Statistical Arbitrage by Pair Trading using Clustering and Machine Learning

Anonymous
Department of XXX
University of XXX
City, Province, Zipcode
[example@email.com]

Anonymous
Department of XXX
University of XXX
City, Province, Zipcode
[example@email.com]

Abstract

In this project, we study the application of machine learning methods to find statistical arbitrage opportunities in the U.S stock market using pair trading strategy. Pair trading is a trading strategy that takes long (buy) position and matched short (sell) position of two stocks with high correlation. The first step in the project is to distinguish the stock pairs with high correlation based on clustering methods. The second step in the project is to apply neural network such as LSTM to predict trading signals of a given stock pair. We use Quantopian for backtesting, financial data, and evaluating the performance of different machine learning methods.

1 Introduction or Problem statement

1.1 Topic

Pair trading is one of the most popular trading strategies since 1980 for finding statistical arbitrage opportunities in the stock market. The logic behind pairs trading is to trade pairs of stocks belonging to the same industry or having similar characteristics, such that their historical returns track each other and are expected to continue to do so in the future.

Pairs trading is still widely used by investment companies. In this project we aim to use machine learning methods to improve on this trading algorithm. The goal of the project is to exploit inefficiencies in the market which are increasingly harder to find. We hope to find arbitrage by incorporating more machine learning models in the clustering process and prediction process of our trading algorithm.

1.2 Contribution

We will experiment with getting our time-series data to be stationary by using fractional-differencing, which is better at preserving memory of the time-series data than first-differencing. Memory is typically lost by first-differencing the data, yet almost all papers still do this. Another novelty we incorporate in our paper involves using non-symmetric upper/lower barriers and stop loss in our trading strategy for determining when mean-reversion of the stock spreads will occur. In addition, we will broaden our search for stock pairs by investigating a larger universe of stocks than most other papers.

In order to make improvements on other existing papers and related works, we will perform purged K-fold Cross-validation, which drops the first and last few observations of the training set so as to avoid serial correlation in the time-series data. Also, if necessary, we will combine NLP and

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

ML in the stock clustering process to find the pair stocks. Most papers will only do one or the other, so we can provide a new perspective on the application of machine learning methods to pair trading.

2 Method

We will use the following machine learning methods including:

- Clustering (e.g. PCA)
- Prediction (e.g. Neutral Network, LSTM), compare to Feedforward Neural Network.

3 Related work

Below we list five references related to pair trading, compare any similarities and differences between their work and our project.

1. *Statistical Arbitrage Trading with Implementation of Machine Learning, Hakon Andersen & Hakon Tronvoll*
   - Look for arbitrage opportunities in the Norwegian Market.
   - PCA and density-based clustering to cluster stocks (DBSCAN).
   - Use cointegration to identify mean-reversion and weak stationarity
   - Conclusion: Pairs trading does not provide excess return or favorable Sharpe ratio.
   - Compare to unsupervised machine learning model
2. *Pair Trading: Clustering Based on Principal Component, Rafael Govin Cardoso*
   - PCA for clustering compared to clustering by industry groups.
   - Data from 3 emerging markets: Brazil, South Africa, and India.
   - Use AIC to determine the number of lag terms
3. *Pairs Trading, Convergence Trading, Cointegration, Daniel Herlemont*
   - Pair selection using Co-integration
   - Dickey-Fuller test for mean reversion
   - Only consider pairs in the same sector
   - Boundaries are 2-rolling standard deviations, position is opened when the boundary is hit twice. Close position when ratio hits the mean.
4. *Cluster-Based Statistical Arbitrage Strategy, Anran Lu, Atharva Parulekar, Huanzhong Xu*
   - Assign stock-ETF pairs a score based on cointegration test (Johansen's test).
   - PCA/K-means for clustering, but cointegration statistics seem to outperform.
   - Use LSTM to predict trading signals. Compare performance to AR-1.
   - 40-day trading window
   - Data from Quantopian and Yahoo Finance
5. *Pairs Trading Using Machine Learning: An Empirical Study, R.W.J. van der Have*
   - Pair selection using Cointegration
   - Pairs Trading using Neural Network
   - Performance Metrics used: Sharpe Ratio, Sortino Ratio, Max-Drawdown
   - Data, all ETFS on the NYSE from Datastream database.
   - Restrict pairs within sectors
6. Our Analysis
   - We consider stocks in the U.S Russell 3000
   - We will include a feature selection process during clustering
   - In addition we will incorporate elements of NLP to aid in clustering
   - We will incorporate cross-validation to train our LSTM prediction model
   - Our data will come from Quantopian, Yahoo Finance, and Bloomberg
   - We will use 10 years worth of data, with 2-year testing period and 8-year train-validation set.

4 Plan of Experiments

4.1 Dataset and Software

We will be using real-word financial time-series data from Quantopian, Yahoo Finance, and Bloomberg. The data contains both stock price data including High, Low, Adj.Close, Open, and volume data. We also have access to company fundamental data such as earnings per share, price volatility, and enterprise value to find valid, robust eligible pairs.

Our plan is to implement the model in Python using packages such as `numpy` and `scikit-learn`. We will also use the Quantopian backtesting platform for evaluating our model.

4.2 Challenges

Our trading strategy is well-known and there is a good chance other traders in the market have already exploited it.

Since the advent of machine learning, statistical arbitrage through pairs trading has become increasingly harder to achieve. We will have to push the limits of our data and models in order to make financial gains.

4.3 Evaluation

One intuitive way to evaluate a trading strategy is by the following indicators:

- Sharpe Ratio: $SR_i = \frac{R_i - R_f}{\sigma_i}$. Here, R_i denotes the return of a stock pair, and R_f the risk-free rate. The standard deviation of a pair is denoted with σ_i . The Sharpe ratio is measures the risk premium per unit of risk

- Annualized Return of Portfolio

- Standard Deviation of Portfolio

- alpha: $\alpha_i = E[R_i] - R_f - \beta_i (E[R_m] - R_f)$. Where $\beta_i = \frac{Cov(R_i, R_m)}{\sigma_M^2}$ and $E[R_m]$ denotes the expected market return. The alpha tells us how much better or worse our pairs trading strategy performed relative to its benchmark.

- Max Drawdown

We will be comparing the indicators above through different models / strategies with the base level which is set in advance.

5 Plan of Project

5.1 Division of Work

Two people will work on clustering and filtering the universe of stocks. They will also work on feature-selection.

Two people will work on applying machine learning algorithms on prediction and trading pairs of stocks such that financial gain can be achieved on a test set.

5.2 Feasibility Analysis

The backtesting platform for our trading strategy is already built on Quantopian and data is readily available which does not require much processing and cleaning. In this case we are able to focus primarily on clustering and prediction. The most challenge we would expect that statistical arbitrage would be difficult to achieve, especially now compared to 5 years ago.