MS&E 448: Group 6

Grant Avalon | Irene Jeon
Michael Becich | Sreyas Misra
Vincent Cao | Liezl Puzon

Multi-factor Statistical Arbitrage Model

Overview

1. Background

2. Data Inputs

3. Methods

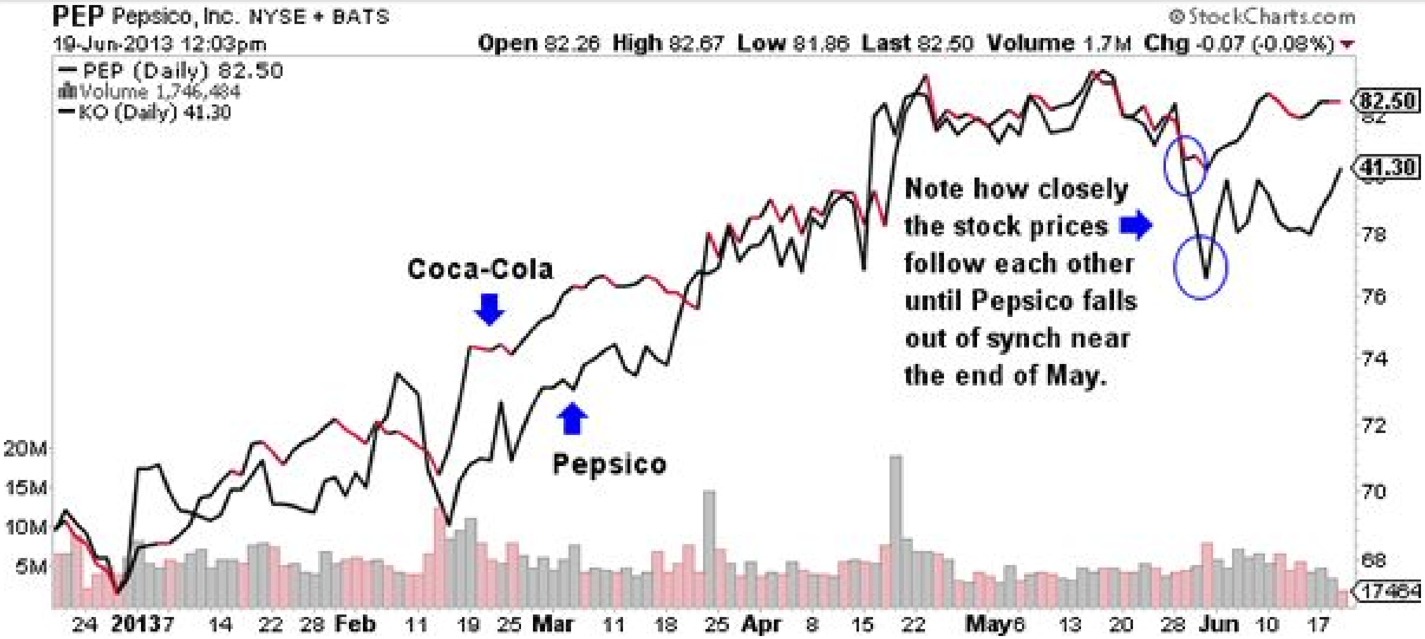
4. Trading Algorithm

5. Preliminary Results

6. Discussion

Background: Statistical Arbitrage

Stat. Arb. exploits "mispricings" between mean-reverting pairs or baskets of stocks.



Classic stat arb. identifies pairs of stocks based on how their prices stay together.

Can we pair stocks using not just stock prices/returns but also stock fundamentals?

Multi-Factor Statistical Arbitrage

- Using only price/returns data creates unstable clusters that are exposed to market risks and don't persist well over time.
- By incorporating other stock time-series data like fundamentals (P/E ratio, revenue growth, etc.), we can create stabler stock clusters.
- Use a modified O-U process to model mean-reversion in case pairs cease to be cointegrated

Background: Model Design

PCA is performed twice: once for returns, once for fundamental factors

Lower-Dimensionality Reduction

$$\mathbf{x}_j \approx \bar{\mathbf{x}} + \sum_{i=1}^{i=k} g_{ji}\mathbf{e}_i$$

Highest Variance 1st PC

$$\mathbf{w}_{(1)} = \arg\max \left\{ \frac{\mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w}}{\mathbf{w}^T\mathbf{w}} \right\}$$

$$r_i = \sum_{k=1}^{K} B_{k(i)} * PC_k + \varepsilon_i$$

Regressed

$$\varepsilon_i = \sum_{f=1}^{F} B_{f(i)} * PC_f + residual_i$$

Data Inputs: Incorporating Time-Varying Data

So far, we studied the S&P 500 stock index with time series data going back 5 years.

S&P 500 Stock Log Returns

Return Time Series (5Y)

$$\begin{array}{cc} \text{N = 446 Stocks} & \begin{bmatrix} \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix} \end{array}$$

(Google Finance) Python scraper [1]

S&P 500 Fundamental Factors

Factor Time Series (5Y)

$$\begin{array}{cc} \text{N*F Stacked Factors} & \begin{bmatrix} \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix} \end{array}$$

Not Yet Implemented!*

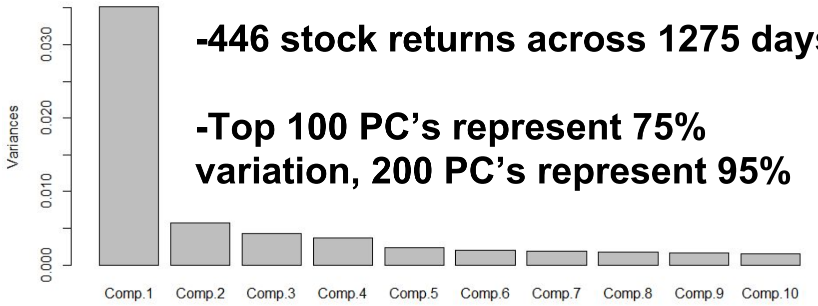
****Could be broken into separate PCA's if difficult to meaningfully normalize***

[1] https://github.com/liezl200/stockScraper

Principal Component Analysis (PCA) & K-means Clustering

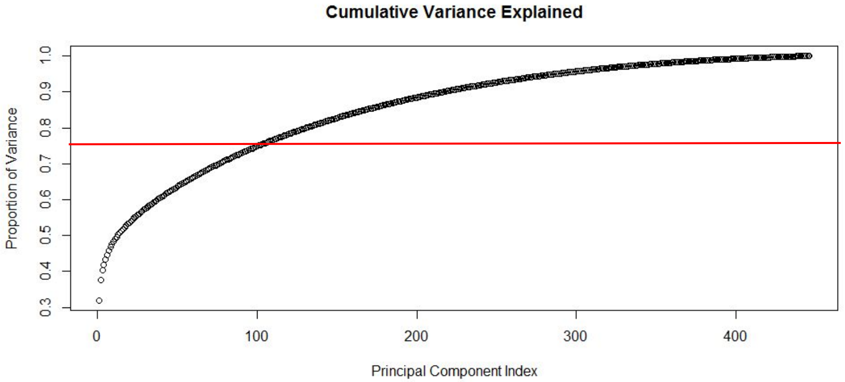
To reduce dimensionality in noisy system and pre-process groups by largest-variance PC's

PCA (Accounting for Variance)

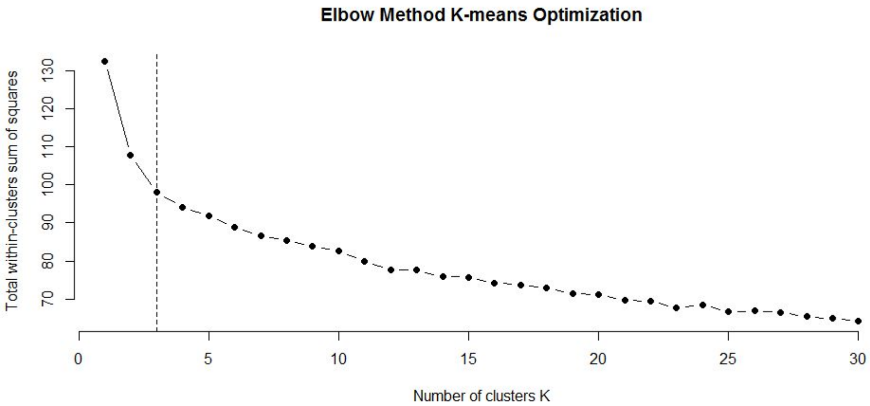


-446 stock returns across 1275 days

-Top 100 PC's represent 75% variation, 200 PC's represent 95%



K-means (Elbow Method for Optimal K)

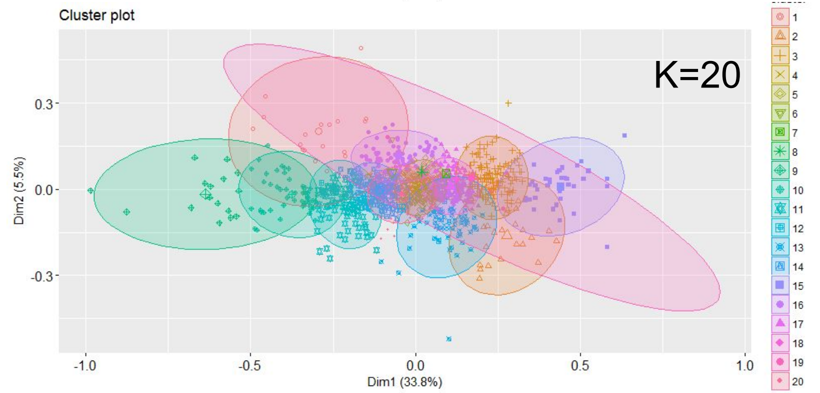


-Elbow Method recommends K=3 for lowest error (SSE) drop

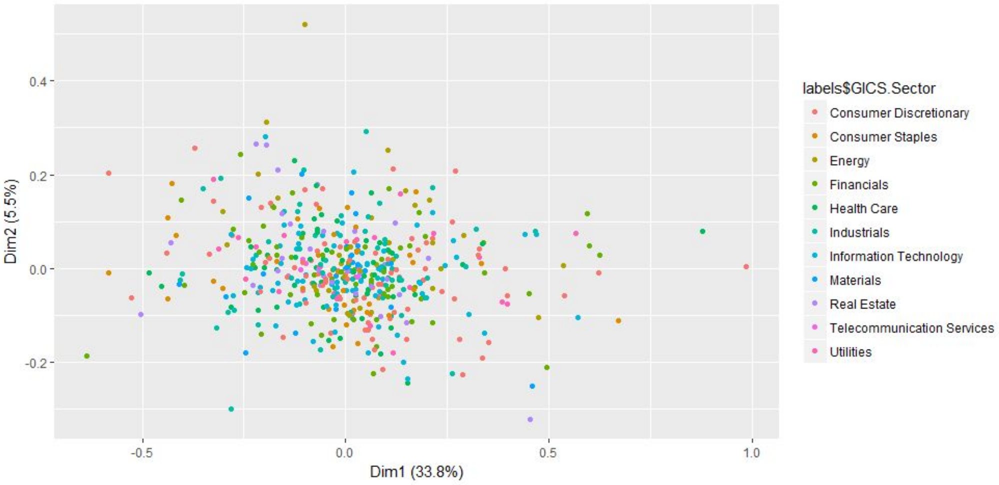
-Not enough specificity to differentiate sectors of market (K=20 used)

Clustering Results

Pairwise PC-analysis revealed cluster separation, but poor correlation to industry sectors



Correspondence to GICS Sector



Trading Algorithm: Co-integrated Stock Pairs*

Pairs were identified such that each pair was in the same K-means cluster

For each cluster i where 1 < size i < 30...

IF stock 1 and 2 individually pass Augmented Dickey Fuller Test *< checks if both stocks are integrated*

*AND IF pair(1,2) passes Engel-Granger Test bidirectionally** * *< checks if the pair is co-integrated*

THEN Stock 1 and 2 are pairs with reversion half life ln(2)/B

** Done using MATLAB econometrics toolbox*
***performs test with both stocks as regressor*

Our best pairs have the smallest min(E-G p-value) and fastest reversion speeds.

Trading Algorithm: Execution

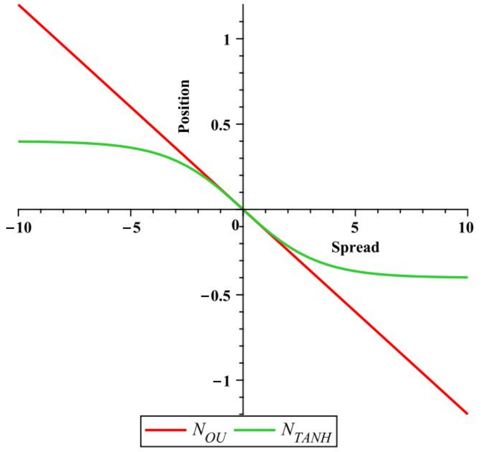
Mean-Reversion was modelled as an O-U process

For each cointegrated pair...

Calculate parameters of O-U Process through Maximum Likelihood Estimation

Using said parameters and current mispricing, find proportion of portfolio of optimal position

If mispricing goes beyond certain threshold, begin unwinding position



$$dX_t = \alpha\left(\mu - X_t\right)dt + \sigma\, dW_t$$

$$N_{OU} = \left(\frac{-k(S-\overline{S}) - rS}{\sigma^2}\right)W.$$

Unwinding partially protects from the risk that our pair ceases to be cointegrated.

Trading Algorithm: Trade Conditions

For each cointegrated pair…

trade if these conditions are met:

- Trade N minutes before closing each day (N = 30 minutes)
- Only run the trading logic at 3:30PM Eastern Time, which 30-minutes before market closes
- If spread is within a certain range, allocate capital to pairs trade

Preliminary Results

Top 5 pairs to examine

- Top 5 cointegrated pairs (smallest p-value && largest Beta):
  - JPM and PBCT [JP Morgan (*Financial*) & People's United Financial (*Info Tech*)]
  - BCR and XRAY [Bard (*Health Care*) & Dentsply Sirona (*Health Care*)]
  - BBBY and SPLS [Bed, Bath & Beyond (*Consumer Discretionary*), Staples (*Consumer Discretionary*)]
  - SCHW and HBAN [Charles Schwab (*Financials*) & Huntington Bancshares (*Financials*)]
  - BCR and SYK [Bard (*Health Care*) & Stryker Corporation (*Health Care*)]

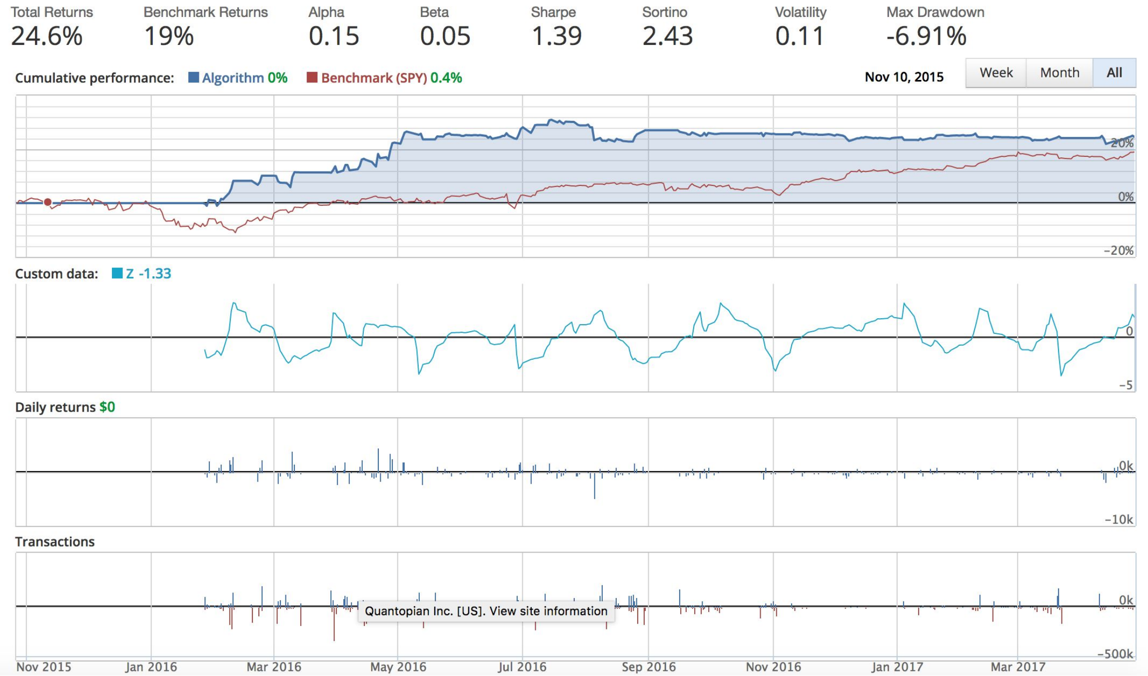
X1 <chr>	X2 <chr>	X3 <dbl>	X4 <dbl>	X5 <dbl>
'JPM'	'PBCT'	0.0010000	-39.91300	6.63600
'BCR'	'XRAY'	0.0010000	-110.56000	5.39840
'BBBY'	'SPLS'	0.0010000	10.36700	4.02260
'SCHW'	'HBAN'	0.0010000	-10.47900	3.85430
'BCR'	'SYK'	0.0010000	-21.95500	2.09440

Even though our clusters aren't very industry correlated, our pairs are very similar companies.

Performance Results

For the pair with smallest p-value and largest Beta

Performance graph for JPM and PBCT (2015-10-27 to 2017-04-27, 60 day trailing window)



Discussion

Important questions to answer by further fine-tuning

Returns + Fundamentals

Types of factors?

What time-range?

Weighted sampling?

PCA + Clustering

Cluster size?

What do our PCs actually represent?

How many PCs should we be using?

What types of clusters should be eliminated?

Cointegrated Pairs

Types of Cointegration tests?

What mean-reversion speeds are best?

Trading Algorithm

Look-back windows?

How to work trades in?

Other indicators to initiate unwinding (social media volatility)

Trading signal thresholds?

Future Directions

How can we improve this algorithm?

Immediate Next Steps:

- Improve PCA/K-means clustering (silhouette scores) to better match industry sectors
- Determine optimal time-intervals to re-cluster data
- Generalizing this algorithm into a class to pair trade more than one pair
- Condense the stocks in the S&P 500 to look at more interesting ones
- Figure out how to scrape fundamental factor data
  - Which factors to choose to get most meaningful results

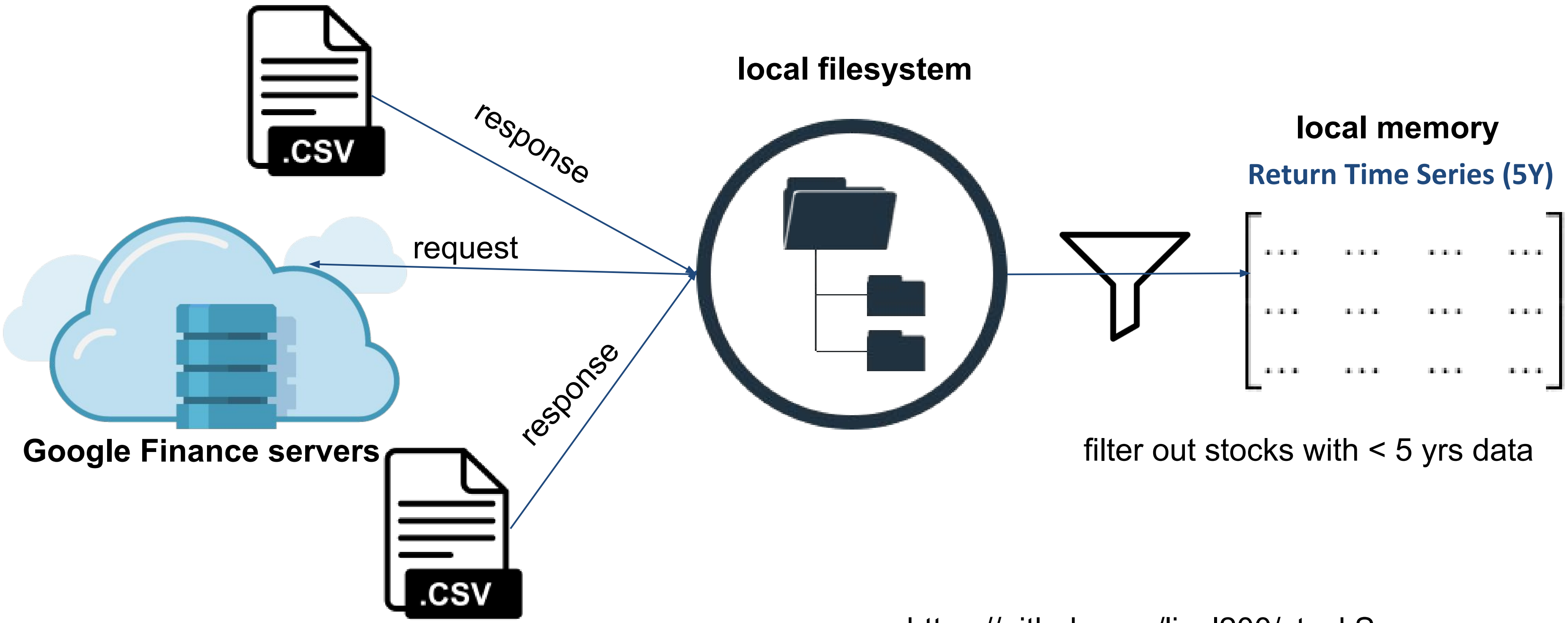
Thank you! Questions?

References:

- http://ieeexplore.ieee.org/document/6007312/?reload=true
- https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1617662
- https://www.math.nyu.edu/faculty/avellane/AvellanedaLeeStatArb071108.pdf
- https://www.linkedin.com/pulse/statistical-arbitrage-strategy-r-jacques-joubert
- https://cran.r-project.org/web/packages/egcm/egcm.pdf
- https://arxiv.org/pdf/1405.2384.pdf
- http://www.ewp.rpi.edu/hartford/~youneh/INVII/Week%204/capm_2up.pdf
- https://www.quantopian.com/posts/pair-trade-with-cointegration-and-mean-reversion-tests
- https://www.quantopian.com/posts/statistical-arbitrage-on-returns-using-pca

Data Scraper: Software Architecture

So far, we studied the S&P 500 stock index with time series data going back 5 years.



https://github.com/liezl200/stockScraper

Discussion/Analysis

How to use results to build a dynamic trading strategy