Projet PI2

Synthesis of the concepts related to the article used as a basis for the project

The Cointeration Alpha :
Enhance Index Tracking and Long-Short Equity Market Neutral Strategies
By: Carol Alexander & Anca Dimitriu

Table of contents

1. Introduction and some definitions

- Index tracking strategy aims to replicate a benchmark accurately in terms of returns and volatility
- For the Index tracking strategy: the portfolio optimisation is based on cointegration rather than correlation.
- The standard in portfolio management and risk measurement is the correlation analysis of asset returns

Remark: Correlation analysis is valid only for stationary variables.

a) Definitions

- **Stationarity:** In the most intuitive sense, stationarity means that the statistical properties of a process generating a time series do not change over time. It does not mean that the series does not change over time, just that the *way* it changes does not itself change over time. The algebraic equivalent is thus a linear function, perhaps, and not a constant one; the value of a linear function changes as x grows, but the *way* it changes remains constant

- A variable has **a stochastic trend** if its difference has a stationary invertible ARMA(p,q) representation plus a deterministic component.

- **ARMA:** auto regressive moving average

- **Statistical arbitrage:** group of trading strategies which utilize mean reversion analyses to invest in diverse portfolios of up to thousands of securities for a very short period of time.

- **Mean reversion:** is a theory used in finance that suggests that asset prices and historical returns eventually will revert to the long-run mean or average level of the entire dataset

- **Alpha** : The excess return of an investment relative to the return of a benchmark index. Its two main sources are usually credited to be a successful stock selection and market timing.

- **Pairs trade** is a trading strategy that involves matching a long position with a short position in two stocks with a high correlation.

- A strategy is said to be **market neutral** if it generates returns which are independent of the relevant market return. The advantages of market neutral long-short equities investing are perceived to be independence of the market direction

2. Cointegration

a) Definition

Cointegration is a statistical property of time series introduced into economic analysis, notably by Engle and Newbold in 1974.

Cointegration is an econometric technique for testing the correlation between non-stationary time series variables. The notion of cointegration implies the notion of integration. If two or more series are themselves non-stationary, but a linear combination of them is stationary, then the series are said to be cointegrated.

Cointegration tests identify scenarios in which two or more non-stationary time series are integrated together in such a way that they do not deviate from the long-term equilibrium. The tests are used to identify the degree of sensitivity of two variables to the same average price over a given period of time. We will return to the tests in another section.

From a technical point of view:

- Its key characteristics (i.e. mean reverting tracking error, enhanced weights stability and better use of the information contained in stock prices) allow a flexible design of various trading strategies, from index tracking to long-short market neutral.
- The aim of the cointegration analysis is to detect any **stochastic trend** in the price data and use these common trends for a dynamic analysis of correlation in returns
- In the area **of equity markets**, cointegration analysis has frequently targeted two objectives: to estimate the degree of co-movement in stocks within a given market index and to identify the economic fundamentals generating this type of behaviour.

Example of cointegration in finance:

Cointegration is used in finance. For example, a stock market index and the price of its associated futures contract evolve over time. They follow approximately a random walk. Testing the hypothesis that there is a statistically significant relationship between the futures price and the cash price could now be done by finding a cointegrating vector. (If such a vector has a low order of integration, it may mean an equilibrium relationship between the original series, which are said to be cointegrated of an order less than one).

b) Differences between correlation and cointegration

Often the terms correlation and cointegration are misused because they are not really the same thing. They are often confused with synonyms.

Correlation and cointegration are terms used in regression analysis. In similar terms, both are commonly used in currency trading to calculate the relationship between two or more variables over a specific time period.

When two variables move in the same direction, they are said to be positively correlated. If they move in opposite directions, the correlation is said to be negative.
Cointegration is used to determine the degree of sensitivity of two variables to the same average price over a given period. Thus, cointegration does not reflect whether pairs move in the same or opposite directions, but can tell you whether the distance between them remains the same over time.

Correlation is easier to identify than cointegration. However, cointegration is considered the most reliable regression analysis tool.

Use in finance:

Correlation is a term that novice traders may be more familiar with, but cointegration is what advanced traders will pay attention to.

For day traders it means that the movements of these variables are not related. However, in the longer term, the variables may follow a common mean value.
Since cointegration allows for the identification of variables that would not move too far apart in the long run and return to a mean distance between them, the concept of cointegration is used for hedging.

Here too, the degree of cointegration must be calculated. The higher the degree of cointegration between two variables, the greater the probability that they maintain a stable or constant distance. Another variable is the time it takes for two cointegrated variables to return to the mean.

c) Cointegration methods

There are methods for testing cointegration. They are used to identify long-term relationships between two or more sets of variables.

- Engle-Granger Two-Step Method

The two-step Engle-Granger method begins by creating residuals based on static regression and then tests the residuals for the presence of unit roots. It uses the Augmented Dickey-Fuller's test (ADF) or other tests to check the stationarity of units in time series. If the time series is integrated, the Engle-Granger method will indicate the stationarity of the residuals. We will return to the ADF test later.

The Engle-Granger method has limitations. First, if there are more than two variables, the method can show more than two cointegrating relationships. Another limitation is that it is a single-equation model. Some of its drawbacks have been taken into account in recent cointegration tests, such as the Johansen and Phillips-Ouliaris tests. The Engle-Granger test can be determined using STAT or MATLAB.

- Johansen Test

The Johansen test is used to test cointegration relationships between several non-stationary time series data.

This test can be compared to the Engle-Granger test. The Johansen test allows more than one cointegrating relationship to be established. However, it is subject to asymptotic properties.

This means that the sample size must be large because a small sample size would produce unreliable results. Using the test to find the cointegration of multiple time series avoids the problems created when errors are carried over to the next step.

The Johansen test comes in two main forms.

First of all, "Trace tests" evaluate the number of linear combinations in a time series, i.e. K must be equal to the value K0, and the assumption that the K value is greater than K0.
It is illustrated as follows:
$H0 : K = K0$
$H0 : K > K0$

When we use the "Trace tests" to test cointegration in a sample, K0 is set to zero. This allows us to test whether the null hypothesis will be rejected. If it is rejected, we can infer that there is a cointegrating relationship in the sample. Therefore, the null hypothesis must be rejected to confirm the existence of a cointegrating relationship in the sample.

Next, the Maximum Eigenvalue test consists of having an eigenvalue defined as a non-zero vector. When a linear transformation is applied to it, it is modified by a scalar factor. The Maximum Eigenvalue test is similar to Johansen's "Trace tests". The main difference between the two is the null hypothesis.

It is illustrated as follows:
$H0 : K = K0$
$H0 : K = K0 + 1$

In a scenario where K=K0 and the null hypothesis are rejected, this means that there is only one possible outcome of the variable to produce a stationary process. However, if K0 = m-1, 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.

3. Integration

➢ Why the series as to be integrated?

Cointegration tests identify scenarios where two or more non-stationary time series are integrated together in a way that they cannot deviate from equilibrium in the long term.

We say that **two series are co-integrated if a linear combination has a lower level of integration.**

Stock prices & Stock market indexes are usually integrated of order 1, cointegration exist one there exists at least one stationary linear combination of them.

Order of integration: I(d)

- **Integration of order 0**

A time series is integrated of *order 0* if it admits a moving average representation with:

$$\sum \left| \boldsymbol{Bk^2} \right| < \infty$$

Where b is the possibly infinite vector of moving average weights (coefficients or parameters).

- **Integration of order d**

A time series is integrated of *order d* if: $(\boldsymbol{1}-\boldsymbol{L})^{\boldsymbol{d}*\boldsymbol{X_t}}$ is a stationary process where L is the backshift operator (operates on an element of time series to produce the previous element).
In other words, a process is integrated to *order d* if taking repeated differences d times yields a stationary process.

- **Constructing an integrated series:**

An *I(d) process* can be constructed by summing an I(d-1) process:

  - Suppose $X_t = I(d-1)$
  - Let Z_t be a series and $\boldsymbol{Z_t} = \sum \boldsymbol{Xt}$
  - Z_t is $I(d) \Rightarrow$ We can demonstrate it by observing its first differences are $I(d-1)$:
    - $\Delta \boldsymbol{Z_t} = \boldsymbol{X_t}$
    - And $\boldsymbol{X_t} \sim \boldsymbol{I}(\boldsymbol{d}-\boldsymbol{1})$

This implies that the autocovariance is decaying to 0 sufficiently quickly. This is a necessary, but not sufficient condition for a stationary process.
Therefore, all stationary processes are $I(0)$, but not all $I(0)$ processes are stationary.

4. Stationarity

Co-integration is mainly used for non-stationary variables while correlation analysis is only valid for stationary variables.

This requires a prior analysis of the trend in prices and other level financial variables, which are generally found to be at a level of order one or higher. Taking the first difference in log(price) is the standard procedure for ensuring stationarity and makes any other inference based on returns.
However, this procedure has the disadvantage that valuable information is lost:

- Removing variables prior to analysis removes any possibility of detecting common price trends.
- When the variables of a system are integrated in different orders, and therefore require different numbers of differentiations becoming stationary, the interpretation of the results becomes difficult.

In contrast, the objective of cointegration analysis is to detect any stochastic trends in the price data and to use these common trends for dynamic correlation analysis in returns.
When applied to share prices and stock market indices, which are generally found to be first-order integrated, co-integration exists when there is at least one stationary linear combination of these.

Conclusion:

Stationarity is used in order to know if the **residuals are stationary** (a test is performed) which is a necessary hypothesis for the good functioning of the method.

- **Stationarity:** In the most intuitive sense, stationarity means that the statistical properties of a process generating a time series do not change over time. It does not mean that the series does not change over time, just that the *way* it changes does not itself change over time. The algebraic equivalent is thus a linear function, perhaps, and not a constant one; the value of a linear function changes as x grows, but the *way* it changes remains constant

5. ADF Test (Augmented Dickey-Fuller)

Sources:

- *https://fr.wikipedia.org/wiki/Test_augment%C3%A9_de_Dickey-Fuller*
- *https://help.xlstat.com/s/article/test-de-racine-unitaire-dickey-fuller-et-de-stationarite-dune-serie-chronologique?language=fr*
- *ADF Test : https://www.quantstart.com/articles/Cointegrated-Time-Series-Analysis-for-Mean-Reversion-Trading-with-R/*
- *Article*

a) Definitions

The Dickey-Fuller Augmented Test (ADF test) is a **statistical test** to determine whether a **time series** is **stationary**.

A time series (time-related series, e.g. Y_t(t=1,2, 3,...)) is said to be stationary (in the **weak sense**) if its **statistical properties** (expectation, variance, auto-correlation) **do not vary over time**.

Example:

- *Temporary stationary series: White noise / A series with a normal law independent of t, N~(μ,σ²)*
- *Non-stationary series: Random walk (Because a random walk has a unitary root)*

It is interesting to note that a non-stationary series can be **stationary in difference** ↔ Y_t is not stationary, but **Y_t-Y_{t-1} is stationary.**

Example of a stationary series in difference: Random walk

A time series can also be **trend stationary**. That is, it has an additive component that is a function of time. It will have to be compensated to keep only the stationary component of the series.

b) Stationarity tests

Stationarity tests are designed to check whether a series is stationary or not.
2 different types of tests:

- KPSS test :
  We define the null hypothesis H0: "The series is stationary".
  And the H1 hypothesis: "The series is not stationary".
- Unit root tests :
  - **Augmented Dickey-Fuller (*ADF*)** & Dickey-Fuller test,
  - Phillips-Perron Test (*PP*)

In these tests the null hypothesis **H0 is defined: "The series is not stationary".**
And the H1 hypothesis: "The series is stationary".

If the series is not stationary, it means that it was generated by a process with a unit root.

The null hypothesis for the test is that the data are non-stationary. We want to REJECT the null hypothesis for this test, so we want a p-value of less than 0.05 (or smaller).

It is possible to perform this test with different software:

- With EXCEL tools: Follow this link: Test ADF with Excel
- With Python: Follow this link: Test ADF with Python
- With R: Follow this link: Test ADF with R
  (Function adf.test with R : RDocumentation/adf/test)

c) ADF Formula

During this **project** we will focus on the **ADF test** which is testing if γ=0 in this model of the residues:

$$\Delta\hat{\varepsilon}_t = \gamma\hat{\varepsilon}_{t-1} + \sum_{i=1}^{p}\alpha_i\Delta\hat{\varepsilon}_{t-i} + u_t$$

6. Two applications of cointegration based trading strategies

a) Index tracking

1) Stocks selection

- The first stage, stock selection, can be the result of proprietary selection models, technical analysis or just stock picking skills of a portfolio manage
- The second stage of index tracking concerns determining the portfolio holdings in each of the stocks selected in the previous stage.

$$\log(\text{index}_t) = c_1 + \sum_{k=1}^{n} c_{k+1} * \log(P_{k,t}) + \varepsilon_t$$

2) Back-Test

- **ADF test:**
  Result: once the minimum calibration period for ensuring cointegration is used, increasing it does not necessarily improve the cointegration results

- **Returns:**

Assuming that the portfolio weights $w_{k,T}$ are estimated at time T, the price of the portfolio at time $T+x$, $x <= 10$, can be computed based on the prices $P_{k,T}$ and $P_{k,T+x}$ of the n stocks in the portfolio as follows:

$$\pi_{T+x} = \pi_{T-1}\sum_{k=1}^{n}\frac{W_{k,T}}{P_{k,T}}P_{k,T+x} \qquad (5)$$

The portfolio returns were further estimated as the first difference in log prices of the portfolio.

- **Transaction costs:**

In the framework of our strategy, the transaction costs were incurred on each portfolio re-balancing, i.e. every 10 trading days. However, in order to avoid creating artificial jumps in the returns series, the transaction costs were equally distributed to all the daily returns during the non-trading period. If the portfolio weights w_{kT} are estimated at time T, the transaction costs at time T can be computed as follows:

$$TC_T = 0.002\sum_{k=1}^{n} abs(w_{k,T} - w_{k,T-10})P_{k,T} \tag{6}$$

The transaction costs decrease significantly with the number of stocks in the portfolio and also, but less obviously, with the number of years in the calibration period.

- **Volatility of tracking portfolio returns:**
  - Smaller number of stocks portfolios display higher volatility of the excess returns
  - The calibration period appears not to have a big impact on the volatility of the tracking portfolio returns.

- **Correlation of tracking portfolios returns with market returns:**
  - The tracking error is not correlated with the benchmark returns

- **Skewness and excess kurtosis:**

[5]The skewness and excess kurtosis were computed as:

$$sk = \frac{n}{(n-1)(n-2)}\sum_{i=1}^{n}\left(\frac{TE_i - \overline{TE}}{\sigma_{TE}}\right)^3$$

$$excesskurt = \frac{n(n+1)}{(n-1)(n-2)(n-3)}\sum_{i=1}^{n}\left(\frac{TE_i - \overline{TE}}{\sigma_{TE}}\right)^4 - 3\frac{(n-1)^2}{(n-2)(n-3)}$$

[6] The Sharpe ratio was computed as the average annual excess return of an investment strategy over the risk free rate divided by the annualised standard deviation of returns.

[7] The information ratio is simply the average annual return of an investment strategy divided by its annualised standard deviation.

  - The tracking errors generated by different portfolios appear to have different degrees of non-normality, but generally they have small positive skewness and excess kurtosis.

- **Sharpe ratios:**

Formula and Calculation of Sharpe Ratio

$$Sharpe\ Ratio = \frac{R_p - R_f}{\sigma_p}$$

where:

R_p = return of portfolio

R_f = risk-free rate

σ_p = standard deviation of the portfolio's excess return

- Provided that our tracking portfolios have generated average returns very close to market index returns with similar volatilities, the Sharpe ratios generally stay in the same range.

3) Summary

The cointegration index tracking strategy has the following features:

• To ensure cointegration, a **minimum number of stocks** in the portfolio (in our case 20 stocks, 0.67% of the index basket) and a minimum calibration period (in our case of 3 years) are required;

• The tracking portfolios have **similar returns and volatility with the market index**, and are highly correlated with the latter;

• The **excess returns** from the index tracking (i.e. tracking errors), are uncorrelated with the market, have low volatility and slightly leptokurtic distributions with positive skewness;

• The **periods of significant market decline**, such as the Asian and Russian crises and the burst of the technology bubble are generating the largest part of the excess returns on the tracking portfolios;

• The **overall performance of the tracking strategy** is dependent on the portfolio selection method used, the number of stocks and calibration period. Special attention should be given to the stock selection method, especially to the amount of trades required to rebalance the portfolio, as the transaction costs may erode the returns of the tracking portfolios.

From the analysis of tracking portfolios with different number of stocks, we have found that most of the performance measures have favoured the 30-stocks tracking portfolio. However, the slight under-performance of the 20-stocks portfolios as compared to the market index will be of further use when designing the short part of the market neutral strategy. Of the stock selection methods considered, the annual re-ranking and frequency-based re-ranking with an indicator function estimated over 3-years appear to provide the best results in terms of returns and consistency.

b) Long-Short equity strategy

- The characteristics of a successful long-short equity strategy are usually recognised to be steady pattern in returns, low volatility and market neutrality.
- The fact that long-short equity strategies ensure a more efficient use of information than long only strategies is the result of not restricting the weights of the undervalued assets to zero.

Formula used to construct the data:

There are two differences between the actual index and the reconstructed one: the value of the divisor and the constituent stocks (both of which change periodically in the actual index but not in the reconstructed index).

$$\text{actual\_DJIA}_t = \text{divisor}_t \sum_{k=1}^{30} \text{stock\_price}_{k,t} \tag{1}$$

$$\text{reconstructed\_DJIA}_t = \text{divisor}_T \sum_{k=1}^{30} \text{stock\_price}_{k,t} \tag{2}$$

The use of a reconstructed index instead of the actual one is justified by our interest in the current structure of the index: that is, we compare the performance of portfolios comprising the stocks currently included in DJIA with a market index constructed from the same stocks.

Having constructed the simple tracking strategy, a natural extension for exploiting the tracking potential of the cointegrated portfolios would be to replicate 'plus' and 'minus' benchmarks.

The results of the long-short strategies are highly dependent on the stock selection method used and on the spread between the 'plus' and 'minus' benchmarks tracked.

1) The cointegration test:

The new cointegration regressions can be written as:

$$\log(\text{index\_plus}_t) = a_1 + \sum_{k=1}^{n} a_{k+1} * \log(P_{k,t}) + \varepsilon_t \tag{7}$$

$$\log(\text{index\_minus}_t) = b_1 + \sum_{k=1}^{n} b_{k+1} * \log(P_{k,t}) + \varepsilon_t \tag{8}$$

We note that the stock weights are not restricted to be positive in the tracking portfolios above; in fact it is likely that we shall take some short positions in the portfolios tracking both 'plus' and 'minus' benchmarks.

The portfolios remain fairly cointegrated with the benchmarks tracked, even if the latter diverge significantly from the actual market index.

2) Back-Test

- **Returns/transaction costs:**

The new cointegration regressions can be written as:

$$\log(\text{index\_plus}_t) = a_1 + \sum_{k=1}^{n} a_{k+1} * \log(P_{k,t}) + \varepsilon_t \tag{7}$$

$$\log(\text{index\_minus}_t) = b_1 + \sum_{k=1}^{n} b_{k+1} * \log(P_{k,t}) + \varepsilon_t \tag{8}$$

We note that the stock weights are not restricted to be positive in the tracking portfolios above; in fact it is likely that we shall take some short positions in the portfolios tracking both 'plus' and 'minus' benchmarks.

- From the long-short strategies analysed, the ones producing the most consistent positive results are the ones tracking small spreads, even if the magnitude of these returns is reduced.
- By contrast, the strategies tracking large spreads are generating less consistent and less frequent positive returns, which also have a higher magnitude.

- **Volatility of tracking portfolio returns:**

  - Considering the results of the long-short strategies in respect of volatility, a reasonable conclusion would be that some of the most aggressive strategies, tracking benchmarks that are quite far from the market index, display high volatility. But this is almost never greater than the market volatility and the more conservative strategies have much lower volatility than the market.

  - The most consistent positive returns, with low volatility and no significant correlation with the market are generated by strategies tracking narrow spreads between the 'plus' and the 'minus' benchmark.

- **Correlation of tracking portfolios returns with market returns:**

In order to examine the correlation of the long-short strategy returns with the market returns, we can write the returns on the 'plus' and 'minus' portfolios separately as follows:

$$R_{+,t} = \alpha_+ + \rho_+ \frac{\sigma_+}{\sigma_{B+}} R_{B+,t} + \varepsilon_{+,t} \quad (10)$$

$$R_{-,t} = \alpha_- + \rho_- \frac{\sigma_-}{\sigma_{B-}} R_{B-,t} + \varepsilon_{-,t} \quad (11)$$

where $\rho_{+/-}$ are the correlation coefficients between the 'plus' and respectively 'minus' portfolio returns with the 'plus'/'minus' benchmark returns, σ_+ / σ_{B+} and σ_- / σ_{B-} are the relative volatilities of the 'plus' respectively 'minus' portfolio returns to the 'plus'/'minus' benchmark returns, and ε_+ and ε_- are the tracking errors of the 'plus'/'minus' portfolios.

  - The stock selection method and the spread between the benchmarks tracked are influencing the level of correlation, but the relationship is not straightforward. Also, there is a slight decrease in correlations as the number of stocks in the tracking portfolios increases.

- **Skewness and excess kurtosis:**

The results on skewness and excess kurtosis, the daily returns to the long-short strategies appear to have only a small degree of non-normality. This is far lower than the non-normality that is usually identified for hedge fund returns.

- **Sharpe ratios:**

The Sharpe ratios produced by the annual re-ranking and frequency-based re-ranking are considerably larger than the ones displayed by the daily re-ranking strategies.

3) Summary

The results obtained from back-testing prove that, when setting up a long-short market neutral strategy based on cointegration, the following parameters have a significant impact on the strategy's success:

1. **Stock selection method** – in terms of returns, the stock selection method is critical to the success of the long-short strategy. As shown by the negative results of the daily re-balancing stock selection method, a high variability of the stock weights may generate huge transaction costs affecting the strategy's potential to generate returns. Based on the same grounds, the other two stock selection methods employed (annual and frequency-based re-ranking) have proved to be equally appropriate, considering that they were generating similarly low transaction costs in the simple tracking strategies.

2. **Benchmarks to be tracked by the 'plus'/'minus' portfolios** – as shown by the back-test results, the spread between the benchmarks tracked in the long-short strategy cannot be increased without a corresponding increase in the volatility and kurtosis of the returns, and also a potential reduction of the strategy skewness. The Sharpe ratios, as a measure of the trade-off between returns and volatility, favour low spreads between the benchmarks tracked. In addition to an increase in volatility, the cointegration relationship will weaken as the reconstructed benchmarks diverge from the market index. Therefore, it is essential to test each portfolio for cointegration with its underlying benchmark.

3. **Number of stocks in each portfolio** – as displayed by the residuals' stationarity tests, in order to identify a cointegration relationship, a minimum number of stocks is required in the tracking portfolio. Apart from the minimum number of stocks, the long-short strategies appear to provide best results with close to the maximum number of stocks in each 'plus'/'minus' portfolio. A particularly successful strategy appeared to be the combination of 30-stocks in the 'plus' portfolio with 20 or 25-stocks in the 'minus' portfolio, a combination that is able to exploit the difference in the cointegration regressions intercepts.

4. **Calibration period** – as showed by the cointegration tests and implied by the theory, a minimum number of years is required to construct a cointegration relationship. Beyond this number, the effect of the calibration period on the returns of different stock selection strategies is not uniform. Still, the best results were obtained for longer calibration periods.