0,1]$, $Z_n(r) \xrightarrow[d]{} B(r)$, a vector Brownian motion with covariance matrix

$$\Omega = \sum_{\ell=-\infty}^{\infty} \mathbb{E}[\Delta Y_t \Delta Y_{t-\ell}]. \tag{16.5}$$

For a proof see Davidson (1994, Theorems 31.5 and 31.15). Interestingly, Theorem 16.4 employs exactly the same assumptions as for Theorem 14.15 (the CLT for mixing processes). This means that we obtain the stronger result (the FCLT) without stronger assumptions.

The covariance matrix Ω appearing in (16.5) is the **long-run covariance matrix** of ΔY_t as defined in Section 14.13. It is useful to observe that we can decompose the long-run variance as $\Omega = \Sigma + \Lambda + \Lambda'$ where $\Sigma = \operatorname{var}[\Delta Y_t]$ and

$$\Lambda = \sum_{\ell=1}^{\infty} \mathbb{E}\left[\Delta Y_t \Delta Y'_{t-\ell}\right].$$

This decomposes the long-run variance of ΔY_t into its static (one-period) variance Σ and a sum of covariances Λ. The matrix Λ is not symmetric.

16.5 Orders of Integration

Take a univariate series Y_t. Theorems 16.3 and 16.4 showed that if ΔY_t is stationary and mean zero then the level process Y_t, suitably scaled, is asymptotically a Brownian motion with variance ω^2. For this theory to be meaningful this variance should be strictly positive definite. To see why this is a potential restriction suppose that $Y_t = a(\mathrm{L})e_t$ where the coefficients of $a(z)$ are absolutely convergent and e_t is i.i.d. $(0, \sigma^2)$. Then $\Delta Y_t = b(\mathrm{L})e_t$ where $b(z) = (1-z)a(z)$ so $\omega^2 = b(1)^2\sigma^2 = 0$. That is, ΔY_t has a long-run variance of 0. We call the process ΔY_t **over-differenced**, since Y_t is strictly stationary and does not require differencing to achieve stationarity.

To meaningfully differentiate between processes which require differencing to achieve stationarity we use the following definition.

Definition 16.2 Order of Integration.

1. $Y_t \in \mathbb{R}$ is **Integrated of Order** 0, written $I(0)$, if Y_t is weakly stationary with positive long-run variance.

2. $Y_t \in \mathbb{R}$ is **Integrated of Order** d, written $I(d)$, if $u_t = \Delta^d Y_t$ is $I(0)$.

$I(0)$ processes are stationary processes which are not over-differenced. $I(1)$ processes include random walks and unit root processes. $I(2)$ processes require double differencing to achieve stationarity. $I(-1)$ processes are stationary but their cumulative sums are also stationary, and are therefore over-differenced stationary processes. Many macroeconomic time series in log-levels are potentially $I(1)$ processes. Economic time series which are potentially $I(2)$ are log price indices, for their first difference (inflation rates) are potentially non-stationary proceses. In this textbook we focus on integer-valued orders of integration but fractional d are also well-defined. In most applications economists presume that economic series are either $I(0)$ or $I(1)$ and often use the shorthand "integrated" to refer to $I(1)$ series.

The long-run variance of ARMA processes is straightforward to calculate. As we have seen, if $\Delta Y_t = b(\mathrm{L})e_t$ where e_t is white noise with variance σ^2, then $\omega^2 = b(1)^2\sigma^2$. Now suppose $a(\mathrm{L})\Delta Y_t = e_t$ where $a(z)$ is invertible. Then $b(z) = a(z)^{-1}$ and $\omega^2 = \sigma^2/a(1)^2$. For an ARMA process $a(\mathrm{L})\Delta Y_t = b(\mathrm{L})e_t$ with invertible $a(z)$, then $\omega^2 = \sigma^2 b(1)^2/a(1)^2$. Hence, if ΔY_t satisfies the ARMA process $a(\mathrm{L})\Delta Y_t = b(\mathrm{L})e_t$ then Y_t is $I(1)$ if $a(z)$ is invertible and $b(1) \neq 0$.

Consider vector processes. The long-run covariance matrix of $\Delta Y_t = \Theta(\mathrm{L})e_t$ is $\Omega = \Theta(1)\Sigma\Theta(1)'$. The long-run covariance matrix of $\boldsymbol{A}(\mathrm{L})\Delta Y_t = e_t$ is $\Omega = \boldsymbol{A}(1)^{-1}\Sigma\boldsymbol{A}(1)^{-1\prime}$. It is conventional to describe the vector ΔY_t as $I(0)$ if each element of ΔY_t is $I(0)$ but this allows its covariance matrix to be singular. To exclude the latter we introduce the following.

Definition 16.3 The vector process Y_t is **full rank** $I(0)$ if its long-run covariance matrix Ω is positive definite.

16.6 Means, Local Means, and Trends

Theorem 16.4 shows that $Z_n(r) \xrightarrow[d]{} B(r)$. The continuous mapping theorem shows that if a function $f(x)$ is continuous[2] then $f(Z_n) \xrightarrow[d]{} f(B)$. This can be used to obtain the asymptotic distribution of many statistics of interest. Simple examples are $Z_n(r)^2 \xrightarrow[d]{} B(r)^2$ and $\int_0^1 Z_n(r)dr \xrightarrow[d]{} \int_0^1 B(r)dr$. The latter produces the asymptotic distribution for the sample mean as we now show.

Let $\overline{Y}_n = n^{-1}\sum_{t=1}^n Y_t$ be the sample mean. For simplicity assume $Y_0 = 0$. Note that for $r \in \left[\frac{t}{n}, \frac{t+1}{n}\right)$,

$$\frac{1}{n^{1/2}} Y_t = Z_n(r) = n\int_{t/n}^{(t+1)/n} Z_n(r)dr.$$

Taking the average for $t = 0$ to $n-1$ we find

$$\frac{1}{n^{1/2}}\overline{Y}_n = \frac{1}{n^{3/2}}\sum_{t=0}^{n-1} Y_t = \sum_{t=0}^{n-1}\int_{t/n}^{(t+1)/n} Z_n(r)dr = \int_0^1 Z_n(r)dr.$$

This is the integral (or average) of $Z_n(r)$ over $[0,1]$.

The continuous mapping theorem can be applied[3]. The above expression converges in distribution to the random variable $\int_0^1 B(r)dr$. This is the average of the Brownian motion over $[0,1]$.

[2] With respect to the uniform metric ρ.

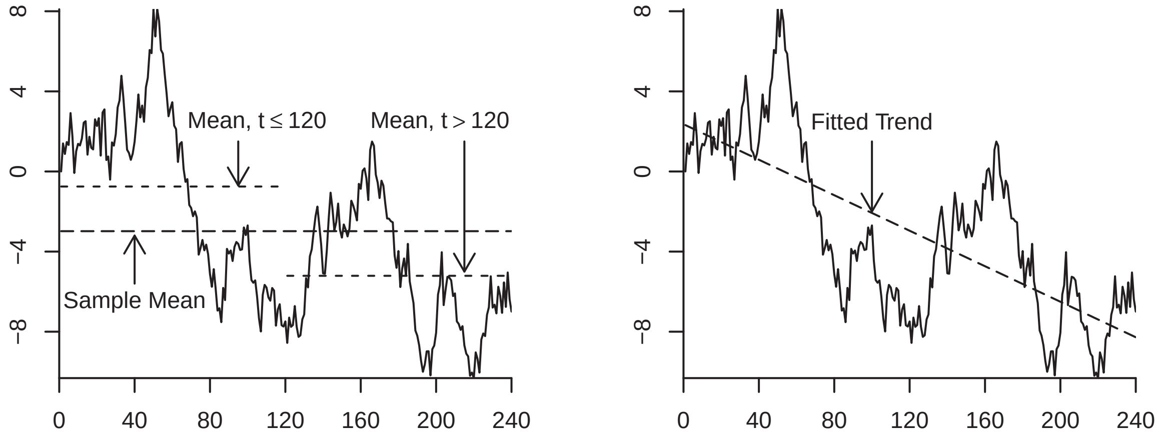
[3] The integral $f(g) = \int_0^1 g(r)dr$ is a continuous function of g with respect to the uniform metric. (Small changes in g result in small changes in f.)

Now consider sub-sample means. Let $\overline{Y}_{1n} = (n/2)^{-1}\sum_{t=0}^{n/2-1} Y_t$ and $\overline{Y}_{2n} = (n/2)^{-1}\sum_{t=n/2}^{n-1} Y_t$ be the sample means on the first-half and second-half of the sample, respectively. By a similar analysis as for the full-sample mean

$$\frac{1}{n^{1/2}}\overline{Y}_{1n} = \frac{2}{n^{3/2}}\sum_{t=0}^{n/2-1} Y_t = 2\int_0^{1/2} Z_n(r)dr \underset{d}{\longrightarrow} 2\int_0^{1/2} B(r)dr$$
$$\frac{1}{n^{1/2}}\overline{Y}_{2n} = \frac{2}{n^{3/2}}\sum_{t=n/2}^{n-1} Y_t = 2\int_{1/2}^{1} Z_n(r)dr \underset{d}{\longrightarrow} 2\int_{1/2}^{1} B(r)dr$$

which are the averages of $B(r)$ over the regions $[0,1/2]$ and $[1/2,1]$. These are distinct random variables. This gives rise to the prediction that if Y_t is a unit root process, sample averages will not be constant (even in large samples) and will vary across subsamples.

Furthermore, observe that the limit distributions were obtained after dividing by $n^{1/2}$. This means that without this standardization the sample mean would not be bounded in probability. This implies that the sample mean can be (randomly) large. This leads to the rather peculiar property that sample means will be large, random, and non-informative about population parameters. This means that interpreting simple statistics such as means is treacherous when the series may be a unit root process.



(a) Fitted Means

(b) Fitted Trend

Figure 16.1: Random Walk with Fitted Mean, Sub-Sample Means, and Trend

To illustrate, Figure 16.1(a) displays a simulated random walk with $n = 240$ observations. Also plotted is the sample mean $\overline{Y}_n = -2.98$, along with the sub-sample means $\overline{Y}_{1n} = -0.75$ and $\overline{Y}_{2n} = -5.21$. As predicted, the mean and sub-sample means are large, variable, and uninformative regarding the population mean.

Now consider a linear regression of Y_t on a linear time trend. The model for estimation is

$$Y_t = \beta_0 + \beta_1 t + e_t = X_t'\beta + e_t$$

where $X_t = (1,t)'$. Again for simplicity assume that $Y_0 = 0$. Take the least squares estimator $\widehat{\beta}$. Theorem

14.36 shows that

$$\frac{1}{n^2}\sum_{t=1}^{n} t \to \int_0^1 r dr = \frac{1}{2}$$
$$\frac{1}{n^3}\sum_{t=1}^{n} t^2 \to \int_0^1 r^2 dr = \frac{1}{3}.$$

Define $D_n = \begin{bmatrix} 1 & 0 \\ 0 & n \end{bmatrix}$. We calculate that

$$D_n^{-1}\frac{1}{n}\sum_{t=1}^{n} X_t X_t' D_n^{-1} = \begin{bmatrix} \frac{1}{n}\sum_{t=1}^{n} 1 & \frac{1}{n^2}\sum_{t=1}^{n} t \\ \frac{1}{n^2}\sum_{t=1}^{n} t & \frac{1}{n^3}\sum_{t=1}^{n} t^2 \end{bmatrix} \to \begin{bmatrix} 1 & \int_0^1 r dr \\ \int_0^1 r dr & \int_0^1 r^2 dr \end{bmatrix} = \int_0^1 X(r)X(r)' dr$$

where $X(r) = (1, r)$.

An application of the continuous mapping theorem with Theorem 16.1 yields

$$D_n^{-1}\frac{1}{n^{3/2}}\sum_{t=1}^{n} X_t Y_t = \int_0^1 X(r) Z_n(r) dr \underset{d}{\longrightarrow} \int_0^1 X(r) B(r) dr.$$

Together we obtain

$$\begin{aligned} D_n n^{-1/2}\widehat{\beta} &= D_n n^{-1/2}\left(\sum_{t=1}^{n} X_t X_t'\right)^{-1}\left(\sum_{t=1}^{n} X_t Y_t\right) \\ &= \left(D_n^{-1}\frac{1}{n}\sum_{t=1}^{n} X_t X_t' D_n^{-1}\right)^{-1}\left(D_n^{-1}\frac{1}{n^{3/2}}\sum_{t=1}^{n} X_t Y_t\right) \\ &\underset{d}{\longrightarrow} \left(\int_0^1 X(r)X(r)' dr\right)^{-1}\left(\int_0^1 X(r)B(r) dr\right). \end{aligned}$$

This shows that the estimator $\widehat{\beta}$ has an asymptotic distribution which is a transformation of the Brownian motion $B(r)$. For compactness we often write the final expression as $\left(\int_0^1 XX'\right)^{-1}\left(\int_0^1 XB\right)$.

To illustrate, Figure 16.1(b) displays the random walk from panel (a) along with a fitted trend line. The fitted trend appears large and substantial. However it is purely random, a feature only of this specific realization, is uninformative about the underlying parameters, and is dangerously misleading for prediction.

16.7 Demeaning and Detrending

A common preliminary step in time series analysis is demeaning (subtracting off a mean) and detrending (subtracting off a linear trend). With stationary processes this does not affect asymptotic inference. In contrast, an important property of unit root processes is that their behavior is altered by these transformations.

Take demeaning. The demeaned version of Y_t is $Y_t^* = Y_t - \overline{Y}_n$. An important observation is that Y_t^* is invariant to the initial condition Y_0, so without loss of generality we simply assume $Y_0 = 0$.

The normalized process is

$$Z_n^*(r) = \frac{1}{\sqrt{n}} Y_{\lfloor nr \rfloor} - \frac{1}{\sqrt{n}}\overline{Y}_n = Z_n(r) - Z_n(1) \underset{d}{\longrightarrow} B(r) - \int_0^1 B \overset{\text{def}}{=} B^*(r).$$

$B^*(r)$ is **demeaned Brownian motion**. It has the property that $\int_0^1 B^*(r)dr = 0$.

Take linear detrending. Based on least squares estimation of a linear trend the detrended series is $Y_t^{**} = Y_t - X_t'\widehat{\beta}$ where $X_t = (1,t)'$. Like the demeaned series the detrended series is invariant to Y_0. The associated normalized process is

$$\begin{aligned} Z_n^{**}(r) &= \frac{1}{\sqrt{n}} Y_{\lfloor nr \rfloor} - \frac{1}{\sqrt{n}} X'_{\lfloor nr \rfloor}\widehat{\beta} \\ &= Z_n(r) - X(\lfloor nr \rfloor / n)' D_n \frac{1}{\sqrt{n}}\widehat{\beta} \\ &\xrightarrow[d]{} B(r) - X(r)'\left(\int_0^1 XX'\right)^{-1}\left(\int_0^1 XB\right) \overset{\text{def}}{=} B^{**}(r). \end{aligned}$$

$B^{**}(r)$ is the continuous-time residual of the Brownian motion $B(r)$ projected orthogonal to $X(r) = (1,r)'$. We call $B^{**}(r)$ **detrended Brownian motion**.

There is another method of detrending through first differencing. Suppose that $Y_t = \beta_0 + \beta_1 t + Z_t$. The first difference is $\Delta Y_t = \beta_1 + \Delta Z_t$. An estimator of β_1 is the sample mean of ΔY_t:

$$\overline{\Delta Y}_n = \frac{1}{n}\sum_{t=1}^n \Delta Y_t = \frac{Y_n - Y_0}{n}.$$

The normalization $Z_0 = 0$ implies $Y_0 = \beta_0$ so an estimator of β_0 is Y_0. The detrended version of Y_t is $\widetilde{Y}_t = Y_t - Y_0 - (t/n)(Y_n - Y_0)$. The associated normalized process is

$$\widetilde{Z}_n(r) = Z_n(r) - \frac{\lfloor nr \rfloor}{n} Z_n(1) \xrightarrow[d]{} B(r) - rB(1) \overset{\text{def}}{=} V(r).$$

$V(r)$ is called a **Brownian Bridge** or a **tied-down Brownian motion**. It has the property that $V(0) = V(1) = 0$. It is also a detrended version of $B(r)$ but is distinct from the linearly detrended version $B^*(r)$.

We summarize the findings in the following theorem.

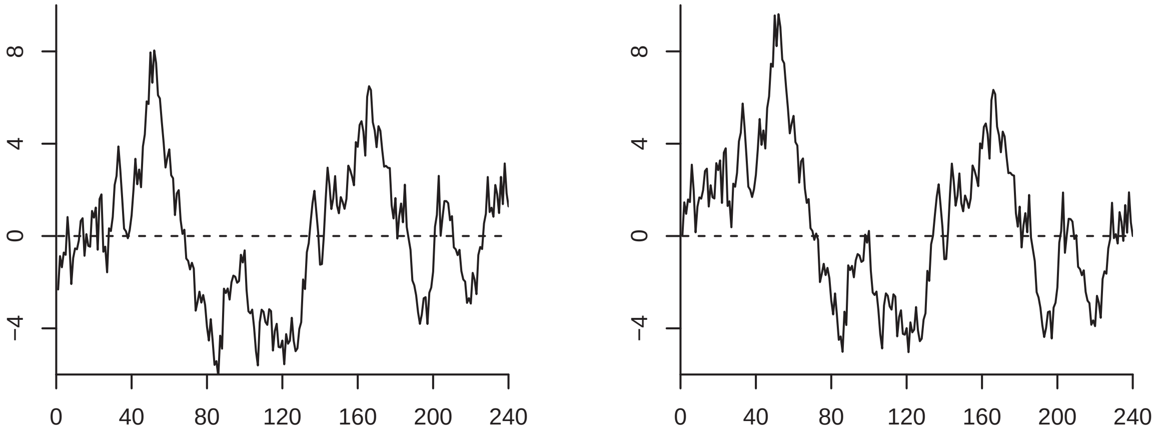
Theorem 16.5 Under the conditions of either Theorem 16.3 or Theorem 16.4, then as $n \to \infty$

1. $Z_n^*(r) \xrightarrow[d]{} B^*(r)$
2. $Z_n^{**}(r) \xrightarrow[d]{} B^{**}(r)$
3. $\widetilde{Z}_n(r) \xrightarrow[d]{} V(r)$.

To illustrate, Figure 16.2 displays two detrended versions of the series from Figure 16.1. Panel (a) shows the linear detrended series Y_t^*. Panel (b) shows the first-difference detrended series $\widetilde{Y}_t$. They are visually similar to one another and to Figure 16.1 except that the strong linear trend has been removed.

16.8 Stochastic Integrals

The distribution of the least squares estimator in the regression model $Y_t = X_t'\beta + e_t$ requires the distribution of the sample moments $n^{-1}\sum_{t=1}^{n-1} X_t e_{t+1}$. When X_t is non-stationary the limit distribution is non-standard and equals a **stochastic integral**.



(a) Linearly Detrended Series

(b) First Difference Detrended Series

Figure 16.2: Detrended Random Walk

It may help to recall the definition of the Riemann-Stieltijes integral. Over the region $[0,1]$ the integral of $g(x)$ with respect to $f(x)$ is

$$\int_0^1 g(x)df(x) = \lim_{N\to\infty}\sum_{i=0}^{N-1} g\left(\frac{i}{N}\right)\left(f\left(\frac{i+1}{N}\right)-f\left(\frac{i}{N}\right)\right).$$

A stochastic integral is the case where the function f is random and is defined as a probability limit.

Definition 16.4 The **stochastic integral** of vector-valued $X(r)$ with respect to vector-valued $Z(r)$ over $[0,1]$ is

$$\int_0^1 XdZ' = \int_0^1 X(r)dZ(r)' = \operatorname*{plim}_{N\to\infty}\sum_{i=0}^{N-1} X\left(\frac{i}{N}\right)\left(Z\left(\frac{i+1}{N}\right)-Z\left(\frac{i}{N}\right)\right)'.$$

Now consider the following setting. Let (X_t, e_t) be vector-valued sequences where e_t is a MDS with finite covariance and X_t is non-stationary. Assume that for some scaling sequence D_n the scaled process $X_n(r) = D_n^{-1}X_{\lfloor nr\rfloor}$ satisfies $X_n(r) \xrightarrow[d]{} X(r)$ for some deterministic or stochastic process $X(r)$. Examples of X_t sequences include the partial sum process constructed from e_t or another shock, a detrended version of a partial sum process, or a deterministic trend proceses. We desire the asymptotic distribution of $\sum_{t=1}^{n-1} X_t e'_{t+1}$. Define the partial sum process for e_t as $S_n(r) = n^{-1/2}\sum_{t=1}^{\lfloor nr\rfloor} e_t$. From Theorem 16.1, $S_n \xrightarrow[d]{} B$. We calculate that

$$\frac{1}{\sqrt{n}}D_n^{-1}\sum_{t=0}^{n-1} X_t e'_{t+1} = \sum_{t=0}^{n-1} X_n\left(\frac{t}{n}\right)\left(S_n\left(\frac{t+1}{n}\right)-S_n\left(\frac{t}{n}\right)\right)' = \int_0^1 X_n dS'_n.$$

The equalities hold because $S_n(r)$ and $X_n(r)$ are step functions with jumps at $r = t/n$. Since $X_n(r)$ and $S_n(r)$ converge to $X(r)$ and $B(r)$, by analogy we expect $\int_0^1 X_n dS_n$ to converge to $\int_0^1 X dB$. This is true, but rather tricky to show since the stochastic integral is not a continuous function of $B(r)$. A general statement of the conditions has been provided by Kurtz and Protter (1991, Theorem 2.2). The following is a simplification of their result.

Theorem 16.6 If $(e_t, \mathscr{F}_t)$ is a martingale difference sequence, $\mathbb{E}\left[e_t e_t'\right] = \Sigma < \infty$, $X_t \in \mathscr{F}_t$, and $(X_n(r), S_n(r)) \xrightarrow[d]{} (X(r), B(r))$ then

$$\int_0^1 X_n dS_n' = \frac{1}{\sqrt{n}} D_n^{-1} \sum_{t=1}^{n-1} X_t e_{t+1} \xrightarrow[d]{} \int_0^1 X dB'$$

where $B(r)$ is a Brownian motion with covariance matrix Σ.

The basic application of Theorem 16.6 is to the case $X_n(r) = S_n(r)$. Thus if $S_t = \sum_{i=1}^t e_t$ and e_t is a MDS with covariance matrix Σ then

$$\frac{1}{n} \sum_{t=1}^{n-1} S_t e_{t+1}' \xrightarrow[d]{} \int_0^1 B dB'.$$

We can extend this result to the case of serially correlated errors.

Theorem 16.7 If Z_t satisfies the conditions of Theorem 16.4 and $S_t = \sum_{i=1}^t Z_t$ then

$$\frac{1}{n} \sum_{t=1}^{n-1} S_t Z_{t+1}' \xrightarrow[d]{} \int_0^1 B dB' + \Lambda$$

where $B(r)$ is a Brownian motion with covariance matrix $\Omega = \Sigma + \Lambda + \Lambda'$, $\Sigma = \mathbb{E}\left[Z_t Z_t'\right]$, and $\Lambda = \sum_{j=1}^{\infty} \mathbb{E}\left[Z_{t-j} Z_t'\right]$.

The proof is presented in Section 16.22.

16.9 Estimation of an AR(1)

Consider least squares estimation of the AR(1) parameter α in the model $Y_t = \alpha Y_{t-1} + e_t$. The centered estimator is $\widehat{\alpha} - \alpha = \left(\sum_{t=1}^{n-1} Y_t^2\right)^{-1} \left(\sum_{t=1}^{n-1} Y_t e_{t+1}\right)$. We use the scaling

$$n(\widehat{\alpha} - \alpha) = \frac{\frac{1}{n} \sum_{t=1}^{n-1} Y_t e_{t+1}}{\frac{1}{n^2} \sum_{t=1}^{n-1} Y_t^2}.$$

We examine the denominator and numerator separately under the assumption $\alpha = 1$.

Similarly to our analysis of the sample mean the denominator can be written as an integral. Thus

$$\frac{1}{n^2}\sum_{t=1}^{n-1} Y_t^2 = \frac{1}{n}\sum_{t=1}^{n-1}\left(\frac{1}{n^{1/2}}Y_t\right)^2 = \int_0^1 Z_n(r)^2 dr \underset{d}{\longrightarrow} \int_0^1 B(r)^2 dr = \sigma^2 \int_0^1 W(r)^2 dr.$$

The convergence is by the continuous mapping theorem[4]. The final equality recognizes that if $B(r)$ has variance σ^2 then $B(r)^2 = \sigma^2 W(r)^2$ where $W(r)$ is standard Brownian motion. For conciseness we often write the final integral as $\int_0^1 W^2$.

For the numerator we appeal to Theorem 16.6.

$$\frac{1}{n}\sum_{t=1}^{n-1} Y_t e_{t+1} = \int_0^1 Z_n dS_n \underset{d}{\longrightarrow} \int_0^1 BdB = \sigma^2 \int_0^1 WdW.$$

This limiting stochastic integral is quite famous. It is known as Itô's integral.

Theorem 16.8 Itô's Integral $\int_0^1 WdW = \frac{1}{2}\left(W(1)^2 - 1\right)$.

If you are not surprised by Itô's integral take another look. The derivative of $\frac{1}{2}W(r)^2$ is $W(r)dW(r)$. Thus by standard calculus and $W(0) = 0$ you might expect $\int_0^1 WdW = \frac{1}{2}W(1)^2$. The presence of the extra term $-1/2$ is surprising. This arises because $W(r)$ has unbounded variation.

The random variable $W(1)^2$ is χ_1^2 which has expectation 1. Therefore the random variable $\int_0^1 WdW$ is mean zero but skewed.

The proof of Theorem 16.8 is presented in Section 16.22.

Returning to the least squares estimation problem we have shown that when $\alpha = 1$

$$n(\widehat{\alpha} - 1) \underset{d}{\longrightarrow} \frac{\frac{\sigma^2}{2}\left(W(1)^2 - 1\right)}{\sigma^2 \int_0^1 W^2} = \frac{\int_0^1 WdW}{\int_0^1 W^2}.$$

Theorem 16.9 Dickey-Fuller Coefficient Distribution If $Y_t = \alpha Y_{t-1} + e_t$ with $\alpha = 1$, and $(e_t, \mathscr{F}_t)$ is a strictly stationary and ergodic martingale difference sequence with a finite variance, then

$$n(\widehat{\alpha} - 1) \underset{d}{\longrightarrow} \frac{\int_0^1 WdW}{\int_0^1 W^2}.$$

The limit distribution in Theorem 16.9 is known as the **Dickey-Fuller Distribution** due to the work of Wayne Fuller and David Dickey. Theorem 16.9 shows that the least squares estimator is consistent for $\alpha = 1$ and converges at the "super-consistent" rate $O_p\left(n^{-1}\right)$. The limit distribution is non-standard and is written as a function of the Brownian motion $W(r)$. There is not a closed-form expression for the distribution or density of the statistic. Most commonly it is calculated by simulation.

[4] The function $g(f) = \int_0^1 f(x)^2 dx$ is continuous with respect to the uniform metric.

The density of the Dickey-Fuller coefficient distribution is displayed[5] in Figure 16.3(a) with the label "No Intercept". You can see that the density is high skewed with a long left tail. You can see that most of the probability mass of the distribution is over the negative region. This has the implication that the density has a negative mean and median. Hence the asymptotic distribution of the least squares estimator is biased negatively. This has the practical implication that when $\alpha = 1$ the least squares estimator is biased away from one.

We can also examine the limit distribution of the t-ratio. Let $\widehat{e}_t = Y_t - \widehat{\alpha} Y_{t-1}$ be the least squares residual, $\widehat{\sigma}^2 = n^{-1} \sum \widehat{e}_t^2$ the least squares variance estimator, and $s(\widehat{\alpha}) = \widehat{\sigma} / \sqrt{\sum Y_t^2}$ the classical standard error for $\widehat{\alpha}$. The t-ratio for α is $T = (\widehat{\alpha} - 1) / s(\widehat{\alpha})$.

Theorem 16.10 Dickey-Fuller T Distribution Under the assumptions of Theorem 16.9

$$T = \frac{\widehat{\alpha} - 1}{s(\widehat{\alpha})} \xrightarrow[d]{} \frac{\int_0^1 W dW}{\left(\int_0^1 W^2\right)^{1/2}}.$$

The limit distribution in Theorem 16.10 is known as the **Dickey-Fuller T distribution**. Theorem 16.10 shows that the classical t-ratio converges to a non-standard asymptotic distribution. There is no closed-form expression for the distribution or density so it is typically calculated using simulation techniques. The proof is presented in Section 16.22.

The density of the Dickey-Fuller T distribution is displayed in Figure 16.3(b) with the label "No Intercept". You can see that the density is skewed but much less so than the coefficient distribution. The distribution appears to be a "fatter" version of the conventional student t distribution. An implication is that conventional inference (confidence intervals and tests) will be inaccurate. We discuss testing in Section 16.13.

16.10 AR(1) Estimation with an Intercept

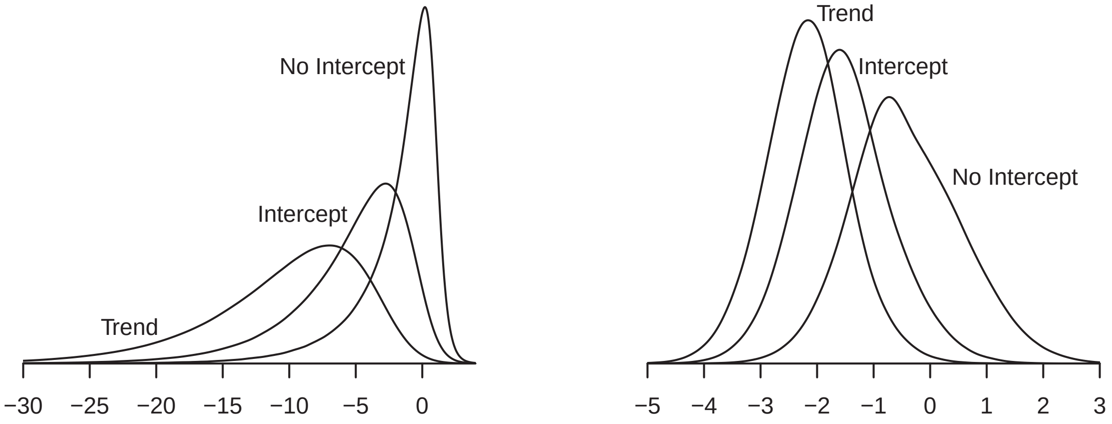
Suppose that Y_t is a random walk and we estimate an AR(1) model with an intercept. The estimated model is $Y_t = \mu + \alpha Y_{t-1} + e_t$. By the Frisch-Waugh-Lovell Theorem (Theorem 3.5) the least squares estimator $\widehat{\alpha}$ of α can be written as the simple regression using the demeaned series Y_t^*. That is, the normalized estimator is

$$n(\widehat{\alpha} - 1) = \frac{\frac{1}{n} \sum_{t=1}^{n-1} Y_t^* e_{t+1}}{\frac{1}{n^2} \sum_{t=1}^{n-1} Y_t^{*2}}$$

where $Y_t^* = Y_t - \overline{Y}$ with $\overline{Y} = \frac{1}{n} \sum_{t=1}^{n-1} Y_t$. By Theorems 16.5.1 and 16.6 the calculations from the previous section show that

$$n(\widehat{\alpha} - 1) \xrightarrow[d]{} \frac{\int_0^1 W^* dW}{\int_0^1 W^{*2}}.$$

[5] The densities in Figure 16.3 were estimated from one million simulation draws of the finite sample distribution for a sample size $n = 10{,}000$. The densities were estimated using nonparametric kernel methods (see Chapter 17 of *Introduction to Econometrics*).



(a) Dickey-Fuller Coefficient Density

(b) Dickey-Fuller T Density

Figure 16.3: Unit Root Distributions

This is similar to the distribution in Theorem 16.9. This is known as the Dickey-Fuller coefficient distribution for the case of an included intercept.

Similarly if we estimate an AR(1) model with an intercept and trend the estimated model is $Y_t = \mu + \beta t + \alpha Y_{t-1} + e_t$. By the Frisch-Waugh-Lovell Theorem this is equivalent to regression on the detrended series Y_t^{**}. Applying Theorems 16.5.2 and 16.6, we find

$$n\left(\widehat{\alpha}-1\right) \xrightarrow[d]{} \frac{\int_0^1 W^{**} dW}{\int_0^1 W^{**2}}.$$

This is known as the Dickey-Fuller coefficient distribution for the case of an included intercept and linear trend.

Similar results arise for the t-ratios. We summarize the results in the following theorem.

Theorem 16.11 Under the assumptions of Theorem 16.9, for the case of an estimated AR(1) with an intercept

$$n\left(\widehat{\alpha}-1\right) \xrightarrow[d]{} \frac{\int_0^1 W^* dW}{\int_0^1 W^{*2}}$$

$$T \xrightarrow[d]{} \frac{\int_0^1 W^* dW}{\left(\int_0^1 W^{*2}\right)^{1/2}}.$$

For the case of an estimated AR(1) with an intercept and linear time trend

$$n\left(\widehat{\alpha}-1\right) \xrightarrow[d]{} \frac{\int_0^1 W^{**} dW}{\int_0^1 W^{**2}}$$

$$T \xrightarrow[d]{} \frac{\int_0^1 W^{**} dW}{\left(\int_0^1 W^{**2}\right)^{1/2}}.$$

The densities of the Dickey-Fuller coefficient distributions are displayed in Figure 16.3(a), labeled as "Intercept" for the case with an included intercept, and "Trend" for the case with an intercept and linear time trend. The densities are considerably affected by the inclusion of the intercept or intercept and trend. The effect is twofold: (1) the distributions shift substantially to the left; and (2) the distributions substantially widen. Examining the "trend" version we can see that there is very little probability mass above zero. This means that the asymptotic distribution is not only biased downward, the realization is nearly always negative. This has the practical implication that the least squares estimator is almost certainly less than the true coefficient value. This is a strong form of bias.

The densities of the Dickey-Fuller T distributions are displayed in Figure 16.3(b). The effect of detrending on the T distributions is quite different from the effect on the coefficient distirbutions. Here we see that the primary effect is a location shift with only a mild impact on dispersion. The strong location shift is a bias in the asymptotic T distribution, implying that conventional inferences will be incorrect.

16.11 Sample Covariances of Integrated and Stationary Processes

Let (X_t, u_t) be a sequence where X_t is non-stationary and u_t is mean zero and strictly stationary. Assume that for some scaling sequence D_n the scaled process $X_n(r) = D_n^{-1} X_{\lfloor nr \rfloor}$ satisfies $X_n(r) \xrightarrow[d]{} X(r)$ where $X(r)$ is continuous with probability one. Consider the scaled sample covariance

$$C_n = \frac{1}{n} D_n^{-1} \sum_{t=1}^{n} X_t u_t.$$

Theorem 16.12 Assume that $X_n(r) = D_n^{-1} X_{\lfloor nr \rfloor} \xrightarrow[d]{} X(r)$ where $X(r)$ is almost surely continuous. Assume u_t is mean zero, strictly stationary, ergodic, and $\mathbb{E}\left|u_t\right| < \infty$. Then $C_n \xrightarrow[p]{} 0$ as $n \to \infty$.

The proof is presented in Section 16.22.

16.12 AR(p) Models with a Unit Root

Assume that Y_t satisfies $a(\mathrm{L})\Delta Y_t = e_t$ where $a(z)$ is a $p-1$ order invertible lag polynomial and e_t is a stationary MDS with finite variance σ^2. Then Y_t can be written as the AR(p) process

$$Y_t = a_1 Y_{t-1} + \cdots + a_p Y_{t-p} + e_t \tag{16.6}$$

where the coefficients satisfy $a_1 + \cdots + a_p = 1$. Let $\widehat{a}$ be the least squares estimator of $a = (a_1, ..., a_p)$. We now describe its sampling distribution.

Let B be the $p \times p$ matrix which transforms $(Y_{t-1}, ..., Y_{t-p})$ to $(Y_{t-1}, \Delta Y_{t-1}, ..., \Delta Y_{t-p+1})$, for example when $p = 3$ then $B = \begin{bmatrix} 1 & 0 & 0 \\ 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix}$. Make the partition $B^{-1\prime} a = (\rho, \beta)$ where $\rho \in \mathbb{R}$ and $\beta \in \mathbb{R}^{p-1}$. Then the AR(p) model can be written as

$$Y_t = \rho Y_{t-1} + \beta' X_{t-1} + e_t \tag{16.7}$$

where $X_{t-1} = (\Delta Y_{t-1}, ..., \Delta Y_{t-p+1})$. The leading coefficient is $\rho = a_1 + \cdots + a_p = 1$. This transformation separates the regressors into the unit root component Y_{t-1} and the stationary component X_{t-1}.

Consider the least squares estimators $(\widehat{\rho}, \widehat{\beta})$. They can be written under the assumption of a unit root as

$$\begin{pmatrix} n\left(\widehat{\rho}-1\right) \\ \sqrt{n}\left(\widehat{\beta}-\beta\right) \end{pmatrix} = \begin{pmatrix} \frac{1}{n^2}\sum_{t=1+p}^{n} Y_{t-1}^2 & \frac{1}{n^{3/2}}\sum_{t=1+p}^{n} Y_{t-1}X_{t-1}' \\ \frac{1}{n^{3/2}}\sum_{t=1+p}^{n} X_{t-1}Y_{t-1} & \frac{1}{n}\sum_{t=1+p}^{n} X_{t-1}X_{t-1}' \end{pmatrix}^{-1} \begin{pmatrix} \frac{1}{n}\sum_{t=1+p}^{n} Y_{t-1}e_t \\ \frac{1}{\sqrt{n}}\sum_{t=1+p}^{n} X_{t-1}e_t \end{pmatrix}.$$

Theorems 16.4 and the CMT show that

$$\frac{1}{n^2}\sum_{t=1+p}^{n} Y_{t-1}^2 \xrightarrow[d]{} \omega^2 \int_0^1 W^2$$

where ω^2 is the long-run variance of ΔY_t which equals $\omega^2 = \sigma^2 / a(1)^2 > 0$.

Theorem 16.12 shows that

$$\frac{1}{n^{3/2}}\sum_{t=1+p}^{n} X_{t-1}Y_{t-1} \xrightarrow[p]{} 0.$$

Theorems 16.4 and 16.6 show that

$$\frac{1}{n}\sum_{t=1+p}^{n} Y_{t-1}e_t \xrightarrow[d]{} \omega\sigma \int_0^1 W dW.$$

The WLLN and the CLT for stationary processes show that

$$\frac{1}{n}\sum_{t=1+p}^{n} X_{t-1}X_{t-1}' \xrightarrow[p]{} \boldsymbol{Q}$$

$$\frac{1}{\sqrt{n}}\sum_{t=1+p}^{n} X_{t-1}e_t \xrightarrow[d]{} \mathrm{N}(0, \Omega)$$

where $\boldsymbol{Q} = \mathbb{E}\left[X_{t-1}X_{t-1}'\right]$ and $\Omega = \mathbb{E}\left[X_{t-1}X_{t-1}'e_t^2\right]$. Together we have established the following.

Theorem 16.13 Assume that Y_t satisfies $a(\mathrm{L})\Delta Y_t = e_t$ where $a(z)$ is a $p-1$ order invertible lag polynomial and $(e_t, \Im_t)$ is a stationary MDS with finite variance σ^2. Then

$$\begin{pmatrix} n(\widehat{\rho}-1) \\ \sqrt{n}(\widehat{\beta}-\beta) \end{pmatrix} \underset{d}{\longrightarrow} \begin{pmatrix} a(1)\dfrac{\int_0^1 WdW}{\int_0^1 W^2} \\ \\ \mathrm{N}(0,V) \end{pmatrix} \tag{16.8}$$

where $V = \boldsymbol{Q}^{-1}\Omega\boldsymbol{Q}^{-1}$.

This theorem provides an asymptotic distribution theory for the least squares estimators. The estimator $(\widehat{a},\widehat{\beta})$ is consistent, the coefficient $\widehat{\beta}$ on the stationary variables is asymptotically normal, and the coefficient $\widehat{a}$ on the unit root component has a scaled Dickey-Fuller distribution.

The estimator of the representation (16.6) is the linear transformation $B'(\widehat{\rho},\widehat{\beta}')'$, and therefore its asymptotic distribution is the transformation B' of (16.8). Since the unit root component converges at a faster $O_p(n^{-1})$ rate than the stationary component it drops out of the asymptotic distribution. We obtain

$$\sqrt{n}\,(\widehat{a}-a) \underset{d}{\longrightarrow} \mathrm{N}(0,\boldsymbol{G}V\boldsymbol{G}') \tag{16.9}$$

where, in the $p=3$ case

$$\boldsymbol{G} = \begin{bmatrix} 1 & 0 \\ -1 & 1 \\ 0 & -1 \end{bmatrix}.$$

The asymptotic covariance matrix $\boldsymbol{G}V\boldsymbol{G}'$ is deficient with rank $p-1$. Hence this is only a partial characterization of the asymptotic distribution; equation (16.8) is a complete first-order characterization. The implication of (16.9) is that individual coefficient estimators and standard errors of (16.6) have conventional asymptotic interpretations. This extends to conventional hypothesis tests which do not include the sum of the coefficients. For most purposes (except testing the unit root hypothesis) this means that asymptotic inference on the coefficients of (16.6) can be based on the conventional normal approximation and can ignore the possible presence of unit roots.

16.13 Testing for a Unit Root

The asymptotic properties of the time series process change discontinuously at the unit root $\rho = a_1+\cdots+a_p = 1$. It is therefore of standard interest to test the hypothesis of a unit root. We typically express this as the test of $\mathbb{H}_0 : \rho = 1$ against $\mathbb{H}_1 : \rho < 1$. We typically view the test as one-sided as we are interested in the alternative hypothesis that the series is stationary (not that it is explosive).

The test for $\mathbb{H}_0$ vs. $\mathbb{H}_1$ is the t-statistic for $a_1+\cdots+a_p = 1$ in the AR(p) model (16.6). This is identical to the t-statistic for $\rho = 1$ in reparameterized form (16.7). Since the latter is a simple t-ratio this is the most convenient implementation. It is typically called the **Augmented Dickey-Fuller** statistic. It equals

$$\mathrm{ADF} = \frac{\widehat{\rho}-1}{s\left(\widehat{\rho}\right)}$$

where $s\left(\widehat{\rho}\right)$ is a standard error for $\widehat{\rho}$. This t-ratio is typically calculated using a classical (homoskedastic) standard error, perhaps for historical reasons, and perhaps because the asymptotic distribution of ADF is invariant to conditional heteroskedasticity. The statistic is called the ADF statistic when the estimated model is an AR(p) model with $p > 1$; it is typically called the Dickey-Fuller statistic if the estimated model is an AR(1).

The asymptotic distribution of ADF depends on the fitted deterministic components. The test statistic is most typically calculated in a model with a fitted intercept or a fitted intercept and time trend, though the theory is also presented for the case with no fitted intercept, and extends to any polynomial order trend.

Theorem 16.14 Assume that Y_t satisfies $a(\mathrm{L})\Delta Y_t = e_t$ where $a(z)$ is a $p-1$ order invertible lag polynomial and $(e_t, \Im_t)$ is a stationary MDS with finite variance σ^2. Then

$$\mathrm{ADF} \underset{d}{\longrightarrow} \frac{\int_0^1 U dW}{\left(\int_0^1 U^2\right)^{1/2}} \overset{\text{def}}{=} \xi$$

where W is Brownian motion. The process U depends on the fitted deterministic components:

1. Case 1: No intercept or trend. $U(r) = W(r)$.
2. Case 2: Fitted intercept (demeaned data). $U(r) = W(r) - r\int_0^1 W$.
3. Case 3: Fitted intercept and trend (detrended data). $U(r) = W(r) - X(r)'\left(\int_0^1 XX'\right)^{-1}\left(\int_0^1 XW\right)$ where $X(r) = (1, r)'$.

Let Z_α satisfy $\mathbb{P}[\xi \le Z_\alpha] = \alpha$. The test "Reject $\mathbb{H}_0$ if ADF $< Z_\alpha$" has asymptotic size α.

Asymptotic critical values are displayed in the first three columns of Table 16.1. The ADF is a one-sided hypothesis test so rejections occur when the test statistic is less than (more negative than) the critical value. For example, the 5% critical value for the case of a fitted intercept is −2.86. This means that if the ADF t-ratio is more negative than −2.86 (for example ADF = −3.0) then the test rejects the hypothesis of no unit root. But if the ADF t-ratio is greater than −2.86 (for example ADF = −2.0) then the the test does not reject the hypothesis of a unit root.

In most applications an ADF test is implemented with at least a fitted intercept (the second column in the table). Many are implemented with a fitted linear time trend (which is the third column). The choice depends on the nature of the alternative hypothesis. If $\mathbb{H}_1$ is that the series is stationary about a constant mean then the case of a fitted intercept is appropriate. Example series for this context are unemployment and interest rates. If $\mathbb{H}_1$ is that the series is stationary about a linear trend then the case of a fitted trend is appropriate. Examples for this context are levels or log-levels of macroeconomic aggregates.

The ADF test depends on the autoregressive order p. The issue of selection of p is similar to that of autoregressive model selection. In general, if p is too small than the model is misspecified and the ADF statistic has an asymptotic bias. If p is too large than the test coefficient $\widehat{\rho}$ is imprecisely estimated, reducing the power of the test. Since $\widehat{\rho}$ is the sum of the p estimated AR coefficients in the levels model the imprecision can be sensitive to the choice of p. A reasonable selection rule is to use the AIC-selected AR model. Improved rules have been studied by Ng and Perron (2001).

We have argued that it is better to report asymptotic p-values rather than "accept/reject". For this calculation we need the asymptotic distribution function but this is not available in closed form. A simple approximation is interpolation of the critical values. For example, suppose ADF = −3.0 with a fitted intercept. The two closest critical values are the 10% (−3.13) and 15% (−2.94). Linear interpolation between these values yields

$$p = \frac{0.10 \times (3.0 - 2.94) + 0.15 \times (3.13 - 3.0)}{3.13 - 2.94} = 0.13.$$

Thus the asymptotic p-value is approximately 13%. Reporting a p-value instead of the "decision" of a test improves interpretation and communication.

How should unit root tests be used in empirical practice? The answer is subtle. A common mistake is "We use a unit root test to discover whether or not the series has a unit root." This is a mistake because a test does not reveal the truth. Rather, it presents evidence whether or not $\mathbb{H}_0$ can be rejected. If the test fails to reject $\mathbb{H}_0$ this does not mean that "We have found a unit root". Rather, the correct conclusion is "We cannot reject the hypothesis that it has a unit root". Thus we do not know. If the test rejects $\mathbb{H}_0$ (if the p-value is very small) then we can conclude that the series is unlikely to be a unit root process; its behavior is more consistent with a stationary process. Another common mistake is to adopt the rule: "If the ADF test rejects then we work with Y_t in levels; if the ADF test does not reject then we work with the differenced series ΔY_t." This is a mistake because it assigns a modeling rule to the result of a statistical test while the test is only designed to answer the question if there is evidence against the hypothesis of a unit root. The choice of Y_t versus ΔY_t is a model selection choice not a hypothesis testing decision.

I believe a reasonable approach is to start with a hypothesis based on theory and context. Does economic theory lead you to treat a series as stationary or non-stationary? Is there a reason to believe that a series should be stationary – thus stable in the mean – or is there reason to believe the series will exhibit growth and change? If you have a clear answer to these questions, that should be your starting place, your default. Use the unit root test to help confirm your assumptions rather than to select a modeling approach. If your assumption is that Y_t has a unit root but the unit root test strongly rejects, then you should re-appraise your theory. On the other hand if your assumption is that Y_t is stationary but the unit root test fails to reject the null of a unit root, do not necessarily depart from your theoretical base. Consider the degree of evidence, the sample size, as well as the point estimates. Use all information together to base your decision.

To illustrate application of the ADF test let's take the eight series displayed in Figures 14.1-14.2 using the variables measured in levels or log-levels. The variables and transformations are listed in Table 16.2. For six of the eight series (all but the interest and unemployment rates) we took the log transformation. We included an intercept and linear time trend in each regression and selected the autoregressive order by minimizing the AIC across AR(p) models with a linear time trend. For the quarterly series we examined AR(p) models up to $p = 8$, for the monthly series up to $p = 12$. The selected values of p are shown in the table. The point estimate $\widehat{\rho} - 1$, its standard error, the ADF t statistic, and its asymptotic p-value are shown. What we see is for for seven of the eight series (all but the unemployment rate) the p-values are far from the critical region indicating failure to reject the null hypothesis of a unit root. The p-value for the unemployment rate is 0.01, however, indicating rejection of a unit root. Overall, the results are consistent with the hypotheses that the unemployment rate is stationary and that the other seven variables are possibly (but not decisively) unit root processes.

The ADF test came into popularity in economics with a seminar paper by Nelson and Plosser (1982). These authors applied the ADF to a set of standard macroeconomic variables (similar to those in Table 16.2) and found that the unit root hypothesis could not be rejected in most series. This empirical finding had a substantial effect on applied economic time series. Before this paper the conventional wisdom was that economic series were stationary (possibly about linear time trends). After their work it became more accepted to assume that economic time series are better described as autoregressive unit root processes. Nelson and Plosser (1982) used this empirical finding to make a further and stronger claim. They argued that Keynesian macroeconomic models (which were standard at the time) imply that economic

Table 16.1: Unit Root Testing Critical Values

	ADF			KPSS	
	No Intercept	Intercept	Trend	Intercept	Trend
0.01%	−3.92	−4.69	−5.21	1.598	0.430
0.1%	−3.28	−4.08	−4.58	1.176	0.324
1%	−2.56	−3.43	−3.95	0.744	0.218
2%	−2.31	−3.20	−3.73	0.621	0.187
3%	−2.15	−3.06	−3.60	0.550	0.169
4%	−2.03	−2.95	−3.50	0.500	0.157
5%	−1.94	−2.86	−3.41	0.462	0.148
7%	−1.79	−2.72	−3.28	0.406	0.134
10%	−1.62	−2.57	−3.13	0.348	0.119
15%	−1.40	−2.37	−2.94	0.284	0.103
20%	−1.23	−2.22	−2.79	0.241	0.091
30%	−0.96	−1.97	−2.56	0.185	0.076
50%	−0.50	−1.57	−2.18	0.119	0.056
70%	0.05	−1.15	−1.81	0.079	0.041
90%	0.89	−0.44	−1.24	0.046	0.028
99%	2.02	0.60	−0.32	0.025	0.017

Source: Calculated by simulation from one million replications of samples of size $n = 10,000$.

Table 16.2: Unit Root and KPSS Test Applications

	p	$\widehat{\rho}-1$	ADF	p-value	M	KPSS$_2$	p-value
log(real GDP)	3	−0.017 (.009)	−1.8	0.71	18	0.23	0.01
log(real consumption)	4	−0.029 (.012)	−2.4	0.37	18	0.113	0.12
log(exchange rate)	11	−0.009 (.004)	−2.2	0.49	26	0.31	< .01
interest rate	12	−0.005 (.004)	−1.5	0.52	26	0.56	< .01
log(oil price)	2	−0.013 (.005)	−2.4	0.35	26	0.23	< .01
unemployment rate	7	−0.014 (.004)	−3.4	0.01	26	0.14	0.06
log(CPI)	11	−0.001 (.001)	−1.0	0.95	26	0.55	< .01
log(stock price)	6	−0.010 (.004)	−2.2	0.47	26	0.30	< .01

time series are stationary while real business cycle (RBC) models (which were new at the time) imply that economic time series are unit root processes. Nelson-Plosser argued that the empirical finding that the unit root tests do not reject was strong support for the RBC research program. Their argument was influential and was a factor motivating the rise of the RBC literature. With hindsight we can see that Nelson and Plosser (1982) made a fundamental error in this latter argument. The unit root behavior in RBC models is not inherent to their structure; rather it is a by-product of the assumptions on the technology process. (If exogenous technology is a unit root process or a stationary process then macroeconomic variables will also be unit root processes or stationary processes, respectively.) Similarly the stationary behavior of 1970s Keynesian models was not inherent to their structure but rather a by-product of assumptions about unobservables. Fundamentally, the unit root/stationary distinction says little about the RBC/Keynesian debate.

The ADF test with a fitted intercept can be implemented in Stata by the command `dfuller y, lags(q) regress`. For a fitted intercept and trend add the option `trend`. The number of lags "q" in the command is the number of first differences in (16.7), hence $q = p - 1$ where p is the autoregressive order. The `dfuller` command reports the estimated regression, the ADF statistic, asymptotic critical values, and approximate asymptotic p-value.

16.14 KPSS Stationarity Test

Kwiatkowski, Phillips, Schmidt, and Shin (1992) developed a test of the null hypothesis of stationarity against the alternative of a unit root which has become known as the KPSS test. Many users find this idea attractive as a counterpoint to the ADF test.

The test is derived from what is known as a **local level model**. This is

$$\begin{aligned} Y_t &= \mu + \theta S_t + e_t \\ S_t &= S_{t-1} + u_t \end{aligned}$$

where e_t is a mean zero stationary process and u_t is i.i.d. $(0, \sigma_u^2)$. When $\sigma_u^2 = 0$ then Y_t is stationary. When $\sigma_u^2 > 0$ then Y_t is a unit root process. Thus a test of the null of stationarity against the alternative of a unit root is a test of $\mathbb{H}_0 : \sigma_u^2 = 0$ against $\mathbb{H}_1 : \sigma_u^2 > 0$. Add the auxillary assumption that (e_t, u_t) are i.i.d normal. The Lagrange multiplier test can be shown to reject $\mathbb{H}_0$ in favor of $\mathbb{H}_1$ for large values of

$$\frac{1}{n^2 \widehat{\sigma}^2} \sum_{i=1}^{n} \left(\sum_{t=1}^{i} \widehat{e}_t \right)^2$$

where $\widehat{e}_t = Y_t - \overline{Y}$ are the residuals under the null and $\widehat{\sigma}^2$ is its sample variance. To generalize to the context of serially correlated e_t KPSS proposed the statistic

$$\mathrm{KPSS}_1 = \frac{1}{n^2 \widehat{\omega}^2} \sum_{i=1}^{n} \left(\sum_{t=1}^{i} \widehat{e}_t \right)^2$$

where

$$\widehat{\omega}^2 = \sum_{\ell=-M}^{M} \left(1 - \frac{|\ell|}{M+1} \right) \frac{1}{n} \sum_{t=1}^{n} \widehat{e}_t \widehat{e}_{t-\ell}$$

is the Newey-West estimator of the long-run variance ω^2 of Y_t.

For contexts allowing for a linear time trend the local level model takes the form

$$Y_t = \mu + \beta t + \theta S_t + e_t$$

which has null least squares estimator

$$Y_t = \widetilde{\mu} + \widetilde{\beta} t + \widetilde{e}_t.$$

Notice that $\widetilde{e}_t$ is linearly detrended Y_t. The KPSS test for $\mathbb{H}_0$ against $\mathbb{H}_1$ rejects for large values of

$$\text{KPSS}_2 = \frac{1}{n^2 \widetilde{\omega}^2} \sum_{i=1}^{n} \left(\sum_{t=1}^{i} \widetilde{e}_t \right)^2$$

where $\widetilde{\omega}^2$ is defined as $\widehat{\omega}^2$ but with the detrended residuals $\widetilde{e}_t$.

Theorem 16.15 If Y_t follows Assumption 16.1 then

$$\text{KPSS}_1 \xrightarrow[d]{} \int_0^1 V^2$$

and

$$\text{KPSS}_2 \xrightarrow[d]{} \int_0^1 V_2^2$$

where $V(r) = W(r) - rW(1)$ is a Brownian bridge, and $V_2(r) = W(r) - \left(\int_0^r X(s)ds\right)' \left(\int_0^1 XX'\right)' \int_0^1 X dW$ with $X(s) = (1, s)'$.

The asymptotic distributions in Theorem 16.15 are non-standard and are typically calculated by simulation. The process $V_2(r)$ is known as a **Second-level Brownian Bridge**. The asymptotic distributions are displayed[6] in Figure 16.4. The densities are skewed with a slowly-decaying right tail. The KPSS_2 distribution is substantially shifted towards the origin compared to the KPSS_1 distribution, indicating a substantial effect of detrending.

Asymptotic critical values are displayed in the final two columns of Table 16.1. Rejections occur when the test statistic exceeds the critical value. For example, for a regression with fitted intercept and time trend, suppose that the statistic equals $\text{KPSS}_2 = 0.163$. This exceeds the 4% critical value 0.157 but not the 3% critical value 0.169. Thus the test rejects at the 4% but not the 3% level. An interpolated p-value is 3.5%. This would be moderate evidence against the hypothesis of stationarity in favor of the alternative hypothesis of nonstationarity.

The KPSS statistic depends on the lag order M used to estimate the long-run variance ω^2. This is a challenge for test implementation. If Y_t is stationary but highly persistent (for example, an AR(1) with a large autoregressive coefficient) then the lag truncation M needs to be large in order to accurately estimate ω^2. However, under the alternative that Y_t is a unit root process, the estimator $\widehat{\omega}^2$ will increase roughly linearly with M so that for any given sample the KPSS statistic can be made arbitrarily small by selecting M sufficiently large.

Recall that the Andrews (1991) reference rule (14.51) is

$$M = \left(6 \frac{\rho^2}{\left(1-\rho^2\right)^2} \right)^{1/3} n^{1/3}$$

where ρ is the first autocorrelation of Y_t. For the KPSS test we should not replace ρ with an estimator $\widehat{\rho}$ as the latter converges to 1 under $\mathbb{H}_0$, leading to $M \to \infty$ rendering the test inconsistent. Instead we can

[6] Calculated by simulation from one million simulation draws of samples of size $n = 10{,}000$.

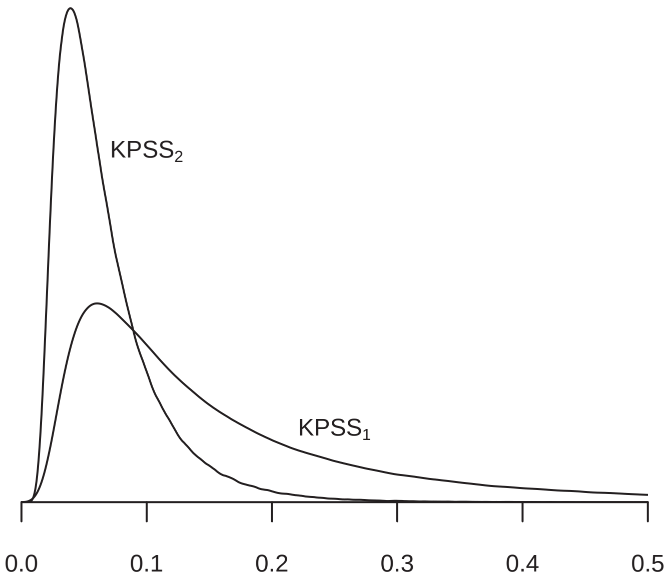


Figure 16.4: Density of KPSS Distribution

use a default rule based on a reasonable alternative. Suppose we consider the alternative $\rho = 0.8$. The associated Andrews' reference rule is $M = 3.1n^{1/3}$. This leads to a simple rule $M = 3n^{1/3}$. An interpretation of this choice is that it should approximately control the size of the test when the truth is an AR(1) with coefficient 0.8 but over-reject for more persistent AR processes.

To illustrate, Table 16.2 reports the KPSS$_2$ statistic for the same eight series as examined in the previous section, using $M = 3n^{1/3}$. For the first two quarterly series $n = 228$ leading to $M = 18$. For the six monthly series $n = 684$ leading to $M = 26$. For six of the eight series (all but consumption and the unemployment rate) the KPSS statistic equals or exceeds the 1% critical value leading to a rejection of the null hypothesis of stationarity in favor of the alternative of a unit root. This is consistent with the ADF test which failed to reject a unit root for these series.

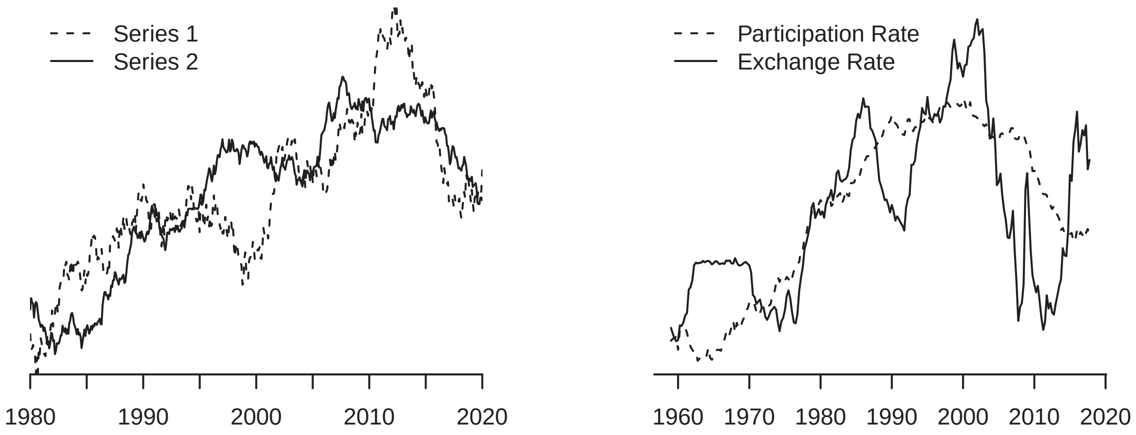
For the consumption series the KPSS statistic has a p-value of 12%, which does not reject the hypothesis of stationarity. Recall that the ADF test failed to reject the hypothesis of a unit root. Thus neither test leads to a decisive result; as a pair the two tests are inconclusive. In this context I recommend staying with the prediction of economic theory (consumption is a martingale) as it is not rejected by a hypothesis test. The KPSS fails to reject stationarity but that does not mean that the series is stationary.

An interesting case is the unemployment rate series. It has KPSS$_2$ = 0.14 with a p-value of 6%. This is borderline significant for rejection of stationarity. On the other hand, recall that the ADF test had a p-value of 1% rejecting the unit root hypothesis. These results are borderline conflicting. To augment our information we calculate the KPSS$_1$ test as the unemployment rate does not appear to be trended. We find KPSS$_1$ = 0.19 with a p-value of 30%. This is clearly in the non-rejection region, failing to provide evidence against stationarity. As a whole, the ADF test (reject unit root), the KPSS$_1$ test (accept stationarity), and the KPSS$_2$ test (borderline reject stationarity), taken together are consistent with the interpretation that the unemployment rate is a stationary process.

The KPSS$_2$ test can be implemented in Stata using the command[7] `kpss y, maxlag(q)`. For the KPSS$_1$ test add the option `notrend`. The command reports the KPSS statistics for $M = 1, ..., q$, as well as asymptotic critical values. Approximate asymptotic p-values are not reported.

16.15 Spurious Regression

One of the most empirically relevant discoveries from the theory of non-stationary time series is the phenomenon of spurious regression. This is the finding that two statistically independent series, if both unit root processes, are likely to fool traditional statistical analysis by appearing to be statistically related by both eyeball scrutiny and traditional statistical tests. The phenomenon was observed[8] and named by Granger and Newbold (1974) and explained using the theory of non-stationary time series by Phillips (1986). The primary lesson is that it is easy to be tricked by non-stationary time-series but the problem disappears if we pay suitable attention to dynamic specification.



(a) Two Unrelated Random Walks

(b) Exchange and Labor Force Participation Rates

Figure 16.5: Plots of Empirical Series

To illustrate the problem examine Figure 16.5(a). Displayed are two time series, monthly for 1980-2018. A casual review of the graphs shows that both series are generally increasing over 1980-2010 with a no-growth period around 2000, and the series display a downward trend for the final decade. A more refined perusal may appear to reveal that Series 2 leads Series 1 by about five years, in the sense that Series 2 reaches turning points about five years before Series 1. A casual observer is likely to deduce based on Figure 16.5(a) that the two time series are strongly related.

[7] The command `kpss` is not part of the standard package, but can be installed by typing `ssc install kpss`.

[8] In numerical simulations.

However the truth is that Series 1 and Series 2 are statistically independent random walks generated by computer simulation, each standardized to have mean zero and unit variance for the purpose of visual comparison. The "fact" that both series are generally upward trended and have "similar" turning points are statistical accidents. Random walks have an uncanny ability to fool casual analysis. Newspaper (and other journalistic) articles containing plots of time series are routinely subject to the tricks of Figure 16.5(a). Economists are also routinely tricked and fooled.

Traditional statistical examination of the series in Figure 16.5(a) can also lead to a false inference of a strong relationship. A linear regression of Series 1 on Series 2 yields a slope coefficient of 0.76 with classical standard error of 0.03. The t-ratio for the test of a zero slope is $T = 26$. The equation R^2 is 0.59. These traditional statistics support the incorrect inference that the two series are strongly related.

Spurious relationships of this form are commonplace in economic time series. An example is shown in Figure 16.5(b), which displays the U.S. labor force participation rate and U.S.-Canada exchange rate, quarterly for 1960-2018. As a visual aid both series have been normalized to have mean zero and unit variance. Both series appear to grow at a similar rate from 1960-2000, though the exchange rate is more volatile. From 2000-2018 they reverse course, with both series declining. The visual evidence is supported by traditional statistics. A linear regression of labor participation on the exchange rate yields a slope coefficient of 0.70 with a clasical standard error of 0.05. The t-ratio for the test of a zero slope is $T = 15$. The equation R^2 is 0.49. The visual and statistical evidence support the inference that the two series are related.

This empirical "finding" that the labor participation and exchange rates are related does not make economic sense. Is this an example of a spurious regression between non-stationary variables? A visual inspection of each series supports the contention that each is non-stationary and may be well characterized as a unit root process. We saw in Sections 16.13 and 16.14 that the ADF and KPSS tests support the hypothesis that the exchange rate is a unit root process. Similar tests reach the same conclusion for labor force participation. Thus the two series are reasonably characterized as unit root processes and these two series could be an empirical example of a spurious regression.

For a formal framework assume that the series Y_t and X_t are random walk processes

$$Y_t = Y_{t-1} + e_{1t} \tag{16.10}$$

$$X_t = X_{t-1} + e_{2t} \tag{16.11}$$

where (e_{1t}, e_{2t}) are i.i.d., mean zero, mutually uncorrelated, and normalized to have unit variance. Let Y_t^* and X_t^* denote demeaned versions of Y_t and X_t. From the FCLT they satisfy

$$\left(\frac{1}{\sqrt{n}} Y^*_{\lfloor nr \rfloor}, \frac{1}{\sqrt{n}} X^*_{\lfloor nr \rfloor}\right) \xrightarrow[d]{} \left(W_1^*(r), W_2^*(r)\right)$$

where $W_1^*(r)$ and $W_2^*(r)$ are demeaned Brownian motions.

Applying the CMT the sample correlation has the asymptotic distribution

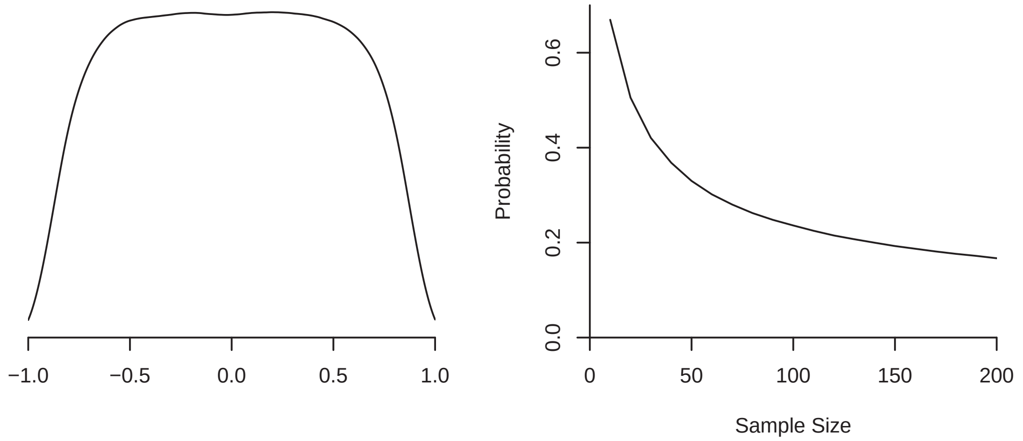
$$\widehat{\rho} = \frac{\frac{1}{n^2}\sum_{i=1}^n Y_i^* X_i^*}{\left(\frac{1}{n^2}\sum_{i=1}^n Y_i^{*2}\right)^{1/2}\left(\frac{1}{n^2}\sum_{i=1}^n X_i^{*2}\right)^{1/2}} \xrightarrow[d]{} \frac{\int_0^1 W_1^* W_2^*}{\left(\int_0^1 W_1^{*2}\right)^{1/2}\left(\int_0^1 W_2^{*2}\right)^{1/2}}.$$

The right-hand-side is a random variable. Furthermore it is also non-degenerate (indeed, it is non-zero with probability one). Thus the sample correlation $\widehat{\rho}$ remains random in large samples.

To understand magnitudes, Figure 16.6(a) displays the asymptotic distributon[9] of $\widehat{\rho}$. The density has most probability mass in the interval $[-0.5, 0.5]$, over which the density is essentially flat. This means that

[9] Calculated by simulation from one million simulation draws of samples of size $n = 10,000$.

the sample correlation has a diffuse distribution. Above we saw that the two simulated random walks had a sample correlation[10] of 0.76 and the two empirical series a sample correlation of 0.70. We can now see that these results are consistent with the distribution shown in Figure 16.6(a) and are therefore uninformative regarding the underlying relationships.



(a) Asymptotic Density of Sample Correlation (b) Coverage Probability of Nominal 95% Interval

Figure 16.6: Properties of Spurious Regression

We can also examine the regression estimators. The slope coefficient from a regression of Y_t on X_t has the asymptotic distribution

$$\widehat{\beta}=\frac{\frac{1}{n^2}\sum_{i=1}^n Y_i^* X_i^*}{\frac{1}{n^2}\sum_{i=1}^n X_i^{*2}} \xrightarrow[d]{} \frac{\int_0^1 W_1^* W_2^*}{\int_0^1 W_2^{*2}}.$$

This is a non-degenerate random variable. Thus the slope estimator remains random in large samples and does not converge in probability.

Now consider the classical t-ratio T. It has the asymptotic distribution

$$\frac{1}{n^{1/2}}T=\frac{\frac{1}{n^2}\sum_{i=1}^n Y_i^* X_i^*}{\left(\frac{1}{n^2}\sum_{i=1}^n X_i^{*2}\right)^{1/2}\left(\frac{1}{n^2}\sum_{i=1}^n \left(Y_i^*-X_i^*\widehat{\beta}\right)^2\right)^{1/2}} \xrightarrow[d]{} \frac{\int_0^1 W_1^* W_2^*}{\left(\int_0^1 W_2^{*2}\right)^{1/2}\left(\int_0^1\left(W_1^*-W_2^*\frac{\int_0^1 W_1^* W_2^*}{\int_0^1 W_2^{*2}}\right)^2\right)^{1/2}}.$$

This is non-degenerate. Thus the t-ratio has an asymptotic distribution only after normalization by $n^{1/2}$, meaning that the unnormalized t-ratio diverges in probability!

To understand the utter failure of classical inference theory observe that the regression equation is

$$Y_t=\alpha+\beta X_t+\xi_t \tag{16.12}$$

with true values $\alpha=0$ and $\beta=0$. This means that the error $\xi_t=Y_t$ is a random walk. The latter is considerably more strongly autocorrelated than allowed by stationary regression theory, invalidating conventional standard errors. The latter are too small by an order of magnitude resulting in t-ratios which are misleadingly large.

[10] Since the variables have been standardized to have a unit variance the sample correlation equals the least squares slope coefficient.

What this means in practice is that t-ratios from spurious regressions are random and large even when there is no relationship. This explains the large t-ratio $T = 26$ for the simulated series and shows that the value $T = 15$ for the empirical series is uninformative. The reason for a large t-ratio is not because the series are related but is rather because the series are unit root processes so classical standard errors mis-characterize estimation variance.

One of the features of the above theory is that it shows that the magnitude of the distortion of the t-ratio increases with sample size. Interestingly, the original Granger-Newbold (1974) analysis was a simulation study which confined attention to the case $n = 50$. Granger-Newbold found the (then surprising) result that t-tests substantially overreject under the null hypothesis of a zero coefficient. It wasn't until the theoretical analysis by Phillips (1986) that it was realized that this distortion *worsened* as sample size increased. These results illustrate the insight – and limitations – of simulation analysis. Using simulation Granger-Newbold pointed out that there was a problem. But by fixing sample size at a single value they did not discover the surprising effect of sample size.

The fact that the t-ratio diverges as n increases means that the coverage of classical confidence intervals worses as n increases. To calibrate the magnitude of this distortion examine Figure 16.6(b). This plots[11] the finite-sample coverage probability of classical nominal 95% confidence intervals for the slope using student t critical values plotted as a function of sample size n. The observations were generated as independent random walks with normal innovations. You can see that the coverage ranges from 0.68 (for $n = 10$) to 0.2 (for $n = 200$). These coverage rates are unacceptably below the nominal coverage level of 0.95.

The above analysis focused on classical t-ratios and confidence intervals constructed with old-fashioned homoskedastic standard errors. This may seem to be an out-of-date analysis as we have made the case that old-fashioned standard errors are not used in contemporary econometric practice. However the problem as described carries over to alternative standard error constructions. The common heteroskedastic standard errors do not fundamentally change the asymptotic distribution. The Newey-West standard errors reduce the under-coverage but only partially. They are designed to consistently estimate the long-run variance of stationary series but fail when the series are non-stationary.

At this point let us collect what we have learned. If we have two time-series which are independent unit root processes, then by time-series plots, correlation analysis, and simple linear regressions it is easy to make the false inference that they are related. Their sample correlations and regression slope estimates will be random, inconsistent, and uninformative.

Our deduction is that it is inappropriate to use simple inference techniques when handling potentially non-stationary time series. We need to be more careful and use better inference methods.

It turns out that a simple modification is often sufficient to fundamentally alter the inference problem. Again, suppose we observe the independent series (16.10)-(16.11). A linear regression model is (16.12) with error $\xi_t = Y_t$. We can write the latter as $\xi_t = Y_{t-1} + e_t$. This means that a correct dynamic specfication of the regression model is

$$Y_t = \alpha + \beta X_t + \delta Y_{t-1} + e_t \tag{16.13}$$

with $\alpha = \beta = 0$ and $\delta = 1$. If equation (16.13) is estimated the error is no longer a random walk and inference on β can proceed conventionally! In this simple example a solution is simply to include the lagged dependent variable Y_{t-1} in the estimated regression. More generally, if a trend component is missing or ΔY_t is serially correlated it is necessary to include the trend terms and/or sufficient lags of Y_t in the estimated regression.

For example, take the simulated random walk series from Figure 16.5(a). Estimating model (16.13) we find $\widehat{\beta} = 0.004$ with a standard error of 0.005. Thus by adding the lagged dependent variable the

[11] Calculated by simulation on a grid of values for n with one million simulation replications.

spurious regression relationship has been broken. Now take the empirical series from Figure 16.5(b). We estimate an analog of (16.13) augmented with a linear trend. The estimate of β in this model is 0.16 with a standard error of 0.12. Once again the spurious regression relationship has been broken by a simple dynamic re-adjustment.

This seems like a straightforward solution. If so, why does the spurious regression problem persist[12] in applied analysis? The reason is partially that non-specialists find that the simple regression (16.12) is easy to interpret while the dynamic model (16.13) is challenging to interpret. One of the tasks of a skilled econometrician is to understand this failure of reasoning, to explain the problem to colleagues and users, and to present constructive useful alternative methods of analysis.

16.16 NonStationary VARs

Let Y_t be an $m \times 1$ time series. Suppose that Y_t satisfies a VAR(p-1) in first differences, thus $\boldsymbol{D}(\mathrm{L})\Delta Y_t = e_t$ where $\boldsymbol{D}(z)$ is invertible and $\Sigma = \operatorname{var}[e_t] > 0$. Then ΔY_t has the long-run covariance matrix $\Omega = \boldsymbol{D}(1)^{-1}\Sigma\boldsymbol{D}(1)^{-1\prime} > 0$. In this case Y_t is a vector $I(1)$ process in the sense that each element of Y_t is $I(1)$ and so are all linear combinations of Y_t.

The model can be written as a VAR in levels as

$$Y_t = \boldsymbol{A}_1 Y_{t-1} + \boldsymbol{A}_2 Y_{t-2} + \cdots + \boldsymbol{A}_p Y_{t-p} + e_t \tag{16.14}$$

where $\boldsymbol{A}_1 + \boldsymbol{A}_2 + \cdots + \boldsymbol{A}_p = \boldsymbol{I}_m$. It can also be written in the mixed format

$$\Delta Y_t = \boldsymbol{A} Y_{t-1} + \boldsymbol{D}_1 \Delta Y_{t-1} + \cdots + \boldsymbol{D}_{p-1}\Delta Y_{t-p+1} + e_t \tag{16.15}$$

where $\boldsymbol{A} = 0$. These are equivalent algebraic representations. Let $d = \operatorname{vec}\left(\left(\boldsymbol{D}_1, ..., \boldsymbol{D}_{p-1}\right)'\right)$.

Let $\left(\widehat{\boldsymbol{A}}, \widehat{d}\right)$ be the multivariate least squares estimator of (16.15). Set $X_t = (\Delta Y_{t-1}, ..., \Delta Y_{t-p+1})$.

Theorem 16.16 Assume that ΔY_t follows the VAR(p-1) process $\boldsymbol{D}(\mathrm{L})\Delta Y_t = e_t$ with invertible $\boldsymbol{D}(z)$, $\mathbb{E}[e_t \mid \mathscr{F}_{t-1}] = 0$, $\mathbb{E}\|e_t\|^4 < \infty$, and $\mathbb{E}\left[e_t e_t'\right] = \Sigma > 0$. Then as $n \to \infty$

$$\begin{pmatrix} n\widehat{\boldsymbol{A}} \\ \sqrt{n}\left(\widehat{d} - d\right) \end{pmatrix} \xrightarrow[d]{} \begin{pmatrix} \Sigma^{1/2}\int_0^1 dWW'\left(\int_0^1 WW'\right)^{-1}\Omega^{-1/2} \\ \\ \mathrm{N}(0, \boldsymbol{V}) \end{pmatrix}$$

where $W(r)$ is vector Brownian motion and

$$\begin{aligned} \boldsymbol{V} &= \left(\boldsymbol{I}_m \otimes \mathbb{E}\left[X_t X_t'\right]\right)^{-1}\Omega\left(\boldsymbol{I}_m \otimes \mathbb{E}\left[X_t X_t'\right]\right)^{-1} \\ \Omega &= \mathbb{E}\left[e_t e_t' \otimes X_t X_t'\right]. \end{aligned}$$

The top component of the asymptotic distribution is a multivariate version of the Dickey-Fuller coefficient distribution. The bottom component is a conventional normal distribution. This shows that the

[12] An amusing exercise is to peruse newspaper/magazine articles for time series plots of historical series. More often than not the displayed series appear to be $I(1)$, and more often than not the article describes the series as "related" based on a combination of eyeball analysis and simple correlation statistics.

coefficient estimator $\widehat{\boldsymbol{A}}$ is consistent at the $O_p(n^{-1})$ rate, converges to a non-standard (biased and non-normal) asymptotic distribution, and the coefficient estimator $\widehat{d}$ has a conventional asymptotic normal distribution.

Parameters of interest, including the coefficients of the levels equation (16.14), impulse response functions, and forecast error decompositions, are linear combination of the estimators $\left(\widehat{\boldsymbol{A}}, \widehat{d}\right)$. For VAR(p) models with $p \geq 2$, unless the linear combination of interest is in the span of $\widehat{\boldsymbol{A}}$, the asymptotic distribution of estimators are dominated by the $O_p(n^{-1/2})$ component $\widehat{d}$. Thus these coefficient estimators have conventional asymptotic normal distributions. Consequently, for most purposes estimation and inference on a VAR model is robust to the presence of (multivariate) unit roots.

There are two important exceptions. First, inference on the sum of levels coefficients $\boldsymbol{A}_1 + \boldsymbol{A}_2 + \cdots + \boldsymbol{A}_p$ is non-standard as the estimator of this sum has the multivariate Dickey-Fuller coefficient distribution. This includes questions concerning the presence of unit roots and many questions concerning the long-run properties of the series. Second, the long-run impulse matrix $\boldsymbol{C} = \boldsymbol{A}^{-1} = \left(\boldsymbol{I} - \boldsymbol{A}_1 - \boldsymbol{A}_2 - \cdots - \boldsymbol{A}_p\right)^{-1}$ is a (non-linear) function of this same sum and thus by the Delta Method is asymptotically a linear transformation of the multivariate Dickey-Fuller coefficient distribution. This means that the least squares estimator of $\boldsymbol{C}$ is non-standard (biased and non-normal). As $\boldsymbol{C}$ is the limit of the CIRF as the horizon tends to infinity this indicates that estimators of the CIRF at long horizons will be non-standard in finite samples. Consequently when a VAR model includes variables which are potentially unit root processes the conventional confidence intervals for the CIRF at long horizons are not trustworthy. This is a widespread issue since macroeconomists routinely estimate VAR models with macroeconomic variables in levels (for example, the Blanchard-Perotti (2002) model presented in Section 15.25).

16.17 Cointegration

A fascinating topic is cointegration. The idea is due to Granger (1981) and was articulated in detail by Engle and Granger (1987). A pair of unit root processes are **cointegrated** if their difference (or some linear combination) is stationary. This means that the pair "hang together" over the long run.

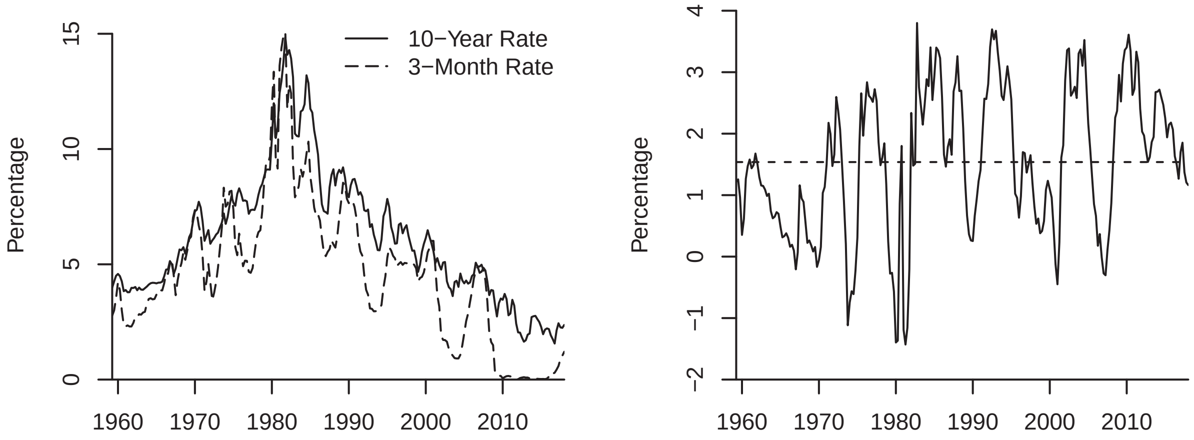
To visualize, examine Figure 16.7(a). This shows two interest rate series. The solid line is the interest rate (quarterly for 1959-2017) on ten-year U.S. Treasury Bonds[13]. The dashed line is the interest rate on 3-month U.S. Treasury Bonds[14]. Over the 59-year period the two series move up and down together. The 10-year rate exceeds the 3-month rate in most time periods. For some periods the two lines pull apart but they always come together again. This indicates that the two time series are tightly tied together. From our unit root analysis we have already determined that the 10-year interest rate is consistent with a unit root process; the same findings apply to the 3-month series. Thus it appears that these are two time series which are individually unit root processes but jointly track each other closely.

To see this further define the **interest rate spread** as the difference between the two interest rates, long (10-year) minus short (3-month). This series is plotted in Figure 16.7(b). The mean of the series is displayed by the dashed line. What we can see is that the spread roughly appears to be mean reverting. With the possible exception of the first decade of the plot we see that that the spread crosses its mean multiple times each decade. The fluctuations appear to be stationary. Applying an ADF unit root test with no trend included to the spread yields ADF $= -4.0$ which is less than the 1% critical value, rejecting the null hypothesis of a unit root. Thus the levels of the two interest rates appear to be non-stationary while the spread is stationary. This suggests that the two interest rate series are cointegrated.

This concept is formalized in the following definition.

[13] From FRED-QD, series *gs10*.

[14] From FRED-QD, series *tb3ms*.



(a) Interest Rates

(b) Interest Rate Spread

Figure 16.7: Cointegration

Definition 16.5 The $m \times 1$ non-deterministic series Y_t is **cointegrated** if there exists a full rank $m \times m$ matrix $[\beta, \beta_\perp]$ such that $\beta' Y_t \in \mathbb{R}^r$ and $\beta'_\perp \Delta Y_t \in \mathbb{R}^{m-r}$ are $I(0)$. The r vectors in β are called the **cointegrating vectors**. The variable $Z_t = \beta' Y_t$ is called the **equilibrium error**.

In the interest rate example of Figure 16.7, there are $m = 2$ series and $r = 1$ cointegrating relationships. Our discussion assumes that the cointegrating vector is $\beta = (1, -1)'$.

The cointegrating vectors β are not individually identified; only the space spanned by the vectors is identified so β is typically normalized. When $r = 1$ a common normalization is to set one non-zero element equal to one. Another common normalization is to set β to be orthonormal: $\beta'\beta = \boldsymbol{I}_r$.

Theorem 16.17 Granger Representation Theorem. If non-deterministic $Y_t \in \mathbb{R}^m$ is cointegrated with $m \times r$ cointegrating vectors β and (16.1) holds, then

1. The coefficients of the Wold representation

$$\Delta Y_t = \theta + \Theta(\mathrm{L}) e_t \tag{16.16}$$

satisfy $\Theta(1) = \beta_\perp \eta'$ and $\theta = \beta_\perp \gamma$ for some full-rank $m \times (m-r)$ matrix η and some $(m-r) \times 1$ γ.

2. The Beveridge-Nelson decomposition of Y_t is

$$Y_t = \beta_\perp \left(\gamma t + \eta' S_t\right) + U_t + V_0 \tag{16.17}$$

where $S_t = \sum_{i=1}^t e_t$, $U_t = \Theta^*(\mathrm{L}) e_t$ is a stationary linear process, and $V_0 = Y_0 - U_0$ is an initial condition.

3. Suppose that (a) all complex solutions to $\det(\Theta(z)) = 0$ are either $z = 1$ or $|z| \geq 1 + \delta$ for some $\delta > 0$; (b) $\beta' \Theta^*(1) \eta_\perp$ is full rank, where $\eta_\perp$ is a full rank $m \times r$ matrix such that $\eta' \eta_\perp = 0$. Then Y_t has the (infinite-order) convergent VAR representation

$$\boldsymbol{A}(\mathrm{L}) Y_t = a + e_t \tag{16.18}$$

where the coefficients satisfy $\boldsymbol{A}(1) = -\eta_\perp \left(\beta' \Theta^*(1) \eta_\perp\right)^{-1} \beta'$. All complex solutions to $\det(\boldsymbol{A}(z)) = 0$ are either $z = 1$ or $|z| \geq 1 + \delta$ for some $\delta > 0$.

4. Under the assumptions of part 3 plus $\sum_{j=0}^\infty \left\| \sum_{k=0}^\infty k \Theta_{j+k} \right\|^2 < \infty$ the VAR representation can be written in error-correction form

$$\Delta Y_t = \alpha \beta' Y_{t-1} + \Gamma(\mathrm{L}) \Delta Y_{t-1} + a + e_t \tag{16.19}$$

where $\Gamma(\mathrm{L})$ is a lag polynomial with absolutely summable coefficient matrices and $\alpha = -\eta_\perp \left(\beta' \Theta^*(1) \eta_\perp\right)^{-1}$.

5. If $\theta = 0$ in the Wold representation (16.16) then $\gamma = 0$ in (16.17) so there is no linear trend in (16.17). The intercept in (16.18) and (16.19) equals $a = \alpha \mu$ where μ is $r \times 1$. Equation (16.19) can be written as

$$\Delta Y_t = \alpha \left(\beta' Y_{t-1} + \mu\right) + \Gamma(\mathrm{L}) \Delta Y_{t-1} + e_t. \tag{16.20}$$

The proof is presented in Section 16.22. The Granger Representation Theorem appears in Engle and Granger (1987). The assumption on $\beta' \Theta^*(1) \eta_\perp$ was introduced by Johansen (1995, Theorem 4.5).

Part 1 shows that the coefficients of the Wold representation sum to a singular matrix in the null space of the cointegrating vectors.

Part 2 gives the Beveridge-Nelson permanent-transitory representation of Y_t. It shows that the trend $\beta_\perp \left(\gamma t + \eta' S_t\right)$ lies in the null space of the cointegrating vectors. Thus there is no trend in the range space of the cointegrating vectors. This shows that the cointegrated vector Y_t can be thought of as possessing r "unit roots and linear trends" and $m - r$ "stationary processes".

Part 3 provides the VAR representation. It shows that the VAR coefficients sum to a singular matrix which is in the range space of the cointegrating vectors.

Part 4 is perhaps the most famous result. It shows that a cointegrated system satisfies equation (16.19) which is called the **error-correction representation**. The error-correction representation is a regression model in stationary transformations as the variables ΔY_t and $\beta' Y_{t-1}$ are stationary. The equation shows that the change ΔY_t relates to past changes ΔY_{t-1} (as in a standard VAR) as well as the equilibrium error $\beta' Y_{t-1}$. The full term $\alpha\beta' Y_{t-1}$ is known as the "error-correction term". It is the key component which governs how the cointegrated relationship is maintained.

Part 5 examines the case of no linear trend. The condition $\theta = 0$ arises when the variables ΔY_t are all mean zero. The theorem (unsuprisingly) shows that this implies that the linear trend does not appear in the Beveridge-Nelson decomposition. More interestingly the theorem shows that this condition implies that the error-correction model can be written to incorporate the intercept.

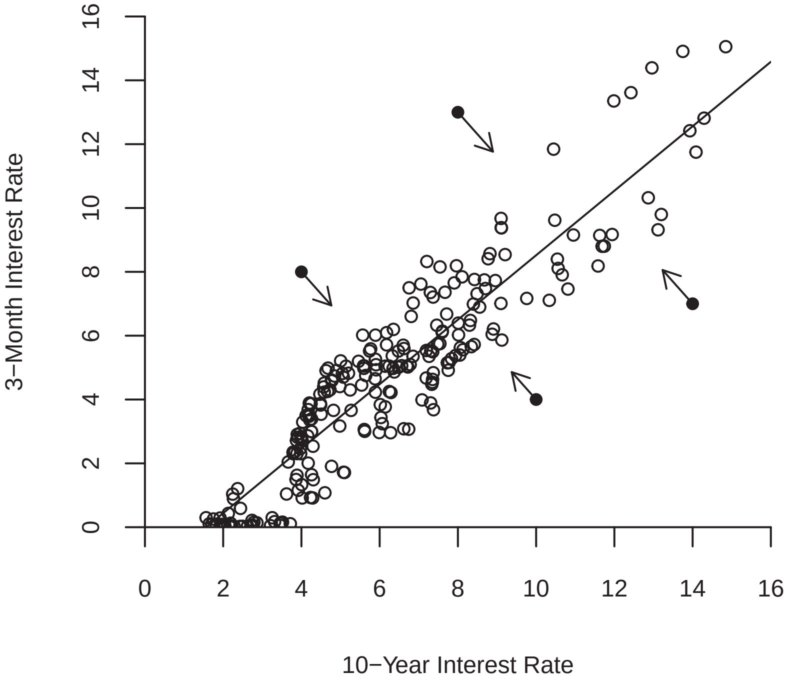


Figure 16.8: Error Correction Effect

To understand the error-correction effect examine Figure 16.8. This shows a scatter plot of the historical values of the two interest rate series from Figure 16.7. Also plotted is an estimate[15] of the linear relation $\beta' Y + \mu$ displayed as the solid line. This is the attractor of the system. For values of Y on this line $\beta' Y + \mu = 0$. For values to the southeast $\beta' Y + \mu < 0$, and for values to the northwest $\beta' Y + \mu > 0$. The components of α dictate how these values impact the expected direction of ΔY. The arrows indicate these directions[16]. When $\beta' Y + \mu > 0$ the error correction decreases the 3-month rate and increases the

[15] From Table 16.4.

[16] From the estimates in Table 16.5.

10-year rate, pushing Y towards the line of attraction. When $\beta' Y + \mu < 0$ the error correction increases the 3-month rate and decreases the 10-year rate, again pushing Y towards the line of attraction. In this particular example the two effects are similar in magnitude so the arrows show that both variables move towards the attractor in response to deviations.

Theorem 16.17 shows that if Y_t is cointegrated then it satisfies a VECM. The reverse is also the case.

Theorem 16.18 Granger Representation Theorem, Part II. Suppose that Y_t satisfies a VAR(∞) model $\boldsymbol{A}(\mathrm{L})Y_t = a + e_t$ with VECM representation

$$\Delta Y_t = \alpha\beta' Y_{t-1} + \Gamma(\mathrm{L})\Delta Y_{t-1} + a + e_t$$

where β and α are $m \times r$ and full rank. Suppose that (a) All complex solutions to $\det(\boldsymbol{A}(z)) = 0$ are either $z = 1$ or $|z| \geq 1 + \delta$ for some $\delta > 0$; (b) $\sum_{j=0}^{\infty} \left\|\Gamma_j\right\| < \infty$; (c) $\alpha'_{\perp}(\boldsymbol{I}_m - \Gamma(1))\beta_{\perp}$ is full rank where $\alpha_{\perp}$ and $\beta_{\perp}$ lie in the null spaces of α and β. Then Y_t is cointegrated with cointegrating vectors β.

The proof is presented in Section 16.22. This result for a finite-order VAR first appeared in Johansen (1995, Theorem 4.2).

The condition that $\alpha'_{\perp}\Gamma(1)\beta_{\perp}$ is full rank is necessary to exclude the (somewhat pathological) possibility that the system is "multi-cointegrated", meaning that a linear combination of $\beta' Y_{t-1}$ and ΔY_{t-1} is of reduced order of integration. Together, Theorems 16.17 and 16.18 show that a VECM representation is necessary and sufficient for a vector time series to be cointegrated.

16.18 Role of Intercept and Trend

The role of intercepts and trends in cointegrating VECMs gives rise to distinct models. We list some major options.

1. **Trend Model 1**. This specification has no intercept or trend terms

$$\Delta Y_t = \alpha\beta' Y_{t-1} + \Gamma(\mathrm{L})\Delta Y_{t-1} + e_t.$$

   This is convenient for pedagogy but is not relevant for empirical applications. In Stata use option `trend(none)`.

2. **Trend Model 2**. This specification is appropriate for non-trended series such as interest rates. In this model the intercept is in the cointegrating relationship

$$\Delta Y_t = \alpha\left(\beta' Y_{t-1} + \mu\right) + \Gamma(\mathrm{L})\Delta Y_{t-1} + e_t.$$

   In Stata use option `trend(rconstant)`.

3. **Trend Model 3**. This is appropriate for series which have possible linear trends. This model has an unconstrained intercept

$$\Delta Y_t = \alpha\beta' Y_{t-1} + \Gamma(\mathrm{L})\Delta Y_{t-1} + a + e_t.$$

   In this model the level series Y_t is the sum of a linear time trend and a unit root process. The equilibrium error $\beta' Y_t$ is stationary so eliminates the linear time trend and the unit root component. In Stata use option `trend(constant)`.

4. **Trend Model 4**. This model extends the VECM model to allow a linear trend in the cointegrating relationship. This model is

$$\Delta Y_t = \alpha\left(\beta' Y_{t-1} + \mu t\right) + \Gamma(\mathrm{L})\Delta Y_{t-1} + a + e_t.$$

   In this model the level series Y_t is the sum of a linear time trend and a unit root process. The equilibrium error $\beta' Y_t$ contains a linear time trend and a stationary process. Thus the cointegrating vector β only eliminates the unit root, not the time trend component. In Stata use option `trend(rtrend)`.

5. **Trend Model 5**. This is a further extension allowing an unconstrained trend term

$$\Delta Y_t = \alpha\beta' Y_{t-1} + \Gamma(\mathrm{L})\Delta Y_{t-1} + a + bt + e_t.$$

   In this model the unconstrained trend induces a quadratic time trend into the levels series Y_t. This is not a typical modeling choice for applied economic time series. In Stata use option `trend(trend)`.

16.19 Cointegrating Regression

If Y_t is cointegrated with a single cointegrating vector ($r = 1$) then it turns out that β can be estimated by a least squares regression of one component of Y_t on the others. This approach may be fruitfully employed when the major focus is the cointegrating vector, the number of variables m is small (e.g. $m = 2$ or $m = 3$), and it is known that the number of cointegrating vectors r is at most one.

Partition $Y_t = (Y_{1t}, Y_{2t})$ and reparameterize β as $(1, -\beta)$. Thus the first component of the cointegrating vector has been normalized to one (this requires that the true value is non-zero) and the remainder multiplied by -1. The coefficient of interest is β. Least squares is fit either to the equation

$$Y_{1t} = \mu + \beta' Y_{2t} + u_{1t} \tag{16.21}$$

(for Trend Models 1 or 2) or to the equation

$$Y_{1t} = \mu + \theta t + \beta' Y_{2t} + u_{1t} \tag{16.22}$$

(for Trend Models 3 or 4).

Define $u_{2t} = \Delta Y_{2t}$, $u_t = (u_{1t}, u_{2t}')'$, and the long-run covariance matrix $\Omega = \Sigma + \Lambda + \Lambda'$ where $\Sigma = \mathbb{E}\left[u_t u_{t-\ell}'\right]$ and $\Lambda = \sum_{\ell=1}^{\infty} \mathbb{E}\left[u_{t-\ell} u_t'\right]$. Partition the covariance matrices conformably with Y, e.g.

$$\Omega = \left[\begin{array}{cc} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{array}\right].$$

Theorem 16.19 If u_t satisfies the conditions of Theorem 16.4 and $\Omega_{22} > 0$ then the least squares estimator satisfies

$$n\left(\widehat{\beta} - \beta\right) \xrightarrow[d]{} \left(\int_0^1 XX'\right)^{-1}\left(\int_0^1 X dB_1 + \Sigma_{21} + \Lambda_{21}\right)$$

where $B(r) = (B_1(r), B_2(r))$ is a vector Brownian motion with covariance matrix Ω and $X(r)$ is determined by the model:

Trend Model 1 or 2 estimated by (16.21): $X = B_2^*$ (demeaned $B_2(r)$).

Trend Model 3 or 4 estimated by (16.22): $X = B_2^{**}$ (detrended $B_2(r)$).

The proof is presented in Section 16.22.

Theorem 16.19 shows that the estimator converges at the superconsistent $O_p(n^{-1})$ rate. This was discovered by Stock (1987) and the asymptotic distribution derived by Park and Phillips (1988). The asymptotic distribution is non-standard due to the serial correlation terms. Take our empirical example. A least squares regression of the 3-month interest rate on the 10-year interest rate yields the estimated equation $\widehat{Y}_{1t} = 1.03 Y_{2t} - 1.71$.

Modifications to the least squares estimator which eliminate the non-standard components were introduced by Phillips and B. E. Hansen (1990) and Stock and Watson (1993). The Phillips-Hansen estimator, known as Fully Modified OLS (FM-OLS), eliminates the non-standard components through first-stage estimation of the serial correlation terms. The Stock-Watson estimator, known as Dynamic OLS (DOLS), eliminates the non-standard components by estimating an augmented regression including leads and lags of ΔY_{2t}.

We are often interested in testing the hypothesis of no cointegration:

$$\mathbb{H}_0 : r = 0$$
$$\mathbb{H}_1 : r > 0.$$

Under $\mathbb{H}_0$, $Z_t = \beta' Y_t$ is $I(1)$ yet under $\mathbb{H}_1$ Z_t is $I(0)$. When β is known $\mathbb{H}_0$ can be tested by appying a univariate ADF test on Z_t. Take the interest rate example. We already conjectured that the interest rate spread is stationary which is the same as the hypothesis that $\beta = 1$ is the cointegrating coefficient. Using this value we computed ADF $= -4.0$ with an asymptotic p-value less than 0.01. Hence we are able to reject the null hypothesis of a unit root in the spread, or equivalently reject the null hypothesis of no cointegration.

When β is unknown, Engle and Granger (1987) proposed testing the null hypothesis of no cointegration by applying the ADF test to the least squares residual $\widehat{u}_{1t}$ from either (16.21) or (16.22). The asymptotic null distribution is different from the Dickey-Fuller distribution since under $\mathbb{H}_0$ the estimated regression is spurious so the least squares estimator is inconsistent. The asymptotic distribution of the statistic was worked out by Phillips and Ouliaris (1990) by combining the theory of spurious regression with Dickey-Fuller distribution theory. Let EG_p denote the Engle-Granger ADF statistic with p autoregressive lags in the ADF regression.

Theorem 16.20 Assume that $(\Delta Y_{1t}, \Delta Y_{2t})$ satisfies the conditions of Theorem 16.4 and $\Omega > 0$. If $p \to \infty$ as $n \to \infty$ such that $p^3/n \to 0$ then

$$\mathrm{EG}_p \underset{d}{\longrightarrow} \frac{\left(\int_0^1 V dV\right)}{\left(\int_0^1 V^2\right)^{1/2} (1+\zeta'\zeta)^{1/2}}$$

where, $V(r) = W_1^*(r) - \zeta' W_2^*(r)$ and $\zeta = \left(\int_0^1 W_2^* W_2^{*\prime}\right)^{-1} \left(\int_0^1 W_2^* W_1^*\right)$, $W(r) = (W_1(r), W_2(r))$ is vector standard Brownian motion, and $W^*(r)$ is demeaned $W(r)$ if (16.21) is estimated or detrended $W(r)$ if (16.22) is estimated.

For a proof see Phillips and Ouliaris (1990).

An unusual feature of this Theorem is that it requires $p \to \infty$ as $n \to \infty$ even if the true process is a finite order AR process because the first stage spurious regression induces serial correlation into the first-stage residuals which needs to be handled in the second stage ADF test. Another unusual feature is the component $1+\zeta'\zeta$ in the denominator. This is due to the variance estimator component which asymptotically is random because of the first stage spurious regression.

Table 16.3: Engle-Granger Cointegration Test Critical Values

	Intercept			Trend		
	$m=2$	$m=3$	$m=4$	$m=2$	$m=3$	$m=4$
0.01%	−5.07	−5.44	−5.81	−5.52	−5.82	−6.12
0.1%	−4.54	−4.93	−5.28	−4.96	−5.28	−5.60
1%	−3.89	−4.29	−4.64	−4.33	−4.66	−4.97
2%	−3.67	−4.07	−4.42	−3.97	−4.44	−4.74
3%	−3.53	−3.93	−4.28	−3.86	−4.30	−4.61
4%	−3.42	−3.82	−4.18	−3.78	−4.20	−4.51
5%	−3.34	−3.74	−4.09	−3.65	−4.12	−4.42
7%	−3.20	−3.61	−3.96	−3.50	−3.98	−4.29
10%	−3.04	−3.45	−3.81	−3.48	−3.83	−4.14
15%	−2.85	−3.26	−3.62	−3.31	−3.65	−3.96
20%	−2.70	−3.11	−3.47	−3.16	−3.50	−3.81
30%	−2.45	−2.86	−3.23	−2.92	−3.26	−3.57
50%	−2.05	−2.47	−2.83	−2.54	−2.87	−3.18
70%	−1.65	−2.08	−2.45	−2.16	−2.49	−2.80
90%	−1.00	−1.48	−1.88	−1.60	−1.94	−2.25
99%	0.09	−0.44	−0.94	−0.69	−1.05	−1.41

Source: Calculated by simulation from one million replications of samples of size $n = 10{,}000$.

The asymptotic critical values[17] are displayed in Table 16.3. The EG test is one-sided, so rejections occur when the test statistic is less than (more negative than) the critical value. The critical values are a function of the number of variables m and the detrending method.

Let's summarize the Engle-Granger cointegration test: The null hypothesis is that the series are not cointegrated, or equivalently that the equilibrium error is $I(1)$. The alternative hypothesis is cointegration. The regression (16.21) or (16.22) is estimated by least squares to obtain the residual. The ADF test is applied to this residual. This is done by fitting an AR(p) model and testing that the sum of the autoregressive coefficients equal one. Critical values are taken from Table 16.3 according to the trend specification (discussed below) and the number of variables m. If the t-statistic is smaller than the appropriate critical value, the null hypothesis of no cointegration is rejected in favor of the hypothesis of cointegrationd. Otherwise, the hypothesis of no cointegration is not rejected.

An important question is which trend model to fit. If the observations are untrended then the intercept regression (16.21) should be fit and the "Intercept" critical values used. If the observations are trended and no constraints are imposed then the trend regression (16.22) should be fit and the "Trend" critical values used. A complication arises in the case of Model 3, which allows the observations to be trended but the trend is excluded from the cointegrating regression. In this cases there are two options. One is to treat the situation as Model 4: estimate regression (16.22) and use the associated critical values. The other option is to estimate (16.21) since the linear trend is not in the cointegrating relationship. In this case the appropriate critical values are from the "Trend" section of the table, but with the row corresponding to $m-1$. This is because one of the unit root processes in regression (16.22) is dominated by a linear trend. For example, if there are $m=3$ variables in the system and (16.21) is estimated, then use

[17] Calculated by simulation from one million simulation draws for a sample of size $n = 10{,}000$.

the critical values for "Trend" and $m = 2$. If there are $m = 2$ variables then use the "Case 3" ADF critical values from Table 16.1.

To illustrate, take the interest rate application. These variables are non-trended so we use model (16.21) with the "Intercept" critical values. The least squares residuals are $\widehat{u}_{1t} = \widehat{Y}_{1t} - 1.03 Y_{2t} - 1.7$. Applying an ADF test with $p = 8$ we obtain EG $= -4.0$. This is smaller than the 1% asymptotic critical value of -3.9 from Table 16.3. We therefore reject the hypothesis of no cointegration, supporting the hypothesis that the pair are cointegrated.

16.20 VECM Estimation

The Granger Representation Theorem (Theorems 16.17 and 16.18) showed that Y_t is cointegrated if (and only if) Y_t satisfies an error-correction model. A VECM(p) model is

$$\Delta Y_t = \alpha \beta' Y_{t-1} + \Gamma_1 \Delta Y_{t-1} + \cdots + \Gamma_{p-1} \Delta Y_{t-p+1} + a + e_t. \tag{16.23}$$

This is a reduced rank regression as introduced in Section 11.11. The standard estimation method is maximum likelihood under the auxilary assumption that e_t is i.i.d. $\mathrm{N}(0, \Sigma)$, described in Theorem 11.7. We repeat this result here for the VECM model.

Theorem 16.21 The MLE for the VECM (16.23) under $e \sim \mathrm{N}(0, \Sigma)$ is given as follows. First, regress ΔY_t and Y_{t-1} on $\Delta Y_{t-1}, ..., \Delta Y_{t-p+1}$ and an intercept to obtain the residual vectors $\widehat{u}_{0t}$ and $\widehat{u}_{1t}$, organized in matrices as $\widehat{\boldsymbol{U}}_0$ and $\widehat{\boldsymbol{U}}_1$. The MLE $\widehat{\beta}$ equals the first r generalized eigenvectors of $\frac{1}{n}\widehat{\boldsymbol{U}}_1'\widehat{\boldsymbol{U}}_0\left(\frac{1}{n}\widehat{\boldsymbol{U}}_0'\widehat{\boldsymbol{U}}_0\right)^{-1}\frac{1}{n}\widehat{\boldsymbol{U}}_0'\widehat{\boldsymbol{U}}_1$ with respect to $\frac{1}{n}\widehat{\boldsymbol{U}}_1'\widehat{\boldsymbol{U}}_1$ corresponding to the r largest eigenvalues $\widehat{\lambda}_j$. This uses the normalization $\widehat{\beta}'\frac{1}{n}\widehat{\boldsymbol{U}}_1'\widehat{\boldsymbol{U}}_1\widehat{\beta} = \boldsymbol{I}_r$. The MLE for the remaining coefficients $\widehat{\alpha}$, $\widehat{\Gamma}_1, ..., \widehat{\Gamma}_{p-1}$, and $\widehat{a}$ are obtained by the least squares regression of ΔY_t on $\widehat{\beta}' Y_{t-1}$, $\Delta Y_{t-1}, ..., \Delta Y_{t-p+1}$, and an intercept. The maximized log-likelihood function is

$$\ell_n(r) = \frac{m}{2}\left(n \log(2\pi) - 1\right) - \frac{n}{2}\det\left(\frac{1}{n}\widehat{\boldsymbol{U}}_0'\widehat{\boldsymbol{U}}_0\right) - \frac{n}{2}\sum_{j=1}^{r}\log\left(1 - \widehat{\lambda}_j\right).$$

This estimation method was developed by Johansen (1988, 1991, 1995) as an extension of the reduced rank regression of Anderson (1951).

The VECM is a constrained VAR so the VECM estimates can be used for any purpose for which a VAR is used. An advantage of the VECM estimation approach is that it provides a coherent model of the system, is computationally straightforward, and can handle multiple cointegrating vectors. A disadvantage is that when there are multiple cointegrating vectors ($r > 1$) then interpretation of the cointegrating space (the space spanned by β) is difficult.

The VECM model assumes that the VAR order p and cointegrating rank r are known. In practice data-based selection rules are used. AIC minimization may be used for selection of p. A simple approach is to select p by estimating unrestricted VAR models. Selection of r is typically done by testing methods; this is reviewed in the next section.

We illustrate with the two interest rate series already introduced. AIC selection on levels VARs selects a VAR(8); we report here a VAR(4) as it yields similar results. This implies a VECM with 3 dynamic lags.

Table 16.4: VECM Cointegrating Vector

	β	s.e.
3-Month	1	
10-Year	−1.01	0.07
Intercept	1.58	0.46

Table 16.5: Vector Error Correction Model

	Δ3-Month$_t$	Δ10-Year$_t$
Z_{t-1}	−0.09	0.07
	(0.04)	(0.03)
Δ3-Month$_{t-1}$	0.37	0.04
	(0.08)	(0.06)
Δ3-Month$_{t-2}$	−0.20	−0.08
	(0.08)	(0.06)
Δ3-Month$_{t-3}$	0.28	0.07
	(0.08)	(0.06)
Δ10-Year$_{t-1}$	0.06	0.21
	(0.07)	(0.08)
Δ10-Year$_{t-2}$	−0.19	−0.09
	(0.12)	(0.08)
Δ10-Year$_{t-3}$	0.10	0.06
	(0.12)	(0.08)

Since interest rates are not a trended series we use Trend Model 2. The estimated model is reported in Tables 16.4 and 16.5.

Table 16.4 reports the estimated cointegrating vector β. The coefficient on the 3-month interest rate is normalized to one. The estimated coefficient on the 10-year rate is near -1 and the estimated intercept is about 1.6. The latter means that the 3-month rate is on average 1.6 percentage points below the 10-year rate. The coefficients of the estimated VECM are reported in Table 16.5, one column for each variable. The first reported coefficient is $\widehat{\alpha}$, the error-correction term. The coefficient for the 3-month rate is negative and that for the 10-year rate is positive and they are of similar magnitude. Thus when the 3-month rate exceeds the 10-year rate by more than the typical 1.6, the 3-month rate tends to fall and the 10-year rate tends to rise, moving the two rates closer to the cointegrating relation. The following six coefficients are the dynamic coefficients of the VECM. We can see that each variable tends to respond mostly to its own lagged changes. The 3-month interest rate has considerably larger coefficients than the 10-year rate indicating that it has stronger serial correlation. The varying signs of the coefficients reveal complicated dynamics..

An asymptotic distribution of the VECM estimator requires a normalization for the cointegrating vectors. A popular choice is $\beta = (\boldsymbol{I}_r, \beta^{*\prime})'$. Johansen (1995, Theorem 13.5) shows that under the assumption that the errors e_t are i.i.d. with covariance matrix Σ, the coefficient estimators $\widehat{\theta} = (\widehat{\alpha}, \widehat{\Gamma})$ satisfy

$$\sqrt{n}\left(\widehat{\theta} - \theta\right) \xrightarrow[d]{} \mathrm{N}\left(0, \Sigma \otimes \boldsymbol{Q}^{-1}\right)$$

where $\boldsymbol{Q} = \mathbb{E}\left[X_t X_t'\right]$ with $X_t = (\beta' Y_{t-1}, \Delta Y_{t-1}, ..., \Delta Y_{t-p+1})$, the regressors given β. This is a classical (homoskedastic) asymptotic distribution for multivariate regression. This result shows that inference on the coefficients θ can proceed using conventional methods. The homoskedastic covariance matrix is due to the assumption that the errors are homoskedastic. If the latter assumption is relaxed then the asymptotic distribution generalizes to the case of an unrestricted covariance matrix.

Johansen (1995, Theorem 13.3) presents the asymptotic distribution of $\widehat{\beta}$. He shows that the asymptotic distribution is normal with a random covariance matrix. The latter is known as a mixed Gaussian distribution. From a practical point of view this means that we can treat the asymptotic distribution as normal since when scaled by an appropriate standard error the asymptotic distribution is standard normal. For brevity we do not present the details.

In Stata use the command `vec` to estimate a VECM with given cointegrating rank r and VAR order p.

16.21 Testing for Cointegration in a VECM

Take the model

$$\Delta Y_t = \Pi Y_{t-1} + \Gamma_1 \Delta Y_{t-1} + \cdots + \Gamma_{p-1} \Delta Y_{t-p+1} + a + e_t. \tag{16.24}$$

The Granger Representation Theorem shows that Y_t is cointegrated with r cointegrating vectors if and only if the rank of Π equals r. Thus testing for cointegration is equal to testing hypotheses on the rank of Π. Write the hypothesis that there are r cointegrating vectors as $\mathbb{H}(r) : \operatorname{rank}(\Pi) = r$.

Cointegration is a restriction on the unrestricted model $\mathbb{H}(m)$. A test for r cointegrating vectors against an unrestricted alternative is a test of $\mathbb{H}(r)$ against $\mathbb{H}(m)$. The likelihood ratio statistic for $\mathbb{H}(r)$ against $\mathbb{H}(m)$ is

$$\mathrm{LR}(r) = 2\left(\ell_n(m) - \ell_n(r)\right) = -n \sum_{j=1}^{m} \log\left(1 - \widehat{\lambda}_j\right) + n \sum_{j=1}^{r} \log\left(1 - \widehat{\lambda}_j\right) = -n \sum_{j=r+1}^{m} \log\left(1 - \widehat{\lambda}_j\right)$$

where $\widehat{\lambda}_j$ are the eigenvalues from the estimation problem (16.21). The test accepts $\mathbb{H}(r)$ for small values of $\mathrm{LR}(r)$; the test rejects $\mathbb{H}(r)$ for large values of $\mathrm{LR}(r)$.

Table 16.6: VECM Cointegration Rank Critical Values: Trend Model 2

$m-r$	1	2	3	4	5	6	7	8	9	10	11	12
0.01%	22.4	37.3	55.7	78.5	105	135	169	208	250	296	347	402
0.1%	17.6	31.5	48.8	70.1	95.7	125	158	196	237	282	332	385
1%	12.8	25.1	41.3	61.3	85.4	113	146	182	222	266	314	366
2%	11.3	23.1	38.7	58.4	81.9	110	141	177	216	260	308	359
3%	10.4	21.9	37.2	56.5	79.8	107	138	174	213	256	304	355
4%	9.71	21.0	36.1	55.2	78.3	105	136	171	210	254	301	352
5%	9.19	20.3	35.2	54.1	77.0	104	135	170	208	251	298	349
7%	8.42	19.2	33.8	52.5	75.0	102	132	167	205	248	295	345
10%	7.57	18.0	32.3	50.6	72.8	99.0	129	163	202	244	290	341
15%	6.60	16.6	30.4	48.3	70.1	95.9	126	159	197	239	285	335
20%	5.89	15.5	29.0	46.5	67.9	93.4	123	156	194	235	281	330
30%	4.86	13.9	26.8	43.7	64.6	89.5	119	151	188	229	274	323
50%	3.45	11.4	23.4	39.4	59.4	83.4	111	143	179	219	263	312
70%	2.39	9.39	20.4	35.5	54.6	77.6	105	136	171	210	253	300
90%	1.35	6.96	16.7	30.4	48.1	69.9	95.7	125	159	197	239	285

Source: Calculated by simulation from one million replications of samples of size $n = 10{,}000$.

The asymptotic distribution theory was developed by Johansen (1988, 1991, 1995).

Theorem 16.22 Assume that the finite-lag VECM (16.24) is correctly specified, the conditions of Theorem 16.18 hold, and the errors e_t are a MDS. Under the hypothesis that Π has rank r

$$\mathrm{LR}(r) \underset{d}{\longrightarrow} \operatorname{tr}\left[\left(\int_0^1 dWX'\right)\left(\int_0^1 XX'\right)^{-1}\left(\int_0^1 XdW'\right)\right]$$

where $W(r)$ is a $m-r$ dimensional standard Brownian motion and $X(r)$ is a stochastic process which is a function of $W(r)$ depending on the trend model.

1. Trend Model 1. $X(r) = W(r)$
2. Trend Model 2. $X(r) = (W(r), 1)$
3. Trend Model 3. $X(r) = (W_1^*(r), r - 1/2)$
4. Trend Model 4. $X(r) = (W^*(r), r - 1/2)$

where $W^*(r) = W(r) - \int_0^1 W$ is demeaned $W(r)$, and $W_1^*(r)$ is the first $m-r-1$ components of $W^*(r)$.

A proof of Theorem 16.22 is algebraically tedious. We provide a sketch in Section 16.22. See Johansen (1995, Chapter 11) for full details.

Table 16.7: VECM Cointegration Rank Critical Values: Trend Model 3

$m-r$	1	2	3	4	5	6	7	8	9	10	11	12
0.01%	15.2	31.5	49.0	71.0	96.9	125	159	196	238	283	333	386
0.1%	10.8	25.9	42.8	63.3	87.7	116	148	185	225	269	318	370
1%	6.63	20.0	35.5	54.7	77.9	105	136	171	210	253	300	351
2%	5.42	18.1	33.2	51.9	74.5	101	132	166	205	247	294	345
3%	4.72	17.0	31.7	50.2	72.5	98.9	128	163	202	244	290	341
4%	4.23	16.2	30.7	48.9	71.0	97.1	127	161	199	241	288	338
5%	3.85	15.5	29.8	47.9	69.8	95.7	126	160	197	239	285	335
7%	3.29	14.5	28.5	46.3	67.9	93.6	123	157	194	236	282	331
10%	2.71	13.4	27.1	44.5	65.8	91.1	120	154	191	232	277	327
15%	2.08	12.2	25.3	42.3	63.2	88.1	117	150	187	227	272	321
20%	1.64	11.2	24.0	40.7	61.2	85.7	114	147	183	224	268	317
30%	1.07	9.75	22.0	38.0	58.0	82.0	110	142	178	218	262	310
50%	0.45	7.68	18.9	34.0	53.1	76.2	103	134	169	208	251	298
70%	0.15	5.96	16.2	30.4	48.5	70.7	96.8	127	161	199	241	287
90%	0.02	4.04	12.8	25.7	42.5	63.3	88.1	117	150	187	227	272

Source: Calculated by simulation from one million replications of samples of size $n = 10,000$.

Theorem 16.22 provides the asymptotic distribution of the LR test for cointegration rank. Because the asymptotic distribution equals the trace of a multivariate Dickey-Fuller distribution the statistic LR is often referred to as the "trace test" or "Johansen's trace test". The asymptotic distribution is a function of the stochastic process $X(r)$ which equals the trend components of Y_t (under the hypothesis of r

cointegrating vectors) projected orthogonal to the other regressors. For Trend Model 2 the intercept is included in the cointegrating relationship so it is a component of $X(r)$. For Trend Model 3 the variables are trended which dominates the other components so appears in the asymptotic distribution. Since the intercept is excluded from the cointegrating relationship the components of $X(r)$ are all demeaned. For Trend Model 4 the linear trend is included in the cointegrating relationship so it is added to the trend components while the intercept is excluded so the $X(r)$ process is demeaned.

The asymptotic distribution is a function only of $m-r$ and the trend specification. Asymptotic critical values[18] are displayed in Tables 16.6-16.8 for $m-r$ up to 12 for Trend Models 2, 3, and 4. These are upper-tailed tests, so the null hypothesis that the cointegrating rank is r is rejected if the test statistic is larger than the appropriate critical value; otherwise the null hypothesis is nor rejected. For example, the hypothesis of no cointegration is the same as $r = 0$. The appropriate critical value is then the column corresponding to the number of variables m. For example, for Trend Model 2 with $m = 4$ variables the 5% critical value is 54.1. If $\text{LR}(r) > 54.1$ the hypothesis of no coingration is rejected (implying that the series are cointegrated); otherwise the hypothesis of no cointegration is not rejected.

How are the test statistics $\text{LR}(r)$ used in practice? When the cointegrating rank is unknown the statistics can be used to determine r. The conventional procedure is a sequential test. Start with $\mathbb{H}(0)$ (the null hypothesis of no cointegration) and the associated statistic LR(0) which has m degrees of freedom. If the test rejects (if LR(0) exceeds the row m critical value) this is evidence that there is at least one cointegrating vector, or $r \geq 1$. Next, take $\mathbb{H}(1)$ (the null hypothesis of one cointegrating vector) and the associated statistic LR(1) which has $m-1$ degrees of freedom. If this test also rejects (if LR(1) exceeds the row $m-1$ critical value) this is evidence that there is at least two cointegrating vectors, or $r \geq 2$. Continue this sequence of tests until one fails to reject.

For example, when there are two variables ($m = 2$) compare the statistic LR(0) against the $m = 2$ critical value. If the test rejects (if the statistic exceeds the critical value) this is evidence that the series are cointegrated. If the test fails to reject the inference is uncertain.

Table 16.8: VECM Cointegration Rank Critical Values: Trend Model 4

$m-r$	1	2	3	4	5	6	7	8	9	10	11	12
0.01%	27.4	44.4	64.6	90.0	117	150	186	226	271	319	372	428
0.1%	22.1	38.1	57.4	81.0	108	139	175	214	258	305	356	412
1%	16.6	31.2	49.4	71.5	97.6	128	162	200	242	288	338	392
2%	14.9	29.0	46.7	68.4	94.0	124	157	195	236	282	332	385
3%	13.9	27.6	45.1	66.4	91.8	121	154	192	233	278	328	381
4%	13.1	26.7	43.9	65.0	90.1	119	152	189	230	275	325	378
5%	12.5	25.9	42.9	63.9	88.8	118	151	187	228	273	322	375
7%	11.7	24.7	41.4	62.1	86.7	115	148	184	225	270	318	371
10%	10.7	23.3	39.8	60.1	84.4	113	145	181	221	266	314	366
15%	9.53	21.7	37.7	57.6	81.5	109	141	177	217	261	309	360
20%	8.70	20.5	36.2	55.7	79.2	107	138	174	213	257	304	356
30%	7.45	18.7	33.8	52.8	75.7	103	134	169	207	250	297	348
50%	5.70	15.9	30.0	48.1	70.2	96.2	126	160	198	240	286	336
70%	4.28	13.5	26.7	43.8	65.0	90.1	119	152	189	231	276	325
90%	2.79	10.5	22.4	38.2	58.0	81.8	110	141	177	217	261	309

Source: Calculated by simulation from one million replications of samples of size $n = 10,000$.

This testing procedure is appealing when m is small (e.g. $m \leq 4$) but is less appealing for large m.

[18] Calculated by simulation from one million simulation draws for a sample of size $n = 10,000$.

With large m the procedure has several challenges. Sequential testing requires multiple testing for which it is difficult to control Type I error. Simultaneously the test can have low power implying that the procedure is likely to "identify" an inappropriately low value of r.

An alternative approach is to use cointegration tests to verify a selected specification. Start with economic modeling to motivate the cointegrating rank r. The likelihood ratio $\mathrm{LR}(r)$ can be used to test this assumption against the unrestricted VAR. If the test rejects $\mathbb{H}(r)$ this is evidence that the proposed model is incorrect.

We illustrate using the interest rate series with a VAR(4) and Trend Model 2. Our starting presumption is that the variables are $I(1)$ and cointegrated, implying that the cointegrating rank is $r = 1$. The value of LR(0) is 31.6. To compute the p-value we use Table 16.6 for Trend Model 2 with $m - r = 2$. The value 31.6 exceeds the 1% critical value of 25.1 so the asymptotic p-value of the test is less than 1%. Thus the null hypothesis of no cointegration is strongly rejected in favor of at least one cointegrating vector. The value of LR(1) is 2.8. The p-value is calculated using $m - r = 1$. The value 2.8 is smaller than the 50% critical value of 3.5 so the p-value is larger than 50%. The statistic does not reject the hypothesis of $\mathbb{H}(1)$. Together the statistics are consistent with the modeling assumption that the series are $I(1)$ and mutually cointegrated.

For a broader application we expand to five Treasury interest rates[19]: 3-month, 6-month, 1-year, 5-year, and 10-year. Our starting presumption is that the series are each $I(1)$ and that the system of variables are cointegrated, so that the cointegrating rank is at least one. If all four spreads are mutually stationary then the system will have four coingrating vectors, thus $r = 4$. However if the the distribution of the spreads change over time the cointegrating rank could be less than four. Thus we expect $1 \leq r \leq 4$ but are uncertain of its precise value.

We report the likelihood ratio tests for cointegration rank in Table 16.9. The LR test for $r = 0$ is 120 which exceeds the 1% critical value of 85.4, and the LR test for $r = 1$ is 68.3 which exceeds the 1% critical value of 61.3, so we safely reject the hypotheses of $r = 0$ and $r = 1$. This suggests that $r \geq 2$. The LR test for $r = 2$ is 33.6 with a p-value of 0.07, which is borderline significant. The tests for $r = 3$ and $r = 4$ are insignificant. In sum, we cannot reject the models $\mathbb{H}(2)$, $\mathbb{H}(3)$, or $\mathbb{H}(4)$. $\mathbb{H}(2)$ is doubtful, but the statistical evidence alone cannot distinguish $\mathbb{H}(3)$ versus $\mathbb{H}(4)$. Our recommendation in this context is to use either $\mathbb{H}(3)$ or $\mathbb{H}(4)$.

Table 16.9: Tests for Cointegrating Rank

	LR(r)	p-value
0	120	< 0.01
1	68.3	< 0.01
2	33.6	0.07
3	10.8	> 0.50
4	2.9	> 0.50

In Stata use `vecrank` to calculate the LR tests for cointegrating rank. The output is a table displaying LR(r) for $r = 0, \ldots, m-1$ along with the asymptotic 5% critical values. The p-value can be calculated from Tables 16.6-16.8.

[19] FRED-MD series TB3MS, TB6MS, GS1, GS5, and GS10.

16.22 Technical Proofs*

Proof of Theorem 16.1. In the text we showed that the limit distributions of S_n coincide with those of B. To appeal to the Functional Central Limit Theorem (Theorem 18.3 of *Probability and Statistics for Economists*) we need to verify that S_n is asymptotically equicontinuous (see Definition 18.7 of *Probability and Statistics for Economists*). For simplicity we focus on the scalar case $e_t \in \mathbb{R}$.

Assume without loss of generality that $\sigma^2 = 1$. Take any $0 < \eta < 1$ and $0 < \epsilon < 1$. Set $\delta \leq \epsilon\eta^4/48^2$. Note that

$$\sup_{|r_2-r_1|\leq\delta} |S_n(r_2) - S_n(r_1)| \leq 2 \sup_{0\leq j\leq\lfloor 1/\delta\rfloor} \sup_{0\leq r\leq\delta} \left|S_n(j\delta + r) - S_n\left(j\delta\right)\right|.$$

Then

$$\begin{aligned}
\mathbb{P}\left[\sup_{|r_2-r_1|\leq\delta} |S_n(r_2) - S_n(r_1)| > \eta\right] &\leq \mathbb{P}\left[\bigcup_{j=0}^{\lfloor 1/\delta\rfloor} \sup_{0\leq r\leq\delta} \left|S_n(j\delta + r) - S_n\left(j\delta\right)\right| > \frac{\eta}{2}\right] \\
&\leq \sum_{j=0}^{\lfloor 1/\delta\rfloor} \mathbb{P}\left[\sup_{0\leq r\leq\delta} \left|S_n(j\delta + r) - S_n\left(j\delta\right)\right| > \frac{\eta}{2}\right] \\
&\leq \left(\frac{1}{\delta} + 1\right)\mathbb{P}\left[\sup_{0\leq r\leq\delta} |S_n(r)| > \frac{\eta}{2}\right] \\
&= \left(\frac{1}{\delta} + 1\right)\mathbb{P}\left[\max_{i\leq\lfloor n\delta\rfloor} \left|\frac{1}{\sqrt{n}}\sum_{t=1}^{i} e_t\right| > \frac{\eta}{2}\right] \\
&\leq 2\left(\frac{1}{\delta} + 1\right)\mathbb{P}\left[\left|\frac{1}{\sqrt{n}}\sum_{t=1}^{\lfloor n\delta\rfloor} e_t\right| > \frac{\eta}{4}\right].
\end{aligned}$$

The final inequality is Billingsley's (B.52) which holds because $\delta < \eta/4\sqrt{2}$ under the assumptions. Our statement (B.52) of Billingsley's inequality assumes that e_t is an i.i.d. sequence; the result can be extended to a MDS sequence.

The CLT implies that $n^{-1/2}\sum_{t=1}^{\lfloor n\delta\rfloor} e_t \underset{d}{\longrightarrow} Z_\delta \sim \mathrm{N}(0,\delta)$. For n sufficiently large the final line is bounded by

$$\frac{3}{\delta}\mathbb{P}\left[|Z_\delta| > \frac{\eta}{4}\right] = \frac{3}{\delta}\mathbb{P}\left[Z_\delta^4 > \frac{\eta^4}{16^2}\right] \leq \frac{3}{\delta}\frac{16^2}{\eta^4}\mathbb{E}\left[Z^4\right] = \frac{48^2}{\eta^4}\delta = \epsilon. \tag{16.25}$$

The first inequality is Markov's, the following equality $\mathbb{E}\left[Z_\delta^4\right] = 3\delta^2$, and the final equality is the assumption $\delta = \epsilon\eta^4/48^2$. This shows that S_n satisfies the definition of asymptotic equicontinuity. ■

Proof of Theorem 16.7. Z_t has the Wold decomposition $Z_t = \Theta(\mathrm{L})e_t$. We add the additional assumption that e_t is a MDS to simplify the proof. By the Beveridge-Nelson decomposition $Z_t = \xi_t + U_t - U_{t-1}$ where $\xi_t = \Theta(1)e_t$ and $U_t = \Theta^*(\mathrm{L})e_t$. Then

$$\begin{aligned}
\frac{1}{n}\sum_{t=1}^{n} S_{t-1}Z_t' &= \frac{1}{n}\sum_{t=1}^{n} S_{t-1}\xi_t' + \frac{1}{n}\sum_{t=1}^{n} S_{t-1}U_t' - \frac{1}{n}\sum_{t=1}^{n} S_{t-1}U_{t-1}' \\
&= \frac{1}{n}\sum_{t=1}^{n} S_{t-1}\xi_t' - \frac{1}{n}\sum_{t=1}^{n-1} Z_tU_t' + o_p(1).
\end{aligned}$$

The first term converges to $\int_0^1 BdB'$ by Theorem 16.6. The Brownian motion has covariance matrix equal to the long-run variance of Z_t, which is Ω. The second term converges in probability to $\mathbb{E}\left[Z_tU_t'\right]$. Making the substitutions $U_t = \xi_{t+1} + U_{t+1} - Z_{t+1}$ and $\mathbb{E}\left[Z_t \xi_{t+1}'\right] = 0$ this can be written as

$$\begin{aligned}\mathbb{E}\left[Z_t U_t'\right] &= \mathbb{E}\left[Z_t \xi_{t+1}'\right] + \mathbb{E}\left[Z_t U_{t+1}'\right] - \mathbb{E}\left[Z_t Z_{t+1}'\right] \\ &= \mathbb{E}\left[Z_t U_{t+1}'\right] - \mathbb{E}\left[Z_t Z_{t+1}'\right] \\ &= \mathbb{E}\left[Z_t U_{t+2}'\right] - \mathbb{E}\left[Z_t Z_{t+2}'\right] - \mathbb{E}\left[Z_t Z_{t+1}'\right] \\ &= \cdots \\ &= -\sum_{j=1}^{\infty} \mathbb{E}\left[Z_t Z_{t+j}'\right] = -\sum_{j=1}^{\infty} \mathbb{E}\left[Z_{t-j} Z_t'\right] = -\Lambda.\end{aligned}$$

The third line makes the substitutions $U_{t+1} = \xi_{t+2} + U_{t+2} - Z_{t+2}$ and $\mathbb{E}\left[Z_t \xi_{t+2}'\right] = 0$, and the substitutions are repeated until infinity. We have shown the result as claimed. ■

Proof of Theorem 16.8. By the definition of the stochastic integral

$$\int_0^1 W dW = \operatorname*{plim}_{N\to\infty} \sum_{i=0}^{N-1} W\left(\frac{i}{N}\right)\left(W\left(\frac{i+1}{N}\right) - W\left(\frac{i}{N}\right)\right). \tag{16.26}$$

Take any positive integer N and any $j < N$. Observe that

$$W\left(\frac{j+1}{N}\right) = W\left(\frac{j}{N}\right) + \left(W\left(\frac{j+1}{N}\right) - W\left(\frac{j}{N}\right)\right).$$

Squaring we obtain

$$W\left(\frac{j+1}{N}\right)^2 - W\left(\frac{j}{N}\right)^2 = 2W\left(\frac{j}{N}\right)\left(W\left(\frac{j+1}{N}\right) - W\left(\frac{j}{N}\right)\right) + \frac{1}{N}\chi_{jN}.$$

where $\chi_{jN} = N\left(W\left(\frac{j+1}{N}\right) - W\left(\frac{j}{N}\right)\right)^2$. Notice that χ_{jN} are i.i.d. across j, distributed as χ_1^2, and have expectation 1. Summing over $j = 0$ to $N-1$ we obtain

$$W(1)^2 = 2\sum_{i=0}^{N-1} W\left(\frac{i}{N}\right)\left(W\left(\frac{i+1}{N}\right) - W\left(\frac{i}{N}\right)\right) + \frac{1}{N}\sum_{i=0}^{N-1}\chi_{iN}^2.$$

Rewriting

$$\sum_{i=0}^{N-1} W\left(\frac{i}{N}\right)\left(W\left(\frac{i+1}{N}\right) - W\left(\frac{i}{N}\right)\right) = \frac{1}{2}\left(W(1)^2 - \frac{1}{N}\sum_{i=0}^{N-1}\chi_{iN}^2\right).$$

By (16.26), $\int_0^1 W dW$ is the probability limit of the right side. By the WLLN this is $\frac{1}{2}\left(W(1)^2 - 1\right)$ as claimed. ■

Proof of Theorem 16.10.

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{t=1}^{n-1}\widehat{e}_{t+t}^2 = \frac{1}{n}\sum_{t=1}^{n-1} e_{t+t}^2 - \frac{1}{n}\frac{\left(\frac{1}{n}\sum_{t=1}^{n-1} Y_t e_{t+1}\right)^2}{\frac{1}{n^2}\sum_{t=1}^{n-1} Y_t^2} = \frac{1}{n}\sum_{t=1}^{n-1} e_{t+t}^2 + o_p(1) \xrightarrow[p]{} \sigma^2.$$

Then

$$T = \frac{\frac{1}{n}\sum_{t=1}^{n-1} Y_t e_{t+1}}{\left(\frac{1}{n^2}\sum_{t=1}^{n-1} Y_t^2\right)^{1/2}\widehat{\sigma}} \xrightarrow[d]{} \frac{\sigma^2\int_0^1 W dW}{\left(\sigma^2\int_0^1 W^2\right)^{1/2}\sigma} = \frac{\int_0^1 W dW}{\left(\int_0^1 W^2\right)^{1/2}}.$$

■

Proof of Theorem 16.12. Pick $\eta > 0$ and $\epsilon > 0$. Pick δ such that

$$\mathbb{P}\left(\sup_{|r-s|\le\delta} |X(r)-X(s)| > \epsilon\right) \le \eta \tag{16.27}$$

which is possible since $X(r)$ is almost surely continuous. Set $N = \lfloor 1/\delta \rfloor$ and $t_k = kn/N$. Write $X_{nt} = D_n^{-1}X_t$. Then

$$C_n = \frac{1}{n}\sum_{k=0}^{N}\sum_{t=t_k}^{t_{k+1}-1} X_{nt}u_t = \frac{1}{n}\sum_{k=0}^{N} X_{n,t_k}\sum_{t=t_k}^{t_{k+1}-1} u_t + \frac{1}{n}\sum_{k=0}^{N}\sum_{t=t_k}^{t_{k+1}-1}\left(X_{nt}-X_{n,t_k}\right)u_t$$

and

$$|C_n| \le \sup_{0\le r\le 1}|X_n(r)|\,A_n + \sup_{|r-s|\le\delta}|X_n(r)-X_n(s)|\,B_n$$

where

$$A_n = \frac{N}{n}\max_{k\le N}\left|\sum_{t=t_k}^{t_{k+1}-1} u_t\right|$$

$$B_n = \frac{1}{n}\sum_{t=1}^{n}|u_t|.$$

Since $X_n \xrightarrow[d]{} X$ and X is continuous,

$$\sup_{0\le r\le 1}|X_n(r)| \xrightarrow[d]{} \sup_{0\le r\le 1}|X(r)| < \infty$$

almost surely. Thus $\sup_{0\le r\le 1}|X_n(r)| = O_p(1)$. Since $X_n \xrightarrow[d]{} X$,

$$\sup_{|r-s|\le\delta}|X_n(r)-X_n(s)| \xrightarrow[d]{} \sup_{|r-s|\le\delta}|X(r)-X(s)| \le \epsilon$$

where the inequality holds with probability exceeding $1-\eta$ by (16.27). Thus for sufficiently large n the left hand side is bounded by 2ϵ with the same probability, and hence is $o_p(1)$.

For fixed N, $A_n \xrightarrow[p]{} 0$ by the ergodic theorem. The assumption that $\mathbb{E}|u_t| < \infty$ implies that $B_n = O_p(1)$. Together, we have shown that

$$|C_n| \le O_p(1)o_p(1) + o_p(1)O_p(1) = o_p(1)$$

as stated. ■

Proof of Theorem 16.17.

Part 1: The definition of cointegration implies that ΔY_t is stationary with a finite covariance matrix. By the multivariate Wold representation (Theorem 15.2), $\Delta Y_t = \theta + \Theta(\mathrm{L})e_t$ with the errors white noise. Premultiplication by β' yields $\beta'\Delta Y_t = \beta'\theta + \beta'\Theta(\mathrm{L})e_t$ which has long-run variance $\beta'\Theta(1)\Sigma\Theta(1)'\beta$ where Σ is the covariance matrix of e_t. The assumption that $\beta' Y_t$ is $I(0)$ implies that $\beta'\theta = 0$ (else $\beta' Y_t$ will have a time trend). This implies θ lies in the range space of $\beta_\perp$, hence $\theta = \beta_\perp\gamma$ for some γ. Also, the assumption that $\beta' Y_t$ is $I(0)$ implies that $\beta'\Delta Y_t$ is $I(-1)$, which implies that its long-run covariance matrix equals zero. This implies that $\beta'\Theta(1) = 0$ and hence $\Theta(1) = \beta_\perp\eta'$ for some matrix η. The assumption that $\beta_\perp'\Delta Y_t$ is $I(0)$

implies that $\beta_\perp'\Theta(1)\Sigma\Theta(1)'\beta_\perp > 0$ which implies that $\Theta(1)$ must have rank $m-r$ and hence so does the matrix η.

Part 2: The Beveridge-Nelson decomposition plus $\Theta(1) = \beta_\perp\eta'$ implies $\Theta(\mathrm{L}) = \beta_\perp\eta' + \Theta^*(\mathrm{L})(1-\mathrm{L})$. Applied to the Wold representation we obtain $\Delta Y_t = \beta_\perp\gamma + \beta_\perp\eta' e_t + \Theta^*(\mathrm{L})\Delta e_t$. Summing we find the stated representation.

Part 3: Without loss of generality assume that $H = [\beta, \beta_\perp]$ is orthonormal. Also define the orthonormal matrix $H_\eta = [\eta_\perp, \overline{\eta}]$ where $\overline{\eta} = \eta(\eta'\eta)^{-1/2}$. Define $X_t = H'Y_t$. The Wold representation implies $\Delta X_t = \begin{pmatrix} 0 \\ \gamma \end{pmatrix} + \boldsymbol{C}(\mathrm{L})e_t$ where using the Beveridge-Nelson decomposition

$$\boldsymbol{C}(\mathrm{L}) = H'\left(\beta_\perp\eta' + \Theta^*(\mathrm{L})(1-\mathrm{L})\right) = \begin{pmatrix} \beta'\Theta^*(\mathrm{L})(1-\mathrm{L}) \\ \eta' + \beta_\perp'\Theta^*(\mathrm{L})(1-\mathrm{L}) \end{pmatrix}.$$

Partition $X_t = (X_{1t}, X_{2t})$ comformably with H. We see that

$$\begin{pmatrix} \Delta X_{1t} \\ \Delta X_{2t} \end{pmatrix} = \begin{pmatrix} \beta'\Theta^*(\mathrm{L})(1-\mathrm{L})e_t \\ \gamma + \eta' e_t + \beta_\perp'\Theta^*(\mathrm{L})(1-\mathrm{L})e_t \end{pmatrix}.$$

Summing the first equation we obtain

$$\begin{pmatrix} X_{1t} \\ \Delta X_{2t} \end{pmatrix} = \begin{pmatrix} \mu \\ \gamma \end{pmatrix} + \boldsymbol{D}(\mathrm{L})H_\eta' e_t \tag{16.28}$$

where $\mu = X_{1,0} - \beta'\Theta^*(\mathrm{L})e_0$ and

$$\boldsymbol{D}(\mathrm{L}) = \begin{pmatrix} \beta'\Theta^*(\mathrm{L}) \\ \eta' + \beta_\perp'\Theta^*(\mathrm{L})(1-\mathrm{L}) \end{pmatrix} H_\eta = \begin{pmatrix} \beta'\Theta^*(\mathrm{L})\eta_\perp & \beta'\Theta^*(\mathrm{L})\overline{\eta} \\ \beta_\perp'\Theta^*(\mathrm{L})\eta_\perp(1-\mathrm{L}) & \left(\eta'\eta\right)^{1/2} + \beta_\perp'\Theta^*(\mathrm{L})\overline{\eta}(1-\mathrm{L}) \end{pmatrix}.$$

This is an invertible matrix polynomial. To see this, first observe that

$$\boldsymbol{D}(1) = \begin{pmatrix} \beta'\Theta^*(1)\eta_\perp & \beta'\Theta^*(1)\overline{\eta} \\ 0 & \left(\eta'\eta\right)^{1/2} \end{pmatrix}$$

which is full rank under the assumption that $\beta'\Theta^*(1)\eta_\perp$ is full rank. This means that $\det(\boldsymbol{D}(z))$ has no unit roots. Second, (16.28) and the definition of X_t imply that

$$\boldsymbol{D}(z) = \begin{pmatrix} 1-z & 0 \\ 0 & 1 \end{pmatrix} H\Theta(z)H_\eta.$$

Since H and H_η are full rank this implies that the solutions to $\det(\boldsymbol{D}(z)) = 0$ are solutions to $\det(\Theta(z)) = 0$ and hence satisfy $|z| \geq 1+\delta$ (because $z \neq 1$) by the assumption on $\Theta(z)$. Together we have shown that $\boldsymbol{D}(\mathrm{L})$ is invertible. Thus (16.28) implies

$$H_\eta\boldsymbol{D}(\mathrm{L})^{-1}\begin{pmatrix} X_{1t} \\ \Delta X_{2t} \end{pmatrix} = a + e_t \tag{16.29}$$

where

$$a = H_\eta\boldsymbol{D}(1)^{-1}\begin{pmatrix} \mu \\ \gamma \end{pmatrix}.$$

(16.29) is a VAR representation for $\left(\beta' Y_t, \beta_\perp'\Delta Y_t\right)$ with all roots satisfying $|z| \geq 1+\delta$. This implies a VAR representation for Y_t which is equation (16.18) with

$$\boldsymbol{A}(z) = H_\eta\boldsymbol{D}(z)^{-1}\begin{pmatrix} \beta' \\ \beta_\perp'(1-z) \end{pmatrix}.$$

By partitioned matrix inversion we calculate

$$\begin{aligned}
\boldsymbol{A}(1) &= H_\eta \boldsymbol{D}(1)^{-1}\begin{pmatrix}\beta' \\ 0\end{pmatrix} \\
&= [\eta_\perp, \overline{\eta}]\begin{pmatrix}\left(\beta'\Theta^*(1)\eta_\perp\right)^{-1} & -\left(\beta'\Theta^*(1)\eta_\perp\right)^{-1}\beta'\Theta^*(1)\eta \\ 0 & \left(\eta'\eta\right)^{-1/2}\end{pmatrix}\begin{pmatrix}\beta' \\ 0\end{pmatrix} \\
&= \eta_\perp\left(\beta'\Theta^*(1)\eta_\perp\right)^{-1}\beta' \\
&= -\alpha\beta'.
\end{aligned}$$

as claimed.

Part 4. Under the assumption $\sum_{j=0}^{\infty}\left\|\sum_{k=0}^{\infty}k\Theta_{j+k}\right\|^2<\infty$, Theorem 15.3 implies that the coefficients $\boldsymbol{A}_k^*=\sum_{j=0}^{\infty}\boldsymbol{A}_{j+k}$ are absolutely summable. We can then apply the Beveridge-Nelson decomposition $\boldsymbol{A}(z)=\boldsymbol{A}(1)+\boldsymbol{A}^*(z)(1-z)$. Applying $\boldsymbol{A}(1)=-\alpha\beta'$ and a little rewriting yields

$$\boldsymbol{A}(z)=\boldsymbol{I}_m(1-z)-\alpha\beta' z-\left(\boldsymbol{I}_m+\alpha\beta'-\boldsymbol{A}^*(z)\right)(1-z).$$

Applied to (16.18) we obtain the stated result with $\Gamma(\mathrm{L})=\boldsymbol{I}_m+\alpha\beta'-\boldsymbol{A}^*(z)$. The coefficients of $\Gamma(\mathrm{L})$ are absolutely summable because the coefficients $\boldsymbol{A}_k^*$ are.

Part 5. The assumption $\theta=0$ direct implies $\gamma=0$. This implies

$$\begin{aligned}
a &= H_\eta \boldsymbol{D}(1)^{-1}\begin{pmatrix}\mu \\ 0\end{pmatrix} \\
&= [\eta_\perp, \overline{\eta}]\begin{pmatrix}\left(\beta'\Theta^*(1)\eta_\perp\right)^{-1} & -\left(\beta'\Theta^*(1)\eta_\perp\right)^{-1}\beta'\Theta^*(1)\eta \\ 0 & \left(\eta'\eta\right)^{-1/2}\end{pmatrix}\begin{pmatrix}\mu \\ 0\end{pmatrix} \\
&= \eta_\perp\left(\beta'\Theta^*(1)\eta_\perp\right)^{-1}\mu \\
&= \alpha\mu
\end{aligned}$$

as claimed. ■

Proof of Theorem 16.18. Write the VECM as $\Gamma^*(\mathrm{L})\Delta Y_t-\alpha\beta' Y_{t-1}=a+e_t$ where $\Gamma^*(z)=\boldsymbol{I}_m-\Gamma(z)$. Set $\overline{\alpha}=\alpha\left(\alpha'\alpha\right)^{-1/2}$ and orthonormal $H=\left[\overline{\alpha},\alpha_\perp\right]$. Assume that $\left[\beta,\beta_\perp\right]$ is orthonormal. Define $Z_t=\beta' Y_t$ and $U_t=\beta_\perp'\Delta Y_t$. Our goal is to show that (Z_t,U_t) is $I(0)$ which is the same as showing that Y_t is cointegrated with cointegrating vectors β.

Premultiplying the VECM model by H' we find the system

$$H'\left(\Gamma^*(\mathrm{L})\Delta Y_t-\alpha\beta' Y_{t-1}\right)=H'a+H'e_t.$$

Using the identity $\boldsymbol{I}_m=\beta\beta'+\beta_\perp\beta_\perp'$ we see that $\Delta Y_t=\beta\Delta Z_t+\beta_\perp U_t$. Making this substitution and setting $\overline{a}=H'a$ $v_t=H'e_t$ we obtain the system

$$\boldsymbol{D}(\mathrm{L})\begin{pmatrix}Z_t \\ U_t\end{pmatrix}=\overline{a}+v_t$$

where

$$\boldsymbol{D}(z)=\begin{bmatrix}\overline{\alpha}'\Gamma^*(z)\beta(1-z)-\boldsymbol{I}_m & \overline{\alpha}'\Gamma^*(z)\beta_\perp \\ \alpha_\perp'\Gamma^*(z)\beta(1-z) & \alpha_\perp'\Gamma^*(z)\beta_\perp\end{bmatrix}.$$

We now show that this is a stationary system. First, note that

$$\boldsymbol{D}(1)=\begin{bmatrix}-\boldsymbol{I}_m & \overline{\alpha}'\Gamma^*(1)\beta_\perp \\ 0 & \alpha_\perp'\Gamma^*(1)\beta_\perp\end{bmatrix}$$

which is full rank under the assumption that $\alpha_{\perp}'\Gamma^*(1)\beta_{\perp}$ is full rank. That means that $\det(\boldsymbol{D}(z)) = 0$ has no solutions $z = 1$. Second, $\boldsymbol{D}(z)$ relates to $\boldsymbol{A}(z)$ by the relationship

$$\boldsymbol{D}(z) = H'\boldsymbol{A}(z)\left[\beta, \beta_{\perp}(1-z)\right].$$

Thus the solutions $z \neq 1$ to

$$\det(\boldsymbol{D}(z)) = \det(H)\det(\boldsymbol{A}(z))\det\left(\left[\beta, \beta_{\perp}(1-z)\right]\right) = 0$$

are all solutions to $\det(\boldsymbol{A}(z)) = 0$, which all satisfy $|z| \geq 1+\delta$ by assumption. Thus $\boldsymbol{D}(z)$ is invertible with summable moving average coefficient matrices. This implies the VAR system for (Z_t, U_t) is stationary.

As discussed above, this shows that (Z_t, U_t) is a stationary process and hence Y_t is cointegrated with cointegrating vector β. ■

Proof of Theorem 16.19. Set $Y_{2t}^* = Y_{2t} - \overline{Y}_2$. The estimator satisfies

$$n\left(\widehat{\beta} - \beta\right) = \left(\frac{1}{n^2}\sum_{t=1}^{n} Y_{2t}^* Y_{2t}^{*\prime}\right)^{-1}\left(\frac{1}{n}\sum_{t=1}^{n} Y_{2t}^* u_{1t}\right).$$

Set $S_t = \sum_{i=1}^{t} u_t$. Theorems 16.4 and 16.5 imply $S_{\lfloor nr \rfloor} \xrightarrow[d]{} B(r)$ and $Y_{2\lfloor nr \rfloor}^* \xrightarrow[d]{} B_2^*(r)$. By the continuous mapping theorem

$$\frac{1}{n^2}\sum_{t=1}^{n} Y_{2t}^* Y_{2t}^{*\prime} \xrightarrow[d]{} \int_0^1 B_2^* B_2^{*\prime}.$$

By Theorem 16.7 and the WLLN

$$\frac{1}{n}\sum_{t=1}^{n} Y_{2t}^* u_{1t} = \frac{1}{n}\sum_{t=1}^{n} Y_{2t-1}^* u_{1t} + \frac{1}{n}\sum_{t=1}^{n} u_{2t}u_{1t} + o_p(1) \xrightarrow[d]{} \int_0^1 B_2^* dB_1 + \Lambda_{21} + \Sigma_{21}.$$

Together we obtain the stated result. ■

Proof of Theorem 16.22 (sketch). For simplicity abstract from the dynamic and trend coefficients so that the unconstrained model is

$$\Delta Y_t = \alpha\beta' Y_{t-1} + e_t.$$

where e_t is a MDS with covariance matrix Σ. We examine two cases in detail. First, the case $\mathbb{H}(0)$ (which is relatively straightforward) and second the case $\mathbb{H}(r)$ (which is algebraically more tedious).

First, take $\mathbb{H}(0)$, in which case the process is $\Delta Y_t = e_t$. The statistic is

$$\begin{aligned}
\mathrm{LR}(0) &= -n\sum_{j=1}^{m}\log\left(1-\widehat{\lambda}_j\right) \simeq n\sum_{j=1}^{m}\widehat{\lambda}_j \\
&= \operatorname{tr}\left[\left(\frac{1}{n}\sum_{t=1}^{n} Y_{t-1}e_t'\right)\left(\frac{1}{n}\sum_{t=1}^{n} e_t e_t'\right)^{-1}\left(\frac{1}{n}\sum_{t=1}^{n} e_t Y_{t-1}'\right)\left(\frac{1}{n^2}\sum_{t=1}^{n} Y_{t-1}Y_{t-1}'\right)^{-1}\right] \\
&\xrightarrow[d]{} \operatorname{tr}\left[\left(\int_0^1 dBB'\right)\left(\int_0^1 BB'\right)^{-1}\left(\int_0^1 BdB'\right)\right] \\
&= \operatorname{tr}\left[\left(\int_0^1 dWW'\right)\left(\int_0^1 WW'\right)^{-1}\left(\int_0^1 WdW'\right)\right]
\end{aligned}$$

where $B(r)$ is a Brownian motion with covariance matrix Σ, and $W(r) = \Sigma^{-1/2}B(r)$ is standard Brownian motion. This is the stated result.

Second, take $\mathbb{H}(r)$ for $1 < r < m$. Define $Z_t = \beta' Y_t$. The process under $\mathbb{H}(r)$ is $\Delta Y_t = \alpha Z_{t-1} + e_t$. Normalize β so that $\mathbb{E}\left[Z_t Z_t'\right] = \boldsymbol{I}_r$. The test statistic is invariant to linear transformations of Y_t so we can rescale the data so that $\mathbb{E}\left[\Delta Y_t \Delta Y_t'\right] = \boldsymbol{I}_m$. Notice that $\Sigma = \mathbb{E}\left[e_t e_t'\right] = \mathbb{E}\left[\Delta Y_t \Delta Y_t'\right] - \alpha \mathbb{E}\left[Z_t Z_t'\right]\alpha' = \boldsymbol{I}_m - \alpha\alpha'$.

The likelihood ratio statistic is

$$\mathrm{LR}(r) = -n \sum_{j=r+1}^{m} \log\left(1 - \widehat{\lambda}_j\right) \simeq \sum_{j=r+1}^{m} \widehat{\rho}_j$$

where $\widehat{\rho}_j = n\widehat{\lambda}_j$ are the $m-r$ smallest roots of the equation $\det\left(S(\rho)\right) = 0$ where

$$S(\rho) = \rho \frac{1}{n^2} \sum_{t=1}^{n} Y_{t-1} Y_{t-1}' - \frac{1}{n} \sum_{t=1}^{n} Y_{t-1} \Delta Y_t' \left(\frac{1}{n} \sum_{t=1}^{n} \Delta Y_t \Delta Y_t'\right)^{-1} \frac{1}{n} \sum_{t=1}^{n} \Delta Y_t Y_{t-1}'.$$

Define a full-rank matrix $H = [\beta, \beta_\perp]$ where $\beta'\beta_\perp = 0$. The roots of $\widehat{\rho}_j$ are the same as those of $\det\left(S^*(\rho)\right) = 0$ where $S^*(\rho) = H'S(\rho)H$, which replaces Y_{t-1} with (Z_{t-1}, X_{t-1}) where $X_t = \beta_\perp' Y_t$. We calculate that

$$\begin{aligned} S^*(\rho) = \rho &\begin{bmatrix} \frac{1}{n^2}\sum_{t=1}^n Z_{t-1}Z_{t-1}' & \frac{1}{n^2}\sum_{t=1}^n Z_{t-1}X_{t-1}' \\ \frac{1}{n^2}\sum_{t=1}^n X_{t-1}Z_{t-1}' & \frac{1}{n^2}\sum_{t=1}^n X_{t-1}X_{t-1}' \end{bmatrix} \\ &- \begin{bmatrix} \frac{1}{n}\sum_{t=1}^n Z_{t-1}\Delta Y_t' \\ \frac{1}{n}\sum_{t=1}^n X_{t-1}\Delta Y_t' \end{bmatrix} \left(\frac{1}{n}\sum_{t=1}^n \Delta Y_t \Delta Y_t'\right)^{-1} \begin{bmatrix} \frac{1}{n}\sum_{t=1}^n Z_{t-1}\Delta Y_t' \\ \frac{1}{n}\sum_{t=1}^n X_{t-1}\Delta Y_t' \end{bmatrix}'. \end{aligned}$$

We now apply the asymptotic theory for non-stationary theory to each component. The process $X_t = \beta_\perp' Y_t$ is non-stationary and satisfies the FCLT $n^{-1}X_{\lfloor nr \rfloor} \xrightarrow[d]{} X(r) \sim BM\left(\beta_\perp'\Omega\beta_\perp\right)$ where Ω is the long-run covariance matrix of ΔY_t. The sum of the errors satisfy $n^{-1/2}\sum_{t=1}^{\lfloor nr \rfloor} e_t \xrightarrow[d]{} B(r) \sim BM(\Sigma)$. The process $X(r)$ is a linear function of $B(r)$.

We find that $\frac{1}{n^2}\sum_{t=1}^n X_{t-1}X_{t-1}' \xrightarrow[d]{} \int_0^1 XX'$, $\frac{1}{n^2}\sum_{t=1}^n X_{t-1}e_t \xrightarrow[d]{} \int_0^1 XdB'$, $\frac{1}{n}\sum_{t=1}^n Z_{t-1}Z_{t-1}' \xrightarrow[p]{} \boldsymbol{I}_r$, $\frac{1}{n}\sum_{t=1}^n \Delta Y_t \Delta Y_t' \xrightarrow[p]{} \boldsymbol{I}_m$, $\frac{1}{n}\sum_{t=1}^n Z_{t-1}\Delta Y_t' \xrightarrow[p]{} \alpha'$, $\frac{1}{n}\sum_{t=1}^n X_{t-1}Z_{t-1}' \xrightarrow[d]{} \zeta$ for some random matrix by Theorem 16.7, and $\frac{1}{n}\sum_{t=1}^n X_{t-1}\Delta Y_t' \xrightarrow[d]{} \zeta'\alpha' + \int_0^1 XdB'$. Together we find that

$$S^*(\rho) \xrightarrow[d]{} \rho \begin{bmatrix} 0 & \\ 0 & \int_0^1 XX' \end{bmatrix} - \begin{bmatrix} \alpha'\alpha & \alpha'\left(\alpha\zeta + \int_0^1 dBX'\right) \\ \left(\zeta'\alpha' + \int_0^1 XdB'\right)\alpha & \left(\zeta'\alpha' + \int_0^1 XdB'\right)\left(\alpha\zeta + \int_0^1 dBX'\right) \end{bmatrix}.$$

Thus $\det\left(S^*(\rho)\right)$ converges in distribution to the determinant of the right-hand-side, which equals (using Theorem A.1.5) $\det\left(\alpha'\alpha\right)$ multiplied by the determinant of

$$\begin{aligned} &\rho \int_0^1 XX' - \left(\zeta'\alpha' + \int_0^1 XdB'\right)\left(\boldsymbol{I}_m - \alpha\left(\alpha'\alpha\right)^{-1}\alpha'\right)\left(\alpha\zeta + \int_0^1 dBX'\right) \\ &= \rho \int_0^1 XX' - \int_0^1 XdB' M_\alpha \int_0^1 dBX' \\ &= \rho \int_0^1 XX' - \int_0^1 XdW'H_1' \int_0^1 H_1 dWX' \\ &= \rho \int_0^1 XX' - \int_0^1 XdW' \int_0^1 dWX' \qquad (16.30) \end{aligned}$$

where $M_\alpha = \boldsymbol{I}_m - \alpha\left(\alpha'\alpha\right)^{-1}\alpha'$ and

$$M_\alpha B(r) \sim BM\left(M_\alpha\left(\boldsymbol{I}_m - \alpha\alpha'\right)M_\alpha\right) = BM\left(M_\alpha\right) = H_1 W(r)$$

where $M_\alpha = H_1 H_1'$, $H_1' H_1 = \boldsymbol{I}_{m-r}$ and $W(r) \sim BM(\boldsymbol{I}_{m-r})$.

The determinant of (16.30) has $m-r$ roots and their sum equals

$$\operatorname{tr}\left[\left(\int_0^1 dWX'\right)\left(\int_0^1 XX'\right)^{-1}\left(\int_0^1 XdW'\right)\right] = \operatorname{tr}\left[\left(\int_0^1 dWW'\right)\left(\int_0^1 WW'\right)^{-1}\left(\int_0^1 WdW'\right)\right]$$

because $X(r)$ is a linear rotation of $W(r)$. This is the stated result. ■

16.23 Exercises

Exercise 16.1 Take $S_t = S_{t-1} + e_t$ with $S_0 = 0$ and e_t i.i.d. $(0, \sigma^2)$.

(a) Calculate $\mathbb{E}[S_t]$ and $\operatorname{var}[S_t]$.

(b) Set $Y_t = (S_t - \mathbb{E}[S_t]) / \sqrt{\operatorname{var}[S_t]}$. By construction $\mathbb{E}[Y_t] = 0$ and $\operatorname{var}[Y_t] = 1$. Is Y_t stationary?

(c) Find the asymptotic distribution of $Y_{\lfloor nr \rfloor}$ for $r \in [\delta, 1]$.

Exercise 16.2 Find the Beveridge-Nelson decomposition of $\Delta Y_t = e_t + \Theta_1 e_{t-1} + \Theta_2 e_{t-2}$.

Exercise 16.3 Suppose $Y_t = X_t + u_t$ where $X_t = X_{t-1} + e_t$ with $(e_t, u_t) \sim I(0)$.

(a) Is Y_t $I(0)$ or $I(1)$?.

(b) Find the asymptotic functional distribution of $\mathrm{n}^{-1/2} Y_{\lfloor nr \rfloor}$.

Exercise 16.4 Let $Y_t = e_t$ be i.i.d. and $X_t = \Delta Y_t$.

(a) Show that Y_t is stationary and $I(0)$.

(b) Show that X_t is stationary but not $I(0)$.

Exercise 16.5 Let $U_t = U_{t-1} + e_t$, $Y_t = U_t + v_t$ and $X_t = 2U_t + w_t$, where (e_t, v_t, w_t) is an i.i.d. sequence. Find the cointegrating vector for (Y_t, X_t).

Exercise 16.6 Take the AR(1) model $Y_t = \alpha Y_{t-1} + e_t$ with i.i.d. e_t and the least squares estimator $\widehat{\alpha}$. In Chaper 14 we learned that the asymptotic distribution when $|\alpha| < 1$ is $\sqrt{n}(\widehat{\alpha} - \alpha) \xrightarrow[d]{} \mathrm{N}\left(0, 1 - \alpha^2\right)$. How do you reconcile this with Theorem 16.9, especially for α close to one?

Exercise 16.7 Take the VECM(1) model $\Delta Y_t = \alpha\beta' Y_{t-1} + e_t$. Show that $Z_t = \beta' Y_t$ follows an AR(1) process.

Exercise 16.8 An economist estimates the model $Y_t = \alpha Y_{t-1} + e_t$ and finds $\widehat{\alpha} = 0.9$ with $s(\widehat{\alpha}) = 0.05$. They assert: “The t-statistic for testing $\alpha = 1$ is 2, so $\alpha = 1$ is rejected.” Is there an error in their reasoning?

Exercise 16.9 An economist estimates the model $Y_t = \alpha Y_{t-1} + e_t$ and finds $\widehat{\alpha} = 0.9$ with $s(\widehat{\alpha}) = 0.04$. They assert: “The 95% confidence interval for α is $[0.82, 0.98]$ which does not contain 1. So $\alpha = 1$ is not consistent with the data.” Is there an error in their reasoning?

Exercise 16.10 An economist takes Y_t, detrends to obtain the detrended series Z_t, applies a ADF test to Z_t and finds ADF $= -2.5$. They assert: “Stata provides the 5% critical value -1.9 with p-value less than 1%. Thus we reject the null hypothesis of a unit root.” Is there an error in their reasoning?

Exercise 16.11 An economist wants to build an autoregressive model for the number of daily tweets by a prominant politician. For a model with an intercept they obtain ADF = −2.0. They assert "The number of tweets is a unit root process." Is there an error in their reasoning?

Exercise 16.12 For each of the following monthly series from `FRED-MD` implement the Dickey-Fuller unit root test. For each, you need to consider the AR order p and the trend specification.

(a) log real personal income: log(*rpi*)

(b) industrial production index: *indpro*

(c) housing starts: *houst*

(d) help-wanted index: *hwi*

(e) civilian labor force: *clf16ov*

(f) initial claims: *claims*

(g) industrial production index (fuels): *ipfuels*

Exercise 16.13 For each of the series in the previous exercise implement the KPSS test of stationarity. For each, you need consider the lag truncation M and the trend specification.

Exercise 16.14 For each of the following monthly pairs from `FRED-MD` test the hypothesis of no cointegration using the Johansen trace test. For each, you need to consider the VAR order p and the trend specification.

(a) 3-month treasury interest rate (*tb3ms*) and 10-year treasury interest rate (*gs10*). Note: In the text we implemented the test on the quarterly series, not monthly.

(b) interest rate on AAA bonds (*aaa*) and interest rate on BAA bonds (*baa*).

(c) log(industrial production durable consumer goods) and log(industrial production nondurable consumer goods) (log of *ipdcongd* and *ipncongd*).