Statistical Arbitrage in Small Cap U.S. Stocks: MS&E448 Final Report

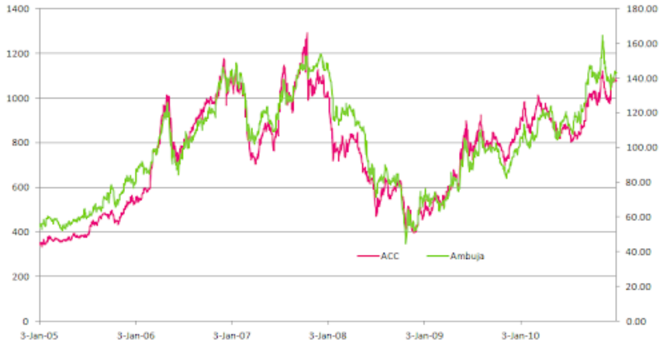
Trishiet Ray, Brian Seabrooks

1 Strategy & Background

1.1 Strategy Overview

We apply modified versions of PCA-based mean-reversion statistical arbitrage techniques described in ”Statistical Arbitrage in the U.S. Equities Market” on Russell2000 equities data from 2018 to 2021. We use a variable medium-term holding period on the order of days to months. On each trading day we compute with a 60-day trading window a standardized returns matrix, a correlation matrix, apply principal component analysis to generate a variable number of risk factors, and finally computed s-scores that act as trading signals. We expected a Sharpe ratio between 1 and 2.5. Our strategy achieved a Sharpe ratio of 1.238 compared to the index’s Sharpe ratio of 0.65, indicating that our strategy has potential for real world alpha generation.

1.2 Mean Reversion & Pairs Trading



We utilize the popular investment strategy of statistical arbitrage, which relies on the principle of mean reversion. This is the idea that the returns of two assets that whose returns are highly correlated are after adjusting for beta (typically because they have similar characteristics or are in the same industry, such as Visa and Mastercard stocks) will revert around a mean value. When assets are overpriced (relative to another stock or index), we short them, and when they are underpriced we buy them. This is known as pairs trading; for each we trade we enter we long one equity and short another.

Let P and Q represents correlated stocks. Let P_t and Q_t represent their corresponding price time series. We model the system as follows:

$$\frac{dP_t}{P_t} = \alpha dt + \beta \frac{dQ_t}{Q_t} + dX_t \tag{1}$$

where the α is a drift term (the change of the average value of a stochastic process), and X_t is a mean reverting process. We assume changes in α is small enough relative to changes in X_t that we can ignore it. We model X_t with a parametric Ornstein–Uhlenbeck (OU) process:

$$dX_i(t) = \kappa_i(m_i - X_i(t))dt + \sigma_i dW_i(t) \tag{2}$$

1.3 Statistical Arbitrage

Our goal in statistical arbitrage is to generate a collection of pair trades of equities relative to factors that explain systematic returns. Here we ignore net positions in the index (which are expected to cancel out) and generate a long/short portfolio of single stocks. The net decomposition of any stock into systematic and idiosyncratic components is expected to look like:

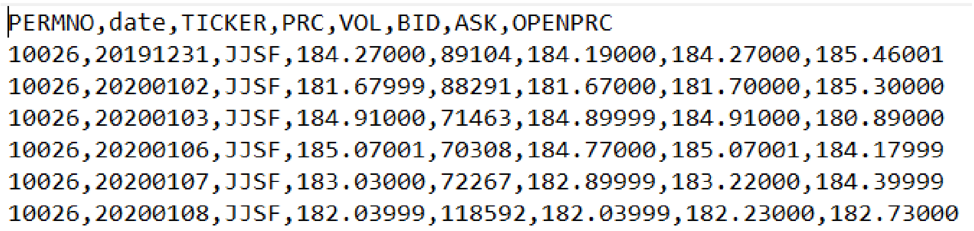
$$\frac{dP_t}{P_t} = \alpha dt + \sum_{j=1}^{n} \beta_j F_t^{(j)} + dX_t \tag{3}$$

where $F_t^{(j)}$ is the return of the jth systematic risk factor associated with the Russel2000 market. This model allows us to separate out the systemic component of returns so that we can focus on the idiosyncratic component, which is modeled by a (pure) mean-reverting process. We used PCA to generate these risk factors, as described below. We expected a Sharpe ratio between 1 and 3 with this strategy.

2 Data

2.1 Dataset

The dataset we were using was the Russell 2000 US equities data from 2018 - 2021 from the Center for Research in Security Prices (CRSP), downloaded via Wharton Research Data Services. We filtered for columns including ticker id, date, ticker name, end-of-day trading price, volume, bid, ask, and open price. An example cross-section of the data file is below:



```
PERMNO,date,TICKER,PRC,VOL,BID,ASK,OPENPRC
10026,20191231,JJSF,184.27000,89104,184.19000,184.27000,185.46001
10026,20200102,JJSF,181.67999,88291,181.67000,181.70000,185.30000
10026,20200103,JJSF,184.91000,71463,184.89999,184.91000,180.89000
10026,20200106,JJSF,185.07001,70308,184.77000,185.07001,184.17999
10026,20200107,JJSF,183.03000,72267,182.89999,183.22000,184.39999
10026,20200108,JJSF,182.03999,118592,182.03999,182.23000,182.73000
```

2.2 Data Issues

A small problem we had with the data was that 72 out of the 2000 tickers were missing values returns values. Upon manually looking up a couple of these tickers, it was determining that they did not go bankrupt but simply had data omissions - we excluded these stocks from our universe. We also tried obtaining obtaining options data for each ticker to incorporate volatility in signal generation, but had issues matching the ids, so we abandoned the idea given time constraints.

3 Investment Universe Selection

3.1 Asset/Index Selection

For our paper, we focus on the US equity market due to availability of data, prior research, and applicability to us as amateur investors in equity. Initially we were working with the S&P500 Index; however, we ended up using the Russell 2000 equities index. We believed that focusing on small-capitalization stocks could potentially provide an edge due to crowding in stat arb strategies on large cap stocks. We exclude stocks missing data in our date range.

4 Modeling

4.1 Standardized Returns Matrix

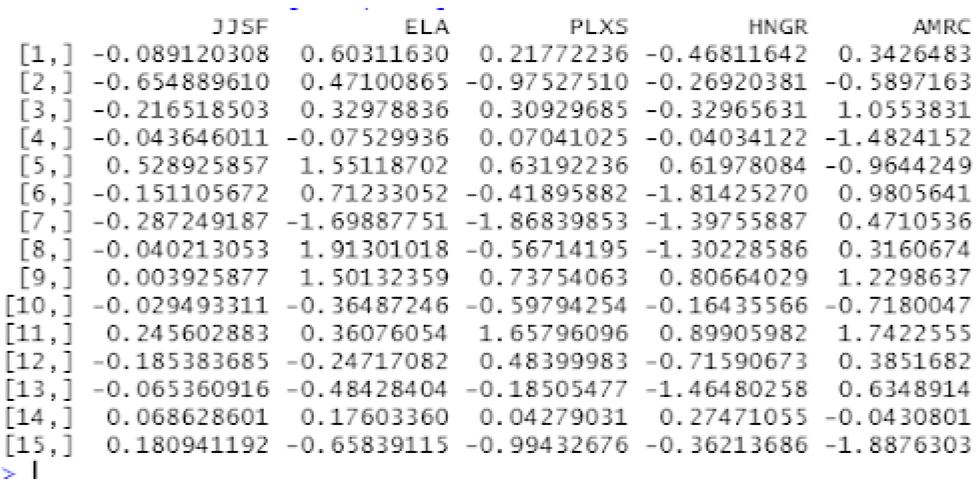
We use historical stock price data on a cross-section of N=1928 stocks for the last M=60 days. On each trading day, we construct a standardized returns matrix by first computing daily returns:

$$R_{ik} = \frac{S_{i(t_0-(k-1)\Delta t)} - S_{i(t_0-k)\Delta t)}}{S_{i(t_0-k\Delta t)}} \tag{4}$$

where S_{it} is the price of stock i at time t adjusted for dividends and $\Delta t = 1/252$ to model one trading day. The standardized returns are given by:

$$Y_{ik} = \frac{R_{ik} - \bar{R}_i}{\bar{\sigma_i}} \tag{5}$$

where $\bar{R}_i$ is the mean return of stock i across the trading window and $\bar{\sigma_i}$ is the standard deviation of the return of stock i across the trading window. A cross-section of a sample standardized returns matrix is shown below:



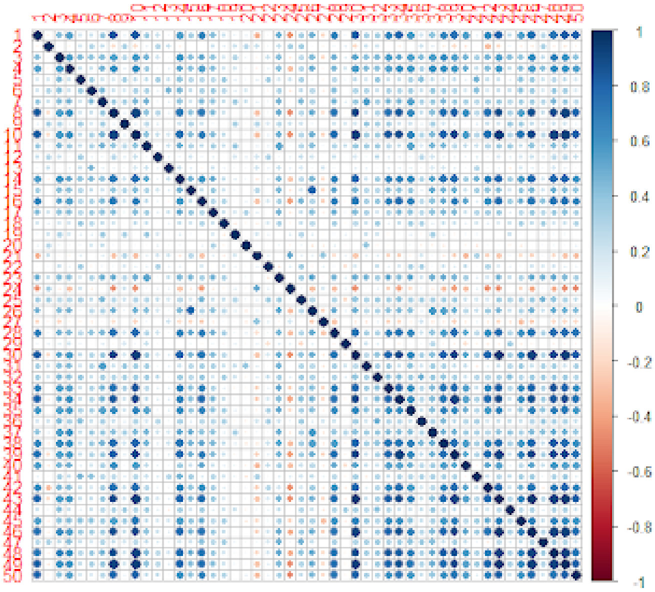
```
             JJSF         ELA        PLXS        HNGR       AMRC
 [1,] -0.089120308  0.60311630  0.21772236 -0.46811642  0.3426483
 [2,] -0.654889610  0.47100865 -0.97527510 -0.26920381 -0.5897163
 [3,] -0.216518503  0.32978836  0.30929685 -0.32965631  1.0553831
 [4,] -0.043646011 -0.07529936  0.07041025 -0.04034122 -1.4824152
 [5,]  0.528925857  1.55118702  0.63192236  0.61978084 -0.9644249
 [6,] -0.151105672  0.71233052 -0.41895882 -1.81425270  0.9805641
 [7,] -0.287249187 -1.69887751 -1.86839853 -1.39755887  0.4710536
 [8,] -0.040213053  1.91301018 -0.56714195 -1.30228586  0.3160674
 [9,]  0.003925877  1.50132359  0.73754063  0.80664029  1.2298637
[10,] -0.029493311 -0.36487246 -0.59794254 -0.16435566 -0.7180047
[11,]  0.245602883  0.36076054  1.65796096  0.89905982  1.7422555
[12,] -0.185383685 -0.24717082  0.48399983 -0.71590673  0.3851682
[13,] -0.065360916 -0.48428404 -0.18505477 -1.46480258  0.6348914
[14,]  0.068628601  0.17603360  0.04279031  0.27471055 -0.0430801
[15,]  0.180941192 -0.65839115 -0.99432676 -0.36213686 -1.8876303
> 
```

4.2 Correlation Matrix

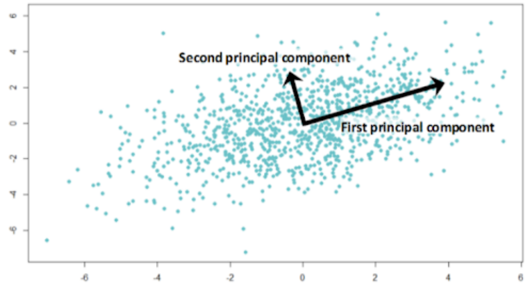
Next, we form a correlation matrix, computing the correlation of each ith stock's return to all other stocks' returns in the trading window, defined as:

$$\rho_{ij} = \frac{1}{M-1} \sum_{k=1}^{M} Y_{ik} Y_{jk} \tag{6}$$

The parameter M, the trading window, is sensitive; we chose it to be 60 days based on prior research so that you get a complete enough picture of the past to form relevant relations, while leaving behind data that is economically irrelevant. The cross-section of a sample generated correlation matrix is depicted below; larger dots indicate larger magnitude of correlation, blue dots indicate positive correlation, and red dots indicate negative correlation.

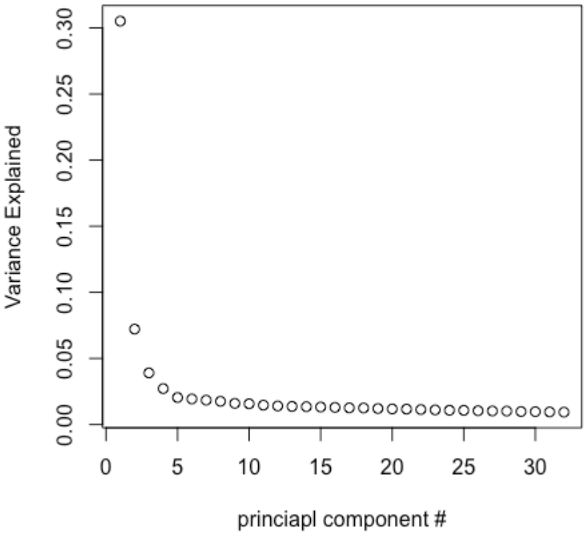


4.3 Principal Component Analysis



The principal component analysis is the process of computing a set of orthonormal eigenvectors for a set of data points; each eigenvector is the direction of a line that minimizes the average squared distance between the points and the line. PCA-based risk factors are known to be economically significant and not biased towards large capitalization equities. The risk factors generated can be viewed as long-short portfolios of industry sectors.

Using PCA to create the factors has advantages: the first is that it allows us to arrive at a set of uncorrelated factors, and the second is that it does not require us to make assumptions about the factors that drive stock returns, allowing us to design our factors more empirically.



Above we demonstrate the variance in the data explained by each our principal components. There are trade-offs in using choosing how many factors to use; using more factors results in a more descriptive risk model with lower variance and high bias, which does lower the opportunity of profit (especially when considering transaction costs). Using less components results in less mean-reversion and higher residual volatility. We use a variable number of principal components; We chose to set a threshold for variance explained as between 65% and 85% and use that number of principal components.

4.4 Factor Model

For each jth eigenvector (ranked by eigenvalue in decreasing order), we then compute the corresponding eigenportfolio:

$$Q_i^{(j)} = \frac{v_i^{(j)}}{\bar{\sigma_i}} \tag{7}$$

where Q_i represents the dollar amount invested in stock i. The corresponding eigenportfolio returns are:

$$F_{jk} = \sum_{i=1}^{N} Q_i^{(j)} R_{ik} \tag{8}$$

These factors F_{jk} represent returns of benchmark portfolios representing systematic risk factors. Now we can break down any stock into a projection on the m factors and the residual term X_t.

Note that:

$$\bar{\beta_j} = \sum_{i=1}^{N} \beta_{ij} Q_i = 0 \tag{9}$$

So the trading portfolio is market-neutral. This means that our portfolio is uncorrelated with the PCA-obtained factors explaining the market return, and is only driven by idiosyncratic returns.

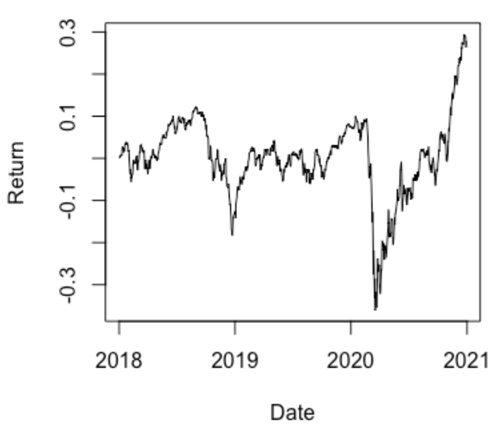
4.5 Parameters

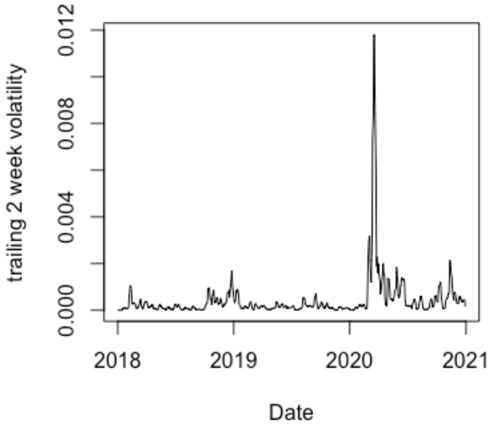
In establishing our parameter values, we used values utilized in prior research works as well as manually played around with values with the data from 2018. Since we are using data from prior to t to estimate the residual process X_t, our simulation technique was out-of sample.

Parameter Name	Value
Look Back Period, M	60 days
Threshold of Variance Explained by PCs	65% - 85%
% of stocks in universe for trading	25% of stocks, w/ prediction errors ≈ 0
OU Lag	1 (AR-1 model)
Proportionality Factor λ	1
S-score cutoff for entering position, s_o	1.25
S-score cutoff for closing position, s_c	0.75
Transaction Cost	0.005%

4.6 Benchmark

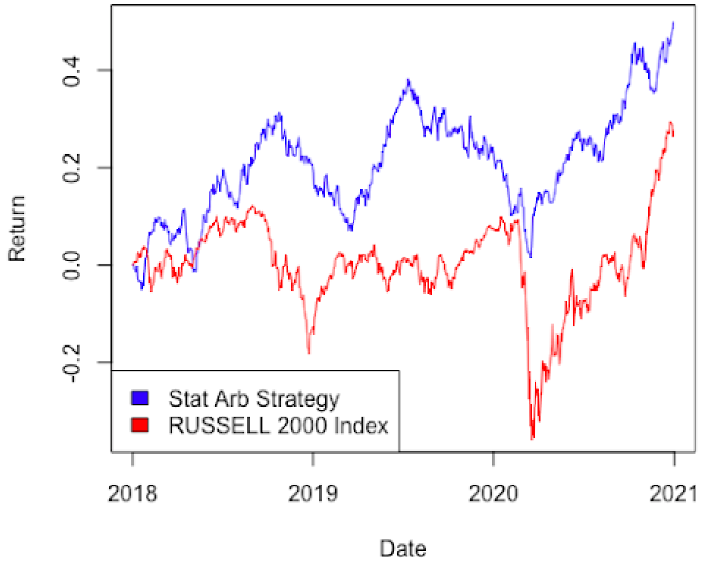
On the left is the return of the Russell2000 index over time; this is our benchmark return that we compare to. On right is a 14 day trailing volatility measure for the index from 2018 - 2021. Note the massive spike in volatility in March of 2020 as COVID-19 affected markets.





4.7 Results & Evaluation

The returns of our strategy are mapped against the benchmark returns below.



We established Sharpe Ratio as our main metric of evaluation.

$$\text{Sharpe Ratio} = \frac{R_p - R_f}{\sigma_p} \tag{10}$$

where R_p is the portfolio return, R_f is the risk free rate, and σ_p is the portfolio return standard deviation. Our strategy was able to achieve a Sharpe ratio of 1.238 vs. that of 0.65 for the Russell2000 index.
We also computed Maximum Drawdown:

$$MDD = \frac{P - L}{P} \tag{11}$$

where P is the peak value and L is the trough value. Our strategy achieved a max drawdown of 0.89 vs. the index's 4.45; a significant improvement which comes from the fact that our peak was much higher during the March COVID crash, which is a significant improvement. However, we do see that our strategy did not perform well during this period; our return crashed with the market, meaning our portfolio was not zero-beta. This implies that our model might not hold up during periods of high volatility.

5 Alpha Model

5.1 Alpha Source & Intuition

Alpha is the portfolio return in excess of market of returns. We aThe source of our alpha relies on a couple of principles.

1. Asset prices can be highly correlated over a period of time, and are expected to remain correlated in the near future. We exploit relationships between stocks to generate our alpha.
2. The returns of two highly correlated assets is expected to revert around a mean. This means that some stocks are temporarily overpriced relatively (which we can long), and others underpriced (for us to short).
3. We can construct factors that drive market returns via the principal component analysis on historical returns data. This then allows us to construct a zero beta portfolio (uncorrelated with the market).

5.2 Signal Generation

On every trading day, we compute s-scores for each stock i, defined as:

$$s_i = \frac{X_i(t) - m_i}{\sigma_{eq,i}} \tag{12}$$

where the equilibrium variance is:

$$\sigma_{eq,i} = \sigma_i \sqrt{\frac{\tau_i}{2}} \tag{13}$$

where τ_i is the mean reversion time. We open a trade when the s-score exceeds a minimum threshold s_o and close the position when a max threshold s_c is reached. Intuitively, we only open trades for extreme deviations from the mean, (indicating an outlier that should revert to the mean), and close trades when they get close to the mean.

5.3 Relative Betting

Our strategy does not take directional positions, but rather relative positions in equities in relation to the index. Our trading signals are generated systematically (as opposed to via individual fundamentals). Our factor model allows a zero beta portfolio, resulting in a low-volatility investment strategy that is hopefully uncorrelated with market returns.

6 Portfolio Construction

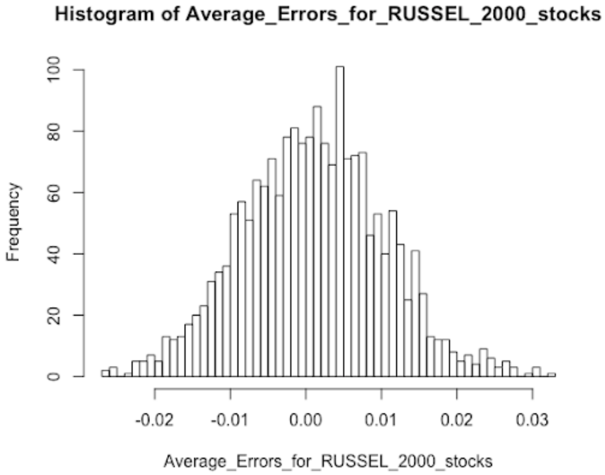
The Profit and Loss equation for our strategy is given as

$$E_{t+\Delta t} = E_t + E_t r \Delta t + \sum_{i=1}^{N} Q_{it} R_{it} - (\sum_{i=1}^{N} Q_{it}) r \Delta t + \sum_{i=1}^{N} Q_{it} D_{it} / S_{it} - \sum_{i=1}^{N} |Q_{i(t+\Delta t)} - Q_{it}| \epsilon \tag{14}$$

$$Q_{it} = E_t \Lambda_t \tag{15}$$

where E_t is equity in the portfolio at t, Q_{it} is the dollar amount invested in stock i at time t, r is the risk free rate, $\Delta t = 1/252$, D_{it} is dividends received on stock i from time t, S_{it} is the price of stock i at time t, ϵ is the transaction cost, and λ_t is a proportionality factor for leverage. In sizing our instruments, we allocate a fixed amount to a portfolio to start (\$1M) and make trades proportional to equity in the portfolio, with no max position.

Additionally we only make trades on the middle 25% of stocks whose returns are reasonably close to the factor model's prediction, in the 25% of the following histogram. These stocks are well described by the factor model but still have idiosyncratic behavior.



7 Risk Management Philosophy

7.1 Inherent Risks

There are multiple types of risk in statistical arbitrage. For example, there is individual stock risk, for a company to exit the market or be marged/acquired. There is a crowding effect, where many market participants employ similar stat arb models - affecting the price of the equities traded. There is risk that the model no longer is descriptive of the market in the future - for example if mean reversion principles break down or systematic risk factors change very quickly. A market movement that seems very unlikely but goes against model prediction can impose heavy losses that are amplied by leverage, potentially causing a margin call and forcing liquidation.

7.2 Risk Measurement & Management

We use our evaluation metrics of the Sharpe ratio and Max drawdown as inherent measures of risk. Inherently our portfolio should be zero-beta, but in periods of large volatility (such as during the March 2020 crash) this seems to break down. To protect against large losses, we added a daily stop loss that closes any position that loses 30% in a day.

8 Execution Discussion

8.1 Real World Considerations

When taking this strategy from class project to real world trading, there would probably have to be a number of changes made. One is the issue of trading frequency would likely have to be adjusted to fit a firm's particular strategy - perhaps this means minimum and maximum holding periods. Another change would have to be that the data used to create the model should look at longer term data, rather than data over the course of just a few recent years. In the real world, we might try to increase the s-score thresholds to only trade on very confident pairs, and use lots of leverage to obtain good returns. Better risk management is definitely needed in a real system. Finally, more fine-tuned parameters would likely lead to better results.

9 Retrospective Discussion

9.1 Challenges

We faced a number of challenges in designing and implementing this project. We both came in with a very limited knowledge of quantitative trading and statistical arbitrage; along the way we learned a lot about implementing realistic trading strategies. Initially, we had issues getting an approved Wharton Research Data Services account for the data. Our lack of experience with R, which is what this is project was coded in, definitely served as a pain point. Dealing with bugs in our code (such as the first row in our returns matrix being all zeros), and translating equations into an actual implementation were all difficult challenges, but we walked away learning a lot!

9.2 Concepts Learned

Some concepts we grew more familiar with include: mean reversion, pairs trading, statistical arbitrage, Ornstein–Uhlenbeck processes, the principal component analysis, standardized returns, correlation matrices, systematic factor models, zero-beta portfolios, signal generation, and backtesting.

9.3 Future Work

If we had more time, we would make some changes and extend upon the work. The parameters we discuss earlier can be continually tuned to produce better returns. Different measures of volatility can be considered by the model; for example volatility measures that consider “fat tails” and “volatility clustering.” This could help to more effectively control the portfolio’s risk. The strategy could incorporate assets outside the RUSSELL 2000 index to hedge portfolio risks. Alternative factor models, such as synthetic ETFs as factors, can be explored. Signal generation could consider not just mean reversion for a given stock, but also other indicators that might be relevant to it it’s short term idiosyncratic returns (such as trading volume). All in all, this project was a great learning experience. We are excited to see where it might go!

10 References

Jolliffe, I. T., Principal Components Analysis, Springer Series in Statistics,Springer-Verlag, Heidelberg, 2002.

M. Avellaneda and J. Lee. Statistical arbitrage in the U.S. equities market. Quantitative Finance, 10:761–782, 2008.

Plerou, V., Gopikrishnan, P., Rosenow, B., Amaral, L. N., Guhr, T. and Stanley, H. E., Random matrix approach to cross correlations in financial data. Phys.Rev., 2002, E 65, 066126.

Pole, A., Statistical arbitrage: Algorithmic trading insights and techniques, Wiley Finance, 2007.

Poterba, J. M. and Summers, L. H., Mean reversion in stock prices: evidence and implications. Journal of Financial Economics, 1988, Vol. 22, 27-59.