Corvinus University of Budapest

Can unsupervised learning enhance pairs trading strategies?

Tud-e a felügyelet nélküli tanulás javítani a pairs trading stratégiákon?

Supervisor:
Milan Csaba Badics

Author:
Richard Petocz
Institute of Finance, Accounting
and Business Law
Department of Finance and Accounting
Year: 2018/19/1
2022

Abstract

Pairs trading is a market-neutral trading strategy which removes the need of fundamental analysis of equities by creating mean reverting pairs with only statistical methods, whose future movements can be predicted, given that their properties stay the same in the out of sample period. In the first half of this study, I examine in the Japanese stock market from Jan 2010 to Dec 2020 which statistical test for cointegration works the best, and which formation and trading period combination and threshold for the trading rules create the best returns and Sharpe ratios. As a result, the Johansen test with a VAR model used to select lags proves to be the best statistical test, outperforming the Engle-Granger test, and the most commonly used 12 months formation and 6 months trading period with a 2 standard deviation threshold-based trading model is the best performing one. In the second half, I compare the standard industry breakdown for preselecting stocks with unsupervised learning, where the statistical factors are determined by principal component analysis and these are clustered into groups by OPTICS density-based algorithm. A low number of principal components and low minimum sample size for OPTICS proves to be the best parameters for clustering, having a 5.09% annual excess return compared to 3.32% of the industry based breakdown, but thanks to the lower number of pairs the volatility is higher as well, 10.41% for PCA and 6.12% for industry breakdown, which is turn results in a lower Sharpe ratio, 0.49 to 0.54. The higher return of the PCA cluster could not be repeated in the US stock market, which indicates that the results found are not consistent, and might only have a higher return in the Japanese stock market thanks to overfitting, which is reinforced by the fact that there is a very low number of pairs selected, 8.3 on average for the trading periods. After including transaction costs, the returns are diminished to 0.43% annual excess return for industry breakdown and 2.14% for PCA, while Sharpe ratios are 0.07 and 0.20.

I Introduction

Pairs trading is a mean reversion strategy, which builds on the promise that two linked assets price will keep their mean reverting behaviour even after discovering them, so we can make a bet on the future movement of the spread (Gatev et al., 2006). In its simplest form, pair trading consists of trading 2 assets, at each time the investor goes long on one of them and simultaneously goes short on the other, while having roughly the same exposure in each leg, thus making the pair market neutral, because under these considerations the net exposure (and the beta to the market as well) is close to zero. With this new synthetic asset, the spread between the two securities we have created something that lets us bet only on the behaviour we have insight on, the mean reversion, and eliminate the direction of the market. In this sense, the created spread incorporates the idea of buying an undervalued security and selling an overvalued one in relative terms to each, without the grueling work of fundamental analysis, relying purely on statistical insight.

One of the main concerns regarding pair trading is whether the equilibrium relationship between the two price series will hold out of sample. To find pairs, we have to conduct a lot of hypothesis testing, which increases the danger of multiple comparisons bias, or in other words, finding false positives, whose relationship was only a coincidence, thanks to doing hundreds or thousands of tests. These false pairs will diverge from their historical mean and cause big losses to the investor, who betted on mean reversion.

One method of trying to decrease this potential danger is to do a preselection of the assets in the investment universe, and only conduct statistical tests on the potential pairs, which we have previously found economicly linked. The most basic of this preselection is the sector or industry breakdown, where we are only searching for pairs within the same sector or industry, with the hypothesis that these assets (in this case equities) are fundamentally linked, and also implying that those within different sectors or industries are not (Gatev et al., 2006; Do and Faff, 2010, 2012; Engelberg et al., 2008).

There are numerous methods to determine the final pairs, but almost all of them consist of some valuation metric, e.g.: top 20/50/100 smallest squared distance (Gatev et al., 2006; Huck, 2013; Perlin, 2009; Do and Faff, 2011), or p-value for cointegration (Huck, 2015; Caldeira and Moura, 2013; Rad et al., 2015), and finding a hedge ratio between the securities. After we have identified the pairs we want to trade, the most common method to use is the threshold-based

trading system, which goes long on the spread, when it has moved some value (usually expressed in standard deviations) below the mean, and goes short on the spread, when it has moved some value above the mean, and close the existing position if the spread has crossed its historical mean from the formation period, in which both cases we can benefit from mean reversion.

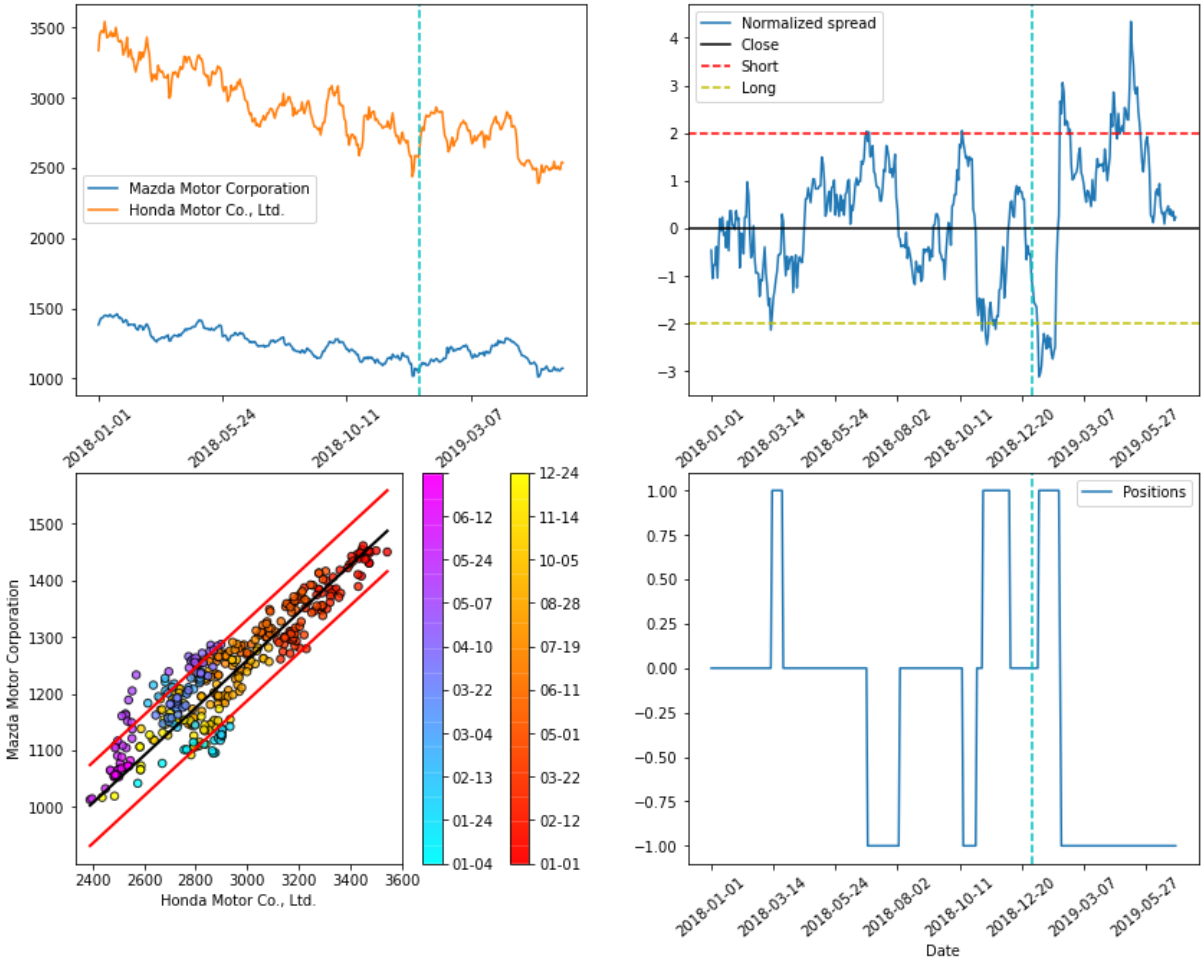

Figure 1: Example of a pair, Source: Own figure

Figure 1 illustrates a pair deemed eligible for trading. The pair was formed from price data from 1 January 2018 to 31 December 2018 from the companies Mazda Motor Corporation and Honda Motor Company, both of them are operating in the Consumer Cyclical sector, Auto Manufacturers industry (as categorized by Yahoo Finance). The top left chart shows the adjusted price series for the two companies in the period, the top right shows the normalized spread between the two assets, using a hedge ratio of 2.38. The bottom left chart depicts a scatter plot of the securities prices, indicating the change of time with different colors, the red and orange colors show the formation period, and the blue and purple colors show the out of sample

trading period. The black line is the estimated equilibrium level, and the two red lines are the opening thresholds. Most of the pairs selected are highly correlated, as in case of the Mazda-Honda pair, but this is not a necessity. The bottom right displays the positions for this pair using the described threshold-based trading model with 2 standard deviations as opening signals and the mean as the closing signal. The cyan blue dashed line indicates the end of the formation and beginning of the trading period on all 3 plots.

The origin of pair trading is different among authors, but it is generally accepted that it was developed in the 1980s and implemented by many finance firms with great success. Vidyamurthy (2004) claims that the strategy was discovered at Morgan Stanley by Nunzio Tartaglia and his group of mathematicians and computer scientists with the mission to create a fully automated statistical approach to trading.

According to Richard Bookstabber (2007), the idea did not come from Tartaglia himself, but from a young programmer called Gerry Bamgerber, who also worked at Morgan Stanley at the time. His task was to report profit and loss (PnL) to the traders of their hedged positions as a single block, so their movements could be better understood. As he investigated these price movements, he noticed that pairs of hedged stocks tend to move together, their prices stayed in a type of equilibrium even after a big trade, so there were some common characteristics between them. He had the idea to think of the two stocks as two legs of the same trading strategy, going long on the underperforming stock and going short on the overperforming one. He had managed to secure himself a trading desk and made $6 million in his first year. After this Morgan Stanley executives gave oversight of the pair trading program to their trusted colleague, Tartaglia, and many know the story only from this point on, leaving Bamgerber's name in the mist. After this, he joined Prinction Newsport for a few years, executing his strategy with success, and 2 years later he left the world of trading in 1987.

Paul Wilmott attributes the discovery of pair trading to himself and his firm, although acknowledges Gerry Bambergers contribution and independent discovery. They were researching short-term mean reversion on stock as a new indicator that can explain price movements. From this they started to look at stock pairs with high positive correlation, with the promise that the deviation from their historical mean would be corrected.

The real origin of the pair trading strategy might never be known, but from these stories it is clear that this systematized trading strategy quickly spread around firms, not to mention that

Tartaglia's team branched off in a few years time to different hedge funds. As computers became widespread, the high computational power required for pair trading was not only attainable by big investment houses. As Andrew Pole writes, pair trading has entered an ice age in the early 2000s, because of the slimming margin caused by more and more institutional money trying to capitalize on the strategy, and has become unprofitable.

After many years of silence, interest in pair trading picked up again, as new statistical methods and machine learning techniques were employed to create mean reverting long-short pairs in order to make profits. This paper aims to answer whether machine learning can enhance the profitability of pairs trading strategies. I am going to use two unsupervised learning techniques to create a statistical factor breakdown for stock instead of the usual industry breakdown. This two techniques are dimensionality reduction to find the factors for assets from their returns, and density based clustering to find groups of assets with similar exposures to these statistical factors.

II Theoretical background

II.1 Literature overview

In the academic literature, pairs trading has gained popularity after an article by Gatev et al. in 1999, where they have established one of the fundamental frameworks for pairs trading, the distance approach. They tested the strategy on data from 1963 to 1998 on the US market, paired stocks that had the smallest squared distance in the formation period, and used different trading rules to access the strategy. Their results show that the strategy could produce significant excess returns which were robust to trading costs, and that pairs trading adds value compared to a simple mean reversion strategy. The same authors reproduced this paper in 2006 to conclude that the strategy is still profitable and robust to transaction costs.

These two articles were used as a benchmark for Do and Faff for their study in 2010, focusing on the profitability of pair trading, using data up to 2008. They found a declining trend for pair trading, where recent time periods perform poorly compared to the older ones. They concluded that overcrowding with institutional investors are not the primary reason for the decline, but the less reliant convergence property of the pairs. Worth mentioning the fact that in their study they found that economic turbulence has a positive effect on the performance of pairs trading. The researchers propose two new aspects for pairs trading: the industry breakdown, matching stocks

only if both of them are in the same industry, classified by Fama and French (1997) and the number of mean crosses, which represent how many times the market players have corrected the mispricing. Both these metrics improved the strategy and showed that the distance approach is not sufficient to capture economic links between assets, which causes more divergent pairs, and this was amended by the industry breakdown and mean crosses that could improve the results significantly.

One of the first instances to use cointegration to create mean reverting long-short market neutral strategy is in Alexander and Dimitriu's paper (2002). They used the Dow Jones Industrial Average to test the algorithm, which showed robust returns with very low volatility and almost no correlation to the market. They have also shown that by increasing the spreads for the long-short portfolio they also increase the volatility, which can not produce enough excess return to increase the Sharpe ratio, so eventually the risk-adjusted return suffers from it.

Huck (2015) uses the S&P 500 and Nikkei 225 indices for pairs trading to compare the distance, stationarity and cointegration pair selection frameworks, where he finds that the cointegration approach is the only one with significant excess returns of 1.5% per month. The study also mixed trading signals with the VIX index to time trades, building on the idea that pairs trading has seen the best performance during market turbulences, but with no success, as the results showed that timing volatility does not increase the profitability of the strategy under the cointegration approach. Another application of cointegration is in Caldeira and Moura (2013) on the Brazilian stock market with an excess return of 16.38% annually, and Sharpe ratio of 1.34, while having the extra constraint of only choosing the top 20 pairs with the highest in-sample Sharpe ratio.

Engelberg et al. (2008) examined the effects of news, liquidity and common information on pairs trading and found that profits are higher when the divergence is caused by news, which temporarily decreases the liquidity for one of the stocks, or when a news effects both shares in the pair, but they react at a different speed to this information, for example because one of them have a bigger share of institutional holdings, thus creating a divergence initially and convergence later, when both stocks have incorporated that news. This can be explained by the different liquidity level of the stocks, which also provides an explanation for the profits, which can be either for providing liquidity or for taking a position in a stock that reacts slower to common news. This phenomenon is stronger for smaller, less liquid stocks that are unlikely to be covered by the same sell-side analyst and held by the same institution.

Idiosyncratic news for only one of the stocks in a pair increases the probability of opening a position by diverging prices, although this price movement is less likely to revert, and these news increase the length of the holding period and decrease the probability of converge, thus lowering the profitability of the overall strategy.

Jacobs and Weber (2015) concluded that pairs trading is persistently profitable across 34 international markets using data from 2000 to 2013, but it is not consistent over time, and it is effected by investor attention, news and limits to arbitrage. The best returns are found in emerging markets, where the average idiosyncratic volatility is large and in countries where there are a lot of eligible pairs, and also the stock market's relative size to the country's economy is a big portion. Their analysis shows that pairs which experience firm-specific news coverage have considerably lower returns compared to the average pair, and on the days when macroeconomic news are released, that are common to stocks, the returns for pairs trading are higher than usual, which is in line with the findings of Engelberg (2008). They found that the investor attention negatively effects the profits for pairs trading both in cross-section and in time series, as it reduces trading opportunities.

Rad et al. (2016) compared three different pairs trading strategies from 1962 to 2014 on the US market, namely, the distance, cointegration and copula approaches, and found that the distance and cointegration methods provide a higher excess return, although the copula approach had in recent years more trading opportunities, compared to the other two methods. One of the reasons for the weaker performance of the copula method can be attributed to the large number of unconverged trades. All three strategies have significant alpha and all of them benefit from turbulent times with high volatility, with the cointegration approach performing best in these times. The returns for all three strategies had a negative correlation to the liquidity factor and no correlation to the market returns, which confirms the market neutrality of pairs trading strategies.

Dunis et al. (2010) examined the effect of high-frequency data on pairs trading, which diminishes the returns for standard strategies caused by higher turnover and more transaction costs, but combining them with in-sample information ratio selection can improve significantly the results. The intraday out-of-sample information ratios for 5 different frequencies averaged around 3, while the daily data sampling produces an information ratio of 1.3 with only the best 5 in-sample pairs traded. These very promising results are overshadowed by the fact that their dataset is only 50 European stocks with 4.5 months of high-frequency data, which is too small

to draw definitive conclusions of their method. In case they had difficulty getting historical intraday data, they should have at least applied the strategy on different markets.

To apply PCA for clustering purposes, Cardoso (2015) used the statistical factors for hierarchical clustering with the distance approach, which has lowered the volatility of the strategy for the cost of lowering the returns as well, with no conclusion to Sharpe ratios, as they have moved in both ways on 3 different markets. Overall, the PCA approach had better performance than the industry group breakdown in terms of volatility, which can be contributed to better market neutrality, as the long-short pairs eliminate each others exposure to the factors, thus lowering the risk of the strategy. The results found in his paper were diminished when transaction costs were subtracted, and he has found them to be the most determinant factor for the profitability, this can probably be accounted for the high turnover of the moving average threshold-based trading strategy he has used.

Multiple machine learning techniques were used to find better ways of trading pairs in Sarmento and Horta (2021). On 208 ETFs their PCA based clustering achieved better Sharpe ratio compared to the industry breakdown or no breakdown at all. They used an LSTM forecasting algorithm for the spread to trade, which could reduce the decline periods caused by diverging pairs by 75% percent, at the expense of lowering the returns as well.

II.2 Main concepts

There are a lot of definitions and theoretically related concepts necessary for pairs trading. The idea that a trading strategy can work purely based on statistical properties, without any insight into the companies fundamentals and their prospects requires a good understanding of the underlying concepts. I will explain the most important topics on the fundamental of how and why pairs trading can be a statistically established and profitable strategy.

II.2.1 Mean reversion

Mean reversion can be defined as the following stochastic differential equation:

$$dX_t = \mu\,(\theta - X_t)dt + \sigma dB_t,$$

where X is the mean reverting process, μ is the intensity of mean reversion, θ is the long term mean of the process, and B is a standard Brownian motion. This is called an Ornstein-

Uhlenbeck process. In every step, there is a new increment to the process, which consists of two parts. The first one is the deterministic part, there is no stochasticity in it, this part of the increment is negatively proportional to how far the process is at time t from the mean θ. The other part is the standard Brownian motion, which is a stochastic process that is only dependent on the standard deviation σ of the process. After considering this, the Ornstein-Uhlenbeck process can be thought of as a Brownian motion with a mean reverting property.

After fitting the parameters of the Ornstein-Uhlenbeck process, we can calculate the half-life, which is $\ln(2)/\mu$. The half-life is a measure of how fast the process is returning to the halfway point between its current value and its mean. Usually in pair trading the smaller the half-life the better, as it was used in Sarmento and Horta (2021). A small multiple of the half-life is an appropriate parameter for the lookback period or rolling window length for trading execution (Chan, 2013).

The augmented Dickey-Fuller test, one of the most popular ways to test for stationarity, builds on a similar idea. The null hypothesis is that the increment of the process is independent of the previous increment, thus making it an autoregressive of order 0 or AR(0) process. The alternative hypothesis is that it is dependent, so it is an AR(1) or higher order process. The coefficient of the previous increment has to be negative in order for the ADF test to be significant. This is in line with the assumption that the spread is mean reverting, because the negative coefficient of the previous increment would keep the process close to its mean, and the fact that it is an AR(1) process also means, that it can not be a simple Brownian motion, the dispersion would be slower, echoing the definition of stationarity by Chan (2013), that the variance is a sublinear function of time, and strengthens the assumption behind the hurst exponent.

The linear model that the Augmented Dickey-Fuller test builds on is the a modified discrete version of the Ornstein-Uhlenbeck stochastic differential equation. From this discrete form, we can calculate the half-life of mean reversion, by using a linear regression with the change in the process (Δx_t) as the dependent variable, and the the value of the process at the previous time instant (x_{t-1}) as the indipendent variable we get μ, which can be interpreted as the speed of mean reversion, and $\ln(2)/\mu$ is the half-life.

II.2.2 Cointegration

There is cointegration between multiple time series when all of them are integrated of order of 1 and there exists at least one linear combination, that is stationary and integrated of order of 0. It is hard to find stationary time series in finance, but with this concept we can create them artificially, and since stationary processes can be easily predicted, since their statistical properties, such as the mean will not change, we can coin a trading strategy that builds on this assumption.

This statistical concept was introduced by Engle and Granger in 1987, where the authors suggested a procedure, called the Engle-Granger cointegration test, which is as follows: (1) test with the augmented Dickey-Fuller test if the time series are stationary, if they are not, then there is unit root in the process, and we can continue with the next step, (2) run an ordinary least squares regression between the two time series, (3) use the residuals from the previous step to test for stationarity with the augmented Dickey-Fuller test (or any other unit root test), (4) if we can reject the null hypothesis of the existence of a unit root in the residuals, then that means that the the residuals are stationaty and the time series are cointegrated. (Engle and Granger, 1987)

An inconvenience in the Engle-Granger method is the choice of the dependent and independent variables for the linear regression, as the two different possible arrangements give different results, the spreads and their p-values can be different, and their there is no way of telling which one will be better ahead of time.

The Johansen test expands the cointegration test to multiple time series, and the hedge ratios are also defined by the test, this eliminates the need to select the dependent and independent time series ahead of time, and gives significance levels for more than one cointegrating linear combination, and orders them, from best to worst, which is very convinient for pairs trading purposes. Since I am only trying to find pairs of stocks, which means that the Johansen test will give 2 weight vectors for the stocks, I am going to use the one with the higher eigenvalue, which indicates a better result. To determine whether that one linear combination is significant, I am going to use the trace statistic.

II.2.3 Hurst Exponent

Hurst exponent is one of the latest propositions to use as a pair selection method. It was introduced by Ramos-Requena et al. in 2017 and in a later article by Balladares et al. in 2021, where Ramos-Requena was also a co-writer. The hurst exponent approach focuses on the idea of mean reversion, picking the pairs which have the strongest mean reverting property for trading. The statistic for mean reversion is derived from the below equation:

$$\{|x_{t+\tau} - x_t|^2\} \sim \tau^{2H}$$

x_t is the value of the process at time *t*, τ is the timeshift, *H* is the hurst exponent and {-} denotes the average. An important definition to understand before delving into what the hurst exponent measures is the quadratic variation. If we set $\tau = 1$ and instead of calculating the average, we take the sum of them, then we get the quadratic variation, which is equal to the time change of the process, t, if the process is random, e. g. Brownian motion. In case of the hurst exponent, we calculate the average of them, so it is equal to the timestep τ, and this makes the hurst exponent for truly random processes 0.5. As we increase τ, more and more memory of the process is represented in the average, and the hurst exponent becomes larger than 0.5 for a time series that is trending, and smaller than 0.5 for a time series that is mean reverting, since the average dispersion of the process under a given τ interval is larger or smaller, then it would be in case of a brownian motion.

The writers of the articles had success in finding profitable pairs, along with the introduction of a new direction for research. The pairs selected with hurst exponent out of the Dow Jones Industrial Average from 2000 to 2015 outperformed the ones selected with the distance and correlation approach, and had low correlation with the standard methods, which shows that this method is fundamentally different from the others.

One of the concerns with this method is that a time series can be mean reverting without returning to its historical mean. The authors applied a rolling window threshold-based trading model with tight stoplosses, which shows that they also encountered this problem, and decided to use rolling window statistics. The combination of the hurst exponent and one of the standard methods could produce better returns, as in Sarmento (2021).

II.3 Statistical factor breakdown

In most of the literature, the researchers are using some kind of breakdown to define pairs, for example, only searching within industry groups (Do and Faff, 2010, 2012; Engelberg et al., 2008). This has the advantage of having less calculations to do, making the process faster, and besides that the lower number of pairs to test is also decreasing the effect of multiple comparisons bias, which is in other words finding false positives. When deciding that a pair is significant, we are committing money to it, and in case of a diverging pair that money is bound to be lost, that is the danger of false positives, while false negatives only have the effect that we could not capitalise on the opportunity, which is much less risky, than the false positives.

Another reason to define a breakdown is to categorize securities into groups of similarly behaving assets that are bound to move together because of some obvious or unexplainable economic link, which is long run, so we can build on the hypothesis, that what connections we found in the past will be extended in the future.

II.3.1 Arbitrage Pricing Theory

The main idea behind pairs trading is to find assets that are related in some way, after a pair has been selected, the investor bets that the prices will move closely together, deviations from the historical equilibrium are short term anomalies, which can be exploited to profit when prices converge again.

Arbitrage Pricing Theory (APT) provides a framework to find closely related assets, without the constraint of being in the same industry. APT defines asset returns as a linear combination of factors and a specific return for every asset.

$$r_X = \sum_{i=1}^{n} ß_X^i \, f^i + r_X^e + \varepsilon_X$$

$$E[r_X] = \sum_{i=1}^{n} ß_X^i \, E[f^i] + r_X^e$$

Where r_X is the return for company X, $ß_X^i$ are the company specific weights of the factors, which can be interpreted as the sensitivities to the factors, f^i are the factors, r_X^e is the company specific component and ε_X is random white noise with zero mean and it is uncorrelated with the factors.

The expected return for company *X* should be the linear combination of the expected returns for the factors with the company specific component.

Given that the factors are uncorrelated with each other, the weights can be expressed as:

$$ß_X^i = \frac{cov(r_X, f^i)}{var(f^i)}$$

For this framework to be eligible for multiple time periods the following assumptions have to be met: (1) r_X^e has to be constant in time for the given time period, as it can be interpreted as a representation of the company fundamentals. The theory assumes, that there is no major change in them, and the deviation from the expected return of the company is captured in ε_X. (2), f^i are exogenous factors representing the state of the market. They are only observable for us, without the ability to change them, an outside process is generating them, which is not explained in the model. (3) $ß_X^i$ are the sensitivities of company *X* to the factors, without a restructuring or fundamental shift in the company's profile these weights are constant for the period.

APT allows relationships between the factors, but they should not be perfectly correlated, as redundant factors could be removed, and a more compact model could explain just as much with fewer factors. The factors are changing through time, but r_X should not have an affect on them, as they are determined outside of the model, capturing the fundamentals of the market. We should not expect the dependence structure between two factors to change with time.

$$E\left[f_t^i\, f_s^j\right] = E\left[f_{t+h}^i\, f_{s+h}^j\right]$$

Where f^i and f^j are factors and *t*, *s*, *h* are time instances bigger than 0. The above expression states that the dependence between the *i* and *j* factors at different times should not change with time shifts. (Rampertshammer, 2007, p. 6-7)

II.3.2 Clustering

Following Sarmento and Horta (2021) for the requirements of clustering and the decision to choose density-based clustering, I can reason the following: the clustering algorithm should meet these 3 criterions, (1) not to cluster all data points, (2) the number of clusters should not be specified, as we do not know the structure of the data, (3) clusters should be able to take any shape. With an algorithm like this, we have a process to cluster companies that is robust to outliers, the number of clusters is data driven, and the clusters shape can take any form.

Partitioning clustering algorithms (e.g., k-means) will not meet any of the criterions above. The number of clusters has to be determined in advance, it does not handle noisy data well, not robust to outliers, and it assumes a convex shape for the clusters by minimizing the squared distances of the points to the cluster means. Using a partitioning algorithm would introduce a lot of bias, and with even the best efforts would overfit the data.

Hierarchical clustering requires the investor to decide the granuality of the clustering, which leads to bringing biases from the investor to the clustering. As a result, this method of clustering could replicate one of the standard search methods, which would defeat the purpose of using machine learning.

Density-based clustering seems to be a good fit for the criterions: it is robust to outliers, the number of clusters are determined by the algorithm, without any external bias, there is no assumption for the shapes of the clusters, the algorithm simply picks the closest points to each other, which is exactly what we want.

The DBSCAN algorithm dissects the data points into connected clusters and outliers based on the Euclidian distances (Ester et al., 1996). Clusters should be recognized by having a typical higher than average density. The DBSCAN algorithm uses two parameters: *minPts*, which is the minimum points a cluster has to contain, and ε, the maximum distance by which two points can be considered connected. The clustering process can be described in 3 steps: (1) find all points that have at least *minPts* in their ε distance, and call these core points. (2) Match these core points if they are in ε distance of each other and make the initial clusters. (3) Assign non-core points if they are in ε distance of a core point to its cluster. By this way, the algorithm has also separated outliers that are not considered connected to the clusters, making the clustering more robust. This is a great achievement, but there is still a characteristic that is not ideal. By choosing only one parameter for ε, it is very sensitive to it and assums that all clusters have the same density, whereas we have no insight to suspect this. We have to find a different algorithm that solves this problem.

OPTICS provides a smart solution to the previous challenge faced with DBSCAN. The core distance is defined as the distance within which a point has *minPts* of other points in its neighbourhood, if a point does not have *minPts* in ε, then it is not a core point. The reachability distance between two points is the maximum of their core-distance and their actual Euclidian distance. OPTICS differs from DBSCAN in the cluster building process, while DBSCAN uses

simply ε to build the initial clusters, as described above, OPTICS creates a priority queue, which can be found in detail in the original paper about OPTICS (Ankerst, Breunig and Sander, 1999).

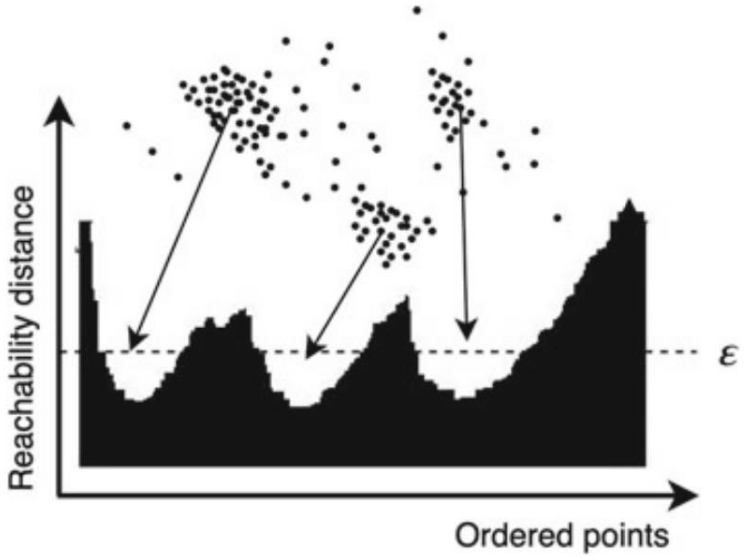

Figure 2: Visualization of the priority queue for OPTICS, Source: Sarmento et al., 2021

The gist of the procedure is that the closest points in space become neighbours in the queue, as seen in Figure 2. Clusters have a lot of points close to each other, with low reachability distances, it is depicted as valleys in Figure 2, the deeper the valley, the closer the points are. From this queue, the result from the DBSCAN algorithm can be extracted by using ε as the maximum distance two points can have to be considered connected, illustrated in Figure 2 as well. However, OPTICS has a more sophisticated method of creating clusters, it defines a minimum steepness to detect cluster edges, thus eliminating the need to define an arbitrary ε for the cluster density, and making the clustering process even more data driven, exactly what we need to exclude as many biases from the creation of clusters.

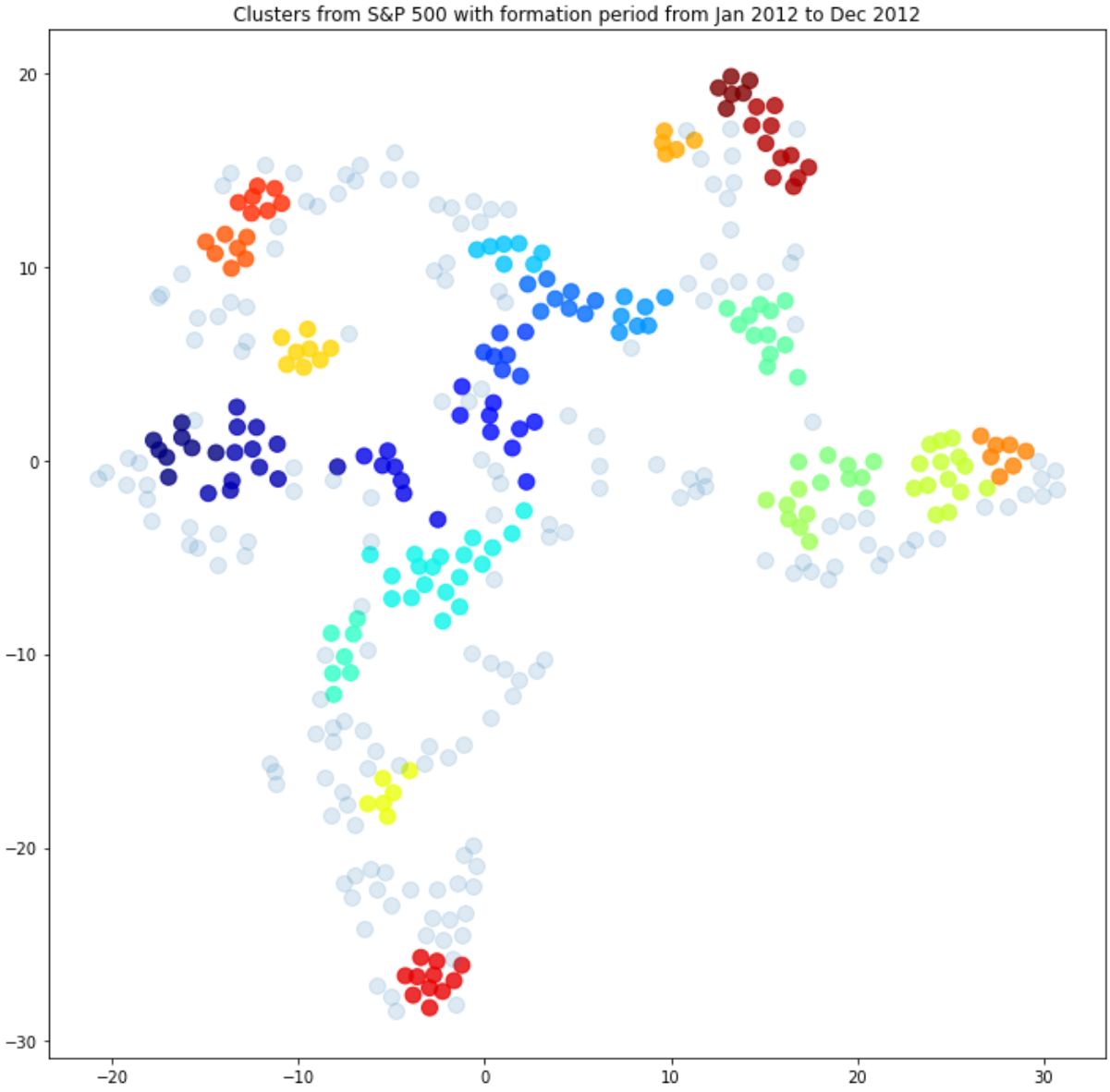

Figure 3: Representation of the output from clustering statistical factors with OPTICS, Source: Own figure

In Figure 3 the clustering output can be seen with the help of t-SNE. For this data there are 3 principle components, to visualize their disposition in space, we need the t-SNE algorithm, which is exactly designed for this purpose (Maaten and Hinton, 2008). Every point represents a company, different colors show different clusters, 23 in total, the light gray points are the outleir companies. From this approximation we can assess what OPTICS is doing. As we can see, the companies in the same clusters are closely together on the plot, and it is hard to find a dense region of companies of minimum sample size, which is 5 in this case, that are classified as outliers.

III Methodology

To have a clear picture of what I am testing on what dataset and why, I am going to describe my processes since the beginning of data gathering to the final analysis, so it is easy to understand and can be replicated to tweak it for further research.

III.1 Data

To access the profitability of the strategy, I have implemented it on two different markets: United States and Japan. If it can create outstanding returns in both of these markets, then the strategy produced in this paper is truely robust to different market environments, and not just a result of overfitting. Using data from different stock markets has already been done to examine the robustness of the trading strategy (Jacobs and Weber, 2015), and these two markets have been in the focus of Huck (2015).

To determine the scope of stocks that can be selected to form pairs, I used two stock markets' respective stock indecies, as it has been done by some researches before (Huck, 2013, 2015; Caldeira and Moura, 2013): S&P 500, which is a market capitalization weighted index of 500 large-cap stocks listed on American exchanges, and Nikkei 225, the stock market index for the top 225 Japanese blue-chip companies on the Tokyo Stock Exchange. The S&P 500 and Nikkei 225 were chosen as well-established benchmarks for trading strategies.

The components of each index were downloaded from Bloomberg as of January 1st from 2010 to 2020, that is 11 years for testing, which includes up to the most recent full year, so the question whether pairs trading is still profitable today should be answered with this dataset. To get the different composition through time was important to avoid look-ahead bias and to test on portfolios that could have been created in the past. The historical price series for the stocks were downloaded from Yahoo Finance from 1 January 2007 to 1 January 2021, using the adjusted prices, because it accounts for stock splits and dividends. The company sector and industry information was also downloaded from Yahoo Finance, which uses the Global Industry Classification Standard (GICS) breakdown, that classifies companies to 11 sectors and 157 industries in total.

I have to mention that Yahoo Finance does not keep historical prices for delisted companies (regardless whether it had gone bankrupt or got acquired), so in this sense, this dataset contains survivorship bias, the tests do not include all pairs that might have been selected in the past.

III.2 Data cleaning

In case of the S&P 500 the distributed share classes were one of the problems that could harm both the pair selection and the clustering phases, because OPTICS would falsely conclude that there are some stocks with almost the same exposure to factors, when in reality they are the same company, and this would adversely effect the clustering algorithm. Tickers GOOG, FOX, NWS, UA and DISCK were removed from the dataset, because all of their companies had a different share class that was added to the index sooner.

Every year from 2010 to 2020, I construct a test dataset from the original dataset that contains 12 months (or the relevant formation period) of data backwards from the year where I have the index's ticker list, this is the formation period, because I test for cointegration and other characteristics on this data. I can only include companies that I have data for, so the intercept of the ticker list for the indecies and those companies still listed on the exchange will be in the test data. In these 12 months long period datasets, there are some missing values, which can be caused by either the stock was not a publicly traded company at the start of the formation period, or there are some randomly missing values, which is more likely on the Japanese stock exchange. To solve this, I remove every company that has 20% or more missing values in the formation period, then forward fill the rest to avoid randomly missing values, and then eliminate those which still have missing values, because at this point it can only be at the start of the period up until some time, so this means the company was not publicly traded throughout the whole formation period.

III.3 Trading period

After having a set of pairs selected for trading, there are lots of different methods of creating a portfolio to trade, e. g.: market-neutral (Gatev et al., 2006), where the investor always opens a position with a net exposure of zero or using the cointegration coefficient to weight the stocks (Rad et al., 2016). The method I am going to use is to use the weight for the pairs obtained from the Johansen test, and leverage every trade to open it with 200% gross exposure, and this will roughly translate to 200% gross exposure throughout the period, the two factors that can

influence this are the open trade ratio at any given time, this can only decrease the gross exposure, and the individual pairs movement, that will move the gross exposure after opening the trades, this effect would cancel out in usual circumstances, except in very strong bull market or bear market, where all of the pairs gross exposure would move to the same direction. 200% gross exposure is realistic for practicing long-short equity hedge funds.

It is a common practice to use a 1 day lag to execute the trades from their discovery (Do and Faff, 2010; Gatev et al., 2006). This makes sense, the signals are created from prices at the market close, so they cannot be executed that day, only later, which is in my case the next market close, due to using daily prices. This practice makes more sense when testing on intraday data, and probably has a smaller impact than in the case of daily closes, but I am going to apply this to make the results more robust, and to follow the practices applied in previous studies, to make the results more comparible.

For trading rules, I am using a basic threshold-based method. I calculate the historical mean and standard deviation (SD) of the spread in the formation period and use these statistics for the trading period. If the normalized spread takes a value smaller than -2, I go long on the spread, similarly, if the value goes above +2, I go short on the spread, and, in both cases I exit the position if the mean is crossed. The most widespread parameter used for threshold-based method is 2 SD for opening positions and 0 SD for closing them (Do and Faff, 2010, 2012; Huck, 2015; Engelberg et al., 2008; Gatev et al., 2006), but there are some papers implementing different parameters, like Caldeira and Moura (2013) with 2 SD for opening and 0.5 SD for closing and Huck (2015) 3 SD for opening and 0 SD for closing.

There is a stoploss implemented at -5/+5 SD from the mean, if the spread crosses one of these levels, the position will be closed, and the pair will not be traded for the rest of the trading period. The number of spreads to buy based on gross exposure explained previously, and trades are placed with 1 day lag to the signals, as it was mentioned above.

III.4 Principal Component Analysis

The factors in APT can be defined or extracted in different ways, in this study I am going to use the statistical approach, because of the mathematical certainty of uncorrelated factors, and the fact that most of the problems the statistical approach suffers from is not relevant in this context.

One of the difficulties is that the factors and the sensitivities to them are unobservable or latent, and can only be estimated, for example, by principle component analysis, which gives a best fit for them based on historical data. Another challenge is the interpretation of the factors, compared to macroeconomic or fundamental factors, the resulting eigenvectors are hard to translate to economic phrases, and forecasting from them is difficult. Since we are looking for pairs of stocks with similar sensitivities to factors, these problems are not relevant, and we can be satisfied with the clusters found based on principle component analysis (PCA), since their economic interpretation is not necessary, the only assumption we make is that they still explain most of the asset returns out-of-sample, so the spread of the stocks with similar sensitivities should be stationary.

PCA transforms correlated observable variables into linearly uncorrelated variables, which are called the principle components. The first principle component defines the most variance in the original dataset as possible, and then the second component, which is orthogonal to the first one, accounts for most of the variance in the rest of the dataset and so on. Each component can be interpreted as a risk or statistical factor. We can use the assets normalized return series for the PCA to create orthogonal factors. These factors are some linear combination of the original variables (Avellaneda and Lee, 2010).

The standard way of applying PCA is to use the proportion of variance explained by the eigenvectors to determine the number of principle components. In a clustering framework, we have to make other considerations, additionally to the pay-off between having more variance explained by more components, and having redundant factors, there is the problem of the curse of dimensionality. With more dimensions, the distances in Euclidian space grow exponentially, and they are less meaningful (Bellman, 1966). With this in mind, I should choose a low number of principal components, so that the curse of dimensionality is avoided.

III.5 Statistical tests

The most popular method is the Engle-Granger cointegration test (Caldeira and Moura, 2013; Rad, Low and Faff, 2015). As it is described above, the Engle-Granger method estimates the hedge ratio between the two assets with an ordinary least squares (OLS) regression that minimizes the function

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$$

which is not symmetric if the independent and dependent variables are swapped, as it was pointed out by Armstrong (2001), and Do et al. (2006). This issue can be solved by doing both regression and going forward with the one that has a better test statistic, and then comparing its p-value to a selected α level (Sarmento and Horta, 2021). To avoid doing two tests for every pair, and prefereably to get a better hedge ratio, I am going to use reduced major axis (RMA) regression, which minimizes the function

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)(X_i - \widehat{X}_i)$$

thus making it symmetric regardless how we choose independent and dependent variables.

For the Engle-Granger test, first I tested both time series for stationarity at 5% significant level with the augmented Dickey-Fuller test, if both of the tests were not significant then I estimated the β with the RMA regression, giving the spread S2 – β * S1. If the spread was significant at 5% with the ADF test at a constant mean and with leg selection using Akaike information criterion (AIC), then I selected it for trading.

Since I have not found a detailed description of the use of the Johansen test in the pairs trading literature, I am going to test the Johansen test both with 1 lag selected and another one where the lag was selected by Akaike information criterion using a vector autoregression model (VAR).

In the first case, the selected lag is 1, and the test statistics are compared to the critical values for 5% significant level. For the two time series there are two hypothesis tests given by the Johansen test, using the trace statistic, the first hypothesis tests if there are 0 or more than 0 cointegrating vectors, and the second tests if there are 1 or more than 1 cointegrating vectors are present. For the two time series to be cointegrated, the first hypothesis has to be significant, and we have to reject the null hypothesis, that there are 0 cointegrating vectors, and we have to fail to reject the second hypothesis, which means that there is only 1 cointegrating vector is present.

In the second case, I use a VAR model and AIC to select a lag length. After this, I test with Ljung-Box test to see if there is autocorrelation in the residuals from the VAR model. If the residuals are not autocorrelated at 5% significant level, then I use the lag length minus one from the VAR model, since the VAR is run on levels, and the VECM in the Johansen test is run on first differences. As the Johansen test is very sensitive for lag selection, it will be a good comparison to see if this added complexity is enhancing the results, or a simple Johansen test with 1 lag is just as good.

IV Empirical results

Before trying to examine which selection breakdown performs better, I am going to test which statistical test selects more profitable pairs, what formation and trading period works best, and what is the best threshold rule for trading.

IV.1 Statistical tests results

	Engle-Granger	Johansen 1 lag	Johansen VAR AIC lag
Annual excess return	-0.49%	1.80%	3.32%
Annual volatility	3.44%	7.08%	6.12%
Sharpe ratio	-0.142	0.254	0.542
Positive months	52.3%	55.3%	59.1%
Max drawdown	-17.49%	-28.94%	-16.00%
Selected pairs	66.9	22.0	21.0
Profitable pairs	29.5 (44.1%)	10.2 (46.2%)	8.9 (42.4%)
Hit stoploss	26.7 (40.0%)	8.4 (37.9%)	9.7 (46.1%)
Average trades	1.02	1.14	1.05

Table 1: Summary of the results with different test, Source: Own table

All of the above results are for opening signal 2 standard deviation from the mean, closing signal is the mean, stoploss level is 5 standard deviations away from the mean, the formation period is 12 months and the trading periods are 6 months, and with industry breakdown and with no transaction costs. Pairs which have a negative β coefficient are removed from trading, because pairs trading is about finding long-short pairs, which eliminate the movement of the market, and trading short term mean reverting anomalies, which will correct itself. For negative coefficient long-long pairs the market movement is not removed, and this mean reverting property is ceased to exist in strong market moves.

Since the VAR lag selection for the Johansen test proved to be useful and added value, as it is better in excess return, volatility, Sharpe ratio, the ratio of positive monthly returns, and also in max drawdown, I am going to use it for the comparison with the Engle-Granger test from here.

The results from the 3 test comparisons can be seen in Table 1 with performance measures and some other statistics about the pairs. The Engle-Granger approach with -0.49% annual excess return has failed to produce a positive excess return for the 11 years from Jan 2010 to Dec 2020 from the components of the Nikkei stock index, while the Johansen test had 3.32% annual excess return for the period and a Sharpe ratio of 0.54. The Johansen test was also better in avoiding big losses, as the -16.0% max drawdown is better then the -17.5% for the Engle-Granger test, and the 59.1% of positive months is also better than the 52.3%. One measure where the Engle-Granger test seems better is the lower volatility of 3.44% compared to the 6.12% annual volatility for the Johansen. This can be explained by the fact that the Johansen method is more strict with the pair selection, only selecting 21 pairs, which is 4.96% of all possible, on average for the trading periods, while the Engle-Granger approach select 67, 15.79% of all pairs. Lastly, although the Engle-Granger method had a better ratio of profitable pairs and on average less pairs hit the stoploss threshold of 5 standard deviations from the mean as a ratio of all selected pairs, it still had a worse return. The average trades in the trading period is very close to each other, with 1.02 for Engle-Granger and 1.05 for Johansen.

The way the pairs are traded in the trading period might explain why the bigger percent of diverging pairs, which has hit the stoploss level are not necessary means lower returns overall. A pair can be divergent, hitting the stoploss and still have a 0% return, because the opening signals are only triggered when the price of the pair is crossing them, thus if a pair never crosses an opening signal, only the stoploss, then it will be counted among the ones that hit the stoploss level, but still have a return of 0%.

IV.2 Time frames

Another crucial aspect of the pairs trading strategy is the question of how long should the formation period be, where the β is estimated and the cointegration is tested and the trading period, where the trading signals of the spread crossing the thresholds are created.

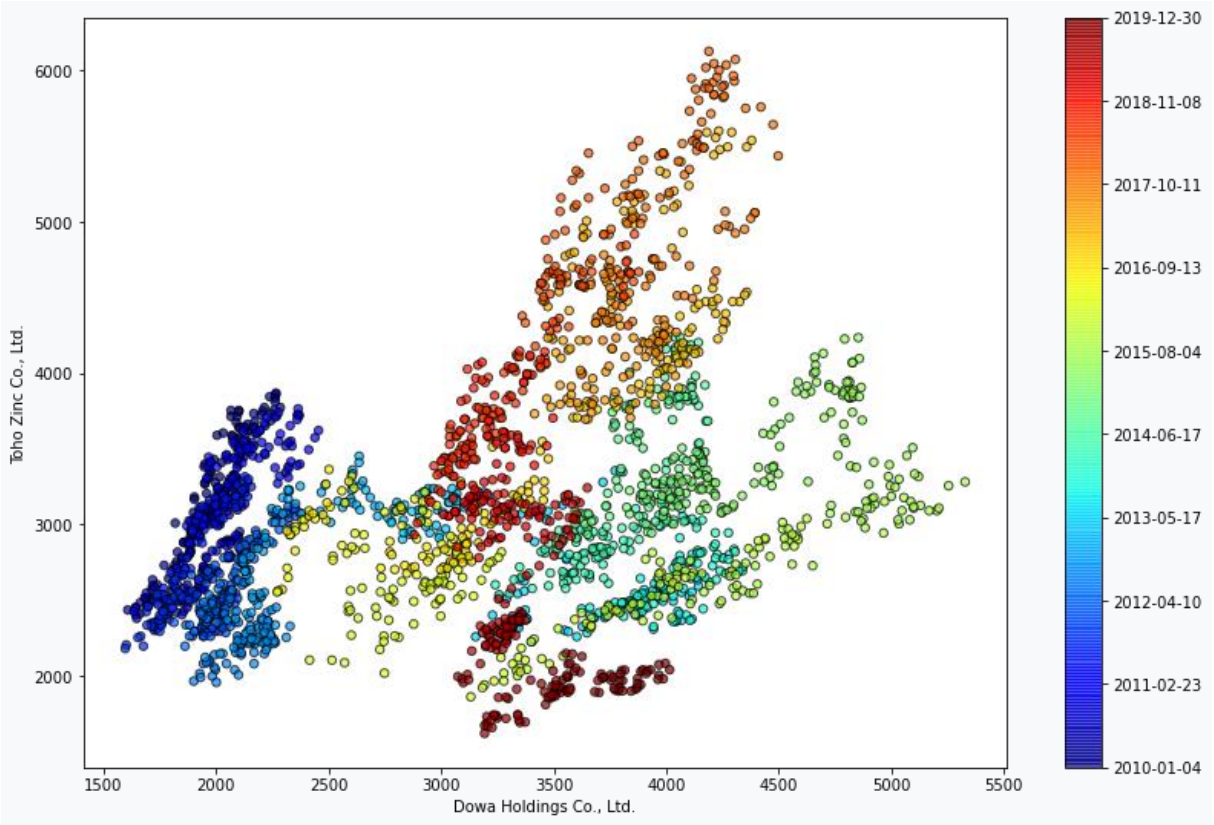

Figure 4: Price dynamics of two companies from the same industry over 10 years, Source: Own figure

Figure 4 shows the price changes over 10 years for the stocks Toho Zinc and Dowa Holdings, both of them are in the Basic Materials sector and Other Industrial Metals and Mining industry. Although these two companies are closely related to each other in an economic sense, since they are operating in the same industry in the country, it is clear that their prices are only gathering around a linear line for a relatively short time, and the hedge ratio between them is changing fast, but there are periods when the prices are moving in sync around a linear line, and if we can estimate the β well there is a short amount of time when we can profit from this anomaly. For example, the first 1.5 years, then from 2013 to 2014 and also from 2016 to 2018 would have been a good period for pairs trading.

Most of the literature uses 12 months for the formation period and 6 months for the trading period (Do and Faff, 2010, 2012; Huck, 2015; Engelberg et al., 2008; Gatev et al.,2006), so I am going to use it as a benchmark to compare other periods to this. Other periods that I am testing include 24 months formation, 6 months trading (Perlin, 2009; Huck, 2013, 2015), 36 months formation, 12 months trading (Sarmento and Horta, 2021) and I am also testing a 6 months formation, 3 months trading period, to see how this shorter time scale effect the results.

	6-3	12-6	24-6	36-12
Annual excess return	-0.92%	3.32%	-3.03%	-1.12%
Annual volatility	5.97%	6.12%	7.39%	5.41%
Sharpe ratio	-0.153	0.542	-0.410	-0.206
Positive months	52.3%	59.1%	43.9%	52.3%
Max drawdown	-32.51%	-16.00%	-30.15%	-19.46%
Selected pairs	22.1	21.0	22.6	23.6
Profitable pairs	8.6 (38.9%)	8.9 (42.4%)	8.5 (37.3%)	9.0 (38.1%)
Hit stoploss	9.5 (43.0%)	9.7 (46.1%)	6.9 (30.5%)	8.6 (36.5%)
Average trades	0.98	1.05	0.69	0.87

Table 2: Summary of the results with different formation and trading periods, Source: Own table

Table 2 shows that only the 12 months formation and 6 months trading periods manage to have positive excess return, so the most popular period selection proves to be the best. Surprisingly, the second most used period, 24 months formation and 6 months trading was the worst in terms of annual excess return, while the 6 formation, 3 trading was the second best, although I have never seen anybody using it, but it also proves that shorter time frames are better for pairs trading, since cointagration and mean reversion to the estimated equilibrium level is only short lived. The ratio of profitable pairs are all close to 38%, except for the 12-6 periods, which had the best with 42.4%, this is not surprising, the time frame that had the best return had the highest ratio of profitable pairs, although it also had the highest ratio of divergent pairs, 46.1%, while longer time frames decreased this ratios to 30.5% for 24-6 and 36.5% for 36-12, which might be explained by the fact, that longer periods used to estimate the hedge ratio results in less divergent pairs, but on average their return is still lower then for shorter periods.

IV.3 Thresholds

The last aspect I am examing before comparing the industy breakdown with the PCA breakdown is the choice of the opening triggers from the mean in standard deviations. This was also tested by Huck and Afawubo (2015), who found that a 3 standard deviation opening signal is better for cointegration than 2 standard deviations. Here I tested 1, 2, 3 SD from the mean and also two methods, which are averaging in the trades, which means that if the lower threshold is hit than only half of the position size is taken, and when it hits the further threshold I invest

the rest of the position, and this fully invested position is closed if the mean or the stoploss level is reached. I used 1-2 SD and 2-3 SD thresholds for the averaging in strategies. The relevance of the averaging in method was pointed out by Ernie Chan (2013), who suggested that using multiple opening triggers might result in a higher Sharpe ratio.

	1	2	3	1-2	2-3
Annual excess return	1.67%	3.32%	1.13%	3.53%	2.70%
Annual volatility	7.50%	6.12%	4.92%	7.05%	5.36%
Sharpe ratio	0.223	0.542	0.230	0.501	0.505
Positive months	50.8%	59.1%	53.8%	61.4%	60.6%
Max drawdown	-17.01%	-16.00%	-20.92%	-15.43%	-18.68%
Selected pairs	21.0	21.0	21.0	21.0	21.0
Profitable pairs	8.9 (42.4%)	8.9 (42.4%)	6.5 (30.7%)	10.3 (48.9%)	9.3 (44.4%)
Hit stoploss	9.7 (46.1%)	9.7 (46.1%)	9.7 (46.1%)	9.7 (46.1%)	9.7 (46.1%)
Average trades	1.64	1.05	0.72	1.38	0.91

Table 3: Summary of the results with different opening triggers, Source: Own table

From Table 3, it is clear that using a single opening trigger, the 2 SD is the best with 3.32% annual excess return, compared to 1.67% for 1 SD and 1.13% for 3 SD. The 1-2 SD made the best annual excess return with 3.53%, but it failed to outperform the original 2 SD in Sharpe ratio, which had 0.54 compared to 0.50 of the averaging in method. Another notable finding is that the volatility is decreasing as the opening trigger is increased, unsurprisingly, since this means less trading, as it can be seen in the average trades, and less pairs being traded means less exposure to the equity markets volatility, and the portfolio remains more in cash. Although the 1-2 SD averaging in strategy has a better return, max drawdown, ratio of positive months and more profitable pairs, I am going to use the original 2 SD opening trigger for the rest of this study, as it provides the best Sharpe ratio, the other metrics are also very close, and since this is the most commonly used in the literature it provides a better basis for comparison with other studies.

IV.4 Breakdowns

Although OPTICS is regarded as a parameterless clustering algorithm, it is far from parameterless, and the pairs trading strategies results can vary by changing some of them. OPTICS do not require to determine the number of clusters we want to find ahead of time, thus many refer to it as ’parameterless’, but we still have to define the minimum steepness as a boundary for the reachability plot when the algorithm defines the clusters, and the minimum

sample size of the clusters. The minimum steepness is not a very sensible parameter, results do not vary much when changing it, so I used 0.01 for the strategy. Minimum sample on the other hand changes the results drastically, both the number of selected pairs for trading and the strategies profitability.

The number of principal components to use for clustering is also an important question, their is a trade-off between allowing more components to explain more of the original variance of the dataset, and on the other hand the more components we use the worse the curse of dimensionality will get (Bellman, 1966), and we might cluster stocks by meaningless noise, which is not desirable when searching for economicly connected assets.

		1 PC	2 PC	3 PC	5 PC
2 min sample	Annual excess return	3.01%	3.76%	5.09%	-4.16%
	Annual volatility	8.37%	9.55%	10.41%	8.82%
	Sharpe ratio	0.359	0.393	0.489	-0.472
	Selected pairs	10.2	7.6	8.3	8.2
	Diff Industry	9.7 (94.6%)	7.0 (91.7%)	7.2 (86.9%)	6.4 (77.4%)
	Diff Sector	8.5 (82.6%)	5.8 (76.1%)	5.1 (61.2%)	4.4 (53.0%)
3 min sample	Annual excess return	0.34%	1.42%	0.02%	-2.76%
	Annual volatility	6.62%	6.76%	8.96%	7.64%
	Sharpe ratio	0.052	0.211	0.002	-0.361
	Selected pairs	18.5	15.5	15.7	11.7
	Diff Industry	17.7 (95.6%)	14.3 (92.4%)	13.7 (87.0%)	9.2 (78.7%)
	Diff Sector	15.3 (82.6%)	11.9 (76.5%)	10.4 (65.9%)	6.3 (53.9%)
4 min sample	Annual excess return	0.59%	-0.18%	0.30%	-0.95%
	Annual volatility	6.02%	6.44%	7.36%	8.32%
	Sharpe ratio	0.098	-0.028	0.041	-0.115
	Selected pairs	23.6	22.1	17.2	14.2
	Diff Industry	22.5 (95.4%)	20.9 (94.4%)	15.1 (88.1%)	12 (84.3%)
	Diff Sector	18.8 (79.8%)	17.5 (79.0%)	11.2 (65.1%)	8.1 (56.8%)
6 min sample	Annual excess return	0.49%	-0.82%	-0.35%	-2.65%
	Annual volatility	5.14%	5.26%	6.59%	8.60%
	Sharpe ratio	0.095	-0.155	-0.053	-0.308
	Selected pairs	35.3	32.9	24.2	19.1
	Diff Industry	34.0 (96.4%)	30.8 (93.8%)	21.9 (90.6%)	16.9 (88.3%)
	Diff Sector	28.6 (80.9%)	25.1 (76.5%)	16.6 (68.8%)	12.1 (63.6%)
9 min	Annual excess return	-0.22%	-1.59%	-3.28%	-3.27%

Annual volatility	4.87%	4.93%	6.29%	9.92%
Sharpe ratio	-0.045	-0.323	-0.522	-0.329
Selected pairs	50.1	48.2	35.2	19.6
Diff Industry	48.1 (96.0%)	45.3 (94.0%)	32.6 (92.8%)	17.6 (89.8%)
Diff Sector	41.3 (82.6%)	37.7 (78.3%)	24.9 (70.8%)	12.3 (62.8%)

Table 4: Summary of the results with different number of principal components and minimum sample sizes, Source: Own table

Table 4 contains the results, before analysing the performance of the different clustering methods, there are some other interesting outcomes to examine. As the number of principal components is increasing, the number of selected pairs decreases, which in most cases results in higher volatility, as expected. The ratio of pairs where the two stocks are from a different industry or different sector is monotonicly decreasing as more principal components are used to define the clusters. This is in line with the fact that as more of the original variance is in the components, the more closely related stocks will get in the same clusters. With a bigger minimum sample size, less clusters are formed, but since there are more possible pairs to test, there are more pairs traded, regardless of the number of principal components.

In terms of annual excess return, the PCA clustering definitely favors small minimum sample sizes, in fact the best was 2 by far, and basicly anything above 4 has failed to make a positive excess return. For small minimum sample sizes, the best number of principal components were 2 and 3, and 5 has always made a negative excess return under any number of minimum sample sizes. The only 2 compositions that had a better excess return than the industry breakdown was 2 minimum sample size and 2 components with 3.76%, and 2 minimum sample size and 3 components with 5.09%, but thanks to their higher volatility, because less pairs were selected, they both failed to give a higher Sharpe ratio than the industry breakdowns 0.54, with 0.39 and 0.49 respectively.

IV.5 Different stock markets

To test if these results, the better excess return of the PCA clustering holds in out of sample dataset with the best results parameters found on the Japanese stock market, I am going to compare the industry breakdown to the PCA clustering breakdown on the same period, Jan 2010 to Dec 2020 on the US stock market, out of the components of the S&P500 index. I am going to use the best statistical test, the Johansen test with lag selection of a VAR model, the

best parameters, 12 months formation and 6 months trading period, no transaction costs, and 2 SD opening triggers and the mean as closing trigger for trading. For the PCA breakdown, 3 principal components and a minimum sample size of 2.

	NIKKEI		S&P	
	Industry	PCA	Industry	PCA
Annual excess return	3.32%	5.09%	-0.71%	-1.40%
Annual volatility	6.12%	10.41%	3.68%	5.32%
Sharpe ratio	0.542	0.489	-0.192	-0.264
Positive months	59.1%	53.0%	49.2%	50.0%
Max drawdown	-16.00%	-23.38%	-12.05%	-22.52%
Selected pairs	21.0	8.3	53.9	16.5
Different Industry	0	7.2 (86.8%)	0	12.8 (77.4%)
Different Sector	0	5.1 (61.2%)	0	8.5 (51.8%)
Profitable pairs	8.9 (42.4%)	3.6 (42.7%)	21.8 (40.5%)	5.6 (34.1%)
Hit stoploss	9.7 (46.1%)	3.7 (44.8%)	25.5 (47.3%)	9.3 (56.2%)
Average trades	1.05	1.06	1.16	1.08

Table 5: Summary of the results from the Japanese and US stock market, Source: Own table

Table 5 shows that the PCA approach failed to deliver better excess return for the US stock market, and also increased the ratio of divergent pairs, which has hit the stoploss level from 47.3% to 56.2%, while decreasing the ratio of positive return pairs from 40.5% to 34.1%. The similarity is that the PCA clustering even out of the components of the S&P500 index gives a smaller number of pairs to trade 16.5 on average instead of 53.9 of the industry breakdown, which is much lower, just like in the Japanese stock market, and also in turn this has incraesed the volatility of the strategy from 3.68% annually to 5.32%. One other interesting thing to note is that the ratio of pairs whose stocks are from a different industry or sector is much lower in the US market than in the Japanese. The fact that the PCA clustering's higher return could not be replicated in a different market, and the fact that the number of pairs to trade is very low, 8.3 on average shows that these results are not consistent, and might only be true for the Japanese stock market in the given period as a result of overfitting.

The result that the less documented Japanese stock market has a better excess return than the US market using pairs trading has already been shown by Huck (2015), and this study also reinforces it. For both breakdowns, this phenomenon still holds true.

IV.6 Transaction costs

So far these results shown before did not include transaction costs. To compare the results with transaction costs, I am going to implement the transaction costs used by Sarmento and Horta (2021), where they estimated the transaction costs based on Do and Faff (2012). For commission 8 bps per trade, for market impact 20 bps per trade, and for shorting cost an annual 1% of the short exposure which is paid daily over the lifetime of the trade.

	No transaction costs		With transaction costs	
	Industry	PCA	Industry	PCA
Annual excess return	3.32%	5.09%	0.43%	2.14%
Annual volatility	6.12%	10.41%	6.15%	10.49%
Sharpe ratio	0.542	0.489	0.070	0.204
Positive months	59.1%	53.0%	54.5%	47.7%
Max drawdown	-16.00%	-23.38%	-23.58%	-25.76%
Selected pairs	21.0	8.3	21.0	8.3
Profitable pairs	8.9 (42.4%)	3.6 (42.7%)	8.4 (40.0%)	3.3 (39.3%)
Hit stoploss	9.7 (46.1%)	3.7 (44.8%)	9.7 (46.1%)	3.7 (44.8%)
Average trades	1.05	1.06	1.05	1.06

Table 6: Summary of the results with and without transaction costs, Source: Own table

From Table 6 it can be seen that having transaction costs the annual excess return decreases dramatically, about 3% for both cases. As a result, since the PCA strategy had a bigger excess return, after transaction costs besides having a better 2.14% annual excess return compared to 0.43% of the industry breakdown, it also has a better Sharpe ratio of 0.20, compared to 0.07. The number of profitable pairs slightly decreased, and the ratio of positive monthly return decreased in both cases by about 5%.

V Conclusion and further research

In the first section of this study, I examined the most popular statistical tests and the most commonly used parameters for pairs trading, and found that the Johansen test with a VAR model for lag selection outperformed the Engle-Granger approach, for the time frames the 12 months formation and 6 months trading period was the most suitable implementation as it is confirmed by the general use of this period selection, for thresholds the 2 standard deviation

from the mean was the best, and even the averaging in method could not significantly improve on this, which is also confirmed by the wide use of this threshold in the literature.

The breakdown with PCA and OPTICS clustering was very sensitive to the parameters of the number of principal components and the minimum sample size. The number of selected pairs decreased as the number of principal components were increased, and also the ratio of pairs with stocks from different industries or sectors were decreasing. As increasing the minimum sample size, more pairs were selected for trading, but their returns were worse when there was only a few pairs selected. In terms of excess returns, the best configuration was a low number of principal components with low minimum sample size, but since there were less profitable pairs selected, the volatility has increased for the strategy.

The best performing parameters for PCA clustering were 3 principal components and a minimum sample size of 2 with 5.09% annual excess return and Sharpe ratio of 0.49, which outperformed the industry breakdown of 3.32% annual excess return, but this result could not be repeated in the US stock market from the components of the S&P500 index, where the industry breakdown outperformed with -0.71% annual excess return to the PCA clustering's -1.40%.

After including transaction costs, the excess returns have diminished by about 3%, with the ratio of positive months also decreased by 5% for both breakdowns. As a result, thanks to the higher excess return of the PCA clustering, after transaction costs it had a higher Sharpe ratio of 0.20 compared to the industry breakdown's of 0.07.

Further research could include trying to compute the covariance matrix (or in this case, since the returns are usually normalized before PCA, the correlation matrix) for the dimensionality reduction to be more descriptive of the data by using exponencial smoothing for the covariances (Pozzi et al., 2012), so the data for the statistical factors are weighted higher for recent price movement and lower for distant ones. By using exponential weighting, the covariance matrix can distinguish genuine from spurios relationships better, while recovering faster after market turbulences, when most of the pairwise correlations are shifted to more positive values. Using a shrinkage on the covariance matrix before PCA could also improve the meaningfulness of the estimated factors (Ledoit and Wolf, 2003). The covariance matrix is subject to estimation error, and the shrinkage pulls together the estimated values, reducing this error and also reducing the

adverse effect of outliers in the data, which can detriment the PCA algorithm. The challenge would be to define a shrinkage constant that works well in this setting.

Evaluating different dimensionality reduction methods to extract the statistical risk factors could yield in interesting results, and this would expand beyond pairs trading. Methods could include matrix factorization, nonlinear techniques like UMAP, or applying deep learning methods, such as autoencoder, which might be better for large datasets, and can be linear or nonlinear, depending on the activation function.

References

Alexander, C., & Dimitriu, A. (2002). The cointegration alpha: Enhanced index tracking and long-short equity market neutral strategies.

Ankerst, M., Breunig, M. M., Kriegel, H. P., & Sander, J. (1999). OPTICS: Ordering points to identify the clustering structure. *ACM Sigmod record, 28(2),* 49-60.

Armstrong, J. S. (2001). *Principles of forecasting: a handbook for researchers and practitioners.* Boston, MA: Kluwer Academic.

Avellaneda, M., & Lee, J. H. (2010). Statistical arbitrage in the US equities market. *Quantitative Finance, 10(7),* 761-782.

Balladares, K., Ramos-Requena, J. P., Trinidad-Segovia, J. E., & Sánchez-Granero, M. A. (2021). Statistical arbitrage in emerging markets: a global test of efficiency. *Mathematics, 9(2),* 179.

Bellman, R. (1966). Dynamic programming. *Science, 153(3731),* 34-37.

Bookstaber, R. (2007). *A demon of our own design.* Hoboken, New Jersey: John Wiley & Sons.

Caldeira, J., & Moura, G. V. (2013). Selection of a portfolio of pairs based on cointegration: A statistical arbitrage strategy.

Cardoso, R. G. (2015). Pair trading: Clustering based on principal component analysis (Doctoral dissertation).

Chan, E. P. (2013). *Algorithmic Trading.* Hoboken, New Jersey: John Wiley & Sons, Inc.

Do, B., & Faff, R. (2010). Does simple pairs trading still work? *Financial Analysts Journal, 66(4),* 83-95.

Do, B., & Faff, R. (2012). Are pairs trading profits robust to trading costs? *Journal of Financial Research, 35(2),* 261-287.

Do, B., Faff, R., & Hamza, K. (2006). A new approach to modeling and estimation for pairs trading. *In Proceedings of 2006 financial management association European conference (Vol. 1, pp. 87-99).*

Dunis, C. L., Giorgioni, G., Laws, J., & Rudy., J. (2010). Statistical Arbitrage and High-Frequency Data with an Application to Eurostoxx 50 Equities. *Liverpool Business School, Working paper*.

Engelberg, J., Gao, P., & Jagannathan, R. (2008). An anatomy of pairs trading: the role of idiosyncratic news, common information and liquidity. *Third Singapore International Conference on Finance*.

Engle, R. F., & Granger, C. W. (1987). Co-integration and error correction: representation, estimation, and testing. *Econometrica: journal of the Econometric Society*, 251-276.

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *kdd Vol. 96, No. 34*, 226-231.

Fama, E. F., & French, K. R. (1997). Industry costs of equity. *Journal of financial economics, 43(2)*, 153-193.

Gatev, E., Goetzmann, W. N., & Rouwenhorst, K. G. (2006). Pairs trading: Performance of a relative-value arbitrage rule. *The Review of Financial Studies, 19(3)*, 797-827.

Huck, N. (2013). The high sensitivity of pairs trading returns. *Applied Economics Letters, 20(14)*, 1301-1304.

Huck, N. (2015). Pairs trading: does volatility timing matter? *Applied economics, 47(57)*, 6239-6256.

Huck, N., & Afawubo, K. (2015). Pairs trading and selection methods: is cointegration superior? *Applied Economics, 47(6)*, 599-613.

Jacobs, H., & Weber, M. (2015). On the determinants of pairs trading profitability. *Journal of Financial Markets, 23*, 75-97.

Johansen, S. (1988). Statistical analysis of cointegration vectors. *Journal of economic dynamics and control, 12(2-3)*, 231-254.

Ledoit, O., & Wolf, M. (2003). Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of empirical finance, 10(5)*, 603-621.

Perlin, M. S. (2009). Evaluation of pairs-trading strategy at the Brazilian financial market. *Journal of Derivatives & Hedge Funds, 15(2)*, 122-136.

Pole, A. (2007). *Statistical Arbitrage.* Hoboken, New Jersey: John Wiley & Sons.

Pozzi, F., Di Matteo, T., & Aste, T. (2012). Exponential smoothing weighted correlations. *The European Physical Journal B, 85(6)*, 1-21.

Rad, H., Low, R. K., & Faff, R. (2016). The profitability of pairs trading strategies: distance, cointegration and copula methods. *Quantitative Finance, 16(10)*, 1541-1558.

Ramos-Requena, J. P., Trinidad-Segovia, J. E., & Sánchez-Granero, M. A. (2017). Introducing Hurst exponent in pair trading. *Physica A: statistical mechanics and its applications, 488*, 39-45.

Rampertshammer, S. (2007). An Ornstein-Uhlenbeck framework for pairs trading.

Sarmento, S. M., & Horta, N. (2021). *A Machine Learning based Pairs Trading Invesment Strategy.*

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of machine learning research, 9(11)*.

Vidyamurthy, G. (2004). *Pairs Trading.* Hoboken, New Jersey: John Wiley & Sons, Inc.

Wilmott, P. (2005). *The Best of Wilmott.* Chichester, West Sussex: John Wiley & Sons Ltd,.