Pairs Trading: Identifying Pairs using Clustering

A PROJECT REPORT SUBMITTED
FOR THE REQUIREMENTS OF THE DEGREE OF

Master of Science in Quantitative Finance

BY

Aditya Khandelwal
Jonathan Wijaya
Low Heng Bong
Shanmugam Praveen Raj
Tan Guan Fong
Vedant Pednekar
Wu Xianyuan

LEE KONG CHIAN SCHOOL OF BUSINESS

SINGAPORE MANAGEMENT UNIVERSITY

SINGAPORE, DECEMBER 2021

Contents

List of Figures

List of Tables

1 Abstract

This paper develops pair trading strategies using various clustering algorithms. Unlike conventional pair trading strategies that identify pairs based on return time series, this study identifies pairs by using clustering. In this study, pairs of assets are found by using different clustering algorithm. Cointegration test, ADF, half life and Hurst exponent are applied to the promising pairs to check for cointegration, stationarity, mean-reversion etc. A synthetic asset is then created from the identified pair with the hedge ratio obtained using Kalman Filter and the asset is backtested and Sharpe ratio and CAGR are calculated. Finally, a comparison of all the clustering algorithms used to identify the pairs, is provided.

Keywords: Pair trading, Clustering, Kalman Filter

2 Introduction

Pair trading is a market-neutral and relative investment strategy which aims to seek out two instruments with historically high positive correlation, whether they are stocks, indices, commodities or cryptocurrencies. After identifying the highly-correlated pairs, traders can then profit from arbitrage that arises from the behaviors of these two instruments, regardless of the market condition. This quantitative method was first developed by a group of quants at Morgan Stanley in the 1980's, and it has been used by major investment banks ever since. In the next section, we will explore this strategy further.

3 Literature Review

3.1 Pair Trading and Cointegration

According to market efficiency, stock price movements reflect all publicly available information. If two securities have same underlying economic phenomena, the assets will respond similarly to the incoming news. Overreaction and herding behaviour often drive prices apart in which any deviation is considered temporary.

By using pair trading strategy, which is a market neutral strategy, profits can be generated by going long or short on the spread under all market conditions. Based on mean reversion trading strategy, pair trading capitalizes on the imbalances between two or more securities, in anticipation of making money when the inequality is corrected in the future (Whistler, 2004) [15]. The trading method can be performed with the following steps:

1. Find two securities that have moved together in the past with high correlation.
2. Calculate the mutual mispricing, which is the spread between the two chosen securities.
3. (Define threshold) Sell the higher-priced security and buying the lower-priced security with the idea that the mispricing will correct itself in the future. (long the spread if the spread cross threshold that triggered a long position, close the position when the spread reverts to zero)

Although the relationship between a pair does not necessarily continue in the future, trading many different pairs at once can create a diversified portfolio to mitigate the risk of individual pairs "falling out of" cointegration, which is first introduced by Robert et al. (1987) [3]. Two time series are cointegrated if they are I(1) series and the spread zt at each time calculated is an I (0) (or stationary) series. Then the mean-reverting and stationary property can be exploited for pair-trading (Vidyamurthy, 2004) [14] .

In this paper, the Augmented Dickey-Fuller Test (ADF) is applied to test for stationarity in time series.

3.2 Unsupervised Learning - Clustering

Clustering is the classification of objects into different groups, or more precisely, the partitioning of a data into subsets (clusters), often according to some pre-defined distance measure, so that the data in each subset shares some common trait with one another. This paper is aiming to find a proper clustering method to help find the most promising pairs. This paper considers four types of clustering algorithms including partitioning methods, model-based methods, hierarchical methods, and density-based methods.

For partitioning algorithms, we have mainly considered K-means and Mean Shift (MS). K-Means first randomly selects k points as cluster centers and assigns objects into their nearest cluster centers, and then calculates the mean of objects in each cluster. The steps are repeated until the k clusters converge (MacQueen, 1967) [10]. MS assigns the data points to the clusters iteratively by shifting points towards the mode, which is the highest density of data points in the region (Duin et al., 2012) [2]. Partitioning methods are very sensitive to noise and outliers, and they are generally only applicable to convex data sets. The selection of the center of mass points is random at the beginning, and different initial centers lead to different results. Therefore, K-means or MS can converge but the outcome they derive is not necessarily the optimal solution.

Gaussian Mixture Model (GMM) is an example of model-based algorithms. It has better flexibility than K-Means algorithm since it is composed of several Gaussian probability distributions that can theoretically fit any type of distribution and is often used to solve situations where data under the same set contains multiple different distributions (Kang et al. 2017; Paalanen et al. 2006) [8] [11]. However, GMM is based on the Expectation Maximum (EM) algorithm, so it is possible to fall into local extremes and requires several iterations.

Hierarchical algorithms include Hierarchical Agglomerative Clustering (HAC) and Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH). HAC is a bottom-up algorithm initially described by Ward (1963) to group objects by minimizing the within-cluster error sum of squares (ESS). Gabarowski et al. (2019) [5] tried to find triangular arbitrage opportunities in the Forex by estimating cross-correlations and considering (q-dependent) cross-correlation coefficients as a measure of the distance between different exchange rate pairs. BIRCH clusters large datasets by first generating a small and compact summary of the large. This condense summary is then clustered instead of clustering the larger dataset (Zhang et al. 1996) [16]. One of the advantages of hierarchical clustering is that it does not require us to specify the number of clusters. However, it has high computational complexity and outliers could have a serious impact on the outcomes. Besides, since it provides a dynamic criterion which can cause bias from investors' choice of desired level of granularity, a hierarchical clustering with an automatic termination criterion is more appropriate (Sarmento, 2020) [12].

Since partitioning and hierarchical algorithms do not perform well on the data distributing with different density and non-globular data, the density-based algorithms that can find density, size and outliers is introduced to solve these problems. One typical algorithm of density-model clustering is Density-Based Spatial Clustering of Applications with Noise (DBSCAN). DBSCAN finds dense set of points and forms a cluster around them. Clusters are formed by measuring the Euclidean distance between the points and the minimum number of points. The biggest advantage of DBSCAN is that it can handle arbitrarily shaped clusters without being affected by the cluster structure. It also can identify noise and is not sensitive to the order of the data. In addition, it does not need to specify the number of clusters k as K-Means does, and it can automatically discover the number of clusters in the data. Nevertheless, it cannot handle the data sets with large difference in the data density and different values of parameters initially set will produce different results.

Ordering Points To Identify the Clustering Structure (OPTICS) is the extension of DBSCAN to solve the problem above. The algorithm can process different distance parameters, meaning that it can construct density-based clusters based on different densities at the same time (Ankerst et al. 1999) [1]. Therefore, we apply OPTICS in the clustering part to generate the stock pairs at the first stage.

4 Methodology

To get the most promising pairs of stocks, we perform the following steps to screen the stock pairs. We apply one of the clustering algorithms to initially select stocks that can be clustered into pairs after reducing the data dimension. Then the dynamically updated hedge ratio is calculated by Kalman-Filter-regression. Based on the regression result, we further test the cointegration, mean reversion of the stock pairs, and select the pairs that are most likely to be profitable. Finally, the pairs trading strategy is employed on the selected pairs.

4.1 Data

This research mainly focuses on the stocks listed in NASDAQ and the New York Stock Exchange (NYSE). We collect the daily adjusted close prices from 2010-01-04 to 2021-12-14 for 11403 stocks and 3117 days. Later, during the data cleaning process, we screen out stocks with more than 0.01% of missing data. The final data set we obtained consists of daily adjusted prices for 3812 stocks over 3117 days. In the clustering section, we split the data set into the training set and the test set at 70% and 30% respectively.

4.2 Pairs Selection

4.2.1 Dimensionality Reduction

Principal component analysis PCA uses an orthogonal transformation to convert potentially linearly correlated variables into linearly uncorrelated variables, also known as principal components, and the new variables can present the characteristics of the data in a smaller dimension. Before diving into the clustering algorithms, the dimension reduction method PCA is applied to extract the systematic risks that affect the assets so that the ones with the same systematic risk can be identified (Sarmento, 2020) [12]. PCA is applied on scaled adjusted returns defined as the return R_i is subtracted by the mean $\bar{R}_i$ and divided by the standard deviation σ_i :

$$Y_i = \frac{R_i - \bar{R}_i}{\sigma_i}$$

where, $R_i = \frac{P_{i,t} - P_{i,t-1}}{P_{i,t-1}}$, $P_{i,t}$ is the adjusted close price of stock i at time t.

4.2.2 Unsupervised Learning Clustering

After selecting the pairs with unconditional mean-reverting behaviour analysing systematic risk with PCA, we then apply unsupervised learning method of clustering to cluster the pairs into different group. OPTICS, DBSCAN, K Means, Mean Shift, Agglomerative Clustering, Gaussian Mixture and BIRCH have been used to create the clusters in this paper.

Once the compact dataset is clustered by the clustering algorithm, it will then be visualized into different clusters, say 10 number of clusters. Each stock in the clusters is then paired with the other stocks from the same group, say 200. After filtering these candidates from thousands of possible combinations, other rules such as ADF Test, Hurst Exponent, half-life of mean-reversion and Sharpe Ratio from back testing are then applied to further reduce the number of best pairs (Ankerst et al. 1999) [1].

4.2.3 Hedge Ratio

When two or more non-stationary series can be combined to make a stationary series, the component series are said to be cointegrated. One of the challenges of pairs trading is to determine the coefficients that define this stationary combination. In pairs trading, that coefficient is called the hedge ratio, and it describes the amount of instrument B to purchase or sell for every unit of instrument A. The hedge ratio can refer to a dollar value of instrument B, or the number of units of instrument B, depending on the approach taken [9]. The principal advantage of using Kalman Filters over another method of dynamically updating our hedge ratios is that the Kalman Filter algorithm reduces the degree of arbitrariness in parameter optimization. To see this, consider that one method of dynamically updating the hedge ratio is to utilize a rolling window of prices that we can run a linear regression on. Not only is the selection of a rolling window arbitrary, but the deletion of the last slice of time (bar) and inclusion of the nearest slice of time places equal and arbitrary importance on every slice of time, no matter how old. Would an exponential weighting scheme on the "freshness" of data work? Perhaps, but that leaves us with yet another arbitrary parameter, the informational decay rate of the exponential weighting scheme. The Kalman Filter dispenses with such arbitrariness. The Kalman Filter is very adept at dynamically adjusting our hedge ratios even in the face of market noise and produces very useful by-products(intercept as spread mean, and forecast error of observed variable as spread standard deviation) while doing so [4]. The advantages of using Kalman filter for the Kalman Filter instead of the other widely used methods are provided in the quoted articles. We have used Kalman Filter for the hedge ratio in this study. Hedge ratio is frozen once the trade is opened and unfrozen once the trade is closed. The spread of the pair is given by $S_t = X_t - \beta Y_t$, where β is the hedge ratio determined by Kalman Filter. Hedge ratio is determined using the python implementation of Kalman Filter: pykalman.

4.2.4 Pairs Selection Criteria

We proceed to select pairs that satisfy the following requirements:

1. The two stocks within the same pairs must be cointegrated. We have used ADF test for cointegration.
2. The spread should mean reverse. The Hurst Exponent is useful to test the mean reversion character. The Hurst value H of the spread must be less than 0.5 to ensure the time series is mean reverting.
3. The spread should have coherence between mean-reversion time and the trading period. The Half-life of the spread calculates the time that the spread evolves to its mean value (Tadi & Kortchemski, I., 2021) [13]. Thus, the half-life of mean reversion is employed to select the pairs with min half life of 1 day and the max half life of 42 days.

4.3 Trading Model/Strategy

As we have already applied the Hurst Exponent to test the pairs for their mean reversion properties, we can apply the z-score strategy for backtesting.

z-score is basically the numerical measurement that describes the value's distance to the mean, by number of unit of standard deviation. In general, a z-score of +2.0 means that the current value is 2 S.D. above the mean and a z-score of –2.0 means that the current value is 2 S.D. below the mean.

Long entry is when the z-score crosses the lower bound usually -2 and closed when the z-score hits zero or the stop loss bound whichever is earlier. Short entry is when the z-score crosses the upper bound usually 2 and closed when the z-score hits zero or the stop loss bound whichever is earlier. Stop loss for long and short positions have been set at z-score -3.0 and 3.0 respectively.

4.4 Environments of Research

All the outcomes of the research are generated by using Python 3.X version with the following libraries. The daily data is derived from Yahoo Finance.

Main Libraries	Usage
Pandas	Data analysis and manipulation
Matplotlib	Visualization
Numpy	Numerical calculations
Scikit-Learn	For clustering
Statsmodels	For time series statistical analysis and cointegration tests
Pykalman	Kalman Filter
Hurst	Hurst Exponent

Table 1: Python Libraries

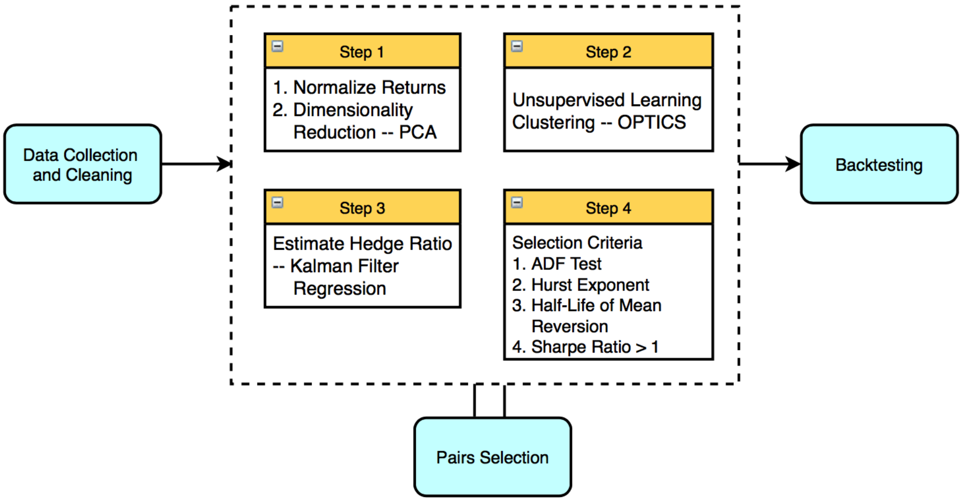


Figure 1: Pairs Selection Process

5 Empirical Results

5.1 Pairs Selection

The pairs selection process is shown in the figure below:

1. Perform PCA on the normalized returns and set the number of principle components as 5 to separate systematic risks.

2. Apply clustering Algorithms.

    - OPTICS: Two parameters for OPTICS clustering have been considered in Python: one is the maximum distance between two neighboring samples, and the other is the number of samples in a neighborhood for a core point, shown as 'eps' and 'min_samples' respectively.

3. Calculate the dynamic hedge ratio β by using Kalman-filter regression and substitute it into $S_t = X_t - \beta Y_t$ to obtain spread time series S_t.

4. Derive the promising pairs with stationary, mean-reversed spread by using the pairs selection criteria:

    - Pass ADF unit root test;
    - Value of Hurst Exponent is between 0 and 0.5;
    - Half-life of mean reversion is set with the min of 1 day and the max of 42 days;
    - Sharpe ratio is more than 1.

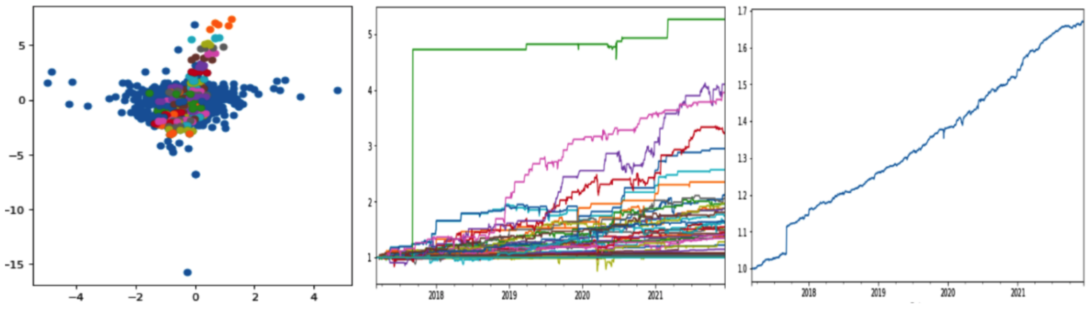
5.2 Analysis of the results

The data set has been split into modelling and backtesting set in the ratio of 70% to 30% respectively.

Traditional approaches like Elbow method (K-Means) and Dendrogram (HAC) are not ideal for time series data. In the analysis part, we mainly use Sharpe ratio and cumulative annual growth rate (CAGR) from our back tests as key measurements of performance.

Hyper-parameters tuning did not achieve any significant difference in the performance on most clustering algorithms. However, they can the be tuned to increase or decrease the number of clusters, which affect both the quantity and quality of the generated pairs. Generally, with more clusters generated, the quantity of the generated pairs decreased but the quality increased. As, with higher number of clusters, the number of instruments within the clusters decrease, which indirectly contributes to having only best correlated stocks in the clusters and decreases the chances of spurious pairs.

Using OPTICS clustering and selection criteria, the final portfolio consists of 50 pairs with Sharpe ratio of 2.74 and CAGR of 7.7%. Overall CAGR is constructing a portfolio that weighs each pair equally. This ensures that no one pair dominates the portfolio performance. This is only used for performance comparison of different clustering algorithms and should not be presumed as the best portfolio construction methodology. No other portfolio construction methodologies were explored for this study.



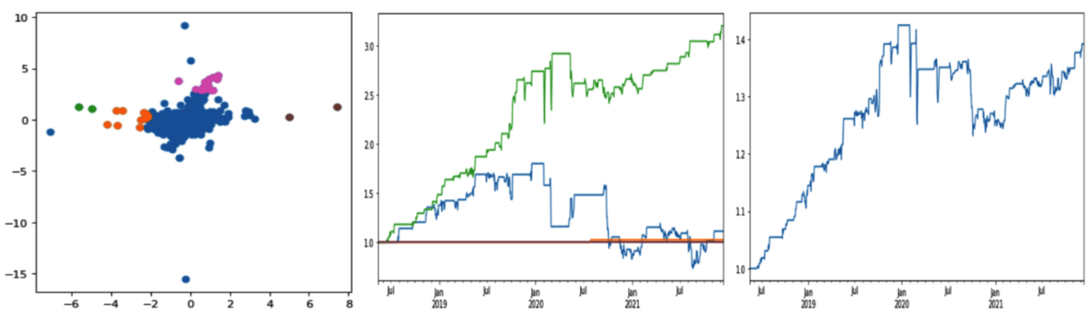
(a) Clustering (b) Return of Individual Pairs (c) CAGR

Figure 2: Performance of Pairs Portfolio Based on OPTICS Clustering

First Ticker	Name	Second Ticker	Name	Performance
MLSS	Milestone Scientific Inc	FENC	Fennec Pharmaceuticals Inc	Sharpe Ratio: 1.67 CAGR: 22%
LXRX	Lexicon Pharmaceuticals Inc	NVAX	Novavax Inc	Sharpe Ratio: 1.59 CAGR: 23%
PDT	John Hancock Premium Dividend Fund	TCCO	Technical Communications Corporation	Sharpe Ratio: 1.4 CAGR: 13%
HYT	BlackRock Corporate High Yield Fund Inc	TCCO	Technical Communications Corporation	Sharpe Ratio: 1.03 CAGR: 17%
MXC	Mexco Energy Corporation	FLNT	Fluent Inc	Sharpe Ratio: 1.14 CAGR: 18%

Table 2: Top 5 Pairs Based on OPTICS Clustering

Using Mean Shift clustering and selection criteria, the final portfolio consists of 6 pairs with Sharpe ratio of 2.41 and CAGR of 4.5%.



(a) Clustering (b) Return of Individual Pairs (c) CAGR

Figure 3: Performance of Pairs Portfolio Based on Mean Shift Clustering

First Ticker	Name	Second Ticker	Name	Performance
NEPH	Nephros Inc	TOMZ	TOMI Environmental Solutions Inc	Sharpe Ratio: 1.06 CAGR: 25%
NEPH	Nephros Inc	RADA	RADA Electronic Industries Ltd	Sharpe Ratio: 0.9 CAGR: 0.5%
ABEO	Abeona Therapeutics Inc	TOMZ	TOMI Environmental Solutions Inc	Sharpe Ratio: 0.06 CAGR: 2%

Table 3: Top 3 Pairs Based on Mean Shift Clustering

Clustering	Generated Pairs	CAGR	Sharpe Ratio	Overall Score
OPTICS	Average	Good	Good	Good
DBSCAN	Average	Average	Good	Average
K-Means	High	Good	Very Good	Good
HAC	High	Good	Very Good	Good
BIRCH	High	Good	Very Good	Poor
GMM	High	Good	Very Good	Good
Mean Shift	Low	Very Good	Average	Wild Card

Table 4: Comparison of the 7 Clustering Methods

6 Conclusion

Most contemporary literature points to OPTICS as the de facto algorithm for pairs trading. However, we believe each clustering has its strengths and weaknesses that could be tailored to solve different kind of problems within pairs trading:

OPTICS performs relatively well in terms of computing speed and back tests. It also identifies a decent number of eligible pairs that are neither too large nor too small for a reasonably sized portfolio. DBSCAN is similar to OPTICS but with comparatively bad performance in terms of CAGR and execution times. With these considerations in mind, OPTICS is best suited for a lightweight pair trading system.

K-MEANS, HAC, BIRCH, and GMM generate larger number of pairs and consume lot more computing resources. That also mean that the pairs are of lesser quality and non-profitable. But with enough dedicated research and filtering, these algorithms could potentially extract very profitable and interesting pairs. These algorithms are akin to brute force approaches that are good for identifying wide range of pairs, both good and bad. They are better suited for a directionless research for pairs identification. On a side note, one needs to consider other three algorithms first as BIRCH is harder to use due to more hyperparameters tuning.

From this study, we found that Mean Shift is capable of identifying those clusters that are closely related for some reason or the other more effectively. We recommend Mean Shift as the clustering algorithm to use to identify closely knit clusters.

References

[1] Mihael Ankerst, Markus M Breunig, Hans-Peter Kriegel, and Jörg Sander. Optics: Ordering points to identify the clustering structure. *ACM Sigmod record*, 28(2):49–60, 1999.

[2] Robert PW Duin, Ana LN Fred, Marco Loog, and Elżbieta Pekalska. Mode seeking clustering by knn and mean shift evaluated. In *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*, pages 51–59. Springer, 2012.

[3] Robert F Engle and Clive WJ Granger. Co-integration and error correction: representation, estimation, and testing. *Econometrica: journal of the Econometric Society*, pages 251–276, 1987.

[4] Oscar Lee Feng Qi. Kalman Filters in Pairs Trading, 2021.

[5] Robert Gebarowski, Paweł Oświecimka, Marcin Watorek, and Stanisław Drożdż. Detecting correlations and triangular arbitrage opportunities in the forex by means of multifractal detrended cross-correlations analysis. *Nonlinear Dynamics*, 98(3):2349–2364, 2019.

[6] Stuart Jamieson. Mean Reversion Pairs Trading with inclusion of a Kalman filter, 2018.

[7] Sofien Kabaar. Creating and Back-Testing a Pairs Trading Strategy in Python, 2020.

[8] Bingyi Kang, Gyan Chhipi-Shrestha, Yong Deng, Julie Mori, Kasun Hewage, and Rehan Sadiq. Development of a predictive model for clostridium difficile infection incidence in hospitals using gaussian mixture model and dempster–shafer theory. *Stochastic environmental research and risk assessment*, 32(6):1743–1758, 2018.

[9] Kris Longmore. Practical Pairs Trading, 2019.

[10] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967.

[11] Pekka Paalanen, Joni-Kristian Kamarainen, Jarmo Ilonen, and Heikki Kälviäinen. Feature representation and discrimination based on gaussian mixture model probability densities—practices and algorithms. *Pattern Recognition*, 39(7):1346–1358, 2006.

[12] Simão Moraes Sarmento and Nuno Horta. Enhancing a pairs trading strategy with the application of machine learning. *Expert Systems with Applications*, 158:113490, 2020.

[13] Masood Tadi and Irina Kortchemski. Evaluation of dynamic cointegration-based pairs trading strategy in the cryptocurrency market. *Studies in Economics and Finance*, 2021.

[14] Ganapathy Vidyamurthy. *Pairs Trading: quantitative methods and analysis*, volume 217. John Wiley & Sons, 2004.

[15] Mark Whistler. *Trading pairs: capturing profits and hedging risk with statistical arbitrage strategies*, volume 216. John Wiley & Sons, 2004.

[16] Tian Zhang, Raghu Ramakrishnan, and Miron Livny. Birch: an efficient data clustering method for very large databases. *ACM sigmod record*, 25(2):103–114, 1996.

Appendix

The Impact of Stop Loss on the Portfolio of Trades – Training Data

At Stop Loss of 3.0: Sharpe Ratio is 4.43 and CAGR is 0.0459.

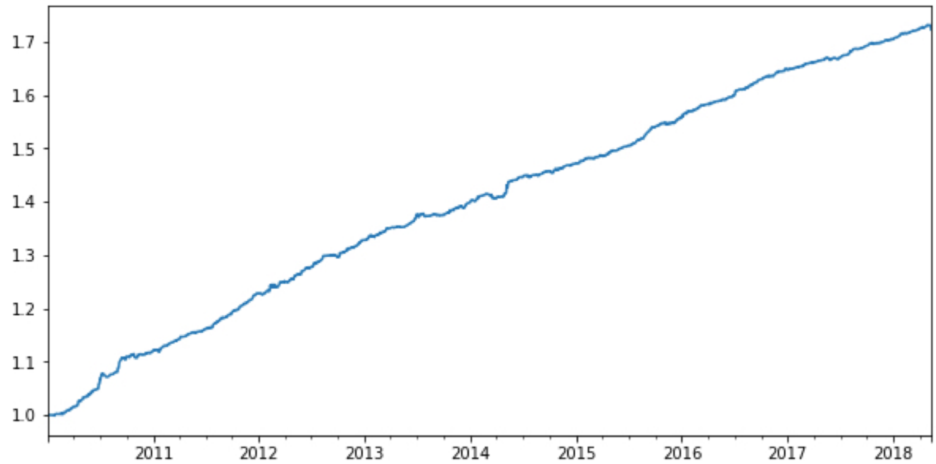


Figure 4: CAGR of the Trading Portfolio When Stop Loss of 3.0

The Impact of Stop Loss on the Overall Portfolio – Training Data

At stop loss of 2.5, Sharpe ratio is 4.01 and CAGR is 0.0346.

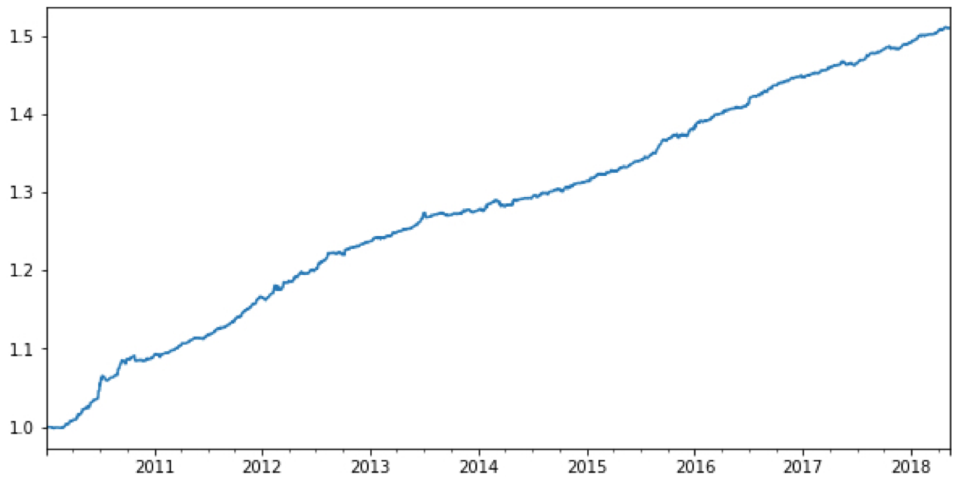


Figure 5: CAGR of the Overall Portfolio When Stop Loss of 2.5

At stop loss of 2.1, Sharpe ratio is 2.99 and CAGR is 0.0117.

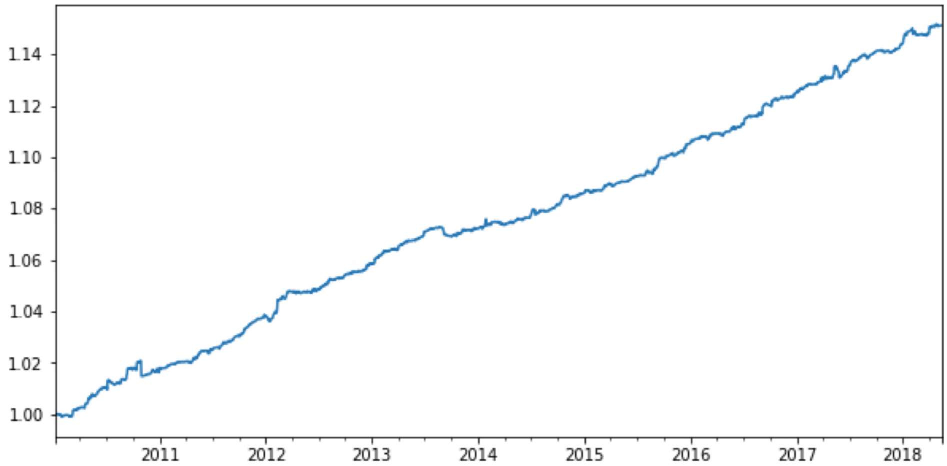


Figure 6: CAGR of the Overall Portfolio When Stop Loss of 2.1

At stop loss of 2.01, Sharpe ratio is 0.61 and CAGR is 0.0013.

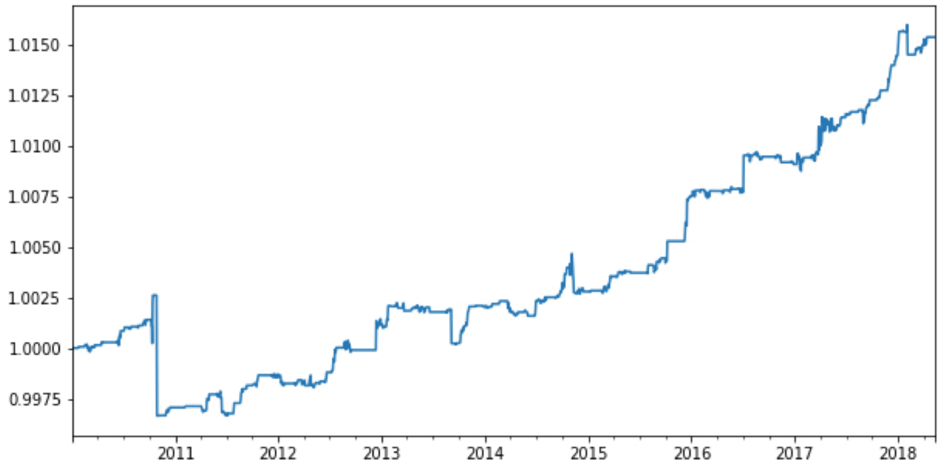


Figure 7: CAGR of the Overall Portfolio When Stop Loss of 2.01

At stop loss of 2.001, Sharpe ratio is 0.43 and CAGR is 0.0001.

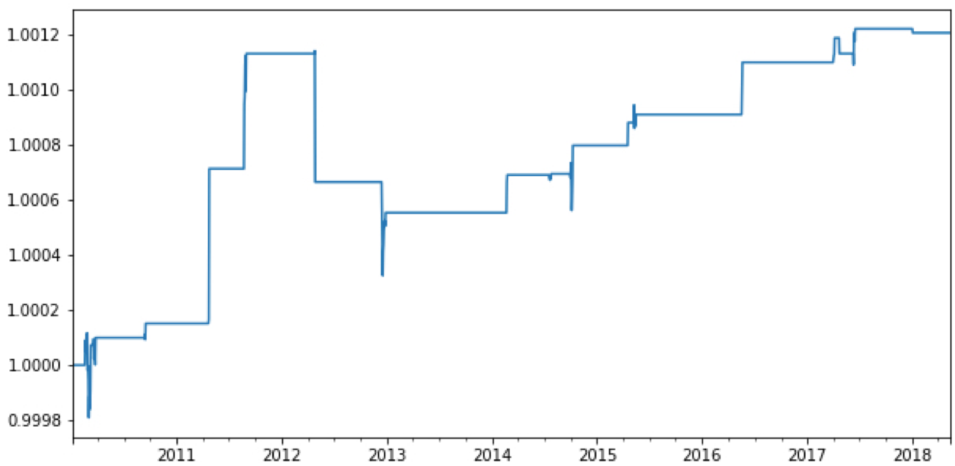


Figure 8: CAGR of the Overall Portfolio When Stop Loss of 2.001

The Impact of Stop Loss on the Portfolio of Trades – Test Data

At stop loss of 3.0, Sharpe ratio is 3.06 and CAGR is 0.0746 for 35 pairs.

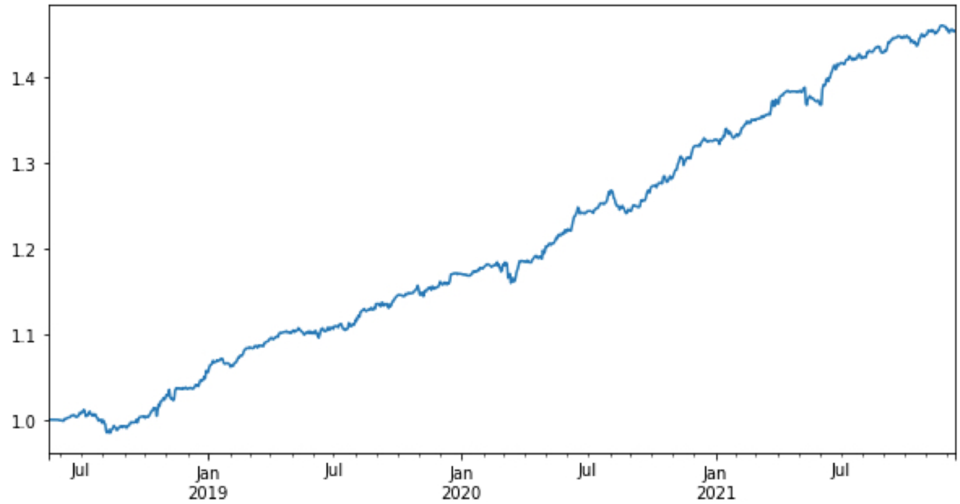


Figure 9: CAGR of the Trading Portfolio When Stop Loss of 3.0

The Impact of Stop Loss on the Overall Portfolio – Test Data

At stop loss of 2.5, Sharpe ratio is 2.41 and CAGR is 0.0518 for 29 pairs.

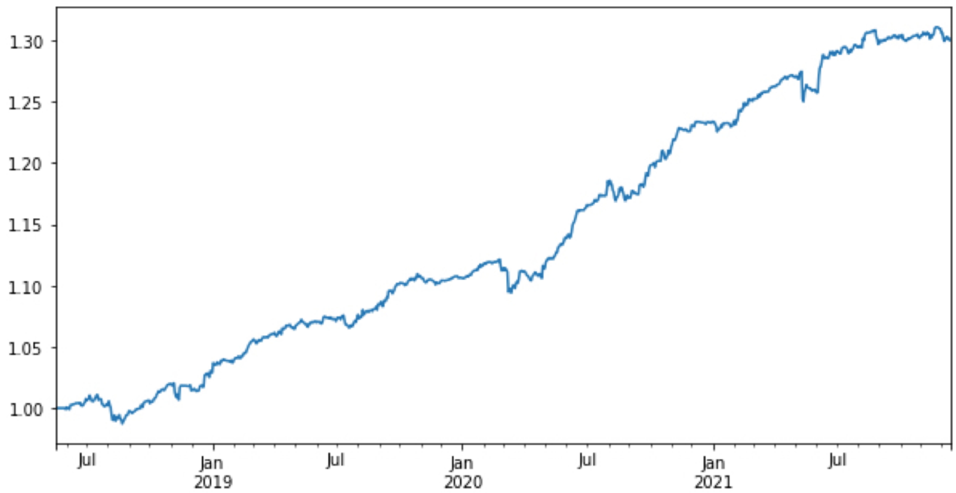


Figure 10: CAGR of the Overall Portfolio When Stop Loss of 2.5

At stop loss of 2.1, Sharpe ratio is 0.62 and CAGR is 0.01 for 9 pairs.

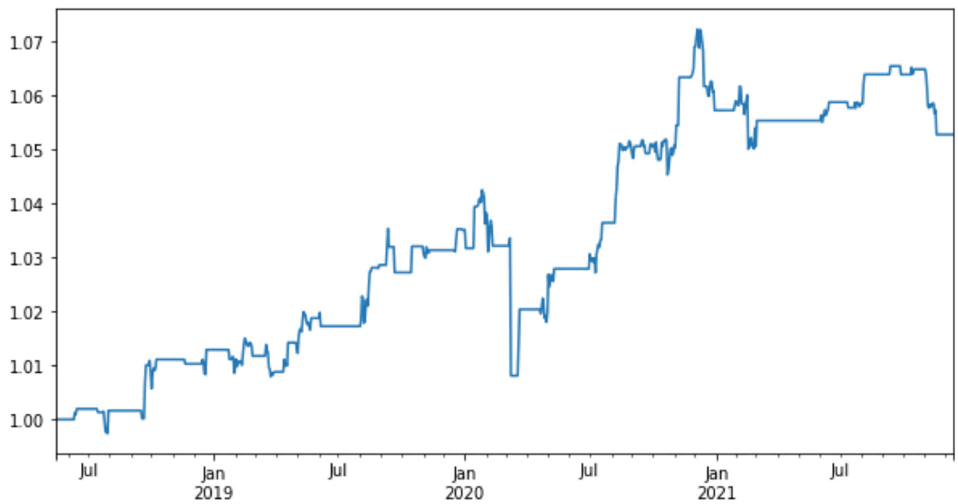


Figure 11: CAGR of the Overall Portfolio When Stop Loss of 2.1

No pairs are found when setting the stop loss at 2.01 or 2.001.