Finding Cointegrated Pairs for Pairs Trading Strategy

Udacity Capstone Project

Table of Contents

Domain Background

Theory behind pairs trading

One of the key objectives of most stock investors is "finding alpha", which could be described as the search for tendencies in stocks that one can successfully bet on without being misled by temporary trends in the market as a whole, colloquially referred to as "beta". Market beta is notorious for misleading investors. If one makes a stock purchase and the price of the stock goes up, it is natural to assume that the investment was a good one, but if the entire market is trending upwards, on average most everyone in the market will reap a benefit, whether they have made wise investments or not. And sooner or later the market will trend downward, and all those who have made investments that are insufficiently market-neutral will lose. Smart investors attempt to avoid this pitfall by finding investments that are rich in "alpha" yet "beta-neutral". They usually do this by going long on some stocks and short on others. If the entire market goes up, they are protected by their long bets. If the entire market goes down, they are protected by their short bets.

One approach is to build a beta-neutral strategy exclusively on fundamental data, going long on stocks with healthy fundamentals and shorting stocks that have weak fundamentals, but this kind of strategy requires a large portfolio to work, so it is only available to investors with large investment budgets.

An alternative for investors with more modest budgets is *pairs trading*. With pairs trading the investor attempts to identify pairs of stocks whose prices change in tandem. The classic example is a pair of stocks from the same industry, such as Coke and Pepsi. If, say, the price of an industrial input like sugar surges and diminishes the profit margins of Coke, it will likely do the same for Pepsi, and so one is likely to see declining profits in both companies that are reflected in the stock price. But in order to be profitable, the pairs trading strategy does not require that the two stocks in the pair to be perfectly coordinated in their price history; it only requires that the difference between them *mean revert*. That means that although the difference between the two prices may vary, it will tend to revert to its historic mean eventually. Mathematically, we say that we are testing to see if the difference between the two stocks is *stationary*, and we test for this using a *cointegration test*, like Augmented Dickey Fuller (ADF).

Once we have identified two stocks that are cointegrated, we need only wait for the difference in their prices to diverge from the historic mean, at which point we short the stock that is higher than expected and go long on the stock that is lower than expected and wait for the price gap between the two stocks to close once again. When the difference between the two stocks reverts to its historic mean, it will not matter whether the high stock came down or the low stock went up. Either way, we make money on at least one of the two bets.

This strategy is market-neutral, because its success depends only on the relative prices of the two stocks, not their absolute prices. The market might trend upward, pulling both stocks upward with it, or downward, pulling both stocks downward. Either way, what matters for the success of the pairs trading strategy is the *difference* between the two stocks, not their

absolute value. As long as the difference reverts to its historic mean, the strategy will make money.

This project addresses the challenge of how to find cointegrated pairs for a pairs trading.

Datasets and Inputs

The dataset I use consists of Wiki Prices data made available to the public for free by Quandl and Intrinio stock fundamentals data made available to me as an Intrinio subscriber.

It contains daily stock price data for the period 2000 – 2008. I chose this period, because it is a period in which there are no dramatic changes in market regime, i.e. no wild swings between bull and bear markets or volatility levels. Obviously, the market regime changed dramatically with the 2008 financial crisis.

The period 2000 – 2008 yields 2,923 timestamped observations and 2485 stocks. This data must be filtered in the following ways:

- Stocks must have price values for the entire date range, and those that lack the full range of values must be removed.
- Non-trading days that have NaN values must be removed for all stocks.
- The dataset has some noise in it. In particular, non-trading days sometimes have tiny unexpected values. Thus if any day is found containing NaN values for almost all stocks, that day will be considered a non-trading day and will be removed from the dataset, even if some small portion of the stocks have a value other than NaN for that day.
- For those tests that use fundamental data, the intersection must be found between the Quandl price data and the Intrinio fundamentals data. Both datasets are supposed to be market-wide, yet they do not contain the exact same set of stocks. Thus stocks present in one set or the other but not both must be removed.
- Opening prices must be subtracting from closing prices to calculate the daily price change and then this must be converted into a daily percentage change to yield daily returns

After filtering the data I was left with 2009 daily returns for 1,797 stocks.

Problem Statement

When trying to find potentially cointegratable pairs, one confronts the following problems:

(1) **Problem of confounding variables**: Sometimes variables appear to be related but their relationship to one another consists solely in the mutual relationship they have to a third variable. This third variable is called a confounding variable, because it "confounds" (or confuses) the proper analysis of the other two variables and their relationship to one another. A good textbook example of this is the relationship of heat and smoke to fire.

Both heat and smoke appear to be strongly correlated, but in truth heat does not produce smoke nor does smoke produce heat. Both are dependent on a third variable, fire. In the stock market, the most obvious example of a confounding variable is the market beta discussed above. Cointegration tests, in particular, can produce spurious results because of market beta, leading the analyst to believe that stocks are cointegrated when they are not.

(2) **Problem with sector pairs**: The most obvious choices for pairs are found by trying pairs in the same industry sector, because the prices of stocks in the same industry do tend to change in tandem, but this strategy is very well known among stock investors and a lot of people are already using it. That tends to erode any arbitrage opportunity. To find less obvious cointegrated pairs you need to cast your net wider and test pairs that straddle industry sectors. But that leads to a different problem: the problem of a large number of features.

(3) **Problem of large numbers of features**: Any test for pairs over large numbers of stocks will confront the problem of the number of features exceeding the number of observations (time slices), because this can lead to ill-conditioned empirical covariance matrices. One can remedy this by increasing the time span over which one analyzes price data, thus increasing the number of observations. But that leads to yet another problem: the problem of using long periods of stock data as though it were homogeneous in its properties when, in fact, it almost never is.

(4) **Problem of using lengthy time series**: Using price series that span more than 10 years carries a risk for the integrity of the data. Stocks can be de-listed, so-called survivorship bias can become a problem, and longer time spans are more likely to straddle market regime changes (bull market to bear market, etc.) and that can undermine statistical inference. Risk management applications often restrict covariance estimations to recent data that spans 3 years or less.

(5) **Problem of multiple comparisons bias**: Finally, since tests for cointegration like Augmented Dickey Fuller use a significance level (typically of 5%) to measure whether a pair should be considered mean reverting or not, these tests suffer from multiple comparisons bias. This problem arises when you run such a large number of tests that on average your tests will produce false positives simply by virtue of the fact that tests of pairs that are not cointegrated will still satisfy the significance level 5% of the time. If you run 10,000 tests, for example, that could yield as many as 500 false positives, and if there are very few cases of real cointegration, the false positives could actually dominate the results.

How do we solve these problems?

Solution Statement

Here is how I intend to address the problems indicated in the previous section.

Solution to the problem of multiple comparisons bias

If it weren't for the problem of multiple comparisons bias, we could do a brute force cointegration test of every pair in the market. To avoid this problem we must forego easy brute force approaches and reduce the number of tests. The technique we use to restrict the number of cointegration tests is clustering. We will only test pairs within clusters, not pairs that straddle clusters. Additionally, we will use partial correlations to add an extra layer of restrictiveness by limiting our cointegration tests to pairs within the same cluster that also exhibit strong partial correlations to one another.

Solution to the problem of confounding variables

One can get an idea of how much two stocks covary by looking at the covariance matrix, but if two stocks covary because of mutual dependence on a confounding variable, the covariance matrix will not reveal it. However, the precision matrix might. The precision matrix is proportional to the partial correlations matrix and *partial correlations*. These numbers can provide a measure of how correlated two stocks are, *independently* of their relationship to all other stocks. They constitute a kind of *ceteris paribus* correlation, i.e. the correlation of the two stocks *all other influences being equal*, which by its nature excludes the influence of confounding variables.

Thus in addition to clustering, we will use conditional correlation to measure how suitable pairs are to be tested for cointegration. This should both increase the percentage of successful cointegration tests by restricting tests to the most interesting candidates as well as help us devise pair trading strategies that are ***beta neutral***.

Solution to the problem of overused sector pairs:

Recall that restricting the search for pairs to industry sectors has certain advantages. Pairs within a sector are more likely to cointegrate and fewer tests are necessary, so cointegration tests are less subject to multiple comparisons bias. But as indicated earlier, the downside of restricting the search to industry sectors is that a large number of investors already use this strategy, which means it is a crowded field and most trading profits have been arbitraged away. So instead of using sectors, I will apply clustering techniques to limit the number of tests yet apply them to the entire market with a view to finding pair candidates that straddle sector boundaries.

I will use sector-wide pair searches for benchmarking purposes only. This should show us what level of success we could expect to achieve with sector-wide searches then allow us to compare that success with the success we achieve when tackling all the stocks in the market at once.

Solution to the problem of large numbers of features:

Large numbers of features can lead to ill-conditioned empirical covariance matrices, and even when there are more observations than features, the empirical covariance matrix can be ill-conditioned if features are highly correlated. Thus we have two objectives in this project that are in tension with one another: (1) the search for cointegrated pairs across a large number of stocks that and (2) the need for a well-conditioned empirical covariance matrix.

We address this problem by regularizing the covariance matrix calculation using Graph-Lasso, which is the approach employed in the SciKit Learn project Visualizing Stock Market Structure.

Evaluation Metrics

We need a metric to measure the success of the techniques we use to narrow the scope of our search for cointegrated pairs. One measure of that success is the number of cointegrated pairs we find divided by the number of pairs tested. The higher the percentage of successful cointegrations, the more effective our methods of choosing candidate pairs.

Benchmark Model

I use two different benchmarks in this project. On the one hand, I cluster on sector data, as indicated above, then calculate the percentage of cointegrated pairs I am able to find and compare that percentage to the percentage of successful cointegrations I can achieve clustering on the entire market.

As a second type of benchmark, I use a project developed by Jonathan Larkin, the former Chief Investment Officer of Quantopian, which addresses the problem of too many features in a different way from the SciKit Learn project. Instead of regularizing the covariance matrix, Larkin uses PCA dimension reduction in a novel way. We cannot achieve stable covariance estimation by reducing the number of stocks, because our purpose is to explore the properties of stocks! So how can we use PCA?

Larkin's project does not treat stocks as features when it does dimension reduction. It transposes the price data, converting the time stamps of the daily returns into column headers and treats each daily return as a separate feature! The project then uses PCA to reduce the number of dimensions from 2009 time slices to 50 principal components.

When used in this way, PCA has aspects in common with other techniques that can do spectral analysis like DFT (Discrete Fourier Transform), or even dimension reduction across time, like DTW (Dynamic Time Warping) and Forecastable Component Analysis (FCA or ForeCA). See the followings posts and articles for further information:

- https://stats.stackexchange.com/questions/82291/time-series-dimensionality-reduction

- https://stats.stackexchange.com/questions/283521/dimensionality-reduction-of-multivariate-time-series
- https://arxiv.org/abs/1205.4591

With only 50 features, ill-conditioned covariance matrices cease to pose a problem. After doing dimension reduction, the Larkin project runs the DBSCAN to cluster stocks using principal components concatenated with stock fundamentals data as inputs instead of using the raw price data as input. The SciKit Learn project also does dimension reduction using Locally Linear Embedding, but only for purposes of 2D visualization, not as part of the clustering algorithm.

Project Design

Download project and all its data here (NOTE: doesn't include this file or the report).

The project is organized as two notebooks:
- one with utilities for downloading and cleaning stock data and
- another with routines to support the main algorithm

The latter notebook consists of:
- a section of helper functions followed by
- a section that does analysis on data for the entire market followed by
- a section that does analysis on data for different industry sectors

Project Results and Conclusions

Conclusions

Here are the results of the project according to approach used.

Project Strategy:

- Covariance estimation and regularization technique: GraphLasso
- Clustering technique: Affinity Propagation
- Data clustered on: covariance matrix
- Stock universe: Entire market (1,797 stocks)
- Number of possible pairs in stock universe: 1,613,706
- Results:
  - Number of cointegrated pairs found within clusters: 1,077
  - Fraction (on average) of pairs tested that cointegrated: 0.151424
  - Percentage of partially correlated pairs that cointegrated:
    - With a minimum partial correlation of 0.01: 25.3% (825 pairs)
    - With a minimum partial correlation of 0.1: 30.6% (15 pairs)

- With a minimum partial correlation of 0.4: 38.5% (5 pairs)

These statistics compare favorably or unfavorably to the benchmark data, depending on how many pairs we want in our portfolios, so there is no single conclusion to draw from the results. There appears to be a tradeoff between the reliability of the cointegrations and the number of stocks, so the best approach will depend on the size of the investor's portfolio and the investor's risk aversion.

The highest rate of cointegration I found in any of the tests I ran was 38.5%. I obtained that using the combination of Affinity Propagation together with a partial correlation requirement of 0.4. But this only yielded 5 cointegrated pairs! If I relax the partial correlation requirement to 0.01, I obtain 825 pairs, but the rate of cointegration falls to 25.3%. As explained above, the lower the cointegration rate, the higher the danger of multiple comparisons bias, which means that we can be much less certain of the quality of those 825 pairs than we can of the 5 pairs.

How does this compare to the benchmark data?

The cointegration success rates of 38.5% and 25.3% correspond roughly to the benchmark scores I obtained for the Energy and Technology sectors:

- Energy (78 stocks) produced 124 pairs with a cointegration rate of 38.4%
- Technology (152 stocks) produced 1,070 pairs with a cointegration rate of 26.36%

Since our goal is to achieve cointegration success comparable with a strategy that uses sector data, these results suggest that our approach may be serviceable.

The second benchmark was Jonathan Larkin's approach, which runs a DBSCAN clustering algorithm on principal components and fundamentals data derived from 1,101 stocks. It produced 1,906 cointegrated pairs, a much larger number than produced using our Project Strategy. However, the rate of cointegration success at 5.9% was very low. Our approach produces results that are clearly superior to those produced by Larkin's algorithm.

Result Data

Project Strategy:

- Covariance estimation and regularization technique: GraphLasso
- Clustering technique: Affinity Propagation
- Data clustered on: covariance matrix
- Stock universe: Entire market (1,797 stocks)
- Number of possible pairs in stock universe: 1,613,706
- Results:
  - Number of cointegrated pairs found within clusters: 1,077
  - Fraction (on average) of pairs tested that cointegrated: 0.151424

  - Percentage of partially correlated pairs that cointegrated:
    - With a minimum partial correlation of 0.01: 25.3% (825 pairs)
    - With a minimum partial correlation of 0.1: 30.6% (15 pairs)
    - With a minimum partial correlation of 0.4: 38.5% (5 pairs)

Benchmark Strategy A (clustering on principal components):

- Covariance estimation and regularization technique: *None*
- Clustering technique: DBSCAN
- Data clustered on: Principle components and stock fundamentals
- Stock universe: Intersection of Quandl and Intrinio data (1,101 stocks)
- Number of possible pairs in stock universe: 26,241
- Results:
  - Number of cointegrated pairs found within clusters: 1,906
  - Fraction (on average) of pairs tested that cointegrated: 0.059298

Benchmark Strategy B (clustering by industry sector):

All tests of sector data use the following:
- Covariance estimation and regularization technique: GraphLasso
- Clustering technique: Affinity Propagation
- Data clustered on: covariance matrix

Basic Materials
- Stock universe: Entire market (21 stocks)
- Number of possible pairs in stock universe: 210
- Results:
  - Number of cointegrated pairs found within clusters: 17
  - Fraction (on average) of pairs tested that cointegrated: 0.21342
  - Percentage of partially correlated pairs that cointegrated (from most restrictive to less restrictive):
    - With a minimum partial correlation of 0.4: 0% (0 pairs)

Cyclical Consumer
- Stock universe: Entire market (111 stocks)
- Number of possible pairs in stock universe: 6,105
- Results:
  - Number of cointegrated pairs found within clusters: 107
  - Fraction (on average) of pairs tested that cointegrated: 0.122808
  - Percentage of partially correlated pairs that cointegrated (from most restrictive to less restrictive):
    - With a minimum partial correlation of 0.4: 50% (2 pairs)

Energy

- Stock universe: Entire market (78 stocks)
- Number of possible pairs in stock universe: 3,003
- Results:
  - Number of cointegrated pairs found within clusters: 124
  - Fraction (on average) of pairs tested that cointegrated: 0.383974
  - Percentage of partially correlated pairs that cointegrated (from most restrictive to less restrictive):
    - With a minimum partial correlation of 0.4: 0% (0 pairs)
    - With a minimum partial correlation of 0.2: 40% (2 pairs)

Financials

- Stock universe: Entire market (315 stocks)
- Number of possible pairs in stock universe: 49,455
- Results:
  - Number of cointegrated pairs found within clusters: 88
  - Fraction (on average) of pairs tested that cointegrated: 0.10068
  - Percentage of partially correlated pairs that cointegrated (from most restrictive to less restrictive):
    - With a minimum partial correlation of 0.4: 0% (0 pairs)
    - With a minimum partial correlation of 0.2: 0% (0 pairs)
    - With a minimum partial correlation of 0.1: 4.5% (22 pairs)

Healthcare

- Stock universe: Entire market (82 stocks)
- Number of possible pairs in stock universe: 3,321
- Results:
  - Number of cointegrated pairs found within clusters: 196
  - Fraction (on average) of pairs tested that cointegrated: 0.106253
  - Percentage of partially correlated pairs that cointegrated (from most restrictive to less restrictive):
    - With a minimum partial correlation of 0.4: 0% (0 pairs)
    - With a minimum partial correlation of 0.2: 0% (0 pairs)
    - With a minimum partial correlation of 0.1: 34.6% (9 pairs)

Industrials

- Stock universe: Entire market (157 stocks)
- Number of possible pairs in stock universe: 12,246
- Results:
  - Number of cointegrated pairs found within clusters: 104
  - Fraction (on average) of pairs tested that cointegrated: 0.112479
  - Percentage of partially correlated pairs that cointegrated (from most restrictive to less restrictive):
    - With a minimum partial correlation of 0.4: 0% (0 pairs)
    - With a minimum partial correlation of 0.2: 11.1% (9 pairs)

Non-cyclicals

- Stock universe: Entire market (62 stocks)
- Number of possible pairs in stock universe: 1,891
- Results:
  - Number of cointegrated pairs found within clusters: 26
  - Fraction (on average) of pairs tested that cointegrated: 0.088492
  - Percentage of partially correlated pairs that cointegrated (from most restrictive to less restrictive):
    - With a minimum partial correlation of 0.4: 0% (0 pairs)
    - With a minimum partial correlation of 0.2: 0% (0 pairs)
    - With a minimum partial correlation of 0.05: 8.1% (8 pairs)

Technology

- Stock universe: Entire market (152 stocks)
- Number of possible pairs in stock universe: 11,476
- Results:
  - Number of cointegrated pairs found within clusters: 1070
  - Fraction (on average) of pairs tested that cointegrated: 0.263577
  - Percentage of partially correlated pairs that cointegrated (from most restrictive to less restrictive):
    - With a minimum partial correlation of 0.4: 0% (0 pairs)
    - With a minimum partial correlation of 0.2: 0% (0 pairs)
    - With a minimum partial correlation of 0.1: 71.4% (5 pairs)

Telecom

- Stock universe: Entire market (14 stocks)
- Number of possible pairs in stock universe: 91
- Results:
  - Number of cointegrated pairs found within clusters: 0
  - Fraction (on average) of pairs tested that cointegrated: 0
  - Percentage of partially correlated pairs that cointegrated (from most restrictive to less restrictive):
    - With a minimum partial correlation of 0.4: 0% (0 pairs)

Utilities

- Stock universe: Entire market (57 stocks)
- Number of possible pairs in stock universe: 1,596
- Results:
  - Number of cointegrated pairs found within clusters: 26
  - Fraction (on average) of pairs tested that cointegrated: 0.022817
  - Percentage of partially correlated pairs that cointegrated (from most restrictive to less restrictive):
    - With a minimum partial correlation of 0.4: 0% (0 pairs)