ORIE 5741 Learning with Big Messy Data · Project Report

Pair Trading Strategy of Russell 3000 Index

Kelvin Jiang, Xiyao Fu, Hengli Zhu
https://github.com/HengliZhu/ORIE-5741-Project

1 Introduction

1.1 Motivation

Pair trading is a market-neutral strategy that enables traders to profit under various market conditions. This approach involves pairing two similarly exposed assets, and betting on the relative performance of one against the other. The primary motivation for pair trading is to hedge against market risk. By holding simultaneous long and short positions in correlated securities, traders can mitigate exposure to systemic market risks. Such risk mitigation is especially valuable during periods of increased market volatility or economic uncertainty, making pair trading an appealing strategy as it focuses on relative rather than absolute market movements.

From the market efficiency perspective, pair trading plays a crucial role by exploiting pricing inefficiencies between related securities. When two historically correlated stocks diverge in price without a fundamental reason, pair traders bet on the reversion of this disparity to the historical norm. This strategy not only corrects pricing discrepancies between similar assets but also contributes to overall market efficiency.

1.2 Key Problem

The key question in pair trading revolves around how to accurately and reliably identify pairs of stocks (or other assets) that exhibit a strong and stable historical correlation and cointegration. This question underscores the necessity of using sophisticated statistical methods to ascertain the genuine correlation and cointegration between the selected assets and predict their future movements. The challenge is compounded by the need to continuously monitor the pairs for stability and adjust strategies as market conditions evolve.

Trading signals are generated by statistical models that analyze the price movements of two correlated assets, identifying when their price ratio or spread diverges significantly from the historical norm. The effectiveness of these signals hinge on precise model calibration, which must account for factors like mean reversion speed, asset volatility, and market liquidity. These models must be dynamic, capable of adjusting to real-time market data to refine their predictions continuously, ensuring that the trading signals remain accurate and relevant, thereby maximizing profitability and minimizing timing-related risks.

1.3 Data description

We choose 3000 equities as the asset class to trade in our project. The datasets are provided by Wharton Research Data Services (WRDS). We mainly obtained the daily stock files from CRSP. The raw data contains 3000 stocks which are constituents of Russell 3000 Index. we obtain their prices, identity info, sector, volume. Using stocks in pair trading allows for identification of unique characteristics, anomalies, and mispricings that may not be apparent when working with ETFs. Plus, large pool of stocks have good flexibility and granularity which make it easier to analyze the underlying dynamics and relationships within the market.

The data collection spans 10 years, from 2014-01-01 to 2023-12-31. We choose stocks from the financial and energy sectors. The financial and energy sectors are closely linked to the overall health of the economy, and sensitive to the interest rate. Also, they are often perceived as riskier and more volatile compared to other sectors. During periods of market stress or uncertainty, these sectors may exhibit similar patterns of heightened volatility and risk aversion. These features lead to high possibility of cointegration and good for pair trading.

1.4 Data Processing

The raw data quality is good. We firstly used a threshold-based approach to remove rows (stocks) where more than 0.01% of the values were missing. For the missing values, we used time-weighted interpolation. This method preserves the temporal structure and market trends. By accurately estimating missing values based on time proximity, time-weighted interpolation ensures the integrity of the stock price time series. This is essential for applying techniques like PCA and k-means clustering, which rely on complete and evenly spaced data to uncover patterns and groupings. Finally, we have 296 stocks in total.

2 Pair Searching

2.1 Data split and processing

To assess the effectiveness and robustness of our pair selection, we divided the processed data into two sets. The training set contains data from 2014-01-01 to 2022-12-31. After finding the pairs, to test our pair's cointegration, we use data from 2014-01-01 to 2023-12-31. For the training dataset, we calculate the percentage change in stock price, which is the return of each stock. We use PERMNO as unique stock identifier. After this, we standardized the return data to ensure that all features have zero mean and unit variance.

2.2 PCA

Principal Component Analysis (PCA) is a statistical method of dimension reduction that is used to reduce the complexity of a data set while minimizing information loss. It transforms a data set in which there are a large number of interrelated variables into a new set of uncorrelated variables, the principal components, and which are ordered sequentially with the first component explaining as much of the variation as it can. Each principal component is a linear combination of the original variables in which the coefficients indicate the relative importance of the variable in the component.[1]

We use stock return data as a score matrix A and a coefficient vector l to build a linear combination on A, resulting in principal components Y with a smaller dimension than A. The principal components Y are independent, and each coefficient vector l must maximize the variance of its corresponding principal component Yi, as shown:

$$\max Var(Y_i) = l_i^T C l_i$$

$$\text{s.t.} \quad ||l_i^T|| = 1, \forall i$$

$$l_i^T l_j = 0, \forall j \neq i$$

2.3 K-means cluster

To group similar stocks together and save cointegration test time, we applied the K-means clustering. K-means begins with k arbitrary centers, typically chosen uniformly at random from the data points. Each point is then assigned to the nearest center, and each center is recomputed as the center of mass of all points assigned to it. These two steps (assignment and center calculation) are repeated until the process stabilizes. [2] We determined the optimal number of clusters using the elbow method and silhouette analysis, and divide all stocks into clusters.

PCA reduces the dimensionality of the data by identifying the principal components that capture the most variance. K-Means clustering can be effectively applied to the reduced-dimensional space obtained from PCA, as it operates on the Euclidean distance between data points. K-Means clustering is relatively robust to the presence of irrelevant or noisy variables. When combined with PCA, K-Means can effectively cluster stocks based on the relevant features while minimizing the impact of less important variables. However, k-means clustering also has some limitations. It assumes that the clusters are spherical and of equal size, which may not always hold true for stock data. Additionally, the results of K-Means can be sensitive to the initial placement of centroids and may converge to suboptimal solutions.

We set the number of clusters at 65, and draw the T-SNE visualization, which is in appendix

2.4 Cointegration

We find an imbalance in the number of stocks in each cluster. To filter the clusters used for cointegration, we choose clusters that have 2-12 stocks, which result in 8 valid clusters. We control the stocks in each cluster in a small number mainly to manage risk and increase cointegration likelihood. Pair trading relies on the assumption that the cointegrated stocks will maintain their long-term equilibrium relationship. However, this assumption may not always hold, particularly in the presence of significant market disruptions or structural changes. By focusing on smaller clusters, the impact of any individual stock or pair on the overall portfolio is limited. This helps to mitigate the risk associated with relying on a single or a few pairs, promoting better risk management practices [3].

We use Engle-Granger two-step approach. The stationarity of the linear combination (yt) is tested using the Augmented Dickey-Fuller (ADF) Test. We regress yt on its lagged values yt−1 and find out whether the coefficient φ is 1 or not. Consider autoregressive process of order 1, AR(1), as shown:

$$y_t = \phi y_{t-1} + \epsilon_t$$

The ADF test checks for the null hypothesis H0: γ = 0 against the alternative hypothesis H1: γ < 0.

The linear combination yt of the two stock price series (S1t and S2t) is constructed as:

$$y_t = S_t^1 - aS_t^2$$

The ADF test is based on the following model:

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \sum_{j=1}^{p} \delta_j \Delta y_{t-j} + \epsilon_t$$

Where Δ is the first difference operator, α is the intercept, β is the coefficient of the time trend, γ is the coefficient of the lagged value yt−1, δj are the coefficients of the lagged difference terms, ϵt is the white noise error term and p is the order of the autoregressive process

If the test statistic from the ADF test is smaller than the critical value at a chosen significance level, the null hypothesis is rejected, indicating that the pair is cointegrated. Cointegration refers to the long-term equilibrium relationship between two or more non-stationary time series, where the series move together over time, maintaining a relatively stable difference. In the context of pair trading, cointegration ensures that the prices of the two stocks converge to their long-term equilibrium, even if they temporarily diverge in the short run. Interestingly, two stocks can exhibit strong cointegration while having a negative correlation. This means that the stocks may move in opposite directions in the short term, but their long-term relationship remains intact. The presence of cointegration allows pair traders to exploit the temporary deviations from the equilibrium by taking long and short positions in the respective stocks, with the expectation that the prices will eventually converge. This convergence property forms the basis of profitable pair trading strategies. Without strong cointegration, the pairs may drift apart permanently, leading to potential losses. Therefore, identifying and selecting stock pairs with robust cointegration is essential for successful pair trading, regardless of their short-term correlation dynamics [4].

We choose pairs that have p-value less than 0.025 in the cointegration test. We identify 7 promising trading pairs after this analysis, as shown in the table in appendix.

3 Trading Strategy

3.1 Rolling Regression

Rolling regression is a statistical technique used to continuously estimate the changing relationship between variables over time. In the context of pair trading, rolling regression is employed to dynamically update the correlation coefficients between the prices of two correlated assets, such as stocks. This method involves calculating the regression parameters over a fixed "window" of recent data points, which then rolls forward through time as new data becomes available. By continuously updating these parameters, traders can capture changes in the relationship due to evolving market conditions or other external factors [5].

In pair trading, rolling regression helps determine the optimal entry and exit points for trades. For example, a trader might use rolling regression to calculate a time-varying hedge ratio, which is used to balance the positions of the paired stocks proportionally to their relative volatilities and prices. This approach helps maintain the market-neutrality of the pair trade over time.

If we assume two stock price X and Y follows linear relationship $Y_t = \alpha + \beta X_t + \varepsilon_t$, we will focus on the change of the regression coefficient β. β is not a constant, we allow β to float within the upper/lower band. When β is exceeding the upper band, this tells us Y_t maybe "too expensive" and X_t maybe "too cheap". A reasonable trading strategy in this case is to short Y_t and long X_t. On the other hand, when β is exceeding the lower band, this tells us Y_t maybe "too cheap" and X_t maybe "too expensive". A reasonable trading strategy in this case is to long Y_t and shortX_t. In both cases, the hedging ratio would equal to the β we calculated from the rolling regression [6].

We applied Score to normalize the price ratios. After normalization, we could standardize the procedure of deciding the upper/lower bond of the signal and easily apply the method to any pairs as long as they have a strong cointegration relationship.

If we simply use the whole period mean and standard deviation to calculate the Score and upper/lower band, the trading signal generated is not significant in frequency and also introduces the issue of using future information. We applied a 5-days and a 60-days rolling windows to run the rolling regression and generate the mean/standard deviation respectively.

$$Zscore\_moving = \frac{(5\ days\ moving\ average - 60\ days\ moving\ average)}{60\ days\ moving\ standard\ deviation}$$

We use a 70/30 train-test split and test over different combinations of rolling windows. But in general, to keep the procedure more robust and standard, we set 5-days and 60-days window in this project.

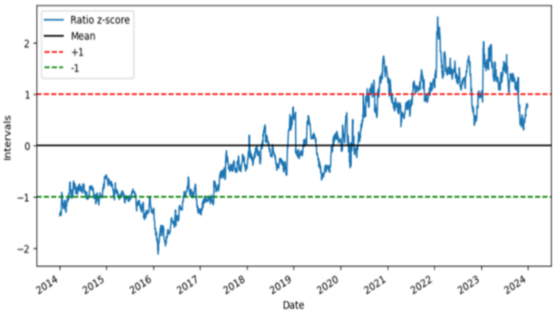


Figure 2 Z-Score Time Series

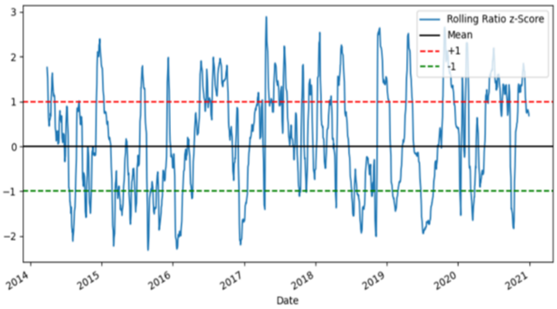


Figure 3 Rolling Z-Score Series

In the example pair of MS and GS, the trading signal is significantly improved in frequency and quality after we run the train-test split and rolling regression. The price ratio is not necessarily normally distributed originally, but with the rolling windows, the frequency histogram is more like a bell shape.

Since we are developing a self-financing strategy, in each trade, we only consider to trade one share of Y and trade β shares of X in the opposite direction to hedge the risk. Multiple trading signals were raised during the period. Every time we use the amount of money from shorting one stock to execute the long position of another stock. Overall, the performance is stable and generates over $800 dollars if we long/short only one share all the time. This profit is significant as we can leverage our position to long/short more shares each time.

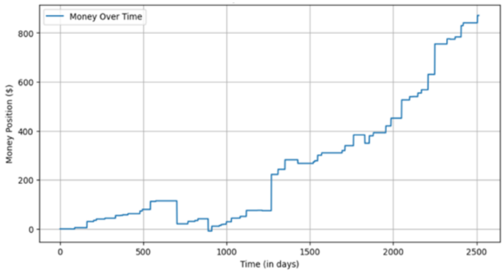


Figure 4 Money Position of MS & GS Pair

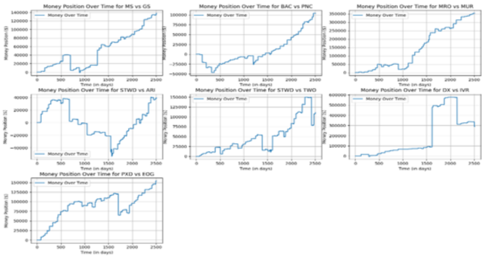


Figure 5 Money Position of All Pairs

For all the seven pairs we identified, all of them generating positive returns during the 10-years period. The plot shows the dollar return of the strategy if we set a \$10,000 value in the long/short position each time. Some of the pairs' P&L experienced significant drawdown within the period, this is mainly because the cointegration between these two stocks was very weak during those years. We should monitor the cointegration closely and set a stop-limit to cut our position when the loss exceeds our predetermined loss tolerance.

3.2 Kalman Filter

3.2.1 Theory of Kalman Filter

The Kalman filter is an advanced recursive filter, also known as an autoregressive filter. By considering the joint distribution of measurements over time, the Kalman filter synthesizes estimates of unknown variables. This approach generally yields more accurate results than methods relying solely on single measurements [7].

Kalman Filter operate using principles from linear algebra and hidden Markov models. They model basic dynamic systems as Markov chains driven by linear operators, which are disrupted by Gaussian (normally distributed) noise. The system's state is represented by a vector of real numbers. With each increment in discrete time, a linear operator acts on this current state vector, generating a new state along with noise. Concurrently, known control information is integrated into the system. Meanwhile, other noise-affected linear operators generate observable outputs from these hidden states [8].

The formula is as below:

$$y(x) = \beta^T x + \varepsilon$$

Here, $\beta^T = (\beta_0, \beta_1, \dots, \beta_n)$, β_0 is the intercept of the Kalman Filter and $\beta_i$is the slope of the i-th variable $x_i$of the vector x. ε is the noise in the regression process and Kalman Filter model assumes that $\varepsilon \sim N(\mu, \sigma^2)$. In this case, because of the number of stocks in each pair, we can denote that $\beta^T = (\beta_0, \beta_1)$ and $x^T = (x_0, x_1)$.

Besides, as an important characteristic of Kalman Filter model, we apply the vector $(\beta_0, \beta_1, \dots, \beta_t)$ to represent the hidden state. And we assume that the slope of time t+1 equals to that of time t plus some random noise. We can utilize the formula below to describe the relationship between the hidden states quantitatively:

$$\beta_{t+1} = I\beta_t + \varepsilon_t$$

Here, β_t is the coefficient vector of time t, I is the identity matrix and ε_t is the random noise of time t.

3.2.2 Trading Strategy Based on Kalman Filter

As an improvement of the Rolling Regression model, Kalman Filter model constructs the regression process of intercepts and slopes between stock A and stock B of each stock pair dynamically as well. As a result, we design the trading signal based on E_t, which is the error of the prediction of y_t, and V_t, which is the variance of the prediction. The trading strategy are constructed as below:

1. When $E_t > \sqrt{V_t}$, we take a short position on 1 share of y and take a long position on β_t^0 shares of x.
2. When $E_t < 0.01\sqrt{V_t}$, we close the position we operate above.
3. When $E_t < -\sqrt{V_t}$, we take a long position on 1 share of y and take a short position on β_t^0 shares of x.
4. When $E_t > -0.01\sqrt{V_t}$, we close the position we operate above.

3.2.3 Back Testing of Kalman Filter Strategy

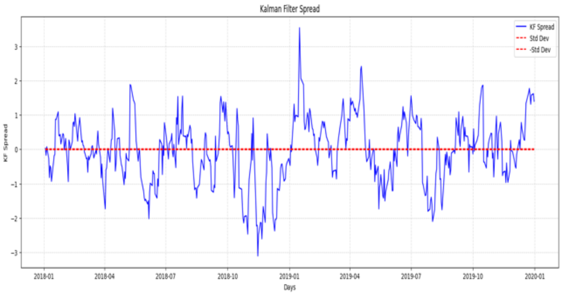


Figure 6 Kalman Filter Spread

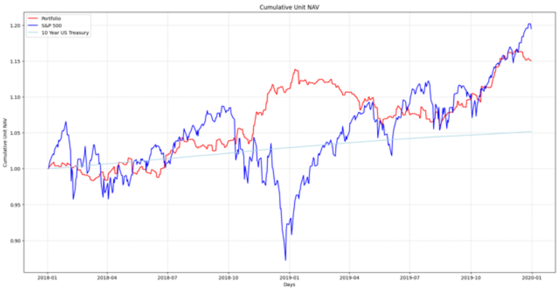


Figure 7 Performance of Portfolio Over Time

With the trading strategy we design above, we calculate the error and variance of the price prediction, compare these two indicators to generate trading signals, and record these signals in the time series data. Besides, in order to better show the performance of the pair trading strategy based on the Kalman Filter. We choose the seven trading pairs selected by clustering and cointegration test, and generate the equal-weighted portfolio of each stock pair. Then we calculate the performance indicators of theses portfolios like cumulative return, volatility and Sharpe Ratio, and compare with S&P 500 Index and 10-year US Treasury. For detail analysis, we take the pair of US Bancorp and Citi as an instance.

	Portfolio	S&P 500	US Treasury
Return	6.13%	5.92%	2.52%
Volatility	1.36%	14.95%	0.03%
Sharpe Ratio	2.31	0.54	NAN
Max Drawdown	1.60%	19.78%	0.00%

Table 1 Performance of Portfolio

In the term of returns, annualized return of equal-weighted portfolio is 6.13%, which is a little bit higher than S&P 500 Index and much higher than 10-year US Treasury. However, when it comes to volatility, the volatility of portfolio is 1.36% and is much lower than S&P 500 Index whose volatility is 14.95%. As a result, the trading strategy of pair trading based on Kalman Filter Model achieved 2.31 on Sharpe Ratio, which is more than four times of that of S&P 500 Index. In addition, it is shown in time series graph that the trading strategy successfully avoids market downside risk and avoids drawdown when the S&P 500 Index is experiencing bad market conditions. In general, considering all trading pairs, the trading strategy of Kalman Filter can achieve return approximately equal to S&P 500 Index but much smaller volatility and much larger Sharpe Ratio as a result.

4 Discussion

4.1 Challenges and Limitations

While the project demonstrates significant strengths, it also faces challenges such as the assumption of spherical and equal-sized clusters in K-Means, which might not hold in real stock data scenarios. The sensitivity of K-Means to initial centroids placement and potential convergence to suboptimal solutions also poses limitations that need addressing in future studies. Besides, the cointegration relationship between pairs may not remain stable over time, reflecting shifts in underlying economic factors, market sentiment, or regulatory environments. This instability can render historical relationships misleading, leading to erroneous pair selections where the expected mean reversion strategy becomes invalid. Additionally, transaction costs play a critical role in the overall profitability of pair trading. Every trade incurs costs such as bid-ask spreads, which can be particularly wide for less liquid stocks, and brokerage fees, which accumulate with increased trading frequency. These costs can significantly diminish the net gains from small price deviations typically exploited in pair trading.

4.2 Future Work

The models tested in this study show potential for scalability and adaptability across different sectors and market conditions. Future research could explore the application of these models in other asset classes or in markets with different characteristics to validate the generalizability of the findings.

4.3 Conclusion

This project not only underscores the viability of pair trading strategies in modern financial markets but also sets a precedent for the application of advanced statistical methods in financial strategies. By demonstrating the practical effectiveness of these models, the study contributes valuable insights into both academic and practical aspects of financial trading.

5 Contribution

Kelvin Jiang: data collection, Rolling Regression, report, slide
Xiyao Fu: data collection, pair search, report, slide
Hengli Zhu: data collection, Kalman Filter, report, slide

Reference

[1] Yang, L. (2015). An Application of Principal Component Analysis to Stock Portfolio Management (Master's thesis, University of Canterbury, Christchurch, New Zealand).

[2] Arthur, D., & Vassilvitskii, S. (2006). k-means++: The Advantages of Careful Seeding. Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, 1027-1035.

[3] KRAUSS, C. Statistical Arbitrage Pairs Trading Strategies: Review and Outlook. Journal of Economic Surveys, [s. l.], v. 31, n. 2, p. 513–545, 2017.

[4] BRUNETTI, M.; DE LUCA, R. Pre-selection in cointegration-based pairs trading. Statistical Methods & Applications: Journal of the Italian Statistical Society, [s. l.], v. 32, n. 5, p. 1611–1640, 2023.

[5] RACICOT, F.-E.; RENTZ, W. F.; KAHL, A. L. Rolling Regression Analysis of the Pastor-Stambaugh Model: Evidence from Robust Instrumental Variables. International Advances in Economic Research, [s. l.], v. 23, n. 1, p. 75–90, 2017.

[6] MCMILLAN, D. G. Forecasting Stock Returns—Historical Mean Vs. Dividend Yield: Rolling Regressions and Time-Variation. Cham: Springer International Publishing, 2018.

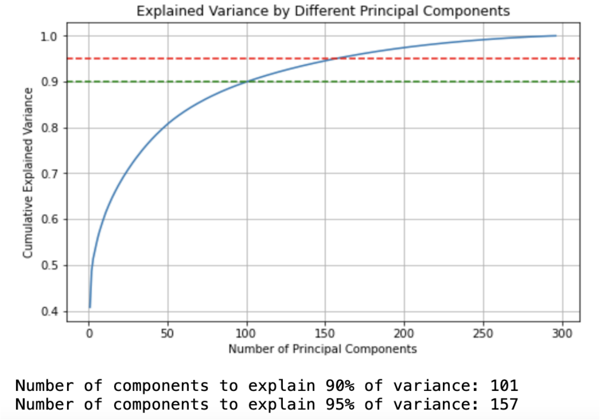
[7] MILSTEIN, A. et al. Neural Augmented Kalman Filtering with Bollinger Bands for Pairs Trading. IEEE Transactions on Signal Processing, Signal Processing, IEEE Transactions on, IEEE Trans. Signal Process, [s. l.], v. 72, p. 1974–1988, 2024.

[8] MOHAMMAD JAVAD NOURAHMADI; MARZIYEH NORAHMADI. Application of Kalman Filter to Estimate Dynamic Hedge Ratio in Pairs Trading Strategy: A Case Study of the Automobile Industry, [s. l.], v. 25, n. 1, p. 63–87, 2023.

Appendix

Asset 1	Asset 2
DX	IVR
STWD	ARI
STWD	TWO
BAC	PNC
MS	GS
MRO	MUR
PXD	EOG

Table 1 Identified Stock Pairs



```
Number of components to explain 90% of variance: 101
Number of components to explain 95% of variance: 157
```

Figure 1 Explained Variance by Different PCs

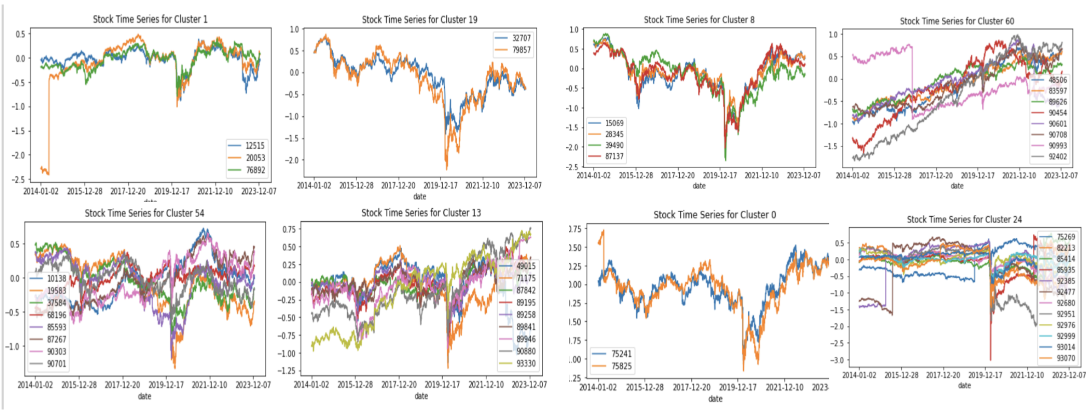


Figure 2 Time Series Graph of Stocks in Each Cluster

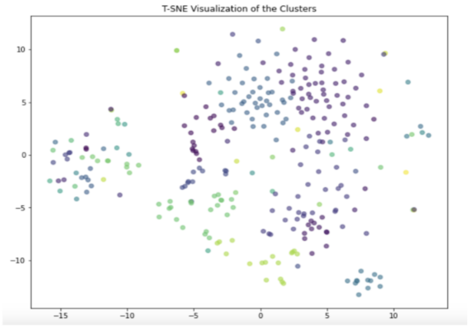


Figure 3 T-SNE Visualization of Clusters

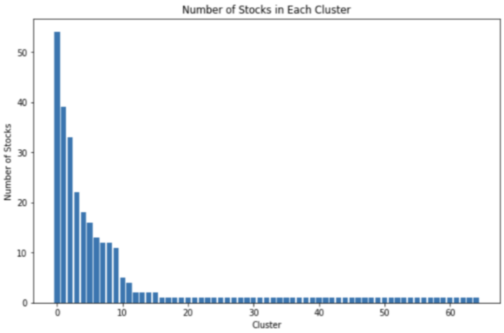


Figure 4 Numbers of Stocks in Each Cluster

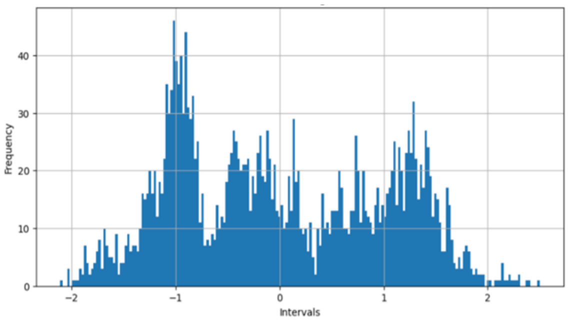


Figure 5 Z-Score Histogram

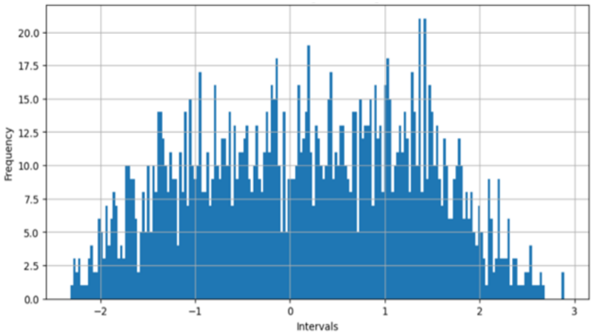


Figure 6 Rolling Z-Score Histogram

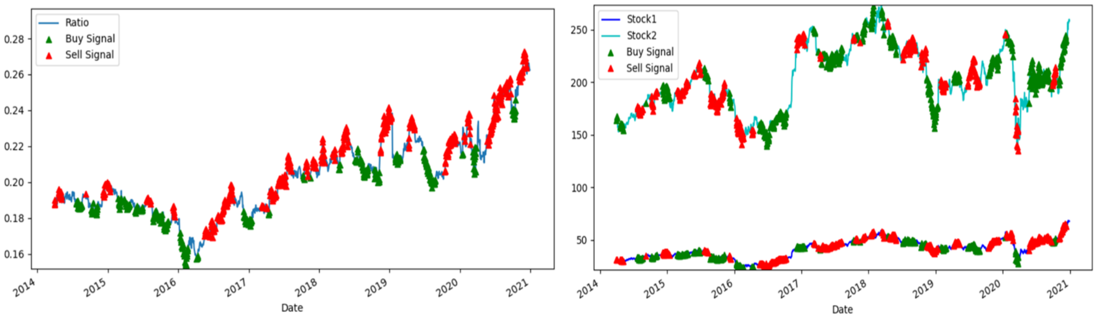


Figure 7 Trading Signal of MS & GS Pair

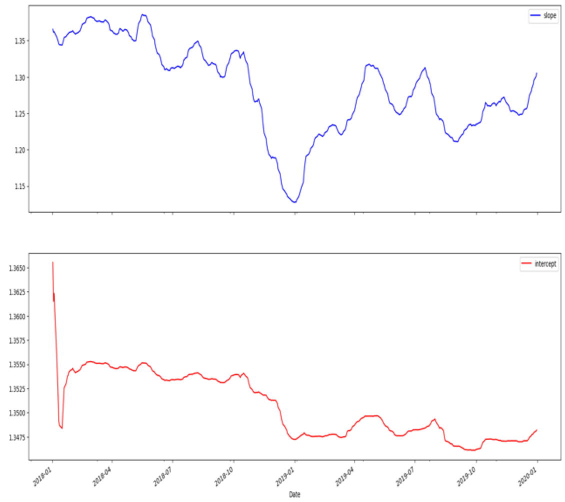


Figure 8 Intercept and Slope of KF Model Over Time