Pair Trading Strategy Optimization with Monte Carlo Simulation

Group Member: Rotem Eshed
Group Number: 29

1 Abstract

The strategy behind pair trading is to identify a pair of securities (or baskets of assets) whose price tends to move together, to track the spread (the "difference" between the assets prices), and once the spread widens s number of standard deviations, long (buy) the asset whose price decreased and short (sell) the asset whose price increased, assuming the assets would maintain their relationship and the prices will revert to their historical trend. In this paper, Monte Carlo simulation was used to simulate a pair of cointegrated stocks, namely Microsoft (MSFT) and Visa (V), to test and identify the optimal standard deviation threshold, considering risk (i.e. the lower the standard deviation threshold to initiate a trade is, the riskier the strategy is) and return. Standard deviation thresholds $s = i$, for $i \in I = \{1, 2, 3\}$ have been tested. With 5000 simulated MSFT and V stock prices (simulated over 252 trading days), in which 1466 pairs out of the 5000 are cointegrated with confidence level $\alpha = 0.05$, it has been found that the stratetegy with the hieghest expected value (i.e. return) is the strategy with standard deviation $s = 1$.

2 Background and Description of Problem

The motivation behind this work is the fact that there are no widely available Monte Carlo simulations related to pair-trading. This paper is an extension of a previous project I worked on where I developed a feasible and proven method to identify virtually all potential pair-trading pairs between and across all exchanges in the world without prior required domain, industry or market knowledge. More specifically, the strategy behind pair trading is to identify a pair of securities whose price is cointegrated, and once it is identified the trader tracks the spread. Once the spread widens, long (buy) the asset whose price decreased and short (sell) the asset whose price increased, assuming the assets would maintain their relationship and the prices will revert to their historical trend (Figure 1). Hence, pair trading is based on mean-reversion strategy. Pair trading requires three steps - (1) pairs formation, (2) co-movement relationship testing, and (3) determination of trading entry and exit rules [7]. My previous project focused on the first two steps and this paper is focused on the third step.

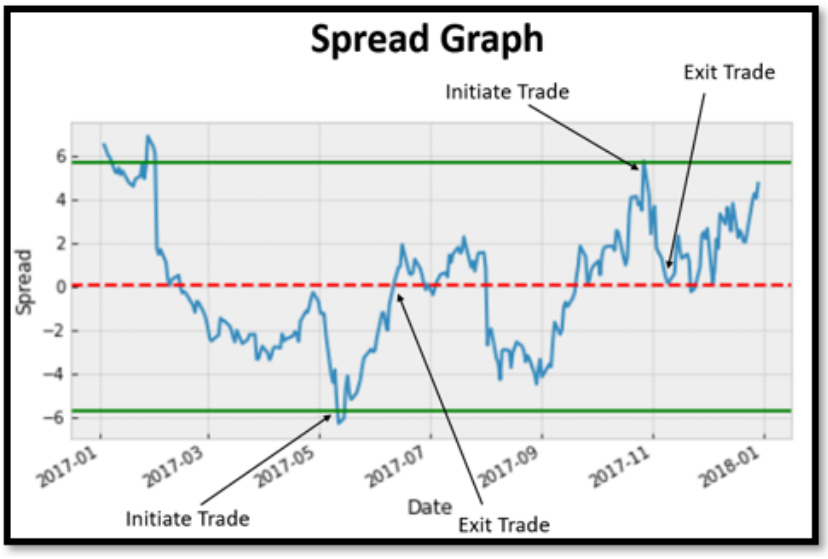

Figure 1: Pair Trading Strategy

2.1 Pair Trading - Steps (1) and (2) and a Summary of Previous Work

The purpose of the previous work was to serve as a proof of concept of finding all worldwide potential pairs of securities 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, it was found which companies were mentioned in financial news together with a company selected by the user and presented the results using a network graph. The assumption was that companies mentioned in the same news articles are potential pairs for co-movement testing. Currently, the pairs formation process begins with running co-movement tests on all possible pairs and eliminating pairs that fail the tests [2, 5]. This method is inefficient, computationally expensive and not feasible for testing stocks within different markets due to the large number of pairs combinations. Another pairs formation method is rule-based which is subjective and requires industry and market expertise [14, 12]. The model developed in the previous work focused on initial pair-formation by eliminating the need of any knowledge required to form pairs for testing and made the actual testing step more efficient by significantly reducing the number of pairs to test. While there are many methodologies to test pairs for co-movement, the two most common are the minimum-distance method [2, 4, 6, 9] and the cointegration method [3, 4]. In the previous and this paper, the cointegration test was used.

In the previous work, for each U.S. stock, financial news articles that mentioned the name/ ticker of the stock asset were collected from mid 2018 through mid 2020. There were around 5 million news articles collected and stored on AWS S3 storage. Each article was processed and stored in a structured relational database with the following schema: article_id, ticker#1, ticker#2 and publication_date, where article_id is the id of a financial news article, ticker#1 is the stock asset associated with the news article and ticker#2 is the stock asset co-occurring with ticker#1 in that news article published on publication_date. The adjacent nodes for a node (stock asset ticker#1) in the graph network are the stock assets highly co-occurring with that node's stock asset ticker#1. It can be found by filtering all the articles for ticker#1, grouping them by ticker#2 and sorting by the number of ticker#2 records. The top 6 tickers are selected and edges are formed in the graph network. Retrieving news data from 2018 to 2020, the dataset contains $\approx$ 29 million (2.6 GB disk size) ticker#1-ticker#2 co-occurrences.

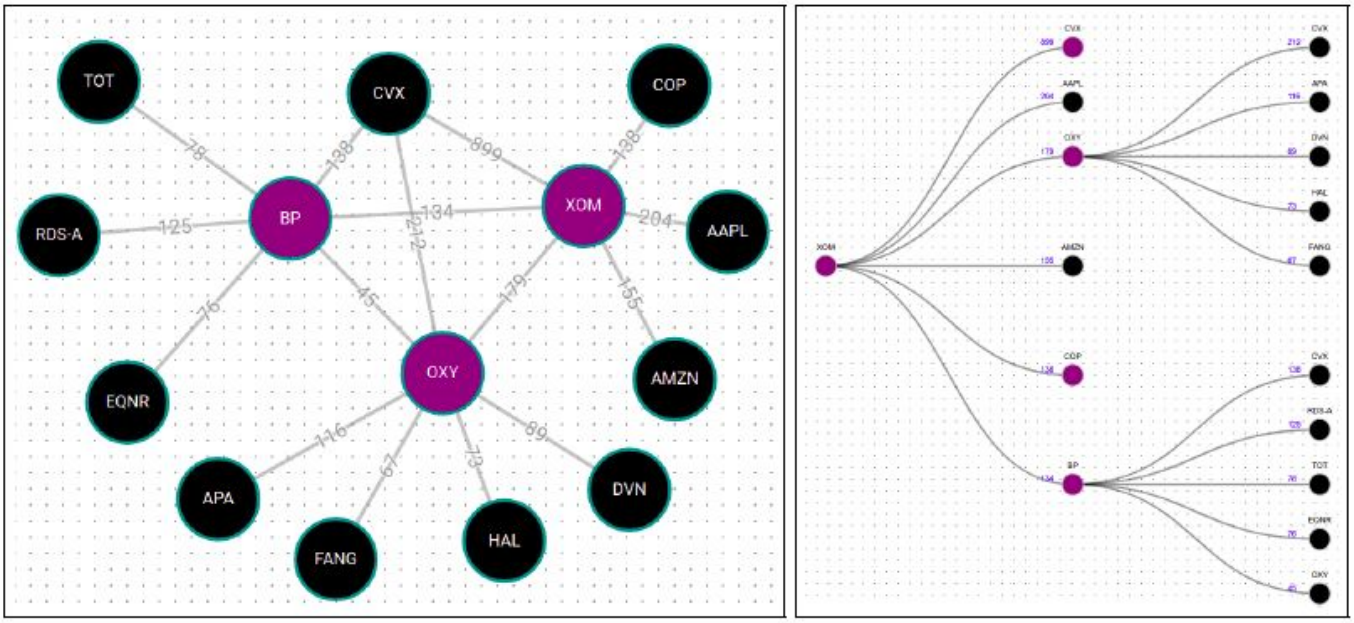

Figure 2: (Left) Ticker Network Graph; (Right) Ticker Network Tree

As mentioned, the network graph of the company tickers is constructed from the data obtained after processing the raw news. 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. The interactive platform allows the user to initiate a network graph by selecting an "initiating" ticker. The tool will return a network graph with the 6 tickers mentioned the most in the news articles together with the initial ticker. The user can further expend the network by clicking on any other nodes/ tickers presented by the graph. As the last step, the user can select from the displayed tickers any two nodes to view line charts comparing stock prices, spread line, cointegration "score" and correlation which will be presented analyzing historical stocks prices within a timeframe determined by the user.

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 (Figure 3) (Cointegration score is 1 minus the p-value of the Augmented Dickey–Fuller hypothesis test with the alternative hypothesis the spread is stationary. See more below). The tool also indicates less "obvious" pairs to test for conintegration. One example is Gaming and Leisure Properties Inc (GLPI) and Atlantica Sustainable Infrastructure PLC (AY). GLPI is a gaming-focused real estate investment trust while 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%.

Figure 3: XOM vs BP: (Left) Network graph; (Center) Co-integration graph; (Right) Price plot

2.2 Pairwise Cointegration Test

To test pairs for cointegration the Engle-Granger test was used 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 (which are referred to as spread) of the linear combination $Y_t = \alpha + X_t + \epsilon$ 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. Note that the linear combination above can be written as $\epsilon = Y_t - \beta X_t - \alpha$. The inclusion of the intercept should be carefully considered. Whether to include it or not depends on the subject matter tested for cointegration and the analyst's own consideration. Since the intercept term will not be traded upon and merely shifts the spread line upward or downward, it is omitted [5, 10, 11].

By design, assuming the linear combination $Y_t = \beta X_t + \epsilon$ holds, the spread's mean is 0 and the spread line will revert to the mean if stationary. To test whether the spread is stationary, the Augmented Dickey Fuller Test (ADF) with no constant and trend, and lag = 1 was performed (note that Akaike information criterion (AIC) and Bayesian information criterion (BIC) can be used to get the optimal lag p, but for simplicity and due to the fact that including lags of order 2, 3, 4.... does not add much value, the model was kept with order 1)[5] which for this paper is formulated as $\Delta\epsilon_t = \delta\epsilon_{t-1} + \phi\Delta\epsilon_{t-1} + v_t$. The Augmented Dickey-Fuller test (as opposed to the Dickey-Fuller test) allows for higher-order autoregressive processes by including the $\sum_{j=1}^{p} \Delta\epsilon_{t-j}$ term to the model (in our test we only include lag $p = 1$). The null hypothesis is that the process is non-stationary. That is, $H_0 : \delta = 0$. Hence we would want to REJECT the null hypothesis and to conclude that the spread is stationary and the pair of stocks are cointegrated. The intuition behind the test is that if the null hypothesis is NOT rejected (i.e. $\delta = 0$), the lagged level (ϵ_{t-1}) provides no relevant information about the current level change ($\Delta\epsilon_t$) besides the one obtained by the lagged change. However, if the null hypothesis is rejected, the lagged level does provide information about the current change and the series exhibits reversion to the mean of 0 [13].

What is Stationarity? On a high level, a stationary time series is one whose statistical properties do not change over time. Since strictly stationary stochastic process is too strict for real life processes, we define a weakly stationary stochastic process as one whose mean and variance do not change over time. That is, they are constant and do not depend on time t, and the auto-covariance of X_t and $X_{t+\tau}$ only depends on lag τ [7].

3 Main Findings

We start with MSFT and V true stock prices from 01/01/2017 through 01/01/2018. During this time period, the pair of stocks are cointegrated at the confidence level $\alpha = 0.05$ (Figures 4 & 5). Note that Figure 5 plots the spread as defined above. That is, the spread is the residuals of the liner combination $Y_t = \alpha + X_t + \epsilon$ (i.e. ϵ_t). We can see in Figure 5 that $\mu_\epsilon = 0$ and that the spread has the mean reversion characteristic. Figure 5 also includes the spread's +/-1, +/-2, and +/-3 standard deviations which will be used in this paper as the thresholds to initiate a trade. That is, we will use Monte Carlo simulation

to generate stochastic processes or simulated prices of MSFT and V based on their true prices from 01/01/2017 through 01/01/2018 and will test and identify the optimal standard deviation threshold to initiate a pair trade. By "optimal," we refer to the highest return for 1 unit invested in each trade, excluding transaction and other costs. In this paper, "1 unit invested" means 1 stock (whether long or short).

We simulated 5000 stock prices of MSFT and V over 252 traded days (which is the number of traded days within a year) based on their true prices from 01/01/2017 through 01/01/2018 (Figures 6 & 8). To generate the 5000 simulated stock prices, we start with the normalized log returns of the true prices, which is:
$r_i = \frac{P_i - P_j}{P_j}$ where r_i is the return at time i, P_i is the price at time i and $j = i - 1$
$r_i = \frac{P_i}{P_j} - \frac{P_j}{P_j}$
$1 + r_i = \frac{P_i}{P_j}$
$log(1 + r_i) = log(\frac{P_i}{P_j})$
$log(1 + r_i) = log(P_i) - log(P_j)$

We find the log returns' μ and σ. We then generate 252 simulated returns which are sampled from $\mathcal{N}(\mu, \sigma)$ and multiply the last true price (i.e. 01/01/2018) by the cumulative product of the (simulated returns + 1). This has been performed 5000 times for both MSFT and V to generate the 5000 simulated prices for each ticker. Note that the prices are simulated from the next day of the true prices last day (i.e. 01/02/2018) and not over the true time interval (i.e. 01/01/2017 through 01/01/2018). Simply put, the simulated prices are simulating the "future" stocks prices from 01/02/2018 through 01/02/2019.

And as expected, based on the Central Limit Theorem, we can see that the means of those 5000 simulated prices are normaly distributed (Figures 7 & 9). The Central Limit Theorem, which is the most-important theorem in the world, states that given $X_1, X_2, ..., X_n$ independent and identically distributed random variables as anything (other than Cauchy distribution) with mean μ and variance σ^2, then the means $\bar{X_1}, \bar{X_2}, ..., \bar{X_N}$ are distributed as $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$ when n is large [1]. Out of the 5000 simulated paired prices, we have 1466 cointegrated pairs at the confidence level $\alpha = 0.05$

3.1 Trading Strategy Example

Let us take one simulated cointegrated pair and execute a trade with a threshold of +/-1 standard deviation. That is, when the spread reaches the +1 standard deviation threshold, we will buy V and sell (short) MSFT. Assuming the spread will revert to its original mean of 0, we will exit the trade when the spread reaches its mean. Similarly, when the spread reaches the -1 standard deviation threshold, we will buy MSFT and sell V and exit the trade when the mean reverts to its mean (Figures 10 & 11). Assuming 1 unit invested, excluding transaction and other costs, this specific trade return 24.29 dollars over 252 days. Let us look at the trading buy/ sell rules to arrive at the 24.29 dollars (Figure 12). We start with row index 1 (because the simulated prices are over 252 trading days only, ignore the first "terminate trade" and last "(-1) buy MSFT-sell V") where we buy V at 107.09 because it reached the +1 standard deviation threshold at this price and exit V at 118.63 because it reverted to its mean at this price, and sell MSFT at 85.4 because it reached the -1 standard deviation threshold at this price and exit at 90.69 because it reverted to its mean at this price where we profited (118.63 - 107.09) + (85.4 - 90.69) = 6.25, etc. See table below for the complete calculation.

Buy V-Sell MSFT	Buy V-Sell MSFT	Buy MSFT-Sell V	Buy MSFT-Sell V	Total Return
(118.63 - 107.09)	(133.46 - 120.6)	(104.06 - 99.6)	(105.35 - 101.82)	
(85.4 - 90.69)	(97.65 - 101.68)	(135.35 - 135.1)	(138.03 - 137.06)	
6.25	8.83	4.71	4.5	24.29

3.2 Risk vs Return

As already mentioned, given two non-stationary series X_t and Y_t, if there exists α and β such that the residuals (which are referred to as spread) of the linear combination $Y_t = \alpha + X_t + \epsilon$ 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. Note that the linear combination above can be written as $\epsilon = Y_t - \beta X_t - \alpha$. That is, if the pair are cointegrated that means that the linear combination $Y_t = \alpha + X_t + \epsilon$ holds and the residuals or the ϵ_t are normally distributed [8]. And since the spread is the ϵ's, it means that when the spread "breaks" the +/-1 standard deviation threshold, the probability the spread will revert to its mean is 68%. In the case of the strategy with +/-2 and +/-3 standard deviation threshold, the probability the spread will revert to its mean is 95% and 99.7%, respectively. But the probability the spread will reach the +/-1 standard deviation is higher than the probability the spread will reach the +/-2 standard deviation, which is higher than the probability the spread will reach the +/-3 standard deviation. Here is where the risk and return come to play. Hence, the strategy with +/-3 standard deviation is the least risky but we should expect the least return, while the +/-1 standard deviation strategy is the riskiest, but we should expect the highest return.

3.3 Results

Below is a table that summarizes the results of the three strategies. The Total Aggregated Returns column refers to the aggregated returns of all 1466 cointegrated pairs, the μ and σ columns refer to the average return and standard deviation of the 1466 cointegrated pairs, respectively.
As expected, we can see the risk-return relationship where as the risk increases so does the return. That is, the Std +/-1 is the riskiest strategy with the highest σ but with the highest return in total and on average. Looking at Figure 13, we can see some negative returns with the standard deviation +/-1 strategy. Similarly, the Std +/-3 is the least risky strategy with the lowest σ but with the lowest return in total and on average (Figures 13, 14 & 15).

Strategy	Total Aggregated Returns	μ	σ
Std +/-1	17072.82	11.65	7.44
Std +/-2	9790.82	6.68	6.56
Std +/-3	444.23	0.3	2.15

4 Conclusion

In this work we used Monte Carlo simulation to test and identify the best strategy out of three statistical arbitrage pair trading strategies, considering the relationship between risk and return. More specifically, we generated 5000 simulated prices for MSFT and V over 252 trading days and ran the three pair trading strategies using 1466 cointegrated pairs (with confidence level $\alpha = 0.05$) out of the 5000 simulated pairs. The general risk-return relationship where the higher the potential return of an investment is, the higher the risk was observed across the simulation to conclude that the highest return with the highest risk strategy is the one with the +/-1 standard deviation threshold, and the lowest return with the lowest risk strategy is the one with the +/-3 standard deviation threshold. The results show that on average, all three strategies had positive returns. One of the advantages of pair trading is that it is market neutral trading strategy where the direction of the overall market does not affect its win or loss.

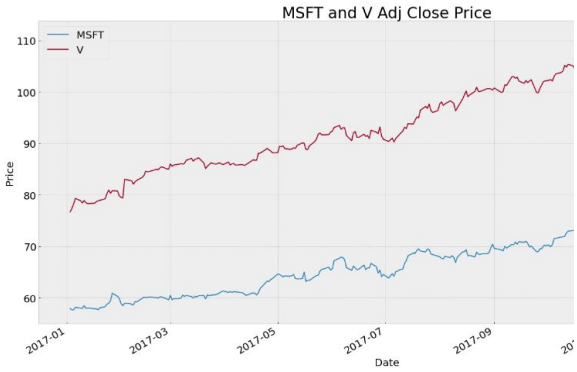

Figure 4: MSFT vs V True Prices from 01/01/2017 through 01/01/2018

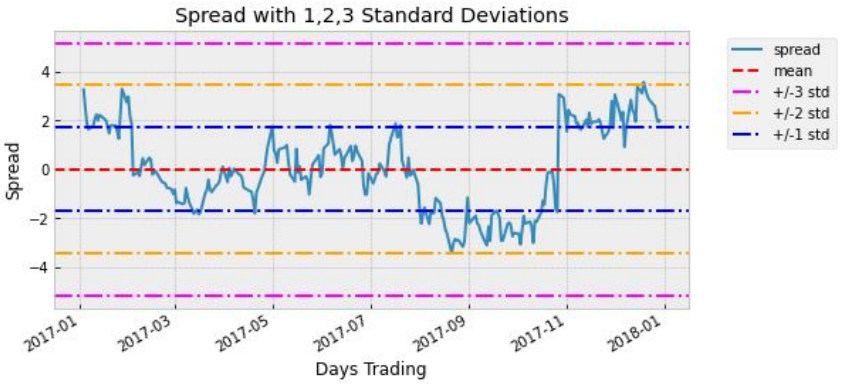

Figure 5: MSFT & V Spread and Cointegration Score

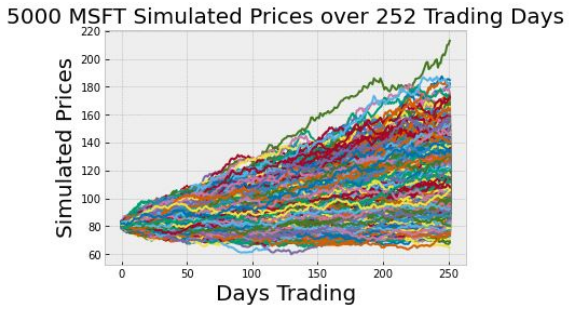

Figure 6: 5000 Simulated MSFT Prices over 252 traded days

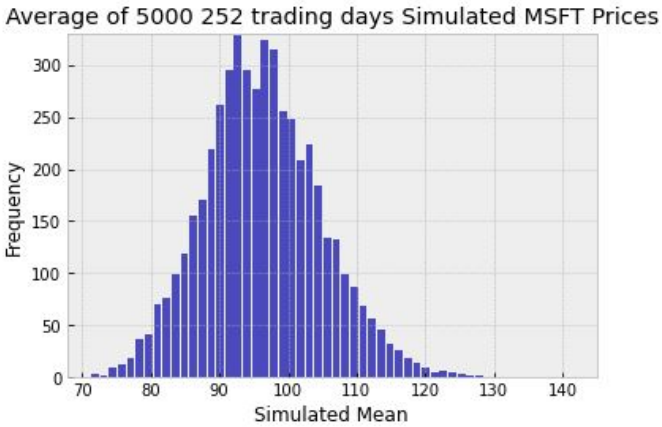

Figure 7: Plot of 5000 MSFT simulated μ's

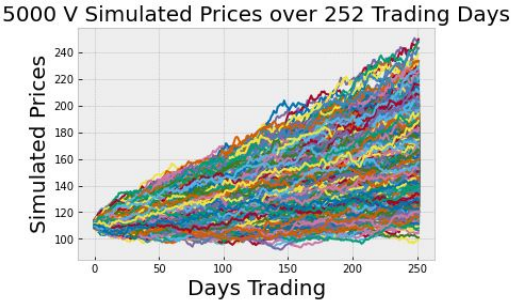

Figure 8: 5000 Simulated V Prices over 252 traded days

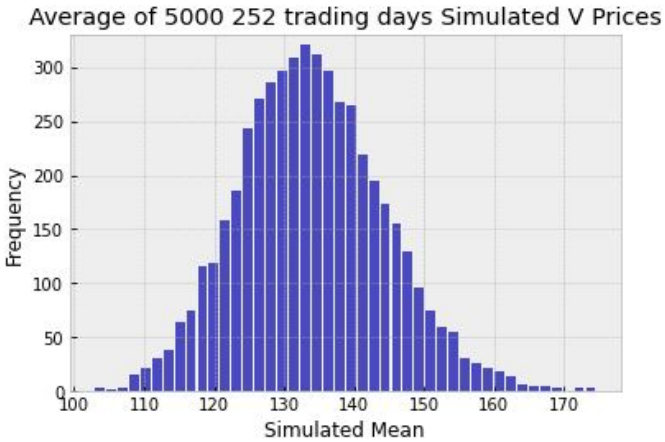

Figure 9: Plot of 5000 V simulated μ's

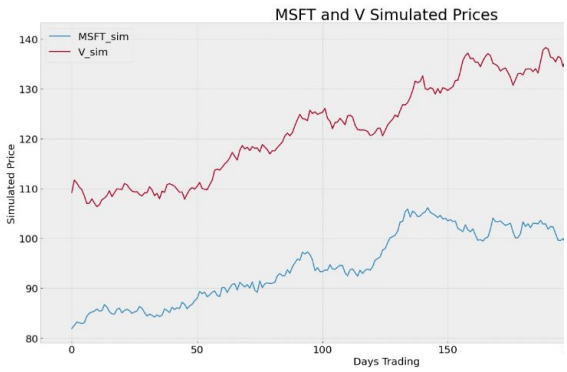

Figure 10: MSFT vs V Simulated Pair Example

Figure 11: MSFT & V Simulated Spread Example

	MSFT_sim	V_sim	spread	(1) buy V-sell MSFT	(-1) buy MSFT-sell V	terminate trade	strategy
0	83.048374	108.507761	0.067949	NaN	NaN	T	terminate trade
1	85.392487	107.080540	3.503516	T	NaN	NaN	(1) buy V-sell MSFT
2	90.683833	118.625639	-0.034157	NaN	NaN	T	terminate trade
3	97.642124	120.596010	5.417308	T	NaN	NaN	(1) buy V-sell MSFT
4	101.672777	133.464052	-0.392769	NaN	NaN	T	terminate trade
5	99.612280	135.355797	-3.899962	NaN	T	NaN	(-1) buy MSFT-sell V
6	104.064788	135.115925	0.735986	NaN	NaN	T	terminate trade
7	101.822352	138.033361	-3.737536	NaN	T	NaN	(-1) buy MSFT-sell V
8	105.348768	137.062465	0.531366	NaN	NaN	T	terminate trade
9	101.572748	138.325761	-4.210751	NaN	T	NaN	(-1) buy MSFT-sell V

Figure 12: Example Buy/ Sell Rules

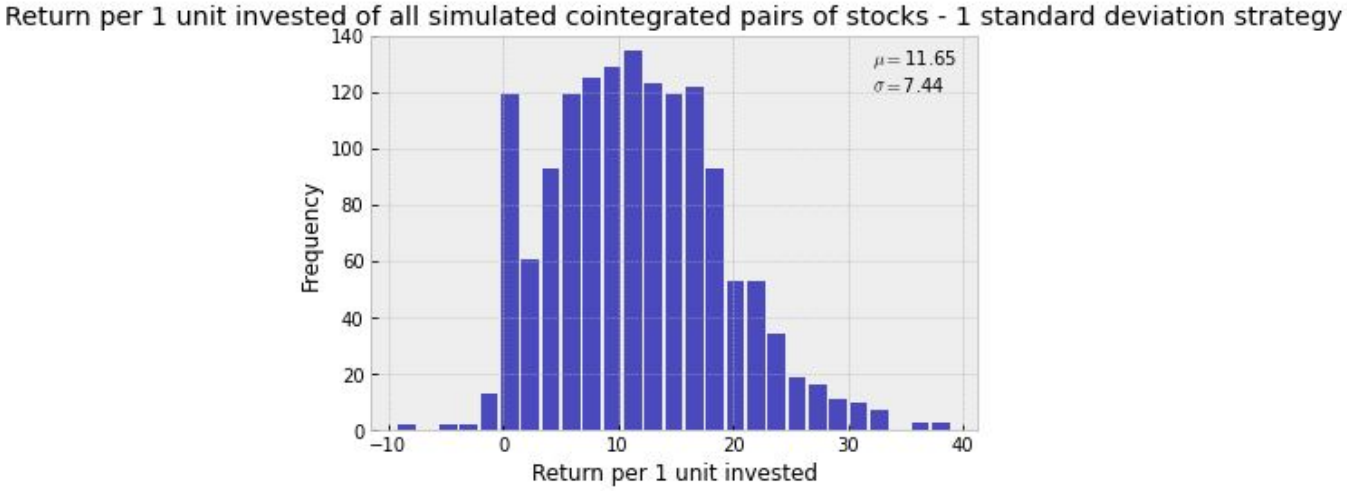

Figure 13: Strategy Std +/-1 Aggregated Result

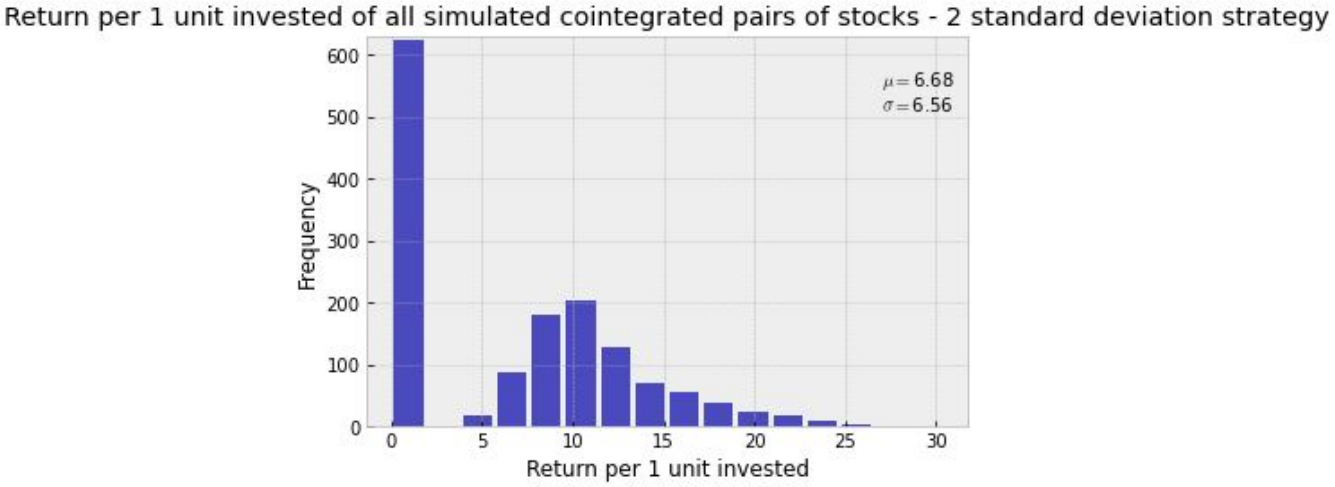

Figure 14: Strategy Std +/-2 Aggregated Result

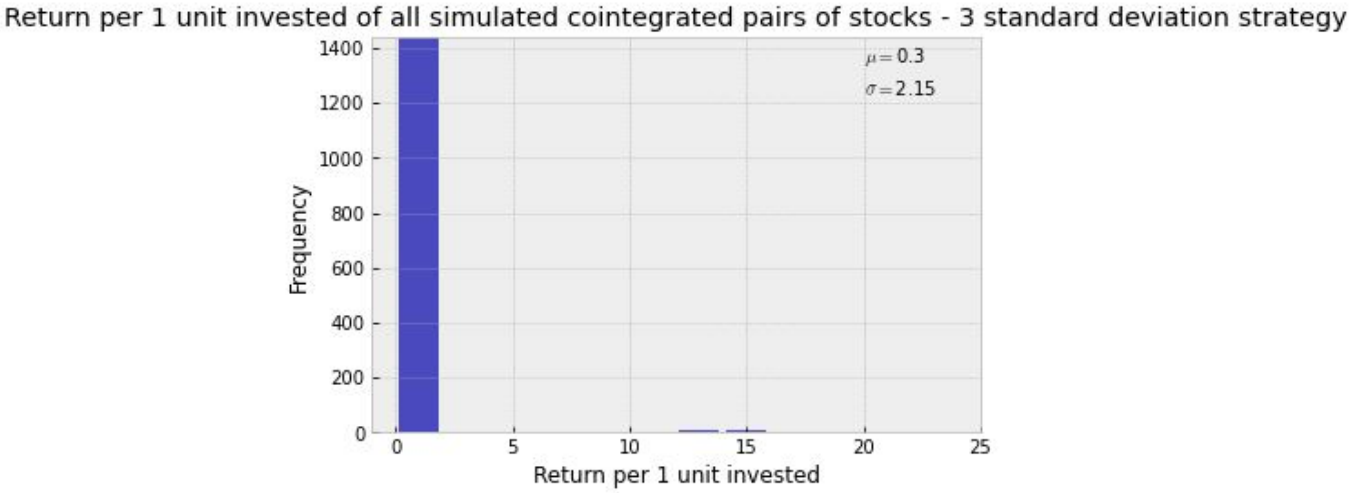

Figure 15: Strategy Std +/-3 Aggregated Result

5 References

[1] Dave Goldsman (2017). Georgia Tech ISYE 6644 Simulation, Module 2.

[2] Gatev, Evan, Goetzmann, William N., and Rouwenhorst, K. Geert (2006). "Pairs Trading: Performance of a Relative-Value Arbitrage Rule". In: *Review of Financial Studies* 19, pp. 797–827. http://www-stat.wharton.upenn.edu/~steele/Courses/434/434Context/PairsTrading/PairsTradingGGR.pdf.

[3] Gulati, Chandra, Lin, Yan-Xia, and McCrae, Michael (Jan. 2006). "Loss protection in pairs trading through minimum profit bounds: a cointegration approach". In: *Faculty of Informatics - Papers (Archive)*, Art. No. 73803. https://ro.uow.edu.au/infopapers/1546/.

[4] Huck, Nicolas and Afawubo, Komivi (Nov. 2014). "Pairs trading and selection methods: is cointegration superior?" In: *Applied Economics* 47, pp. 599–613. https://www.tandfonline.com/doi/abs/10.1080/00036846.2014.975417.

[5] Jansen, S (2020). Time-Series Models for Volatility Forecasts and Statistical Arbitrage - Machine Learning for Algorithmic Trading - Second Edition [Book]. https://www.oreilly.com/library/view/machine-learning-for/9781839217715/Text/Chapter_9.xhtml#_idParaDest-353..

[6] Johnson, Barry (2010). Algorithmic trading & DMA : *an introduction to direct access trading strategies/*. 4Myeloma Press. https://searchworks.stanford.edu/view/9155503.

[7] Mader, H. M., Coles, S. G., Connor, C. B. & Connor, L. J. (2006). *Statistics in Volcanology*. [Book] Special Publications of IAVCEL, 1. 129. Geological Society, London. https://books.google.com/books?hl=en&lr=&id=e5Y_RRPxdyYC&oi=fnd&pg=PA129&dq=stationary+time+series&ots=XuPePbbobp&sig=f0qIO-uhMyAqxHIexCUL1TO2w14#v=onepage&q=stationary%20time%20series&f=false.

[8] Penn State Early College of Science. STAT 462, Applied Regression Analysis, 4.6 Normal Probability Plot of Residuals. https://online.stat.psu.edu/stat462/node/122/.

[9] Rad, Hossein, Low, Rand Kwong Yew, and Faff, Robert (Apr. 2016). "The profitability of pairs trading strategies: distance, cointegration and copula methods". In: *Quantitative Finance 16*, pp. 1541–1558. https://randlow.github.io/2016_QF_PairsTrading.pdf.

[10] Shimul, Shafiun, Abdullah, S, and Siddiqua, Salina (2009). "AN EXAMINATION OF FDI AND GROWTH NEXUS IN BANGLADESH: ENGLE GRANGER AND BOUND TESTING COINTEGRATION APPROACH". In: *BRAC University Journal* 1, pp. 69–76. http://dspace.bracu.ac.bd/bitstream/handle/10361/455/Shafiu.Nahin.Shimul.pdf..?sequence=1.

[11] Skerman, Robert and Della Maggiora, Daniel (2009). "Johansen Cointegration Analysis of American and European Stock Market Indices: An Empirical Study". In: *lup.lub.lu.se*. https://lup.lub.lu.se/student-papers/search/publication/1437434.

[12] Smith, R. Todd and Xu, Xun (2017). "A good pair: alternative pairs-trading strategies". In: *Financial Markets and Portfolio Management* 31, pp. 1–26. https://ideas.repec.org/a/kap/fmktpm/v31y2017i1d10.1007_s11408-016-0280-x.html.

[13] Scheuerell, M. D. Holmes, E. E. and Ward, E. J., (2021) *Applied Time Series Analysis for Fisheries and Environmental Sciences*[Book] . https://atsa-es.github.io/atsa-labs/sec-boxjenkins-aug-dickey-fuller.html.

[14] Vidyamurthy. (2004) *CHAPTER 6 - Pairs Selection in Equity Markets - Pairs Trading: Quantitative Methods and Analysis*[Book]. https://www.oreilly.com/library/view/pairs-trading-quantitative/9781118045701/vidy_9781118045701_oeb_c06_r1.html.