Peter Reinhard Hansen
Department of Economics, Box B
Brown University
64 Waterman Street
Providence, RI 02912
401-863-9864 (phone)
401-863-1970 (fax)
Peter_Hansen@brown.edu
Word count: Approximately 1280 (not including figure caption and math)

Cointegration. When a linear combination of nonstationary variables is stationary, the variables are said to be cointegrated and the vector that defines the stationary linear combination is called a cointegration vector. A time-series is stationary if its distribution does not vary over time. The simples example of a stationary process is $\{\varepsilon_t\} = \ldots, \varepsilon_{-1}, \varepsilon_0, \varepsilon_1, \ldots$, which represents a sequence of independent and identically distributed random variables. The subscript, t, refers to time. If the distribution of a variable depends on t, it is nonstationary. In cointegration analysis the most common form of nonstationarity is that of the integrated variables. The random walk, $X_t = X_{t-1} + \varepsilon_t = X_0 + \sum_{i=1}^{t} \varepsilon_i$, is an example of a nonstationary variable that is integrated of order one. The word integration refers to the cumulation of epsilons.

The concept of cointegration is interesting because it can be applied to uncover relationships that may have theoretical interpretations. For example, the cointegration relations may be defined by the first order conditions of an economic model, and cointegration analysis can be used to estimate economic models, and test certain theoretical hypotheses.

The classical example on cointegration is about a dog that follows its drunk owner. In this example, the positions of the dog and its owner, as a function of time, are two nonstationary processes, because the drunk owner is walking around, taking steps in random directions. However, the two processes are cointegrated because the distance

between dog and owner is stationary.

Another example is a property of interest rates on bonds with different maturities. Individually, they may have large fluctuations, but the difference between them appears to be stationary. This is illustrated in Figure 1, which contains the interest rates on a 3-month Treasury bill and a 1-year Treasury bond for the period January 1970 to April 2002. The thin solid line is the difference between the two interest rates.

$*****$ Figure 1 about here. $*****$

HISTORICAL DEVELOPMENT

Cointegration was introduced by Clive W. J. Granger (1981, 1983), and the statistical analysis of cointegrated processes was formalized by Robert F. Engle and Granger (1987), Søren Johansen (1988, 1991), and Peter C. B. Phillips (1991). Cointegration is related to many concepts of econometrics, such as unit root processes, spurious regression, and common stochastic trends, see Mark W. Watson (1994) for an excellent review. Today there is a voluminous literature on cointegration and applications of cointegration. Recent developments have been on: improving and generalizing existing techniques, such as bias and Bartlett corrections; fractionally cointegrated processes; seasonal cointegration; panel cointegration; nonlinear cointegration; and cointegration in relation to processes with structural changes. Many references can be found the book of James D. Hamilton (1994), which also contains a good introduction to cointegration. Excellent textbooks on likelihood analysis of cointegration are Johansen (1995) and the companion book by Peter R. Hansen and Johansen (1998).

ORDER OF INTEGRATION

For variables to be cointegrated, they must individually be integrated. Formally, a time-series, X_t, is said to be integrated of order d if $(1-L)^d X_t = \psi(L)\varepsilon_t$ is stationary and $\psi(1) \equiv \sum_i \psi_i \neq 0$, where L is the lag-operator such that $\psi(L)\varepsilon_t = \psi_0\varepsilon_t + \psi_1\varepsilon_{t-1} + \cdots$. The notation $X_t \sim I(d)$ is short for "X_t is integrated of order d", and the notation $\Delta^d X_t$ is sometimes used in place of $(1-L)^d X_t$. An alternative definition of the order of integration assumes that $\{\varepsilon_t\}$ is a white noise process and substitutes "covariance-stationary" for "stationary". A white noise process is defined by having expected value zero, $E(\varepsilon_t) = 0$, having finite variance that does not depend on time, $\mathrm{var}(\varepsilon_t) = \sigma^2$, and being uncorrelated across time, $\mathrm{cov}(\varepsilon_s, \varepsilon_t) = 0$ for $s \neq t$. A covariance-stationary process, $\{X_t\}$, has constant mean and variance, and an autocovariance function that only depends on the difference in time, $\mathrm{cov}(X_s, X_t) = \phi(|s-t|)$.

The order of integration, d, can be any real number, but most of the literature is concerned with the unit root processes, which corresponds to $d \in \mathbb{N}$ (the positive integers). The non-integers are referred to as fractionally integrated processes which is related to the long-memory processes.

The order of integration of a vector of variables, $X_t = (X_{1t}, \ldots, X_{pt})'$ is defined as the highest order of integration of the individual elements, X_{1t}, ..., X_{pt}.

If $X_t \sim I(d_x)$, if $Y_t \sim I(d_y)$, and if a and b are non-zero constants, then $aX_t + bY_t$ is, in general, integrated of order $d = \max(d_x, d_y)$. It is the special case where $aX_t + bY_t$ is integrated of order less than $\max(d_x, d_y)$ that characterize cointegration. For two variables to be cointegrated they must be integrated of the same order. The notation $CI(d, b)$ is used for variables that are integrated of order d with a cointegration relation, which is integrated of order $d - b$.

A more general concept of cointegration is multi-cointegration, which is sometimes

called polynomial cointegration. If $(1-L)^d X_t$ and Y_t are cointegrated for some $d \neq 0$, then X_t and Y_t are multi-cointegrated. For example, $X_t \sim I(2)$ and $Y_t \sim I(1)$, then $\Delta X_t = X_t - X_{t-1}$ may cointegrate with Y_t.

EXAMPLE

Let $\{\varepsilon_t\}$ be a sequence of independent and identically distributed variables, and suppose that for $t = 1, 2, \ldots$

$$\begin{aligned} X_t &= \sum_{s=1}^{t}\sum_{i=1}^{s}\varepsilon_i + \sum_{s=1}^{t}\varepsilon_s + \varepsilon_t, \\ Y_t &= 2\sum_{s=1}^{t}\sum_{i=1}^{s}\varepsilon_i + \varepsilon_t, \\ Z_t &= -\sum_{s=1}^{t}\varepsilon_s + \varepsilon_t. \end{aligned}$$

Since $(1-L)^2 X_t = 3\varepsilon_t - 3\varepsilon_{t-1} + \varepsilon_{t-2}$ is stationary and the coefficients satisfy $\sum_i \psi_i = 3 - 3 + 1 = 1 \neq 0$, we see that X_t is integrated of order two. Similarly, it follows that $Y_t \sim I(2)$ and that $Z_t \sim I(1)$, so the vector $(X_t, Y_t, Z_t)'$ is integrated of order two. Further, $2X_t - Y_t = 2\sum_{s=1}^{t}\varepsilon_s + \varepsilon_t$ is integrated of order one, so X_t and Y_t are cointegrated, $CI(2,1)$, with cointegration vector $(2,-1)'$. Similarly, we see that $2X_t - Y_t + 2Z_t = 3\varepsilon_t$ is integrated of order zero, so the three variables X_t, Y_t, and Z_t are cointegrated, $CI(2,2)$, with cointegration vector $(2,-1,2)'$. Also, X_t and Z_t are multi-cointegrated, because the two $I(1)$ variables, ΔX_t and Z_t, are cointegrated, $CI(1,1)$, since $\Delta X_t + Z_t = 3\varepsilon_t - \varepsilon_{t-1}$ is $I(0)$. Similarly it can be seen that Y_t and Z_t are multi-cointegrated.

ANALYSIS OF COINTEGRATED PROCESSES

The classical assumptions in the linear regression model are violated in the presence

of integrated variables, mainly because of the nonstationarity. This can causes the asymptotic normality of parameter estimators to fail, and the statistical analysis of cointegrated variables is therefore more complicated than standard regression analysis.

The two main approaches to cointegration analysis are the regression based approach and the likelihood approach. The former involves standard regression techniques with various modifications, and the latter, which is based on the vector autoregressive model (VAR), is sometimes called the Johansen method.

A cointegration analysis of a system of variables will typically consists of

1. Determining the cointegration rank. This typically involves the use of a Dickey-Fuller type distribution. These distributions have complicated expressions that involve stochastic integrals of Brownian motions. Fortunately, most modern econometrics textbooks, including the book of Hamilton (1994), contains tables of the most common Dickey-Fuller distributions. The trace-test of Johansen (1988, 1991) is a popular test for determining the cointegration rank.

2. Estimating cointegration relations and other parameters. The cointegration relations are not identified without additional restrictions/normalizations.

3. Inference on the cointegration parameters and other parameters. The analysis on cointegration relations is particularly useful because they may be interpreted as structural equations in an economic model. Identified cointegration parameters have a non-normal limiting distribution and are super-consistent. The latter means that small samples can be very informative about the cointegration relations. Inference on other parameters follows traditional results in the standard case.

The likelihood analysis of Johansen (1988) is based on the vector autoregressive (VAR) model, $X_t = \Pi_1 X_{t-1} + \cdots + \Pi_k X_{t-k} + \varepsilon_t$, where X_t is a p-dimensional vector and where $\{\varepsilon_t\}$ is a sequence of independent normally distributed variables with mean 0 and covariance-matrix Ω. This model is rewritten in the form of the error correction model (ECM)

$$\Delta X_t = \Pi X_{t-1} + \Gamma_1 \Delta X_{t-1} + \cdots + \Gamma_{k-1} \Delta X_{t-k+1} + \varepsilon_t, \qquad t = 1, \ldots, T.$$

If the elements of X_t are cointegrated then the $p \times p$ matrix, Π, has reduced rank $r < p$ and the number of common stochastic trends in X_t is given by $p-r$. The reduced rank of Π can be made explicit by expressing the matrix as a product of two matrices, $\Pi = \alpha\beta'$, where α and β are $p \times r$ matrices. This formulation leads to a regression equation that can be estimated by reduced rank regression. An important instrument for the analysis of cointegrated processes is the Granger representation, $X_t = C\sum_{i=1}^{t} \varepsilon_i + C(L)\varepsilon_t + X_0 - C(L)\varepsilon_0$. This representation divides the process into a nonstationary random walk, $C\sum_{i=1}^{t} \varepsilon_i$, a stationary component, $C(L)\varepsilon_t$, and a term that depends on initial values, $X_0 - C(L)\varepsilon_0$. An important result is that $\beta' C = 0$, (an $r \times p$ matrix of zeros), from which it follows that $\beta' X_t = \beta' C(L) \sim I(0)$. So the cointegration relations are given by the columns of β. The relation, $\Delta X_t = \alpha(\beta' X_{t-1}) + \cdots$, shows that α defines how X_t responds on a deviation in $\beta' X_{t-1}$ from its expected value, and α is therefore called the matrix of adjustment coefficients. $\beta' X_{t-1}$ can sometimes be interpreted as the deviation from an equilibrium. In this case α informs about the variables adjustment towards the equilibrium relations, both the direction they take and how fast the variables converge to the equilibrium.

References

Engle, Robert F., and Clive W. J. Granger. (1987). Co-Integration and Error Correction: Representation, Estimation and Testing. *Econometrica*. Vol. 55, pp. 251–276.

Granger, Clive W. J. (1981). Some Properties of Time Series Data and their Use in Econometric Models Specification. *Journal of Econometrics*. Vol. 16, 121–130.

Granger, Clive W. J. (1983). Co-Integrated Variables and Error-Correcting Models. *University of California, San Diego, Discussion Paper*. 1983-13.

Hamilton, James D. (1994). *Time Series Analysis*. Princeton N.J.: Princeton University Press.

Hansen, Peter R. and Søren Johansen. (1998). *Workbook on Cointegration*. Oxford: Oxford University Press.

Johansen, Søren.(1988). Statistical Analysis of Cointegration Vectors. *Journal of Economic Dynamics and Control*. Vol. 12, pp. 231–254.

Johansen, Søren. (1991). Estimation and Hypotheses Testing of Cointegrating Vectors in Gaussian Vector Autoregressive Models. *Econometrica*. Vol. 59, pp. 1551–1580.

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

Phillips, Peter C. B. (1991). Optimal Inference in Cointegrated Systems. *Econometrica*. Vol. 59, pp. 283–306.

Watson, Mark W. (1994). Vector Autoregression and Cointegration, in *Handbook of Econometrics*, ed. by R.F. Engle and D. L. McFadden. Vol. IV, pp. 2843–2915. Amsterdam: Elsevier.

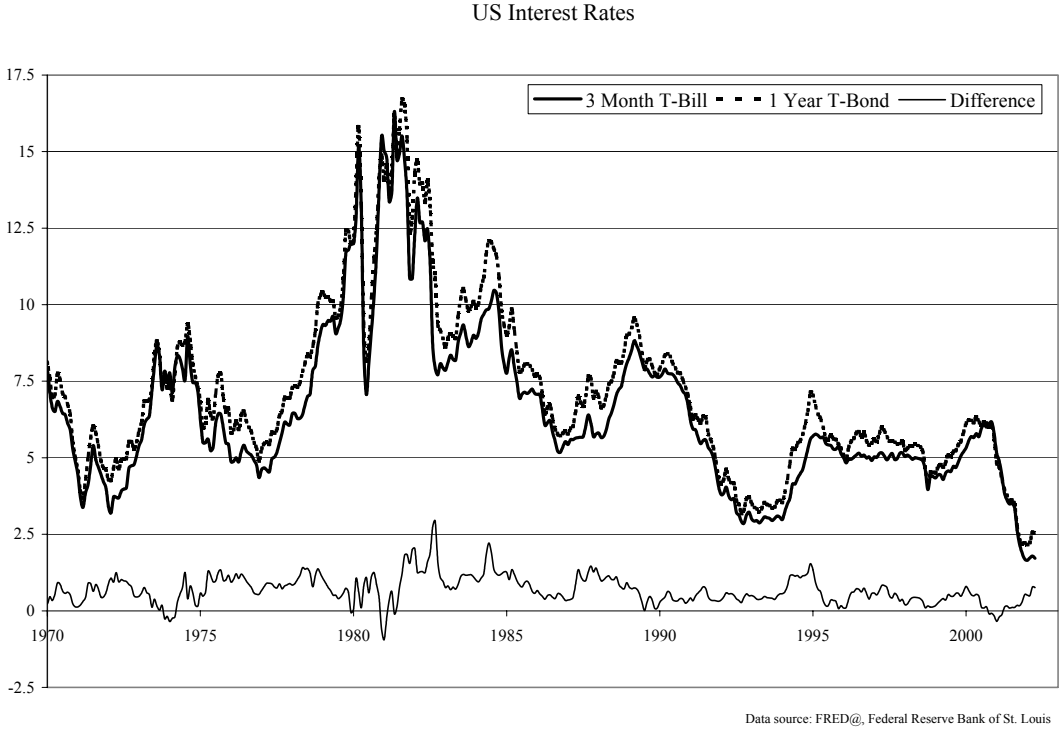

Figure 1: The figure contains the interest rates on a 3-month Treasury bill and a 1-year Treasury bond for the period January 1970 to April 2002. The two interest rates appear to be nonstationary, whereas the difference between them seems to be stationary. So the two interest rates are cointegrated.