co-integration vs correlation

Jamshid Ghorbani

Oct 7, 2025

Cointegration refers to a property by which two (or more) assets, while not being mean reverting individually, may be mean reverting with respect to each other. This commonly happens when the series themselves contain stochastic trends (i.e., they are nonstationary) but nevertheless they move closely together over time in a way that their difference remains stable (i.e., stationary). Thus, the concept of cointegration mimics the existence of a long-run equilibrium to which an economic system converges over time.

The intuitive idea is that, while it may be difficult or impossible to predict individual assets, it may be easier to predict their relative behavior. Mathematically, a multivariate time series y_1,y_2,y_3,..., is cointegrated if some linear combination becomes integrated of lower order, for example, if y_t is not stationary but the linear combination $\omega^T y_t$ is stationary for some weights ω. suppose the multivariate time series y_t denotes the log-prices of some stocks. Such a time series is nonstationary (random walk), by taking a linear combination $\omega^T y_t$ we might be able to obtain a stationary time series. As covered later, this property has remarkable consequences in terms of trading and it forms the basics of pairs trading.

A simple and common way to model cointegration of two time series is as

$$y_{1t} = \gamma x_t + \omega_{1t}$$

$$y_{2t} = x_t + \omega_{2t}$$

where x_t is a stochastic common trend defined as a random walk, $x_t = x_{t-1} + \omega_t$

and the terms $\omega_1 t$, $\omega_2 t$, ω_t,are i.i.d. residual terms, mutually independent, with variances σ_1^2,σ_2^2 and σ^2 ,respectively. The coefficient γ is the key quantity that determines the cointegration relationship. It is important to note that each of the time series, y_1, y_2, is a random walk plus additional noise, therefore nonstationary. However, since they share a common stochastic trend, a simple linear combination of the two can eliminate this trend. The so-called spread is precisely this linear combination without the trend:

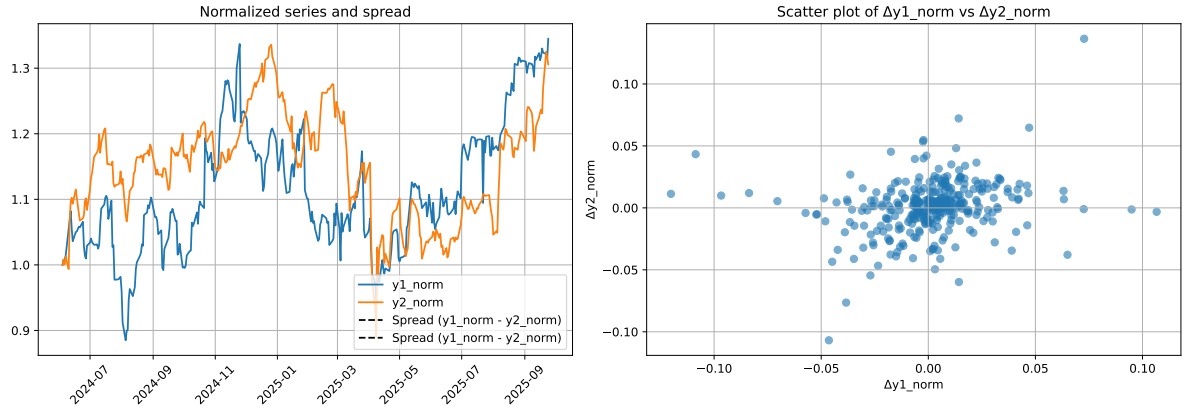
$$z_t = y_{1t} - \gamma y_{2t} = \omega_1 t - \gamma \omega_{2t}$$

which is stationary and mean reverting.

Correlation is a basic concept in probability that refers to how "related" two random variables are. We can use this measure for stationary time series but definitely not with nonstationary time series. In fact, when we refer to correlation between two financial assets, we are actually employing this concept on the returns of the assets and not the price values. the concepts of correlation and cointegration have been introduced, but their similarity and difference may be unclear and confusing. After all, it seems that they both try to capture the concept of similarity of movements of two time series, so superficially they may seem to be similar concepts. However, they are totally different right from their definition. As a matter of fact, the correlation of the differences of the two cointegrated time series in the model can be analytically derived as

$$\rho = \frac{1}{\sqrt{1 + 2\frac{\sigma_1^2}{\sigma^2}}\sqrt{1 + 2\frac{\sigma_2^2}{\sigma^2}}}$$

which can be made as small as desired by properly choosing the variances of the residual terms. That is, we can have two perfectly cointegrated time series with an arbitrarily small correlation, which may be surprising at first. This reveals that cointegration and correlation are two totally different concepts, yet they both attempt to measure the similarity of the movements of two time series. The following examples illustrate this difference.



so initially for the two time series we have a correlation of np.float64(0.26). for checking if they are cointegrated as well or not, we use ADF test.

```
   Augmented Dickey-Fuller Results
=====================================
Test Statistic                 -3.200
P-value                         0.020
Lags                                2
-------------------------------------

Trend: Constant
```

```
Critical Values: -3.45 (1%), -2.87 (5%), -2.57 (10%)
Null Hypothesis: The process contains a unit root.
Alternative Hypothesis: The process is weakly stationary.
```

The plot and the test results show the time series evolution of the variable, highlighting its general trend and short-term fluctuations. The analysis of the two time series reveals a modest short-term correlation of 0.26, suggesting that while the series tend to move in the same general direction, their daily co-movements are relatively weak. However, the Augmented Dickey–Fuller (ADF) test on the series yields a test statistic of –3.200 with a p-value of 0.020, which allows us to reject the null hypothesis of a unit root at the 5% significance level. This indicates that the series is stationary, implying the presence of a long-run equilibrium relationship. Therefore, despite the modest correlation, the results confirm that the series are co-integrated, making them suitable candidates for mean-reversion or pairs trading strategies.

The issue with the Engle-Granger test is that it only measures cointegration between two time series. However, tests such as the Johansen test are used to determine cointegration between several time series.

1 Johansen Test

The Johansen test is used to test cointegrating relationships between several non-stationary time series data. Compared to the Engle-Granger test, the Johansen test allows for more than one cointegrating relationship. However, it is subject to asymptotic properties (large sample size) since a small sample size would produce unreliable results. Using the test to find cointegration of several time series avoids the issues created when errors are carried forward to the next step.

Johansen's test comes in two main forms, i.e., Trace tests and Maximum Eigenvalue test.

Trace tests Trace tests evaluate the number of linear combinations in a time series data, i.e., K to be equal to the value K_0, and the hypothesis for the value K to be greater than K_0. It is illustrated as follows:

$$H_0 : K = K_0$$

$$H_0 : K > K_0$$

When using the trace test to test for cointegration in a sample, we set K_0 to zero to test whether the null hypothesis will be rejected. If it is rejected, we can deduce that there exists a cointegration relationship in the sample. Therefore, the null hypothesis should be rejected to confirm the existence of a cointegration relationship in the sample.

2 Maximum Eigenvalue test

An Eigenvalue is defined as a non-zero vector which, when a linear transformation is applied to it, changes by a scalar factor. The Maximum Eigenvalue test is similar to the Johansen's trace test. The key difference between the two is the null hypothesis.

$$H_0 : K = K_0$$

$$H_0 : K = K_0 + 1$$

In a scenario where $K = K_0$ and the null hypothesis is rejected, it means that there is only one possible outcome of the variable to produce a stationary process. However, in a scenario where $K_0 = m - 1$ and the null hypothesis is rejected, it means that there are M possible linear combinations. Such a scenario is impossible unless the variables in the time series are stationary.