Pairs Identification for Statistical Arbitrage - Pair Trading

Team#037 DataVizards - smeduri7, pbaikani3, sv67, reshed6, nbadri3, pnarula6

Introduction and Motivation

The strategy behind pair trading is to identify a pair of cointegrated securities (e.g. stocks). The assumption is that cointegrated securities will maintain their relationship and the prices will revert to their historical trend. This allows a trader, once the spread (the difference between the prices) widens to long (buy) the asset whose price decreased and short (sell) whose price increased. Pair trading requires three steps - Pairs formation, cointegration testing and trading strategy. **The project is mainly focused on the first two steps.**

Currently, the pairs formation process begins with running cointegration tests on all possible pairs and eliminating pairs that fail the tests. This method is inefficient, computationally expensive and not feasible for testing stocks within different markets. Another pairs formation method is rule-based which is subjective and requires industry and market expertise. We plan to develop a cross-market pairs formation model that is efficient and eliminates industry, domain and market expertise required by the currently used methods.

Target Audience:

The project is mainly targeted towards hedge funds, institutional and retail investors, quantitative analysts and those who deploy automated/algorithmic trading models.

Approach

The purpose of the project is to find potential pairs of stocks using the "cheapest" computational method while eliminating any subject matter expertise required to perform such initial pairs formation. Starting with a list of all U.S. stocks, we find which companies are mentioned in financial news together with a company selected by the user to form a network graph. **The assumption is that companies mentioned in the same news articles are potential pairs for cointegration testing.**

Tickers Network Graph:

The nodes of the network are the company tickers and an edge is formed between two nodes if they appear in the same article. The edge weight corresponds to the number of times the two tickers were co-mentioned in the articles. By starting with an initial ticker selected by the user, we expand the network by a branching factor of b (≈ 5) with highest weights until a depth based on the user interactions with the graph. The user can further select from the displayed tickers any two nodes to view lines chart comparing stock prices, spread line, cointegration "score" and correlation which will be presented analyzing historical stocks prices within a timeframe determined by the user.

Pairwise Cointegration Test:

To test for a cointegration between two stocks we use the two steps Engle-Granger test which is a regression based approach that states that given two non-stationary series X_t and Y_t, if there exists α and β such that the residuals Ɛ (which are referred to as spread) of the linear combination $Y_t = \alpha + \beta X_t + \mathcal{E}$ are stationary, then the variables X_t and Y_t are said to be cointegrated and the state of $Y_t - X_t - \alpha$ is stationary. The first step is to regress one stock on the other and the second step is to test the residuals for stationarity using the Augmented Dickey Fuller (ADF) test. A stationary time series is one whose statistical properties such as mean, variance, autocorrelation, etc. are all constant over time.

Data

Stocks tickers & historical prices:

A complete list of U.S. tickers (~20,000) using a free API was retrieved from EOD Historical Data. The historical stock prices is fetched using the yfinance Python library.

The news article data was collected using a paid API offered by newsfilter.io. The news data for the past 2 years was collected for each company from the list mentioned above in an AWS instance and stored in a S3 bucket. Later this data was processed and stored in a SQL database with the following schema: "article_id, ticker#1, ticker#2, publication_date."

The adjacent nodes for a node (stock asset ticker#1) in the graph network are the stocks co-occurring with that node's stock asset ticker#1. It can be found by filtering all the articles for ticker#1 and grouping them by ticker#2 and sorting by the number of ticker#2 records. The top 5 tickers were selected and edges are formed in the graph network. Retrieving news data for the past 2 years, the dataset contains ~29 million ticker#1-ticker#2 co-occurrence instances (2.6 GB disk size).

Experiments and results:

We evaluate results based on whether a strong cointegration exists within pairs of stocks mentioned in the same article.

Another factor assessing results is whether the tool is able to find cointegrated pairs that are most likely to be overlooked when testing using a rule-based approach.

For example, Gaming and Leisure Properties Inc (GLPI) is a gaming-focused real estate investment trust while Atlantica Sustainable Infrastructure PLC (AY) is a sustainable infrastructure company that owns and manages renewable energy, efficient natural gas, transmission and transportation infrastructures and water assets. However, as indicated by the tool that the two stocks might be cointegrated, their cointegration score (within 01/01/2020 and 11/01/2020) is 98.4%.

We believe it is fair to assume that using the traditional current methodologies to find the cointegrated pair mentioned above will only be possible if cointegration tests would be performed on all possible pairs, which is inefficient and computational "expensive."

As expected, the tool definitely indicates "obvious" pairs for potential cointegration. For example, when analyzing for Exxon Mobil Corporation (XOM), the tool indicates to potentially pair it with British Petroleum (BP). When testing for cointegration, we see that the two stocks are cointegrated with a score of 94.4% between the timeframe of 01/01/2020 - 11/01/2020 and 99.8% between 01/01/2019 - 11/01/2020.