UNIVERSITÀ CATTOLICA DEL SACRO CUORE

Campus of Brescia

Faculty of Mathematics, Physics and Natural Sciences &
Banking, Finance and Insurance Science

Master program in
Applied Data Science for Banking and Finance

Enhancing a Pairs Trading strategy using an Ornstein-Hulenbeck process

Supervisor:
Professor Federico
Mazzorin
Co-Supervisor:
Professor Francesco
Orsini

Dissertation by:
Paolo Bassi
Id Number:
5008285

Academic Year
2021/2022

Contents

Chapter 1

Introduction

The strategy that is analyzed in this work was originally proposed in the paper by Sarmento and Horta (2020), in which one of the most significant features concerns the utilization of machine learning techniques. In pairs trading context there are now a lot of strategies employed, from the simplest and the first one such as the distance approach developed by Gatev et al. (2006) to the more complex and recent which employ some advanced techniques like neural networks. The aim of this thesis is to further investigate the research initiated by the authors. More specifically, the initial objective is to replicate the algorithm they devised, assessing its performance, and to enrich the study considering other periods and metrics, with the aim of implementing a trading strategy that considers the spread as an OU stochastic process.

Moreover, the project will act as a bridge between different major techniques found in the literature, spanning from machine learning approaches, co-integration, and time series. In detail the machine learning approach can be applied to the selection phase to reduce data dimensions into principal components and to group stocks; the cointegration approach relies on cointegration testing to detect stationary spread time series; finally, the time-series approach focuses on finding optimal trading rules for mean-reverting spreads.

1.1 Pairs trading overview

Pairs Trading, which was developed in the 1980s, is a popular investment strategy that has been widely used by hedge funds and institutional investors as a crucial long/short equity investment tool Gatev et al. (2006). Before describing the constituents of a pairs trading strategy, it is important to briefly introduce the concept of statistical arbitrage which is commonly believed to have evolved from pairs trading. Statistical arbitrage refers to a range of investment strategies that share certain key characteristics [Pole (2011)]. Firstly, their trading signals are systematic or rule-based, as opposed to being based on fundamental analysis. Secondly, the trading book is designed to be market-neutral, meaning that it has zero beta with the market. Thirdly, the primary mechanism for generating excess returns is statistical in nature. The goal is to make multiple wagers with positive expected returns, utilizing diversification across stocks, to create a low-volatility investment strategy that is not correlated with the market.

Coming again to pairs trading, the strategy involves two steps. Initially, it necessitates identifying two securities, such as two stocks, that exhibit parallel behavior in their price series, or appear to be correlated. This suggests that both securities are subject to similar risk factors and tend to respond similarly. A pair can be made up of two securities whose price series display an equilibrium relation, and once these pairs are identified, the investor can proceed to the strategy's second step. The strategy relies on the relative value of the two securities meaning that if their price series have moved in proximity in the past, this relationship should persist in the future. Therefore, if a discrepancy arises, it presents an attractive trading opportunity to profit from its correction. To identify such opportunities, the spread between the two securities in the pairs must be continually monitored, along with its mean value and standard deviation. When a statistical irregularity is detected, a market position is established, and the position is closed when the spread eventually corrects.

We will now illustrate how to apply the strategy using an example from this work. Let us assume that we have identified two securities, DTE and PCG, as a potential pair. To verify this relation, we can refer to Figure 1.1, which displays the two-price series from 2006 to 2016.

During the subsequent trading period (here the test period corresponds to 2009), we normalize and monitor the spread, which is defined as $S_t = DTE_t$ - PCG_t, in Figure 1.2. Although it fluctuates around its mean, there are noticeable variations that may trigger a trade, depending on their magnitude. To execute a trade, the investor specifies the long and short thresholds, which

Figure 1.1: Price series of two constituents of a pair during 2006-2016.

determine the minimum deviation required to initiate a long or short position, respectively. A long position is established when the spread is expected to widen, as its current value is below the expected value. This involves buying DTE and selling PCG. Conversely, a short position is established when the spread is expected to narrow, and the opposite transaction is executed. The positions are closed when the spread returns to its expected value.

As shown in Figure 1.2, we can determine the market position according to the green line, which takes on values of -1, 0, or 1, depending on whether the current position is short, outside the market, or long, respectively. For example, at the end of February 2009, we observe the opening of a short position when the spread deviates below the short threshold, followed by its closure when the spread returns to zero.

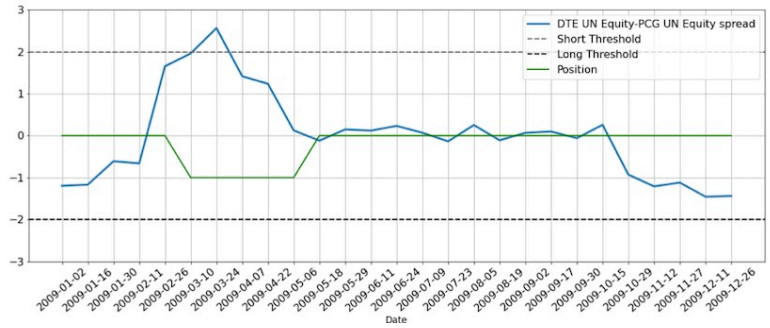


Figure 1.2: Exemplifying a Pairs Trading strategy execution.

In conclusion, this strategy is highly resilient to various market conditions. Regardless of the market's movement (i.e., upward, downward, or sideways), if the asset that the investor purchased is performing relatively better than the one they sold, a profit may be earned.

1.2 Objectives

The primary focus of research in this field is on conventional methods and statistical tools for enhancing the essential elements of this strategy. However, in recent years, Machine Learning techniques have been increasingly employed. The success of a Pairs Trading strategy is heavily reliant on identifying suitable pairs. Nevertheless, with the increasing availability of data, more traders can identify intriguing pairs and swiftly profit from price discrepancies. The challenge is that it can be exceedingly difficult to identify such opportunities. If the investor restricts their search to securities within the same sector, as is typically done, they are less likely to uncover pairs that have yet to be heavily traded. Conversely, if the investor does not impose any constraints on the search space, they may have to examine an excessive number of combinations and are more likely to discover spurious relations. Compared to the distance approach, cointegration provides a more rigorous framework for pairs trading, but besides the obvious issue of finding spurious relations, one of the cointegration's shortcoming is the computational burden required to find pairs, since it would be necessary to perform at least $\frac{n \ x \ (n-1)}{2}$ cointegration tests. To address this challenge, the study of Sarmento and Horta (2020) suggests utilizing unsupervised learning to define the search space. The goal is to cluster pertinent securities (not necessarily from the same sector) and identify profitable pairs within them, which would otherwise be difficult to identify. The authors propose a method for identifying cointegrated pairs in commodity-linked ETFs, asserting that this could achieve best results compare with trading stocks due to their high volatility. This work intends to verify this assumption and investigate other possible outcomes, since there will be considered stocks from $SP500$ composite. Furthermore, different periods of market turmoil will be considered in this work, namely the 2007–2008 financial crisis, the 2009-2010 European debt crisis, and the COVID-19 pandemic period, in order to analyze the robustness of this strategy even in unusual times. Regarding the trading phase, Sarmento and Horta (2020) developed only one strategy based on the threshold-based approach proposed by Gatev et al. (2006) before delving into a forecasting Neural Network-based strategy. This research instead sets up another threshold strategy based on the stochastic spread method used in time series approach. This intent was originated by a suggestion put forward in Krauss (2017), where the author criticized the lack of empirical analysis in finding matching pairs during the formation period, which is usually ignored in the time series approach, since pairs are only assumed to be cointegrated.

However, according to Cummins and Bucca (2012), the stochastic spread method has a major limitation due to the Gaussian properties of the Ornstein-Uhlenbeck process not being consistent with the nature of financial data, such as the assumption of linearity or constant

volatility that the method relies on. Nevertheless, it is also argued that the simplicity of the Ornstein-Uhlenbeck process makes up for this limitation, but at the same time it is still uncertain whether the simplicity of the process is adequate to counterbalance its limitations and make it a useful pairs trading framework that outperforms non-parametric frameworks like the distance method. In this research stage the framework proposed in Zeng and Lee (2014) will be followed: since the application is limited because the model parameters are assumed constant, this work will introduce a novelty by considering different sliding windows in order to analyze the process using evolving parameters and in what measure changing the model parameters would impact on the optimal threshold strategy.

In conclusion this work aims at filling the gap between the cointegration approach and the time-series approach while providing empirical evidence of the latter in terms of tradeoff between simplicity of the assumptions and trading profitability.

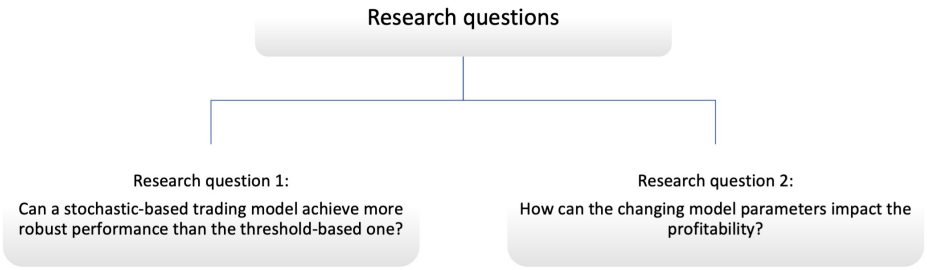


Figure 1.3: Research questions to be answered.

1.3 Thesis Outline

The thesis encompasses a total of seven chapters. Chapter 2 provides an overview of the background and literature regarding Pairs Trading, as well as some key mathematical concepts necessary to understand how this strategy works. Chapter 3 and 4 outline the two proposed implementations in this study. The former chapter presents the proposed pairs selection framework by Sarmento and Horta (2020), while the latter proposes a trading model that employs stochastic mean-reverting spread as proposed in Zeng and Lee (2014) (Research question 1) with a further implementation (Research question 2). Chapter 5 contains information about the methodology. The results are

presented in chapter 6. Lastly, chapter 7 focuses on conclusions and future development.

Chapter 2

Pairs Trading - Background and Literature review

To fully understand the formalities related to Pairs Trading, it is necessary to introduce a set of tools before delving into more specific details. This chapter presents the main econometric techniques for studying time series that will be used throughout the work, highlighting their most important aspects.

2.1 The basics of time series analysis

A time series (also known as a temporal series) is the representation of a discrete stochastic process, and it describes the dynamics of certain phenomena over time. It can be studied to interpret the evolution of a variable, identify any trends and cycles in the data, or attempt to predict future values of the measured characteristic. A time series, Y, is represented by the set of values that a variable assumes over time. In particular, if we consider a time period characterized by T instants, the time series is indicated by $Y = \{Y_1, Y_2, Y_3, ..., Y_T\}$, where Y_t represents the value of the series at time t. In this context, special terminology is usually used to indicate the past values of Y. Specifically, given a time t, the value of Y in the previous period is called the first lag and is indicated by Y_{t-1}, and similarly, the j-th lag is the value taken by Y j instants before, i.e., Y_{t-1}. Finally, the first difference is defined as the variation in the value of Y between time t and time t-1, that is, $\Delta Y_t = Y_t - Y_{t1}$ [Maddala and Kim (1998)].

2.2 Mean-Reversion and Stationarity

We continue by introducing the concept of stationarity. A time series is stationary if its probability distribution does not change over time [Maddala and Kim (1998)], that is when its mean and variance remain generally constant, meaning that the joint distribution of $\left(\mathrm{Y_{t+1}}, \mathrm{Y_{t+2}}, ..., \mathrm{Y_{t+T}}\right)$ does not depend on t for any value of T. If this is not the case, $\mathrm{Y_t}$ is said to be non-stationary. Even if mean reversion and stationarity are not the same mathematical concept, a stationary time series is mean-reverting in nature, and fluctuations around the mean should have similar amplitudes. Additionally, a stationary process can be characterized by its order of integration I(d), where a stationary time series is referred to as an I(0) process, and a non-stationary process is I(1). In detail, a time series that follows a random walk, and whose difference is stationary, is defined as integrated of order 1, or I(1). A series whose first difference follows a random walk is defined as integrated of order 2, or I(2), and so on. If the series does not show any stochastic trend, i.e., it is stationary, it is defined as integrated of order 0, or I(0). Therefore, the order of integration represents the number of times the series must be differenced to become stationary. From a practical standpoint, identifying the order of integration of a series is straightforward: start by performing a stationarity test on the series. If the null hypothesis is rejected, then there is evidence that the series is stationary, and the order of integration is 0. Otherwise, if the null hypothesis is accepted, difference the series and test the hypothesis that $\Delta\mathrm{Y_t}$ has a unit root against the alternative hypothesis that $\Delta\mathrm{Y_t}$ is stationary [Lütkepohl (2005)].

2.2.1 Augmented Dickey-Fuller test

The main test used for this purpose is the Augmented Dickey-Fuller test (ADF) [Dickey and Fuller (1979)], which is a hypothesis test to determine whether a unit root is present in a time series. We can model a consecutive change in time series using the following equation:

$$\Delta y(t) = \lambda y(t-1) + \mu + \beta t + \alpha_1 \Delta y(t-1) + ... + \alpha_\mathrm{k} \Delta y(t-k) + \varepsilon_\mathrm{t} \tag{2.1}$$

where $\Delta y(t) = y(t) - y(t-1)$, μ is a constant, β is the coefficient on a time trend, and k is the lag order of the autoregressive process. The ADF test determines whether λ=0. If the hypothesis is rejected, and $\lambda \neq 0$, the change in a time series at time t is dependent on the value of the series at time $t1$, which means that the series cannot be a simple random walk. The test statistic associated with this is the regression coefficient λ (calculated with $y(t-1)$ as the independent variable and $\Delta y(t)$ as the dependent variable) divided by the standard error (SE) of the regression fit, λ/SE, which has a negative value when mean reversion is expected. This value is then compared with the

critical values corresponding to the distribution of the test statistic, and used to determine whether the hypothesis can be accepted or rejected at a given probability level.

2.2.2 Hurst exponent

The Hurst exponent is a metric that can be utilized to determine the stationarity of a time series. This metric evaluates whether the speed of diffusion of the time series from its initial value is slower than that of a geometric random walk [Hurst (1951)].

The speed of diffusion can be formally defined by the variance as

$$Var(\tau) = \langle \left| z(t+\tau) - z(t) \right|^2 \rangle \tag{2.2}$$

where $z(t)$ is the logarithmic time series value at time t, τ is a time lag, and $\langle - \rangle$ is the average across time. For a price series that follows a geometric random walk, $H = 0.5$, using (2.2) it can be simplified to $\langle \left| z(t+\tau) - z(t) \right|^2 \rangle \sim \tau$. However, as H decreases towards zero, the speed of diffusion decreases, indicating that the price series is more likely to mean-revert. Conversely, as H increases towards 1, the price series is more likely to exhibit a trend. Therefore, the Hurst exponent can be interpreted as an indicator of the degree of mean-reversion or trendiness exhibited by the time series.

2.2.3 Half-life of mean-reversion

The half-life of mean-reversion is a metric that measures the duration it takes for a time series to mean-revert. To compute this metric, the discrete time series (2.1) can be transformed into its differential form, where the changes in prices are now expressed as infinitesimal quantities. This transformation is described by the expression:

$$dy\,(t) = (\lambda y\,(t-1) + \mu)\,dt + d\varepsilon \tag{2.3}$$

This expression is used to model an Ornstein-Uhlenbeck process, where $d\varepsilon$ represents some Gaussian Noise [Chan (2013)]. If λ is greater than 0, the time series will not mean-revert. However, for a negative λ value, the expected value of the time series decays exponentially. This result suggests

that the expected duration of mean reversion is inversely proportional to the absolute value of λ, meaning that a larger absolute value of λ corresponds to a faster mean-reversion.

2.2.4 Cointegration

All the concepts and techniques presented so far lay the foundation for the concept of cointegration. It can happen that two different time series share the same random trend, which causes them to move together, i.e., in the same direction over a long period of time. A formal definition of the concept of cointegration was first provided by economist Clive Granger in his doctoral thesis [Granger (1983)], and later [Engle and Granger (1987)], where it was proposed that a set of variables is considered cointegrated if there exists a linear combination of those variables of order d, resulting in a lower order of integration, $I(d-1)$. Cointegration is confirmed if a set of $I(1)$ variables (non-stationary) can be used to model an $I(0)$ variable (stationary). When considering two time series, y_t and x_t, both of which are I(1), cointegration suggests that there are coefficients, μ and β, that satisfy the equation:

$$y_t - \beta x_t = u_t + \mu \tag{2.4}$$

where u_t is a stationary series. Tests for cointegration are used to identify stable, long-term relationships among sets of variables. The Engle-Granger two-step method and the Johansen test are the most commonly used cointegration tests. The two-step Engle-Granger cointegration test between two time series y_t and x_t used in this study proceeds as follows:

1. Conduct an ADF test to determine if a unit root exists in the series y_t and x_t. If the test result is positive, proceed to step 2.

2. Using Ordinary Least Squares, run the regression specified in equation (2.7) and save the residuals, $\hat{u}_t$.

3. Test the residuals $\hat{u}_t$ for a unit root using ADF test.

4. If the null hypothesis of a unit root in the residuals (null of no-cointegration) is rejected, indicating that the residual series is stationary, the two variables are cointegrated.

However, a significant issue with the Engle-Granger method is that the choice of the

dependent variable can lead to different conclusions [Maddala and Kim (1998)].

As explained in earlier section, a Pairs Trading strategy typically consists of two stages: identifying the pairs and then trading them. In line with this approach, the following two sections will discuss each stage separately.

2.3 Formation phase: pairs selection

The pairs selection stage involves two steps: (i) identifying suitable candidate pairs, and (ii) selecting the most promising ones. In the literature, two methods are commonly recommended for this stage: conducting an exhaustive search for all possible combinations of selected securities or grouping them by sector (or another category) and limiting the pairs to securities within the same sector. While the former may uncover more unusual and intriguing pairs, the latter reduces the risk of identifying false correlations. We will now describe in greater detail the most common methods for selecting pairs, which include the distance approach, the cointegration approach, and the correlation approach.

2.3.1 The minimum distance approach

Gatev et al. (2006) seminal paper introduced the distance approach, which can be considered the baseline for distance-based selection criteria. This method involves selecting pairs that minimize a distance criterion. The authors created a cumulative total return index for each stock and normalized it to the first day of a 12-month formation period. During this period, the Sum of Euclidean Squared Distances (SSD) is calculated between the constructed time series. The authors then recommend ranking the pairs based on the minimum historical SSD. However, according to Krauss (2017), using Euclidean squared distance as a selection metric is sub-optimal from an analytic perspective. As a matter of fact, according to the SSD criterion, an optimal pair would be one that minimizes the distance between stocks. However, this means that a pair with a zero spread across the formation period would be deemed optimal. This is not logically consistent with the notion of a potentially profitable pair, as a good candidate should exhibit high spread variance and strong mean-reversion properties to provide viable trading opportunities.

2.3.2 The cointegration approach

The framework that is most frequently referenced consists of three main stages [Pole (2011)]: Initially, pairs are chosen in accordance with statistical or fundamental similarity metrics. Next, suitability for trading is determined by using the Engle-Granger or the Johansen cointegration test. Lastly, non-parametric techniques are utilized to create optimal entry and exit thresholds. This approach

is widely recognized in the field as one of the more robust in detecting long term relationships, especially compared to the distance approach. Rad et al. (2016) were the first to conduct a large-scale empirical implementation of the cointegration approach, which was subsequently adopted by Huck and Afawubo (2015).

2.3.3 Time series approach

Avellaneda and Lee (2010) have effectively implemented a variation of the approach proposed by Elliott et al. (2005), the authors most frequently referenced in this field [Krauss (2017)], which incorporates state space models and suitable estimation algorithms to parametrically handle mean-reverting spreads. This application provides clear evidence that dynamic trading rules based on time-series analysis can enhance trading returns. Another important work is developed by Bertram (2009), which introduces an optimal statistical arbitrage trading rule for mean-reverting portfolios. We will enter in depth in the development of this type of strategy in the next sections regarding the trading phase.

2.4 Trading phase

2.4.1 Threshold-based trading model

The trading strategy commonly employed is based on the threshold-based trading proposed by Gatev et al. (2006), which uses the spread divergence as a criterion for opening a trade. If the spread between the two price series composing a pair diverges more than two standard deviations from historical data, a trade is initiated. The trade is closed at the end of the trading period when the spread converges to the mean. However, it is possible that the spread may not converge, leading to a potentially large loss. The model can be described formally as follows [Sarmento and Horta (2020)]:

1. During the pair's formation period, calculate the spread's ($S_t = Y_t - X_t$) mean (μ_s) and standard deviation (σ_s) in order to have a normalized spread.

2. Define the model thresholds, including the threshold that triggers a long position (α_L), the threshold that triggers a short position (α_S), and the exit threshold (α_{exit}) that defines the level at which a position should be exited.

3. Monitor the spread's evolution (S_t) and check if any threshold is crossed.

4. If α_L is crossed, initiate a long position in the spread by buying Y and selling X. If α_S is

triggered, initiate a short position in the spread by selling Y and buying X. Exit the position when α_{exit} is triggered and a position is being held.

2.4.2 Time series trading models

Elliott et al. (2005) provide an explicit description of the spread using a mean-reverting Gaussian Markov chain, observed in Gaussian noise. This can be achieved by utilizing a state space model consisting of a state and a measurement equation. In continuous time, the Ornstein-Uhlenbeck (OU) process can be utilized to describe the state process [Krauss (2017)]. The OU process is defined by the equation:

$$dX_t = \theta(\mu - X_t)dt + \sigma dW_t \tag{2.5}$$

where θ is the mean reversion speed rate, μ is the mean of X_t, W_t is the standard Wiener process, and σ is the standard deviation for the Wiener process. The measurement equation is the second component to a state space model. According to this model, a pairs trade is initiated when $y_k \geq \mu + c\frac{\sigma}{\sqrt{2\theta}}$ or when $y_k \leq \mu - c\frac{\sigma}{\sqrt{2\theta}}$, where c is a fixed parameter. Elliott et al. (2005) provide no guidance on how to determine this parameter. The position is reversed at time T, denoting the first passage time result for the Ornstein-Uhlenbeck process. Do et al. (2006) argue that this approach has some advantages. Firstly, the model is fully tractable, meaning that its parameters can be estimated based on maximum likelihood and is optimal in terms of minimum mean squared error. Finally, the approach is fundamentally based on mean-reversion. However, Cummins and Bucca (2012) provide also critiques, arguing that a major limitation lies in the Gaussian nature of the OU-process, which is in conflict with the stylized facts of financial data. Nonetheless, this disadvantage is largely compensated for by analytic simplicity, as this concept may represent a significant improvement compared to nonparametric trading rules. This study also aims at verify this assumption by means of empirical analysis.

Chapter 3

Pairs Selection Structure

3.1 Problem statement

As the practice of Pairs Trading gains more popularity, it becomes more challenging to discover profitable pairs. The shortage of such promising pairs necessitates expanding the search to encompass wider groups of securities, with the hope that by examining a larger collection, the chances of finding a suitable pair will rise. The most straightforward approach typically used is to compile a list of all potential pairs by considering the combination of each security with every other security in the dataset. This results in a total of $\frac{n \ x \ (n-1)}{2}$ possible combinations, where n denotes the quantity of securities that are available, which is clearly an issue in terms of computational cost.

To overcome this problem, the typical and more restrictive practice of solely comparing securities within the same sector is frequently utilized, since it is also particularly straightforward to implement. This significantly decreases the number of required statistical tests, thus diminishing the possibility of discovering false relations. Additionally, there are fundamental grounds to assume that securities within the same sector are more inclined to move in tandem as they are subject to comparable underlying factors.

However, the ease of this approach can also be a drawback; in fact, the more traders become aware of these pairs, the more difficult it is to uncover pairs that have not yet been extensively traded, resulting in a narrower profit margin. This discrepancy inspires the quest for a methodology that lies in between these two scenarios [Sarmento and Horta (2020)]: a proficient pre-partitioning of the universe of assets that does not restrict the pairing of assets to relatively apparent solutions, while

also avoiding excessive search combinations.

3.2 Proposed framework

In this section we will follow Sarmento and Horta (2020) proposed framework, already proven effective, with the further aim to apply it to different set and type of data, time periods, and employ different robustness analysis for what concerns the number of principal components. The authors suggest utilizing an Unsupervised Learning algorithm, with the expectation that it will deduce significant clusters of assets from which to choose pairs. The proposed methodology comprises the following stages:

1. Dimensionality reduction - discover a compressed representation for each security.

2. Unsupervised Learning - implement an appropriate clustering algorithm.

3. Select pairs - establish a set of guidelines for choosing pairs to trade.

The subsequent three sections will delve into the specifics of how each of the aforementioned stages are executed.

3.3 Dimensionality reduction: PCA

The objective is to locate securities that share the same systematic risk-exposure which, as indicated by the Arbitrage Pricing Theory, produce the same anticipated long-term return. Any divergence from the theoretical projected return can, therefore, be interpreted as a pricing error and can be utilized as a basis for conducting trades. To extract the mutual underlying risk factors for each security, it is proposed utilizing Principal Component Analysis (PCA) on the return series, as outlined in Jolliffe (2011).

PCA is a statistical technique that employs an orthogonal transformation to convert a group of observations of potentially interrelated variables into a set of linearly uncorrelated variables, known as the principal components. The transformation is determined in such a way that the first principal component explains the maximum amount of variability in the data. Subsequent components, consequently, possess the highest possible variance under the condition that they

must be perpendicular to the previous components. It is noteworthy that each component can be interpreted as representing a risk factor [Avellaneda and Lee (2010)].

The application of PCA is carried out in the following manner. Initially, the return series for a security i at time t, $R_{i,t}$, is obtained from the security price series P_i,

$$R_{i,t} = \frac{P_{i,t} - P_{i,t-1}}{P_{i,t-1}} \tag{3.1}$$

Next, the return series must be standardized since PCA is sensitive to the relative scaling of the initial variables. This is achieved by subtracting the mean, R_i, and dividing by the standard deviation σ_i, as

$$Y_i = \frac{R_i - \bar{R}_i}{\sigma_i} \tag{3.2}$$

From the normalized return series of all assets, the correlation matrix ρ is calculated, where each entry is determined by

$$\rho_{ij} = \frac{1}{T-1} \sum_{t=T}^{M} Y_{i,t} Y_{j,t} \tag{3.3}$$

The application of PCA on return series is motivated by the fact that a return correlation matrix is more informative for assessing price co-movements, whereas using the price series could lead to the identification of erroneous correlations due to underlying time trends. The subsequent stage involves extracting the eigenvectors and eigenvalues to construct the principal components.

The eigenvectors establish the directions of maximum variance, whereas the eigenvalues quantify the variance along the corresponding direction. Eigenvalues and eigenvectors can be obtained using singular value decomposition (SVD). To that end, the normalized return series for all n securities are stacked in a matrix A. By directly applying the SVD theorem to matrix A, it is decomposed as

$$A = USV^T \tag{3.4}$$

Matrix U is an orthogonal matrix, and its columns represent the left singular vectors. Matrix S is a diagonal matrix and contains the singular values arranged in a descending order along the diagonal, ranging from the eigenvalue linked to the highest variance to the smallest. Matrix V is the transposed orthogonal matrix, and its rows correspond to the right singular vectors.

At this point, we choose the k eigenvectors that correspond to the k directions of maximum variance, where k denotes the number of features required to depict the transformed data. The more eigenvectors that are taken into consideration, the more effectively the data is represented. The matrix that contains the chosen eigenvalues in order of importance is referred to as the feature vector. Finally, the new dataset is obtained by multiplying the original matrix A by the feature vector, resulting in a matrix with size $n \times k$, which is thus reduced to the selected k features. It is essential to establish the number of features, k.

A typical approach involves examining the proportion of the overall variance explained by each principal component and utilizing the number of components that account for a fixed percentage, as outlined in Avellaneda and Lee (2010). Sarmento and Horta (2020) adopted a distinct strategy in light of the curse of dimensionality problem. This term, introduced by Bellman (1966), describes the issue that arises from the exponential increase in volume as additional dimensions are added to Euclidean space. This has a significant impact on the measurement of the distance between similar data points that suddenly become far apart from one another.

As a result, the clustering process becomes less efficient. As per Berkhin (2006), the effect becomes pronounced when exceeding 15 dimensions. Considering this, the number of PCA dimensions is capped at this value and is selected through empirical means. This study employs the restriction explained above on the number of principal components to circumvent the issue of dimensionality, but it further utilizes two distinct metrics to determine the appropriate number. Although it evaluates the explained variance of the components, the primary metric is based on the percentage of pairs that show cointegration in both the validation and testing periods. This is because we want the clustering algorithm able to detect pairs showing a robust cointegration relationship, that is a relation stable in the long run.

Starting from 5 components, the goal is to maximize the metrics as the number increases (10, then 15 components, the limit described previously).

3.4 Unsupervised Learning: OPTICS clustering

Since delving into the different clustering techniques and investigating the choice of appropriate clustering type is beyond the scope of this work, we will consider only the outcome of this steps performed by Sarmento and Horta (2020), where ultimately the OPTICS algorithm is deemed the best, especially in the context of this study. We will briefly describe the reasons. Ordering points to identify the clustering structure (OPTICS), proposed by Ankerst et al. (1999) is a density-based clustering algorithm which offers several advantages: firstly, clusters can exhibit arbitrary shapes, without limit to one global parameter setting. Secondly, it is inherently resistant to outliers since it does not group every point in the dataset. Lastly, it does not necessitate the specification of the number of clusters in advance.

In sum, the algorithm returns the points and the corresponding reachability-distances. The reachability-distance of a point p in relation to a point o can be construed as the minimum distance such that p is directly density-reachable from o. This necessitates that o is a core point, implying that the reachability-distance cannot be less than the core-distance (i.e., the minimum distance ε' between p and a point in its ε-neighborhood, such that p is a core point concerning ε'). The points thus retrieved are sorted such that the nearest spatial points become neighbors in the ordering. Based on this data, a reachability plot can be generated by organizing the ordered points on the x-axis and the reachability-distance on the y-axis.

3.5 Pairs selection criteria

After creating clusters of assets, it is essential to establish specific criteria for selecting pairs to trade while ensuring that the equilibrium of the pairs is maintained. To achieve this, we follow again Sarmento and Horta (2020) who suggest combining methods used in different research. The proposed criteria involve selecting pairs that meet the following four conditions:

1. The constituents of the pair must be cointegrated.

2. The spread Hurst exponent of the pair should indicate a mean-reverting nature.

3. The spread of the pair diverges and converges within appropriate time frames.

4. The spread of the pair reverts to the mean frequently enough.

The use of cointegration methods has been shown to be effective in selecting pairs for trading as they identify more reliable equilibrium relationships compared to other methods. Therefore, a pair is considered suitable for trading only if the two securities that make up the pair are cointegrated by means of the Engle-Granger test because of its simplicity. However, due to the critique for which the choice of the dependent variable may lead to different conclusions [Armstrong (2001)], in the empirical part the Engle-Granger test was run for both possible selections of the dependent variable and the combination that provided the lowest t-statistic was selected, after having passed the p-value step using the 1

A second validation step is recommended to increase confidence in the mean-reverting nature of the pairs' spread and to avoid false positives that might arise due to the multiple comparisons problem. This step enforces that the Hurst exponent associated with the spread of a given pair is less than 0.5, indicating a leaning towards mean-reversion. This criterion is based on the work of Ramos-Requena et al. (2017), who suggest that relying solely on cointegration is too restrictive, therefore, in this work, the Hurst exponent is used as a supplementary check for mean-reverting properties after having found two securities forming a pair cointegrated.

However, a mean-reverting spread alone does not guarantee profitability. The duration of mean-reversion must be coherent with the trading period. For this, on average, the spread should take a period comprises in the trading window of a year (that is between one day and one year, or 252 trading days) to mean-revert. The method suggests filtering out pairs for which the half-life, which can be interpreted as an estimate of the expected time for the spread's mean-reversion, is not consistent with the trading period.

Finally, it is required that each spread crosses its mean at least once per month to ensure sufficient liquidity (that is 12 times per year). Even if there is a negative correlation between the number of mean crosses and the half-life period, as more mean crosses are typically associated with a shorter half-life, these properties are not interchangeable. Imposing this constraint may not only enforce the previous condition but also eliminate pairs that meet the mean-reversion timing requirements but fail to cross the mean, resulting in no opportunities to exit a position.

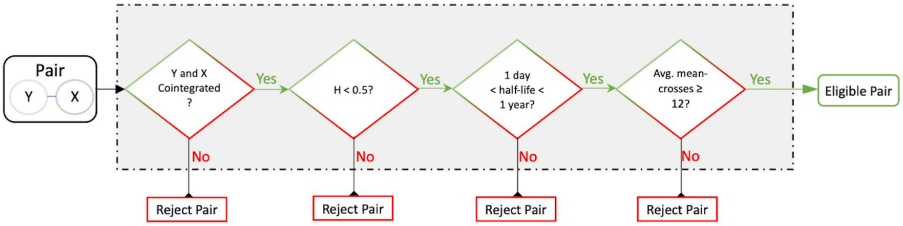


Figure 3.1: Pairs selection rules.

Chapter 4

Trading model structure

4.1 Problem with threshold-based trading model

An issue with the trading model based on thresholds is that the points at which trades are initiated are not precisely defined. The sole condition for entering a trade is when a predetermined threshold is crossed, regardless of the current direction of the spread. This could lead to periods of unfavorable portfolio performance if a currency pair continues to diverge. To avoid this, alternative trading strategies may need to be explored.

4.2 Stochastic spread method

In order to develop a more resilient trading model, it would be beneficial to explore different trading strategy, in particular the ones suggested by the time series approach literature. In this section the framework proposed by Zeng and Lee (2014) will be used, where the primary aim is to identify the most favorable thresholds in relation to the transaction cost and OU process parameters, with the goal of maximizing the long-term expected average profit.

Conversely from the basic threshold-based model, the stochastic spread technique characterizes the mean reversion process of pairs trading as an Ornstein-Uhlenbeck (OU) process. In this section we will recall the main properties of the OU process, while in the next chapter (methodology) we will delve into the details of the strategy employed.

4.2.1 The Ornstein-Uhlenbeck process

The Ornstein-Uhlenbeck (OU) process can be compared to a Wiener process, but with a restrictive force that maintains its proximity to zero [Bibbona et al. (2008)]. Over time, the paths of a Wiener process tend to diverge due to its expanding variance, while OU trajectories remain confined to the vicinity of the origin. The distribution of the Ut process appears to be stationary over time. Additionally, as t approaches infinity, the OU process converges to a stationary distribution that is also Gaussian.

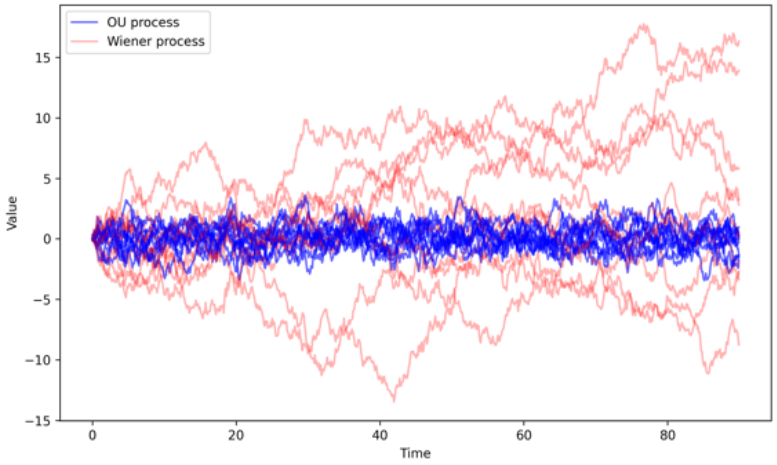


Figure 4.1: Trajectories of an OU (in blue) are compared with trajectories of a Wiener process (in red). The former permits a stationary distribution, while the variance of the latter increases with time.

The Wiener process, denoted as W_t, is the outcome of a (discrete) random walk when the steps are infinitely small and frequent. In meteorology, it is frequently referred to as random walk noise. It was initially introduced as a mathematical representation of the random movement of particles suspended in a fluid, also known as Brownian motion (refer to Stachel et al. (1987)). The velocity of Brownian motion is defined by its derivative, which is a continuous white noise. However, due to its complex mathematical description (it does not follow the standard definition of a stochastic process) and physical inconsistencies, an alternative model, the Ornstein-Uhlenbeck (OU) process, is typically employed to represent the velocities of Brownian particles. The Ornstein-Uhlenbeck (OU) process, denoted as U_t (with $t \geq 0$), is the solution to the Langevin stochastic differential equation (refer to Arnold (1974)),

$$dU_t = \left(-\frac{U_t}{\tau} + \mu\right) dt + \sigma \, dW_t \tag{4.1}$$

which is defined as follows: W_t is a Wiener process, σ (≥ 0) represents the diffusion coefficient, τ ($\geq$

0) is the time constant, and μ is the drift coefficient. For high frequencies (short observation times), the Ornstein-Uhlenbeck (OU) process behaves similarly to a Wiener process (random walk on phase), while for low frequencies (long observation times), it is comparable to white phase noise. The OU process's (two-sided) spectrum is determined when the process attains stationarity as t approaches infinity, and it was originally discovered by the authors in Wang and Uhlenbeck (1945).

The first passage time problem

A first passage time in a stochastic system, is the time taken for a state variable to reach a certain value, usually a boundary or threshold. This section addresses the occurrence of first passage time for an Ornstein-Uhlenbeck (OU) process, U_t, as it crosses two constant absorbing boundaries, S and L. At time $t = 0$, U_t begins at $-S < u_0 < L$, and $T\ (u_0, S, L)$ is the time at which the process first surpasses the interval (S, L). The first passage time $T\ (u_0, S, L)$ is a random variable. The survival probability $P_s\ (u_0, t)$ signifies the likelihood that the process, starting from u_0, has not yet crossed either of the thresholds by time t, or alternatively, that the first passage time has not yet occurred. It is noteworthy that the OU process has a stationary distribution with an expected value of zero, meaning that once it has escaped beyond the barriers, it is highly likely to return to the vicinity of the origin, that means it is showing a mean-revert behavior.

First passage times between two barriers

To approximate first passage times between two symmetric thresholds, S and $-S$, the simulation is halted the first time U_n exceeds S or falls below $-S$, and the first passage time T is derived as $T = nh$, where h represents the discretionary interval. However, the first passage time determined via simulation is typically overestimated. This is due to the fact that although the process is continuous in time, it is observed only at discrete intervals.

There is a possibility that a passage may occur between two points below the threshold (as depicted in Figure 4.2), which cannot be detected using this approach, and the simulation will continue even if a crossing is possible. Consequently, the estimated first passage time is longer than the actual value. Despite this limitation, a technique was devised to account for hidden passages in simulations in the case of a single threshold S. To apply this technique to two symmetric barriers, the crossing of the upper and lower barriers is approximated as two independent processes and summed up the crossing probabilities. Although this technique is not entirely rigorous, it yields good outcomes when the barriers are far enough and the time step is relatively small [Bibbona et al.

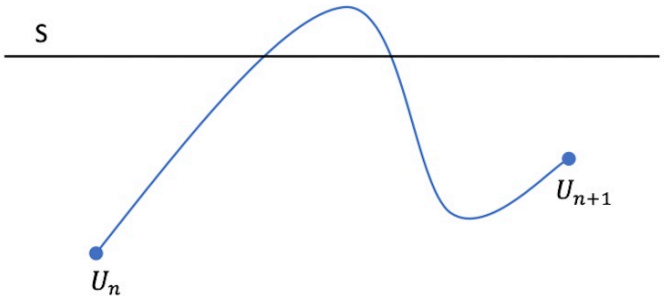


Figure 4.2: The process is continuous, but it is observed at discrete intervals. Therefore hidden passages may occur between two observations.

(2008)].

Chapter 5

Methodology

This study introduces a methodology for each stage involved in a Pairs Trading strategy, namely the application of a stochastic-based model (Research question 1) and a modified version of it (Research question 2). A specific research framework for each stage has been devised, as depicted in Figure 5.1.

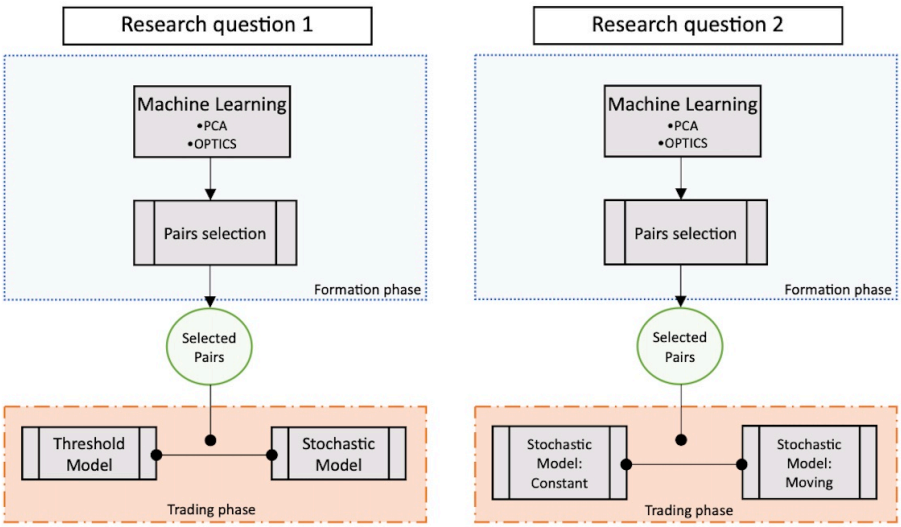


Figure 5.1: Research design overview.

One feasible research approach involves addressing the first research question of this study and subsequently employing the two different windows to generate moving parameters for the stochastic-based model identified to answer the second research question. This chapter aims to

present in detail the techniques used in the work and showed in the literature review. There will be explained the procedures employed, firstly for cleaning the dataset and split in training and validation set, then performing the PCA and the subsequent clustering to form pairs, finally the different trading strategies: the threshold based model, and the other assumes the spread following an OU process. At the end, the metrics used to evaluate both the accuracy of the computation and the profit of the strategy will be shown.

5.1 Dataset

Two datasets have been considered within this work. The first one contains industries classification following the Global Industry Classification Standard ($GICS$), in which stocks belonging to the $S\&P$ 500 are grouped in 11 sectors. The second instead considers the closing prices of stocks belonging to the $S\&P$ 500 (and the NASDAQ Index) during the period from the beginning of January 2005 to the end of July 2022. Usually, the composite ticker will end with country code: IBM US, AAPL US, 1 HK etc. Instead, the Bloomberg exchange ticker will end with Bloomberg allocated exchange code: IBM UN (NYSE), AAPL UW (NASDAQ) etc.

5.2 Data handling

Using Python's Pandas library and some ad hoc function in order to integrate missing data, correct errors or altered data, and remove any further problems, the following datasets have been used in this study. In the first dataset, eventual duplicated tickers were deleted, then sorted by count and grouped by sector. In the second dataset, firstly all dates on which the market was closed were removed from the time series. These dates mainly involved national holidays (Memorial Day, Thanksgiving, President Day, and Christmas) and weekends. Secondly only stocks which are listed on the SP500 (that is, named with UN Equity) will be consider in this work, but since data was downloaded from Bloomberg, it was necessary to delete all the stocks that were also listed on NASDAQ (i.e., with UQ or UW code). Afterwards, it was required checking for outliers and deal accordingly. In this work it was adopted a simpler technique then the others proposed in the literature (such as Hidden Markov Models or other Machine Learning techniques, for which the presence of missing values is another issue), that is the IQR method. IQR stands for Interquartile Range, which is a method for outlier detection by setting:

- Q1 is the first quartile of the data, where 25% of the data lies between minimum and Q1.
- Q3 is the third quartile of the data, where 75% of the data lies between minimum and Q3.

The difference between Q3 and Q1 is called the Inter-Quartile Range or IQR (IQR = Q3 - Q1). To detect the outliers we define a new decision range, and any data point outside this range (i.e. less than the lower bound or more than the upper bound) is considered as outlier and substituted by a missing value. The range is as given below:

- Lower Bound: (Q1 - 1.5 * IQR).
- Upper Bound: (Q3 + 1.5 * IQR).

Lastly, it was necessary to deal with inevitable missing values in the data frame, and in order to do so it was employed a function that deleted series in which missing data were more than 50% of the observations. The remaining missing values have been dealt with some Pandas methods (pandas.DataFrame.ffill and pandas.DataFrame.bfill).

5.3 Data partition

The data needs to be divided into two distinct timeframes: the formation phase and the trading phase. During the formation phase, the data replicates the information accessible to the investor before executing any transactions. Initially, this timeframe is utilized to discover the most attractive pairs of candidates. Then, a reduced segment of the data, referred to as the validation dataset, is utilized to replicate the strategy's performance in the recent past. The trading phase is utilized to mimic the performance of the applied trading model with unobserved data. The periods considered are the 3-year-long formation periods in which the penultimate year is employed to validate the efficiency before executing the strategy on the test set, also defined as trading phase that spans one year based on Do and Faff (2010). This stage will consist of two strategies: the threshold-based model and the stochastic spread model.

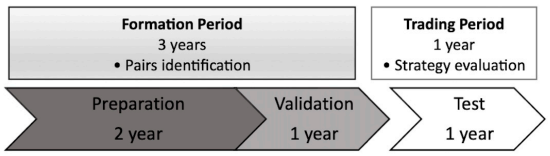


Figure 5.2: Period decomposition.

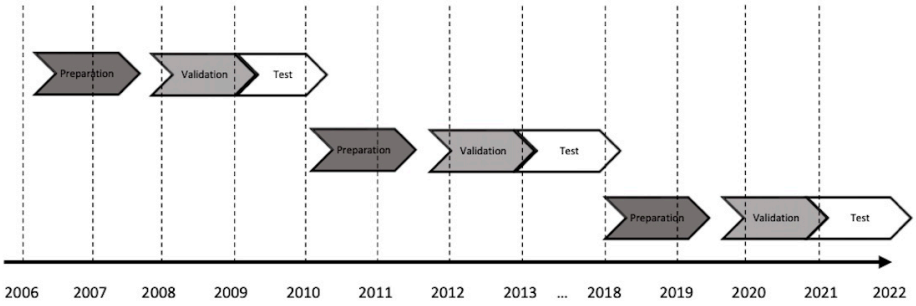


Figure 5.3: Data partition periods.

Dataset partition

	2006-2009	2010-2013	2018-2021
Trading days	1051	1046	1046
Begin	01/01/06	04/01/10	02/01/18
End	31/12/09	31/12/13	30/12/21
N. of samples	1051	1046	1046
Train	519	524	523
Validation	266	262	259
Test	266	260	264
Industries			
Communication Services	20	20	20
Consumer Discretionary	68	72	72
Consumer Staples	31	32	32
Energy	45	47	48
Financials	64	67	67
Health Care	48	50	53
Industrials	68	74	74
Information Technology	35	38	40
Materials	36	37	37
Real Estate	34	34	34
Utilities	33	34	34

Table 5.1: Dataset partitions.

5.4 Trading setting-up

The primary objective of this research stage is to compare the outcomes of various trading techniques with each other. To achieve this, identical configurations, such as the time periods and pairs used, are implemented for both techniques.

5.4.1 Standard threshold-based model

The model parameters utilized in this study are those recommended by Gatev et al. (2006), which have been referenced in numerous studies in the field. The standard deviation (σ_s) and mean (μ_s) of the spread are computed based on the complete formation period. The exit threshold is established as the mean value, signifying that a position is only terminated when the spread returns to its mean.

Parameters	Values
Long threshold	$\mu_s - 2\sigma_s$
Short threshold	$\mu_s + 2\sigma_s$
Exit threshold	μ_s

Table 5.2: Threshold-based model parameters.

Reasearch question 1

5.4.2 Stochastic spread method

One potential approach to enhance the trading strategy is to conduct extensive time series analysis on the spread and exploit the mean-reverting characteristics of the time series. This technique offers several benefits over the framework outlined in the preceding section. For example, the mean-reverting process guarantees that the necessary mean reversion for pairs trading is captured. Additionally, estimating the parameters of the state space model using the Expectation Maximization (EM) algorithm makes the model entirely feasible.

These advantages are expected to result in a framework that generates higher profits than the distance method. Zeng and Lee (2014) provide an analytic framework to maximize the expected profit per unit time in the long run, giving rise to an optimization problem involving the long run expected average profit; this is achieved by finding the optimal thresholds (the right entry and exit points) as functions of the transaction cost and parameters of the OU process. They have formulated a simplified polynomial expression for the expected first-passage time of an OU process with two-sided boundary than the one using the Laplace transform, able to derive the analytic formula for optimal thresholds in a pairs trading strategy. First, we define the spread as the difference between the stock A and B with prices $S_A(t)$ and $S_B(t)$ of the two assets forming a pair as:

$$X_t = S_A(t) - beta * S_B(t) \tag{5.1}$$

For modelling the spread dynamics, we consider the following OU process:

$$dX_t \;=\; \theta(\mu - X_t)dt \;+\; \sigma dW \tag{5.2}$$

where θ is the mean reversion speed rate, μ is the time-dependent mean-reversion level of X_t, W_t is the standard Wiener process, and σ is the standard deviation for the Wiener process.

Similar to Bertram (2010), we can transform equation (5.2) into the dimensionless system (i.e. because Y_τ is not dependent on the model parameters, noticing that the transformation is linear, so that each value of X_t corresponds to a sole value of Y_τ) by $\tau = \theta t$ and $Y_\tau = \frac{\sqrt{2\theta}}{\sigma}(X_t - \mu)$. Hence, we have:

$$dY_\tau \;=\; -Y_\tau d\tau \;+\; \sqrt{2}dW_\tau \tag{5.3}$$

Since Bertram (2010) also demonstrated that the ideal thresholds for maximizing both return per unit time and the Sharpe ratio were symmetrical with respect to the mean, trading signals are produced based on the attainment of a predetermined threshold by Y_τ. We will call $\widetilde{a}$ and $\widetilde{b}$ the thresholds in the real system, whereas a and b, $a > b$, the threshold in the dimensionless system; $\widetilde{c}$ the transaction cost in the real system and cost $c = \widetilde{c}\frac{\sqrt{2\theta}}{\sigma}$ the cost in the dimensionless system, so the net profit for each transaction is $\widetilde{a} - \widetilde{b} - \widetilde{c}$, or $a - b - c$ in the dimensionless system.

The trading cycle consists of two main components: the first part involves initiating and liquidating positions, while the second part involves waiting for the next trading opportunity. Let t_1 and t_2 represent the duration of these two parts, and let τ_1 and τ_2 indicate the corresponding time in the dimensionless system. Following again Bertram (2010), τ_1 denotes the first-passage time from a to b, while τ_2 represents the time required to escape the range $[-a, a]$, and the total time for each trading cycle is $T = \tau_1 + \tau_2$. Mathematically, τ_1 and τ_2 are defined as follows:

$$\tau_1 = inf\Big\{t; Y_t = b \Big| Y_0 = a\Big\} \tag{5.4}$$

$$\tau_2 = inf\Big\{t; \big|Y_t\big| = a \Big| Y_0 = b\Big\} \tag{5.5}$$

Suppose there are N_τ transactions completed in $[0, \tau]$, so the net profit is $NP_\tau = (a - b - c)N_\tau$. By the elementary renewal theorem (which states that asymptotically, the anticipated number of renewals within an interval is proportional to the length of the interval), the expected profit per unit time is given by

$$\mu = \lim_{\tau\to\infty} \frac{E\left[NP_\tau\right]}{\tau} = (a - b - c) \lim_{\tau\to\infty} \frac{E\left[N_\tau\right]}{\tau} = \frac{a - b - c}{E\left[T\right]} \tag{5.6}$$

where $E[T] = E[tau_1] + E[tau_2]$.

Notice that the expected return per unit time in real system is

$$\widetilde{\mu} = \frac{\widetilde{a} - \widetilde{b} - \widetilde{c}}{E\left[\widetilde{T}\right]} = \frac{\sigma\sqrt{\theta}}{\sqrt{2}} \frac{a - b - c}{E\left[T\right]} = \sqrt{\frac{\theta}{2}}\sigma\mu \tag{5.7}$$

The coefficient $\sigma\sqrt{\frac{\theta}{2}}$ is only determined by the prices of the pairs and is (variable) a constant once the (moving) model parameters are known. Hence, optimizing the real return is equivalent to optimizing the return in the dimensionless system. The constant $\sigma\sqrt{\frac{\theta}{2}}$ encodes crucial and intuitive information: a higher mean reversion rate θ implies a greater trading frequency, while a larger implies more significant fluctuations of X_t, both of which result in higher profits per trade. As both time and scale are linearly converted into the dimensionless system, the optimal thresholds can be initially determined in the dimensionless system before being transformed back into the real system.

First-passage times

To determine the optimal thresholds, it is essential to calculate the expected first-passage time for both one-sided and two-sided boundaries.

First-passage time over a one-sided boundary

For one-sided boundary, the expectation is expressed as an infinite sum of polynomials. To summarize, for a $(x) > 0$ and b $(y) > 0$, the expectation of $\mathrm{T_{a,0}}(\mathrm{T_{x,0}})$, the first-passage time from a (x) to 0 is:

$$E\left[T_{x,0}\right] = \frac{1}{2} \sum_{k=1}^{\infty} (-1)^{k+1} \frac{\left(\sqrt{2}x\right)^k}{k!} \Gamma\left(\frac{k}{2}\right) \tag{5.8}$$

and the expectation of $\mathrm{T_{0,b}}(\mathrm{T_{0,y}})$, the first-passage time from 0 to $b(y)$ is:

$$E\left[T_{0,y}\right] = \frac{1}{2} \sum_{k=1}^{\infty} \frac{\left(\sqrt{2}y\right)^k}{k!} \Gamma\left(\frac{k}{2}\right) \tag{5.9}$$

Hence, the expectation $E\left[T_{a,b}\right]$ for the case $a > 0$ can be written as:

$$E\left[T_{a,b}\right] = \begin{cases} E\left[T_{a,0}\right] - E\left[T_{b,0}\right], b > 0 \\ E\left[T_{a,0}\right] + E\left[T_{0,-b}\right], b \leq 0 \end{cases} . \tag{5.10}$$

Utilizing the symmetry of an OU process, we can obtain the expectation value for the case when $a < 0$, using the equations $E[T_{a,b}] = E[T_{-a,0}] + E[T_{0,b}]$ for $b > 0$, and $E[T_{a,b}] = E[T_{-a,0}] - E[T_{-b,0}]$, for $b < 0$. By exploiting the symmetric property of an OU process, the variance of this type of first-passage time between any two points can be computed.

First-passage time over a two-sided boundary

For the first-passage time over a two-sided symmetric boundary, Darling and Siegert (1953) derived the Laplace transform of T, the first-passage time from b to cross the boundary (a, a), after a few simplifications for which we refer to Zeng and Lee (2014), the expectation can be expressed as:

$$E[T_{-a,a,b}] = \frac{1}{2} \sum_{k=1}^{\infty} \frac{(\sqrt{2}a)^{2n} - (\sqrt{2}b)^{2n}}{(2n)!} \Gamma(n) \tag{5.11}$$

Objective function and optimal rule

Having obtained the polynomial expression for the expectation, we can determine the optimal thresholds for pairs trading. The primary objective is to maximize the expected return per unit time, as described in Section 2. The strategy involves taking positions when Y_τ hits the opening threshold a (or a), liquidating positions when it hits the closing threshold b (or b) and waiting for the next opportunity until Y_τ reaches an opening threshold again. Three distinct cases with different values of a and b will be considered, under the assumption that a 0. Since the OU process is symmetric, the case for $a < 0$ will be identical. The objective function is given by

$$f(a, b) = \frac{a - b - c}{E[\tau_1] + E[\tau_2]} \tag{5.12}$$

where $E[\tau_1]$ and $E[\tau_2]$ are explicitly shown by equations (5.10) and (5.11).

The optimization problem is:

$$\max_{a,b} f(a, b) = \frac{a - b - c}{E[\tau_1] + E[\tau_2]} \tag{5.13}$$

And considering three different cases, it is accordingly subjected to three constraints:

Case1: $0 \leq b \leq a$ *subject to* $0 \leq b \leq a - c$.

Case2: $-a \leq b \leq 0$ *subject to* $-a \leq b \leq min\{0,\ a-c\}$.

Case3: $b < -a$ *subject to* $b < -a$.

Among the three cases, there is an ideal rule that yield optimal values of a^* and b^*. The rule in question entails no waiting time between the two trades, as opposed to the common practice which liquidates the position exactly when the spread reverts to the mean at $b^* = 0$ (also called 'Conventional Optimal Rule'). In this context, and consistent with Bertram (2010) it is demonstrated that b^* =-a^* is the global maximal by showing that $f(a^*, b^*) \geq f(a, b)$ for any a, b on the boundary. As this 'New Optimal Rule' reduces transaction costs by half compared to the 'Conventional Optimal Rule', it is reasonable to expect that the 'New Optimal Rule' will outperform the 'Conventional Optimal Rule'.

Compute the log likelihood of the parameters

The Maximum-Likelihood (ML) method will be used to estimate the parameters based on Hu and Long (2007). The log likelihood for the process X_t is given by:

$$L\left(X|\mu,\theta,\sigma\right) = -\frac{n}{2} - \frac{1}{2}\sum_{i=1}^{n}\log\left(1-e^{-2\theta(t_i-t_{i-1})}\right) - \frac{\theta}{\sigma^2}\sum_{i=1}^{n}\frac{X_{t_i}-\mu-\left(X_{t_{i-1}}-\mu\right)e^{-\theta(t_i-t_{i-1})}}{1-e^{-2\theta(t_i-t_{i-1})}} \tag{5.14}$$

Maximizing $L\left(X|\mu,\theta,\sigma\right)$, we get the estimation for the parameters: μ,θ and σ. Assuming that the parameters are constant during the data collection period, we can apply the optimal pairs trading rule.

Optimal threshold finder

In this way as a final step, we transform the transaction cost $\widetilde{c}$ into the dimensionless system cost, $c = \widetilde{c}\frac{\sqrt{2\theta}}{\sigma}$ so the net profit for each transaction is $\widetilde{a} - \widetilde{b} - \widetilde{c}$, or $a - b - c$ in the dimensionless system, to obtain the optimal thresholds as a^* and b^*. Then, we transform back to get the real thresholds as $\widetilde{a^*} = a^*\frac{\sigma}{\sqrt{2\theta}} + \mu$ and $\widetilde{b^*} = b^*\frac{\sigma}{\sqrt{2\theta}} + \mu$. A trading is triggered whenever X_t reaches $\widetilde{a^*}$ or $\widetilde{b^*}$. For instance, if $\mathrm{Y}_{\tau 1} = a$ $(a > 0)$, we would sell 1 dollar worth of stock A and buy β dollars of stock B, and if $\mathrm{Y}_{\tau 2} = b$ $(b < a)$, we would close out our positions and realize a profit. In this case the trading strategy is: Starting from a transaction cost $c = 0.00471$ [Zeng and Lee (2014)] dollars for

each dollars invested, this it is transformed into the dimensionless system obtaining the optimal thresholds as $a^* = -b^*$ and $b^* = -a^*$. Then, it is transformed back to get the real thresholds. The strategy setting is summarized as follows:

Parameters	Values
Long threshold	$\widetilde{b^*} = b^* \frac{\sigma}{\sqrt{2\theta}} + \mu$
Short threshold	$\widetilde{a^*} = a^* \frac{\sigma}{\sqrt{2\theta}} + \mu$
Short exit threshold	$\widetilde{b^*} = b^* \frac{\sigma}{\sqrt{2\theta}} + \mu$
Long exit threshold	$\widetilde{a^*} = a^* \frac{\sigma}{\sqrt{2\theta}} + \mu$

Table 5.3: Stochastic-based model strategy.

Considering again the pair example from section 1.1, DTE and PCG, the trading phase involving the stochastic strategy with fixed parameters is shown in figure 5.4.

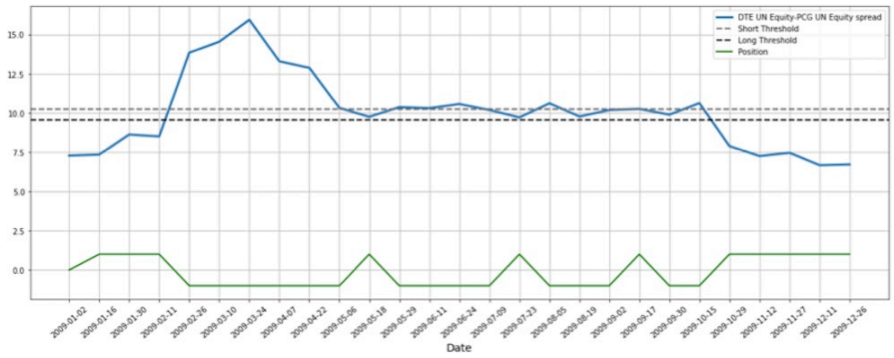


Figure 5.4: Exemplifying a Pairs Trading stochastic strategy execution (constant).

In this example we can see two dashed lines representing short threshold $\widetilde{a^*}$ and long threshold $\widetilde{b^*}$, a blue line for the spread between stocks, and a green line indicating the positions during the trading period. It is worth noticing how in this strategy there is no clear exit from a trade; conversely from the standard threshold-model where trades are closed when the spread reaches the 0 level, that is it reverts to its mean, in this model once a positions is closed, another one is opened in a continuous context as shown from the position-green line where values are only -1, for short, or 1, for long positions. Consequently, the number of trades using this strategy will be higher than the standard one but only the results will give us an insight on how this approach could affect profitability.

Reasearch question 2

Moving model parameters

One possible improvement was originally proposed by Zeng and Lee (2014), who left the analysis of changing parameters for future studies. However, they did not provide any indications on what measures to adopt when considering moving parameters. Therefore, the idea was to consider two metrics that span at most the duration of a trading phase, which is one year. In this part we will investigate two different sliding windows, one considering 126 days (6 months) and the other one 252 days (a year).

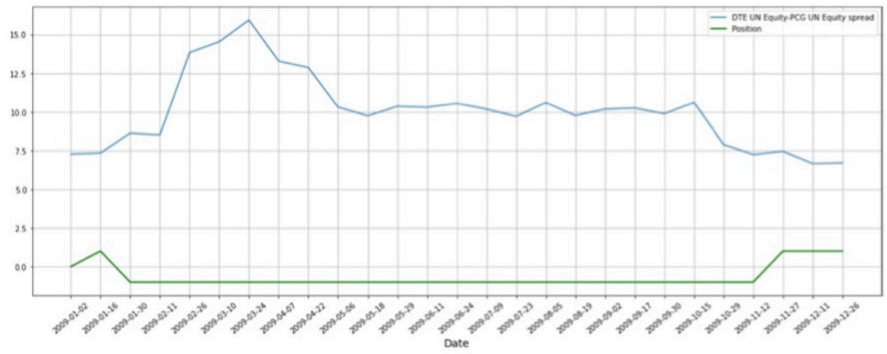


Figure 5.5: Exemplifying a Pairs Trading stochastic strategy execution (126d).

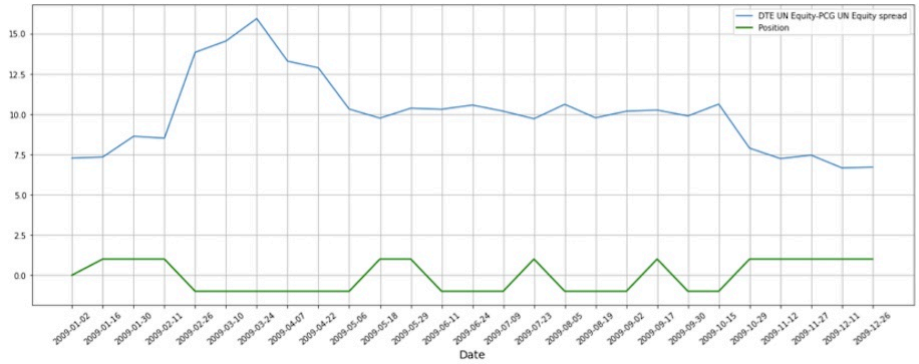


Figure 5.6: Exemplifying a Pairs Trading stochastic strategy execution (252d).

Considering the shortest window, the number of trades has drastically decreased, remaining opened at a short position over almost the trading period. On the other hand, the 252-day-window shows a behavior similar to the one using constant parameters, with only a different position at the end of May 2009. In the following table are displayed the strategy parameters, whereas in cells

regarding moving window there are presented the parameters at the beginning and at the end of the considered periods.

In this way both model parameters and the threshold won't remain constant over the trading phase but will change accordingly with price time series, giving a more realistic framework for the analysis. The two approaches are displayed in the figures below. Since parameters are moving, they are not shown in the pictures to ease the visualization. They are presented in table 5.4.

	Constant	**126 days**	**252 days**
Theta, θ	0,017425383	0,025838531	0,018987622
		...	...
		0,027592268	0,017514422
Mu, μ	9,90677612	11,82620579	10,11539736
		...	...
		7,923524397	9,952181678
Sigma, σ	0,44427107	0,509808152	0,447952658
		...	...
		0,366779379	0,448976294
$\widetilde{a^*}$	10,24696337	12,15385106	10,44822622
		...	...
		8,182119607	10,29407549
$\widetilde{b^*}$	9,566588865	11,49856051	9,782568529
		...	...
		7,664929187	9,610287904
a^*	0,142947368	0,146098595	0,144790193
		...	...
		0,165624378	0,142521614
b^*	-0,142947368	-0,146098595	-0,144790177
		...	...
		-0,165624378	-0,142521599

Table 5.4: Stochastic-based model parameters.

From this table we can see that parameters from 126 days window have a greater range, whereas 252 days window are less variable, indicating how the constant parameters are a sort of average of these lasts. Recalling Zeng and Lee (2014): "If the thresholds are narrow, then the time it needs to complete a trade is small, but so is the profit in each trade. On the other hand, if thresholds are too wide, the profit in each trade is larger, but so is the total time needed to complete

a trade." The latter assumption could be partially verified at this step, since the 126-day-window spans over a larger range of parameters, resulting in a long trade as depicted in figure 5.5. Results from the chapter 6 will give us further insights on how changing the parameters will affect strategy profitability.

5.5 Test portfolios

Three test portfolios are designed to simulate probable trading scenarios. Portfolio 1 includes all pairs identified during the formation period. Portfolio 2 utilizes feedback obtained from running the strategy in the validation set and only selects pairs with positive results [Sarmento and Horta (2020)]. Finally, Portfolio 3 corresponds to a situation where the investor is interested in investing in pairs which were identified in the training set, and still cointegrated in the test set. The objective of this last analysis is to evaluate the general behavior of pairs, as selecting pairs based solely on their performance on the validation set provides no assurance that the top-performing pairs will exhibit the same conduct in the test set.

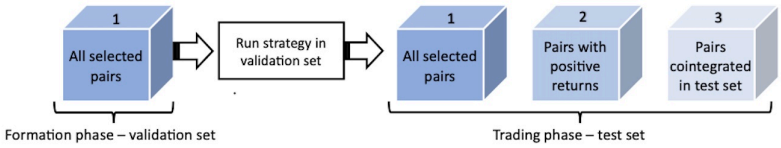


Figure 5.7: Test portfolios.

5.6 Trading simulation

5.6.1 Portfolio construction

The analysis carried out in this study involves portfolios of different sizes. Nonetheless, it is assumed that all pairs within the portfolio have uniform weights. This approach enables the computation of portfolio returns by taking the average performance of all pairs, without the need to consider the relative proportions of the initial investment.

In this step the central idea regards how to allocate capital for each pair. In principle, if the long and short positions of a pair hold the same value, the profits from the short position can fully cover the expenses of entering the long position, resulting in a self-financing portfolio.

However, in practice, this assumption is invalid due to the requirement of collateral when borrowing the security being shorted, making a zero initial investment impractical. Thus, the mandatory investment corresponds to the collateral, which it is assumed to be equivalent to the value of the security being shorted. This amount permits the investor to enter a long position by utilizing the earnings from short positions, essentially with the same value. This leveraging technique is typical in hedge funds to enhance the absolute return in the portfolio.

In order to simplify the computations, it is advantageous to consider a one dollar investment in each pair. This strategy is commonly employed by many researchers in the field (Caldeira and Moura (2013), Rad et al. (2016), Avellaneda and Lee (2010)). Unlike some studies that aim to attain dollar-neutrality, such as Gatev et al. (2006) and Dunis et al. (2010), which allocate \$1 to the long position and \$1 to the short position, the approach adopted in this research respects the cointegration ratio between the two securities [Sarmento and Horta (2020)].

As a consequence, the value invested in X should be times the value invested in Y. To ensure a \$1 initial investment, it is necessary to guarantee that neither the long position nor the short position costs more than \$1. Formally, the condition being imposed is that max(leg1, leg2) = \$1, where leg1 and leg2 represent the capital invested in each leg of the pair. Based on this premise, we establish a framework illustrated in Figure 5.5.

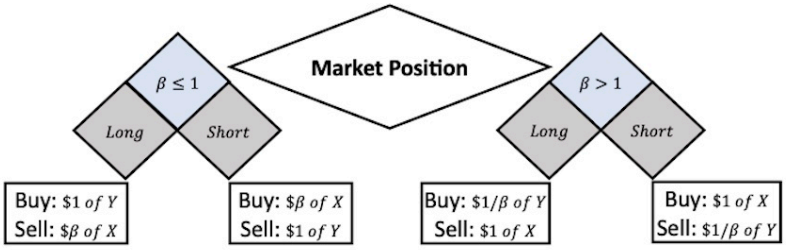


Figure 5.8: Market position definition.

Throughout the trading period, we assume that all the capital earned by a pair is reinvested in the subsequent trade [Sarmento and Horta (2020)]. For example, if a pair generates a 5% return during the first trade, the initial capital for the second trade will be \$1.05, rather than the original \$1. This approach simplifies the computation of the ultimate return.

5.6.2 Transaction costs

The transaction costs considered in this study are derived from the in-depth estimates of all the associated expenses provided by Do and Faff (2012). The costs considered can be categorized into three components, as outlined in the following Table.

	COMMISSION COSTS	MARKET IMPACT	RENTAL COSTS
DESCRIPTION	Fee charged when executing a transaction.	Indicator that reflects the cost that a transaction may cause due to the prevailing liquidity condition on the counter.	Loan fee for the short position.
CHARGE	8 bps	20 bps	1% per year

Table 5.5: Transaction costs considered.

The fee is expressed in relation to the position size. Commission and market impact costs must be adjusted to accommodate both assets in the pair. The costs associated with a single transaction are estimated by Do and Faff (2012) and computed as illustrated in Figure (5.9).

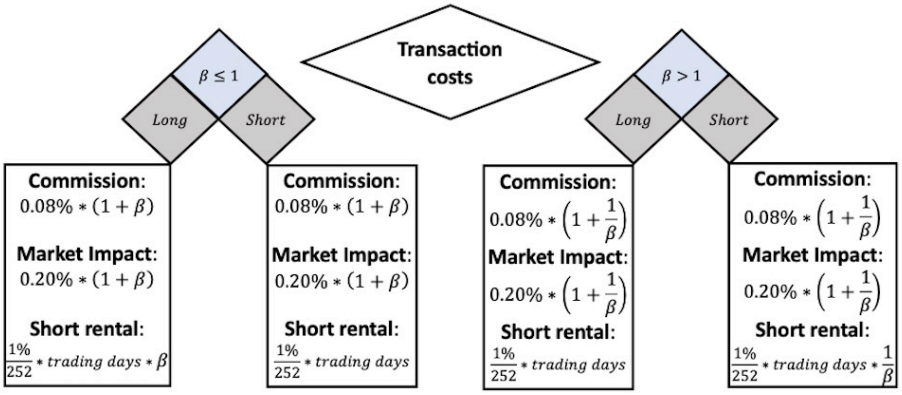


Figure 5.9: Calculation of transaction costs.

5.7 Evaluation metrics

The subsequent section provides a description of the financial evaluation metrics that are most appropriate for analyzing the proposed strategies. These metrics include the Return on Investment

(ROI), the Sharpe Ratio (SR), and the Maximum Drawdown (MDD).

5.7.1 Return on Investment

The Return on Investment (ROI) is computed by dividing the net profit by the initial investment, which is always $1 in this research. The portfolio returns are averaged over all the pairs selected for trading using the return on committed capital (RCC) approach, which accounts for the opportunity cost of committing capital to a strategy, even if it does not trade. The returns presented are based on a leveraged position, with the initial capital corresponding to the initial gross exposure in the unleveraged case, resulting in slightly reduced returns.

By definition the ROI is defined as:

$$ROI = \frac{Net\ Profit}{Initial\ Investment}\ x\ 100 \tag{5.15}$$

5.7.2 Sharpe Ratio

The Sharpe Ratio is a risk-adjusted metric for evaluating Return on Investment, which was first introduced by Nobel laureate William F. Sharpe [Sharpe (1994)]. It seeks to quantify the excess return per unit of risk in an investment. The annual Sharpe ratio is calculated using the following formula,

$$SR_{year} = \frac{R^{port} - R_f}{\sigma_{port}}\ x\ annualization\ factor. \tag{5.16}$$

where R^{port} is the expected daily portfolio returns, which can be computed as the mean value of the portfolio returns using,

$$R_t^{port} = \sum_{i=1}^{N} \omega_i R_t^i \tag{5.17}$$

where R_t^i and ω_i represent respectively the daily returns and the weight of the i-th pair in the portfolio of size N, according to the assumption of an equal weighted portfolio $\omega_i = \frac{1}{N}$. The risk-free rate, R_f, is the expected rate of return of a hypothetical investment with no financial loss risk. The interest that could be earned by investing the same amount of cash with no risk is subtracted from the returns generated by a given strategy. The risk-free rate is commonly set equal to the 3-month US government treasury bill rate. Table 5.6 shows the risk-free annualized rates considered for each testing period, which are obtained by averaging the 3-Month Treasury bill rate during the corresponding period [ris].

Period	R_f	Period	R_f
Jan 2006 - Dec 2006	0.0481	Jan 2014 - Dec 2014	0.00033
Jan 2007 - Dec 2007	0.0435	Jan 2015 - Dec 2015	0.00053
Jan 2008 - Dec 2008	0.0137	Jan 2016 - Dec 2016	0.0032
Jan 2009 - Dec 2009	0.0015	Jan 2017 - Dec 2017	0.0093
Jan 2010 - Dec 2010	0.0014	Jan 2018 - Dec 2018	0.0194
Jan 2011 - Dec 2011	0.0005	Jan 2019 - Dec 2019	0.0206
Jan 2012 - Dec 2012	0.0009	Jan 2020 - Dec 2020	0.0037
Jan 2013 - Dec 2013	0.0006	Jan 2021 - Dec 2021	0.0004

Table 5.6: Risk-free rates considered per test period.

To ensure coherence with (5.18), these values must be converted to the expected daily returns. The volatility of the portfolio, σ_{port}, is calculated using,

$$\sigma_{port} = \sqrt{\sum_{i=1}^{N}\sum_{j=1}^{N}\omega_i cov\left(i,j\right)\omega_j} = \sum_{i=1}^{N}\sum_{j=1}^{N}\omega_i\sigma_{i,j}\omega_j \tag{5.18}$$

that depends not only on the standard deviation of each security but also on the correlation among them.

The annualization factor enables the estimated daily Sharpe Ratio to be expressed in annual terms. This is a common practice and is consistent with the duration of the testing periods. In this approach, another metrics will also be considered, that is the serial correlation of the portfolio returns is measured, and a scale factor is applied accordingly [Lo (2002)]. The scale factors are described in appendix A.

One drawback of the Sharpe Ratio is that the volatility measurement penalizes all volatility the same, whether it is upside (large positive returns) or downside volatility. However, large losses and large gains are not equally undesirable. The decision to use this ratio is motivated by two factors [Sarmento and Horta (2020)]. First, the Sharpe Ratio is widely used in the Pairs Trading literature and is the preferred metric in similar research work [Gatev et al. (2006), Avellaneda and Lee (2010), Caldeira and Moura (2013)], which makes it easier to compare results. Second, the Sharpe Ratio is independent of leverage.

5.7.3 Maximum Drawdown

The Maximum Drawdown (MDD) measures the maximum decline in a time series observed from a peak before a new peak is reached. In this study, the Maximum Drawdown is calculated based on the account balance during the trading period [Sarmento and Horta (2020)]. Specifically, if $X(t)$, where $t \geq 0$, represents the total account balance, the MDD is computed using the following formula:

$$MDD\,(T) = \max_{\tau \in (0,T)} \left[\max_{t \in (0,\tau)} \frac{X\,(t) - X(\tau)}{X(t)} \right] \tag{5.19}$$

5.8 Software and Hardware

All analyses were conducted using Python. Specifically, data cleaning was performed using Pandas, a library for data manipulation and analysis. The econometric analysis of the time series was carried out using the Statsmodels, which provides classes and functions for both estimating many statistical models and performing statistical tests. The sci-kit learn library was employed for the implementation of PCA and the OPTICS algorithm. NumPy was used to compute mathematical functions. Finally, Matplotlib has been exploited to create plots. All simulations were conducted on a local environment using a standard CPU (Processor: 2 GHz Quad-Core Intel Core i5; Memory: 16 GB).

Chapter 6

Results

6.1 Dataset cleaning

Before starting the analysis, the datasets were analyzed and checked as explained in Section 5.2.

To recap, at the beginning the two datasets considered contained 1015 tickers and many doubled segments.

Table 6.1 showcases the results of processing the original first dataset. The first column shows the name of the 11 unique sectors, while the second column displays the number of stocks per sector not eliminated in the first stage of data handling.

Concerning the second dataset, the one containing stock prices, after having performed all the steps described in Section 5.2, that is dealing with missing data and outliers in the price series, the final number of unique tickers is 511, which is a great reduction from the original total of 1015.

Industry	Not eliminated
Consumer Discretionary	94
Industrials	94
Financials	93
Health Care	75
Information Technology	66
Energy	58
Materials	49
Consumer Staples	47
Utilities	45
Communication Services	40
Real Estate	38

Table 6.1: Results after data preprocessing.

6.2 Formation phase

6.2.1 PCA and OPTICS clustering for pair selection

This section outlines the proposed approach for assessing the framework for selecting pairs by showcasing the quantity of pairs chosen by the system through the OPTICS clustering technique. The outcomes are derived from utilizing five main components to depict the minimized data, as will be clarified subsequently.

6.2.2 Pairs selection rules

In this subsection, the aim is to delve further into the particulars of the pair selection process. The suggested selection approach encompasses four criteria that were previously illustrated upon (section to be added), illustrating the number of pairs filtered out by each condition. There are two main aspects we aim to examine in more detail. Firstly, we intend to investigate the proportion of pairs that passed the cointegration test. This result can serve as a proxy to estimate the actual impact of the multiple comparisons problem. Secondly, we are also interested in verifying the fraction of identified mean-reverting pairs that do not satisfy the two remaining imposed conditions to assess the importance of establishing those constraints. This is achieved by demonstrating the number of pairs that were filtered out by each condition.

	Formation period	2006-2009			2010-2013			2018-2021		
	N. of components	5 PC	10 PC	15 PC	5 PC	10 PC	15 PC	5 PC	10 PC	15 PC
Eliminated pairs per stage	Cointegration	362	134	35	313	185	153	1236	673	173
	Hurst exponent	0	0	0	0	0	0	0	0	0
	Half-life	5	4	3	4	4	5	0	0	0
	Mean-crosses	0	0	0	0	1	0	0	0	0
	Not eliminated	59	41	11	88	49	27	68	35	28
	Total	**426**	**179**	**49**	**426**	**239**	**185**	**1304**	**708**	**201**

Table 6.2: Pairs selection progress.

Table 6.1 displays the pairs that were removed at each stage of the selection process. It is important to note that the steps are executed sequentially, with each row representing the pairs eliminated from the subset resulting from the previous selection condition. Therefore, it is expected that most pairs will be eliminated initially, and this is indeed what we observe. Surprisingly, the Hurst exponent plays no role in eliminating pairs that passed the cointegration test but did not exhibit a Hurst exponent below 0.5 in their spread, indicating a lack of a mean-reverting process. The results presented also suggest that some pairs are ineligible because their convergence duration is incompatible with the trading period, and thus, they do not meet the half-life condition. However, it is clear that the mean-crossing criterion is satisfied for the entire subset of pairs that met the previous conditions, indicating that enforcing this final rule is unnecessary in this situation.

Number of principal components

The first step in applying the proposed approach for clustering stocks is to determine an appropriate number of principal components to reduce the original data dimension. This process must balance two desirable but incompatible features. On one hand, a greater number of principal components may result in more accurate representations since the transformed data explains a larger portion of the original variance. On the other hand, increasing the number of components can introduce more components that express random price fluctuations that should be disregarded. Additionally, it contributes to the curse of dimensionality, as explained in section 3.3.

Table 6.3: Clustering statistics when varying the number of principal components.

	2006-2009			2010-2013			2018-2021		
N. of principal components	5	10	15	5	10	15	5	10	15
Number of clusters	36	17	11	47	29	22	34	28	21
Average cluster size	4,91667E+15	4,76471E+15	3,45455E+16	4,31915E+15	4,27586E+15	3,89655E+15	6,11765E+15	5,25000E+00	4,33333E+15
Pairs to evaluate	426	179	49	405	239	185	1304	708	201
Pairs selected (unique tickers)	59 (72)	41 (50)	11 (19)	88 (100)	49 (57)	27 (37)	68 (73)	35 (46)	28 (36)
Total% variance explained	22,00%	24,00%	25,00%	12,00%	13,00%	13,00%	21,00%	24,00%	23,00%
% (N.) of pairs still cointegrated in the test set	40,67% (24)	60,97% (25)	54,54% (6)	14,77% (13)	14,28% (7)	3,7% (1)	23,52% (16)	25,71% (9)	17,85% (5)

As considered in Sarmento and Horta (2020) where the number of features were deemed acceptable up to 15 to protect from the curse of dimensionality, in this work the robustness of these findings was further explored by delving into the application of different principal components: out of the three periods considered, in two occasions it was established that the optimal number of principal components was 5, and only between 2006 and 2009 the optimal number was 10 since it improved both the explained variance and the robustness of cointegration relationships between pairs, two metrics added as a novelty in this research to compute the number of features. In order to make a comparison between the results that will be found, we will perform the trading strategies by considering both 5 and 10 components, whereas we will discard the 15 component-approach since it restricts too much the set of possible pairs without giving lots of advantages neither in terms of variance explained nor number of pairs cointegrated in the long term.

In conclusion and opposite to Sarmento and Horta (2020) findings, at least in this work "the impact of varying the number of principal components within the experimented interval is" not "very small."

Clusters composition (2006-2008, 5PC)

In this section, the objective is to validate the clusters formed by the OPTICS algorithm and gain insight into their composition. Detailing the process for all the periods considered would be very lengthy, thus it will be displayed in this section just the January 2006 to December 2008 period, while the remaining periods will be included in appendix B.

To visually represent the clustering results, it would be desirable to generate a depiction of the formed clusters in a two-dimensional space. However, the data is described in ten dimensions. To address this issue, the T-distributed Stochastic Neighbor Embedding algorithm (t-SNE) was used. T-SNE is a nonlinear technique that reduces the dimensionality of data and is well-suited for embedding high-dimensional data in a low-dimensional space of two (or three) dimensions for

visualization purposes. The algorithm models each high-dimensional object using two (or three) dimensions in such a manner that similar objects are represented by nearby points and dissimilar objects are represented by distant points with high probability. Every cluster is depicted using a range of colors. The assets that are not grouped are represented by smaller circles in a shade of grey. The circles were not labeled to enhance the clarity of the visualization.

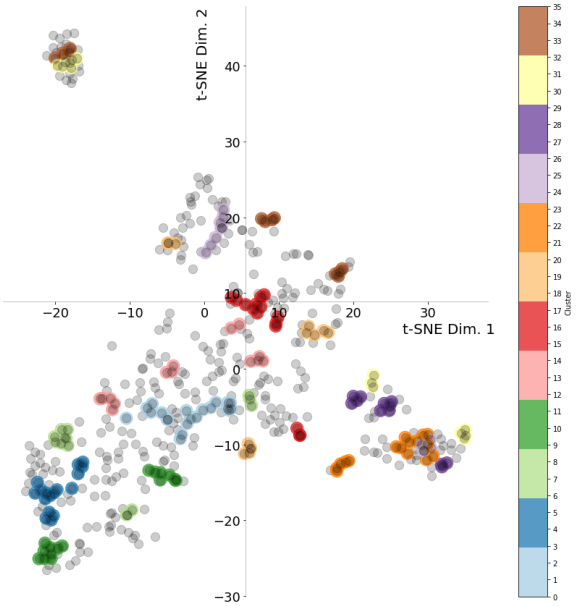
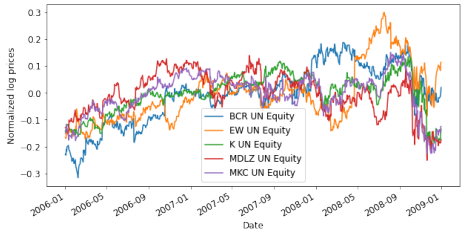
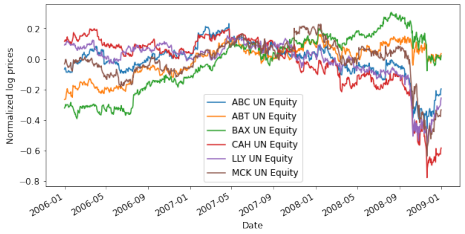


Figure 6.1: Application of t-SNE to the clusters generated by OPTICS with 5PC during Jan 2006 to Dec 2008.

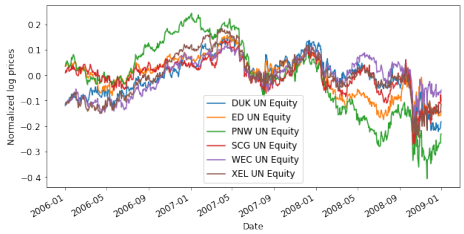
To gain a deeper understanding of the securities that comprise each cluster, the corresponding price series are examined.



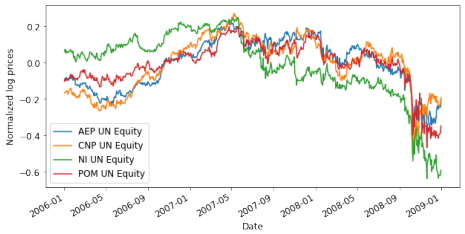
(a) Normalized prices of stocks grouped in Cluster 0.



(b) Normalized prices of stocks grouped in Cluster 1.



(c) Normalized prices of stocks grouped in Cluster 2.



(d) Normalized prices of stocks grouped in Cluster 3.

Figure 6.2: Price series composition of first four clusters formed from Jan 2006 to Dec 2009.

Figure 6.2 displays the logarithm of the price series for each stock. The price series shown are obtained by subtracting the mean of the original price series to aid visualization. Starting with Figure 6.2(a), the first observation is that the OPTICS clustering capabilities go beyond selecting stocks within the same segment. In cluster 0, we observe stocks from the Health Care (BCR, EW), and the Consumer Staples sector (K, MDLZ, and MKC). Even though they do not all belong to the same category, there is a visible connection among the identified price series. Figure 6.2(b) on the other hand demonstrates that the OPTICS-based approach can also effectively group securities into categories (ABC, ABT, BAX, CAH, LLY, and MCK from Health Care). The same goes for figure 6.2(c) and figure 6.2(d), where all stocks belong to Utilities.

In conclusion, we can affirm that the formed clusters have achieved their intended purpose of combining the desired attributes, specifically the ability to group subsets of assets from the same sector while avoiding the formation of clusters that solely contain assets from different categories.

6.3 Trading phase

We will begin by disclosing the outcomes of the threshold-based trading model proposed in subsection 5.4.1, during the validation periods specified in Table 6.4. The analysis using 5 principal components is presented in this section with the corresponding results of the strategies, while the information for the 10 principal components analysis is included in appendix C. It is important to exercise caution while analyzing the results presented in this table, owing to an implicit bias that arises from evaluating the strategy during the same period used to identify the pairs being tested. Consequently, the results may not accurately represent the actual performance of the model. On one hand, the results may seem more satisfactory, given the extra confidence in the robustness of the pairs being used. On the other hand, the results may decline, as there is no elimination of pairs that have not been profitable in the past, due to the lack of available records. Nevertheless, the validation records are useful in supporting the construction of test portfolios, which use these results as a heuristic for selecting pairs [Sarmento and Horta (2020)].

Validation Period	2008	2012	2020
SR(Scale Factors)	6.83(6.80)	8.01(7.97)	4.78(4.76)
ROI	50.01%	22.88%	42.42%
MDD	0.92%	0.47%	1.47%
N. of pairs	59	88	68
% of profitable pairs	94.91%	94.31%	95.58%
N. trades	366	406	371
positive	308	342	309
negative	58	64	62

Table 6.4: Validation results for threshold method.

Table 6.4 indicates that the strategy is profitable in every scenario, with a satisfactory margin. This implies the success of the pairs selection process in selecting profitable pairs up to this point. What is important to notice is the different results obtained using different components in the dimensionality reduction step; this led to the conclusion that, since the choice of these features is arbitrary, one should be careful and develop a thorough analysis on how this could impact profitability before taking investment decisions.

Below it is represented Table 6.5 which displays the results for the test periods of 1 year, that is the trading phase. For each test period, it is shown the outcomes obtained when using any of the three test portfolios implemented, as explained in Section 5.5. To recap, portfolio number 1 comprises the entire set of pairs identified during the formation period, portfolio number 2 comprises only the pairs that showed satisfactory performance during the validation period, while portfolio

number 3 comprises the pairs that are cointegrated both during the validation and the test period. To present the information more concisely, the average over all years and portfolios is included in the rightmost column.

Test Period	2009			2013			2021			AVG
Test Portfolio	1	2	3	1	2	3	1	2	3	-
SR(Scale Factors)	5.07(5.58)	4.93(5.42)	3.48(4.24)	8.11(8.08)	8.15(8.11)	5.29(5.27)	7.12(6.42)	7.22(6.51)	3.80(3.79)	**5,90(5,93)**
ROI	19,10%	19,84%	24,79%	7,85%	8,24%	14,12%	11,17%	10,39%	13,61%	**14,35%**
MDD	0,85%	0,88%	2,50%	0,18%	0,20%	0,51%	0,29%	0,27%	1,17%	**0,76%**
N. of pairs	59	56	24	88	83	13	68	65	16	**52**
% of profitable pairs	83,05%	83,92%	95,83%	72,72%	74,69%	100,00%	82,35%	81,53%	93,75%	**85,32%**
N. trades	73	69	34	100	95	24	86	83	31	**66**
positive	67	64	33	84	81	21	81	78	30	**60**
negative	6	5	1	16	14	3	5	5	1	**6**

Table 6.5: Test results for threshold method.

It is verified the overall profitability as well as a steady decrease, in fact the SR across all tested environments has slightly lowered by 9%, whereas the ROI has decreased by a fair amount, from an average of 25% to 14%. However, this outcome provides additional support to the notion that the procedure for selecting pairs is robust. It is noteworthy that the resulting highest percentage of profitable pairs occurred in all the periods using portfolio 3. This highlights the reliability of the cointegration in detecting profitable pairs. On the other hand, concerning the maximum drawdowns, the findings suggest that using the cointegrated pairs from the validation period (portfolio 3) in these strategies leads to a significant increase in the maximum drawdown amplitude of the portfolio. This is particularly noticeable in 2009 and 2021 when the maximum drawdown has almost increased by an average of 250%.

Finally, the method used to select pairs, namely the OPTICS algorithm, has a tendency to group pairs that appear more stable [Sarmento and Horta (2020)], however this is not completely verified here since the system has identified pairs that showed great performance during validation but then experienced a decline during the test period. This observation is supported by the analysis of the downward progression of the profitability indicators from the validation to the test results, particularly considering ROI. However, it exhibits good consistency regarding the percentage of profitable pairs in the portfolio even with a decline of 10%, averaging at almost 95% profitable pairs in the validation, and at 85% in the test period.

These results lead to the research stage number 1, in which the goal is to use a different trading strategy with the purpose of enhancing the standard model by means of stochastic approach.

6.4 Stochastic-based trading model performance

This section is dedicated to examining the outcomes that enable us to respond to the first research inquiry: "Can a stochastic-based trading model achieve more robust performance than the threshold-based one? ". To achieve this, we assess the efficiency of the stochastic-based model, which was introduced in Chapter 4, against the existing threshold-based model.

The primary objective of implementing a pairs trading approach based on stochastic methods is to enhance the resilience of the portfolio by minimizing the portfolio decline in value during trading phase. To assess the efficacy of the proposed method, the selected pairs as well as the analysis periods are the same as the ones employed in the standard procedure.

6.4.1 Constant parameters

The outcomes of using constant parameters are tabulated in Table 6.6 and Table 6.7 while the results obtained from using the standard threshold-based model were displayed in Table 6.4 and Table 6.5. The results from the validation set are presented just for comprehensiveness, and the primary goal is to identify the pairs capable of producing profits during the validation period instead of evaluating the strategy itself.

Validation Period	2008	2012	2020
SR(Scale Factors)	1.74(1.91)	6.73(6.70)	4.20(4.19)
ROI	45.73%	24.80%	43.95%
MDD	2.41%	1.07%	2.18%
N. of pairs	59	88	68
% of profitable pairs	91.52%	94.31%	89.70%
N. trades	759	853	893
positive	563	617	637
negative	196	236	256

Table 6.6: Validation results for stochastic method (constant).

Test Period	2009			2013			2021			AVG
Test Portfolio	1	2	3	1	2	3	1	2	3	-
SR(Scale Factors)	5,33(5,31)	5,06(5,04)	4,67(5,14)	5,41(5,38)	5,35(5,33)	4,09(3,69)	5,48(5,45)	5,66(5,64)	3,28(3,27)	**4,92(4,91)**
ROI	53,64%	57,57%	60,24%	19,30%	20,29%	28,58%	24,03%	24,58%	22,38%	**34,51%**
MDD	2,02%	2,55%	6,09%	0,73%	0,70%	1,49%	1,14%	1,13%	2,14%	**2,00%**
N. of pairs	59	54	24	88	83	13	68	61	16	**52**
% of profitable pairs	94,91%	98,14%	100,00%	93,18%	93,97%	100,00%	89,70%	91,80%	81,25%	**93,66%**
N. trades	517	453	285	636	582	157	639	593	203	**452**
positive	375	337	207	454	425	107	445	416	139	**323**
negative	142	116	78	182	157	50	194	177	64	**129**

Table 6.7: Test results for stochastic method (constant).

Regarding the test outcomes that replicate a pragmatic trading setting, it is apparent that the portfolio profitability is significantly increased when utilizing the stochastic-based models, as opposed to the standard trading model. This is clearly detectable in the ROI and the percentage of profitable pairs, with a rise of 140% and almost 10% respectively. Nonetheless, this improvement comes at a cost of a slightly poorer Sharpe ratio and a more marked drawdown performance, this last almost triplicated.

6.4.2 Moving parameters

This section regards the examination of the outcomes that allow to respond to the second research inquiry: "How can the changing model parameters impact the profitability?". To achieve this, two different testing windows are considered as introduced in Chapter 5. Below are displayed results for the validation period using moving parameters.

Validation Period	2008		2012		2020	
Rolling window (n. days)	6m (126)	1y (252)	6m (126)	1y (252)	6m (126)	1y (252)
SR(Scale Factors)	4.73(5.20)	1.75(1.74)	6.15(6.13)	6.69(6.67)	3.63(3.99)	4.11(4.09)
ROI	48.14%	46.71%	22.11%	24.60%	38.85%	43.13%
MDD	2.99%	2.30%	1.05%	0.85%	3.16%	1.88%
N. of pairs	59	59	88	88	68	68
% of profitable pairs	88.13%	93.22%	93.18%	94.31%	91.17%	88.23%
N. trades	715	790	866	872	819	902
positive	519	576	622	617	574	643
negative	196	214	244	255	245	259

Table 6.8: Validation results for stochastic method (moving parameters).

Both trading results are presented in Table 6.9 and 6.10.

Test Period	2009			2013			2021			AVG
Test Portfolio	1	2	3	1	2	3	1	2	3	-
SR(Scale Factors)	4,36(4,34)	4,27(4,25)	3,92(3,90)	5,29(5,27)	5,29(5,27)	3,45(3,44)	6,25(6,22)	5,56(5,54)	3,76(3,75)	**4,68(4,66)**
ROI	49,94%	53,95%	54,85%	19,57%	20,85%	25,37%	27,57%	25,38%	26,14%	**33,74%**
MDD	5,16%	5,50%	9,05%	1,21%	1,29%	3,79%	1,38%	1,80%	1,57%	**3,42%**
N. of pairs	59	52	24	88	82	13	68	62	16	**52**
% of profitable pairs	93,22%	96,15%	100,00%	92,04%	95,12%	92,30%	94,11%	93,54%	93,75%	**94,47%**
N. trades	459	382	252	612	586	148	667	603	212	**436**
positive	327	279	179	433	418	98	449	399	139	**302**
negative	132	103	73	179	168	50	218	204	73	**133**

Table 6.9: Test results for stochastic method (126-day-window).

Test Period	2009			2013			2021			AVG
Test Portfolio	1	2	3	1	2	3	1	2	3	-
SR(Scale Factors)	5,32(5,30)	5,03(5,01)	4,53(4,51)	5,68(5,66)	5,49(5,47)	3,92(3,54)	5,63(5,61)	5,54(5,52)	3,09(3,08)	**4,91(4,85)**
ROI	55,58%	58,28%	59,22%	20,25%	20,47%	27,95%	24,84%	24,60%	21,61%	**34,76%**
MDD	2,11%	2,77%	6,63%	0,75%	0,85%	1,90%	1,13%	1,17%	2,18%	**2,17%**
N. of pairs	59	55	24	88	83	13	68	60	16	**52**
% of profitable pairs	96,61%	100,00%	100,00%	92,04%	92,77%	100,00%	92,64%	93,33%	87,50%	**94,99%**
N. trades	539	490	304	626	587	149	684	622	233	**470**
positive	389	358	222	458	433	106	475	433	157	**337**
negative	150	132	82	168	154	43	209	189	76	**134**

Table 6.10: Test results for stochastic method (252-day-window).

6-month window (126 days)

Results derived from 126-day-window show how this strategy doesn't give many improvements in terms of profitability, in fact Sharpe ratio and return on investments are similar to the previously found ones. On the other hand, as suspected in section 5, since the thresholds were wider using this moving window, the total time needed to complete a trade would be longer, resulting in a lower number of trades; at the same time the maximum drawdown is also increased at 3.42% starting from the 2% detected using constant parameters.

1 year window (252 days)

Finally, a moving window corresponding to a trading year of 252 days has been used. Results are more promising than the previous ones since Sharpe ratio, ROI and maximum drawdown have all been improved. Using this setting of parameters also enlarges the number of trades occurring in a year, while at the same time it gives no remarkable increases in terms of performance than the model with constant parameters. This could be due to the fact that the model parameters might have already been well-estimated in the first stage, or that it should be tested long windows, though resulting in a longer trading phase (more than a year), which is seldom employed in pairs trading.

Chapter 7

Conclusions and future improvements

7.1 Conclusions

Let's begin by revisiting the first part of this work, where it was considered the selection method provided by Sarmento and Horta (2020) using Unsupervised Learning to discover more promising pairs. The results give us interesting outcomes in terms of profitable pairs, confirming the robustness of the selection strategy. In order to determine the appropriate choice in implementing dimensionality reduction, various principal components were tested, leading to two conclusions: firstly, in alignment with previous research, the optimal number of features was found to be 5, as it resulted in the highest profitable indicators across all the employed strategies. Secondly, based on this analysis, the curse of dimensionality could be observed even at 10 instead of 15, as the profitability is reduced when using 10 components compared to 5. As a result, the OPTICS clustering algorithm has proven to be powerful, even if affected by the arbitrary choice of feature during the dimensionality reduction step. This strategy can generate high average portfolio Sharpe ratios across all the strategy employed, demonstrating greater consistency concerning the proportion of profitable pairs in the portfolio, with an average of more than 90% profitable pairs. Finally, it also achieves stable portfolio drawdowns, as it maintains the maximum drawdown values within an acceptable range most of the time. Therefore, we may conclude that Unsupervised Learning has the potential to identify more promising pairs, not necessarily belonging to the same sector.

Turning to the first research question that this study aims to address: "Can a

stochastic-based trading model achieve more robust performance than the threshold-based one?". The results indicate that, with respect to profitability performance, the stochastic approach has been successful in generating higher returns, particularly in terms of returns on investments. In some cases, the returns have even doubled, and a significant percentage of profitable pairs have been identified during both the validation and test periods. However, this comes at the cost of a higher drawdown in the portfolio, indicating that this strategy may not be suitable for a conservative investor.

Regarding the second question addressed in this study, namely "How can the changing model parameters impact profitability?", the results indicate that this variation does not lead to a significant improvement in performance compared to using constant parameters. However, it does not result in a deterioration of performance either.

On the other hand, if we consider Zeng and Lee (2014), where the investigation of "how the changing model parameters impact on the optimal thresholds" was suggested, we can conclude that, apart from the obvious variation due to the fact that thresholds are a function of the model parameters, the use of moving parameters has a real effect on the optimal threshold and the strategy's outcome. Specifically, when using a window of 6 months, the threshold range becomes larger than when using constant parameters, resulting in a lower number of trades and a higher portfolio drawdown.

Lastly the strategy has proven to be more profitable than the standard one and also robust, since it was applied to three periods of market turmoil (financial crisis, sovereign debt crisis, and COVID-19), showing high performance returns both in validation and test period, nonetheless.

7.2 Future improvements

There are several potential avenues for future research that could be pursued as a continuation of this study. One possible direction is to delve deeper into the proposed stochastic-based trading model, while another option is to devise alternative schemes for a Pairs Trading strategy:

1. Improving stochastic model:

    - In order to limit the maximum drawdown, add a stop-loss system.

- Model parameters are a function of the transaction cost, therefore try another type of model, perhaps free of these constraints.

- Try longer windows to compute moving parameters; note that in this case also the test period should be modified accordingly.

2. Improving pairs trading strategy:

   - Consider the survivorship bias in selecting stocks.

   - Optimize the choice of principal components in the selection phase.

   - Test portfolio 3 only utilizes cointegration as a proxy for selecting a suitable class to invest in during the trading phase. Since the pairs were selected through a multi-step approach, it may be beneficial to incorporate additional parameters such as Hurst exponent and minimum half-life in order to construct a test portfolio.

Appendix A

Sharpe Ratio Scale Factors

This supplementary section presents a table that outlines the scale factor correction technique introduced by Lo (2002), which has been utilized in the present study. The correction factor is reliant on two variables: the autocorrelation value of the returns, denoted as ρ, and the aggregation value, represented as q. For the purpose of this research, q has been set at 250, since it involves the aggregation of daily data for a period of one year. Moreover, the parameter ρ is assessed for the returns of each portfolio.

ρ (%)	Aggregation Value, q									
	2	3	4	6	12	24	36	48	125	250
90	1.03	1.05	1.07	1.10	1.21	1.41	1.60	1.77	2.67	3.70
80	1.05	1.10	1.14	1.21	1.43	1.81	2.14	2.42	3.79	5.32
70	1.08	1.15	1.21	1.33	1.65	2.19	2.62	3.00	4.75	6.68
60	1.12	1.21	1.30	1.46	1.89	2.55	3.08	3.53	5.63	7.94
50	1.15	1.28	1.39	1.60	2.12	2.91	3.53	4.06	6.49	9.15
40	1.20	1.35	1.49	1.75	2.36	3.27	3.98	4.58	7.35	10.37
30	1.24	1.43	1.60	1.91	2.61	3.65	4.44	5.12	8.23	11.62
20	1.29	1.52	1.73	2.07	2.88	4.04	4.93	5.68	9.14	12.92
10	1.35	1.62	1.86	2.25	3.16	4.45	5.44	6.28	10.12	14.31
0	**1.41**	**1.73**	**2.00**	**2.45**	**3.46**	**4.90**	**6.00**	**6.93**	**11.18**	**15.81**
–10	1.49	1.85	2.16	2.66	3.80	5.39	6.61	7.64	12.35	17.47
–20	1.58	1.99	2.33	2.90	4.17	5.95	7.31	8.45	13.67	19.35
–30	1.69	2.13	2.53	3.17	4.60	6.59	8.10	9.38	15.20	21.52
–40	1.83	2.29	2.75	3.48	5.09	7.34	9.05	10.48	17.01	24.11
–50	2.00	2.45	3.02	3.84	5.69	8.26	10.21	11.84	19.26	27.31
–60	2.24	2.61	3.37	4.30	6.44	9.44	11.70	13.59	22.19	31.50
–70	2.58	2.76	3.86	4.92	7.45	11.05	13.77	16.04	26.33	37.43
–80	3.16	2.89	4.66	5.91	8.96	13.50	16.98	19.88	32.96	47.02
–90	4.47	2.97	6.47	8.09	12.06	18.29	23.32	27.61	46.99	67.65

Figure A.1: Scale factors for time-aggregated Sharpe ratios when returns follow an AR(1) process. Source:[Lo (2002)]

Appendix B

t-SNE Visualization

This supplementary section presents a visual representation of the utilization of t-SNE on the clusters formed during the other two distinct periods not showed previously in the work: January 2010 - December 2012 and January 2018 - December 2020. The subsequent analysis in subsection 6.2.1 pertains to the period spanning January 2006 to December 2008.

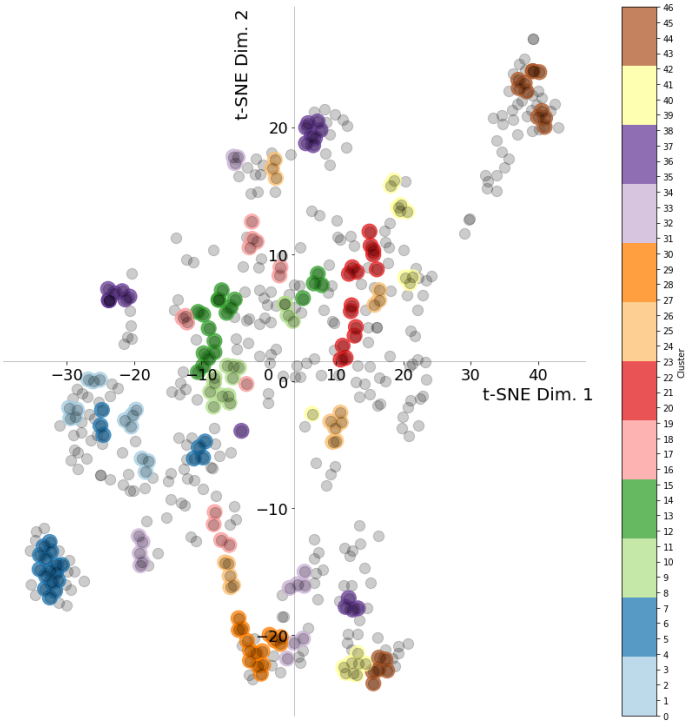


Figure B.1: OPTICS clustering in Jan 2010-Dec 2012 formation period.

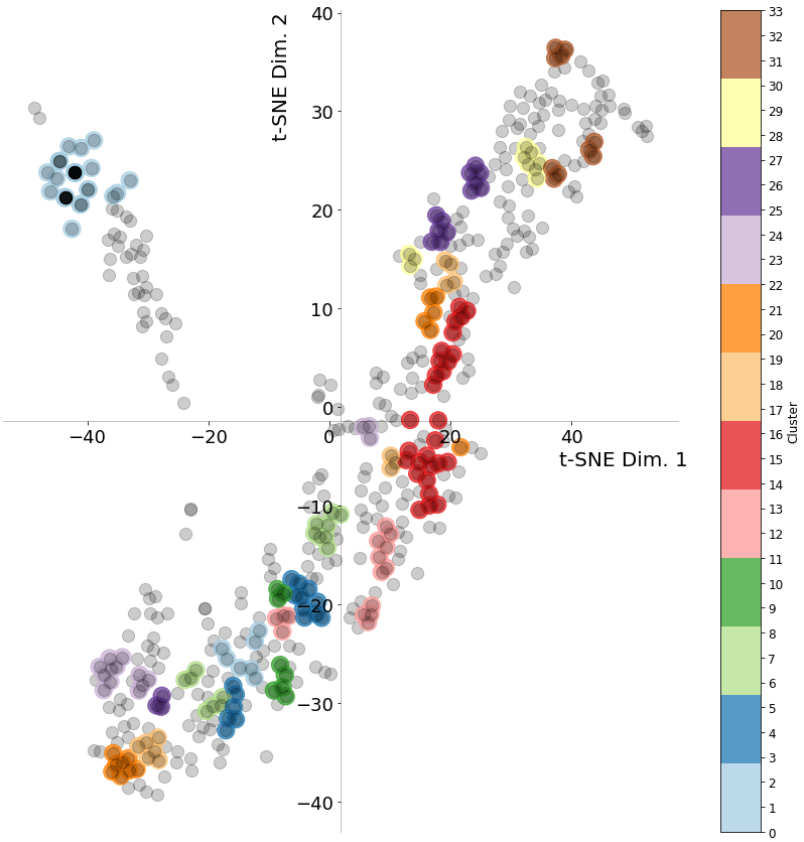


Figure B.2: OPTICS clustering in Jan 2018-Dec 2020 formation period.

Appendix C

Strategy's trading results using 10 principal components

C.1 Threshold-based model results with 10 PC

Validation Period	2008	2012	2020
SR(Scale Factors)	3.19(3.51)	6.69(6.66)	5.75(5.18)
ROI	39.20%	18.54%	28.72%
MDD	2.86%	0.60%	1.36%
N. of pairs	41	49	35
% of profitable pairs	90.24%	95.91%	94.28%
N. trades	243	261	174
positive	202	218	139
negative	41	43	35

Table C.1: Validation results for threshold method (10 PC).

Test Period	2009			2013			2021			AVG
Test Portfolio	1	2	3	1	2	3	1	2	3	-
SR(Scale Factors)	3,64(3,62)	3,38(3,71)	2,40(2,39)	4,62 (4,60)	4,64(4,62)	3,07(2,50)	4,81(4,79)	4,64(4,62)	2,67(2,66)	**3,76(3,72)**
ROI	25,93%	23,84%	19,87%	4,80%	5,04%	9,41%	7,82%	7,95%	5,58%	**12,25%**
MDD	1,91%	1,91%	3,67%	0,28%	0,29%	0,67%	0,36%	0,37%	0,71%	**1,13%**
N. of pairs	41	37	25	49	47	7	35	33	9	**31**
% of profitable pairs	82,92%	83,78%	80,00%	55,10%	57,44%	71,42%	85,71%	87,87%	100,00%	**78,25%**
N. trades	55	49	32	46	45	9	45	43	13	**37**
positive	52	47	30	36	36	8	39	37	11	**33**
negative	3	2	2	10	9	1	6	6	2	**5**

Table C.2: Test results for threshold method (10 PC).

C.2 Stochastic-based model results with 10 PC

C.2.1 Constant parameters

Validation Period	2008	2012	2020
SR(Scale Factors)	3.66(3.65)	5.89(5.31)	4.31(3.88)
ROI	58.17%	20.07%	28.20%
MDD	4.93%	0.70%	2.35%
N. of pairs	41	49	35
% of profitable pairs	90.24%	93.87%	91.42%
N. trades	570	533	546
positive	436	391	366
negative	134	142	180

Table C.3: Validation results for stochastic method (constant parameters and 10 PC)

Test Period	2009			2013			2021			AVG
Test Portfolio	1	2	3	1	2	3	1	2	3	-
SR(Scale Factors)	3,78(4,16)	3,47(3,81)	3,24(3,57)	5,31(5,29)	5,22(5,20)	2,23(2,22)	3,90(3,88)	3,80(3,79)	2,25(2,24)	**3,68(3,79)**
ROI	54,06%	49,91%	43,90%	16,70%	17,06%	16,88%	13,16%	14,13%	9,52%	**26,15%**
MDD	4,18%	5,09%	4,90%	0,87%	0,96%	5,95%	1,17%	1,18%	2,44%	**2,97%**
N. of pairs	41	37	25	49	46	7	35	32	9	**31**
% of profitable pairs	90,24%	94,59%	92,00%	93,87%	93,47%	100,00%	82,85%	84,37%	88,88%	**91,14%**
N. trades	411	365	264	320	308	56	271	256	67	**258**
positive	293	266	188	266	218	39	176	166	43	**184**
negative	118	99	76	94	90	17	95	90	24	**78**

Table C.4: Test results for stochastic method (constant parameters and 10 PC).

C.2.2 Moving parameters

Validation Period	2008		2012		2020	
Rolling window (n. days)	6m (126)	1y (252)	6m (126)	1y (252)	6m (126)	1y (252)
SR(Scale Factors)	3.09(3.40)	3.83(4.21)	4.76(4.74)	5.55(5.01)	4.01(3.99)	4.50(4.06)
ROI	48.23%	60.15%	15.30%	18.96%	29.99%	28.13%
MDD	4.10%	4.85%	0.93%	0.87%	2.08%	1.84%
N. of pairs	41	41	49	49	35	35
% of profitable pairs	85.36%	85.36%	85.71%	91.83%	88.57%	91.42%
N. trades	529	560	471	541	487	517
positive	386	435	338	379	327	343
negative	143	125	133	162	160	174

Table C.5: Validation results for stochastic method (moving parameters and 10 PC).

126-day-window

Test Period	2009			2013			2021			AVG
Test Portfolio	1	2	3	1	2	3	1	2	3	-
SR(Scale Factors)	3,29(3,62)	3,15(3,46)	2,76(3,04)	6,40(6,37)	6,61(6,59)	3,19(3,18)	4,62(4,61)	4,32(4,31)	2,36(2,60)	**4,07(4,19)**
ROI	47,49%	54,06%	39,06%	20,90%	21,95%	24,15%	18,06%	18,69%	10,31%	**28,30%**
MDD	6,17%	7,12%	6,06%	0,75%	0,66%	3,12%	1,96%	2,38%	2,48%	**3,41%**
N. of pairs	41	35	25	49	42	7	35	31	9	**30**
% of profitable pairs	82,92%	85,71%	84,00%	89,79%	90,47%	100,00%	91,42%	90,32%	88,88%	**89,28%**
N. trades	424	382	262	363	334	76	246	218	65	**263**
positive	295	270	186	259	243	57	158	139	36	**183**
negative	129	112	76	104	91	19	88	79	29	**81**

Table C.6: Test results for stochastic method (moving parameters, 126-day-window and 10 PC).

252-day-window

Test Period	2009			2013			2021			AVG
Test Portfolio	1	2	3	1	2	3	1	2	3	-
SR(Scale Factors)	3,67(3,66)	3,44(3,79)	2,97(3,26)	6,16(6,14)	5,89(5,87)	3,23(3,22)	3,32(3,31)	3,18(3,17)	1,84(1,84)	**3,74(3,80)**
ROI	50,49%	50,61%	40,19%	19,12%	18,77%	24,70%	11,86%	12,37%	8,11%	**26,25%**
MDD	3,95%	4,82%	4,91%	0,83%	0,85%	4,44%	1,39%	1,40%	2,52%	**2,79%**
N. of pairs	41	35	25	49	45	7	35	32	9	**31**
% of profitable pairs	90,24%	94,28%	88,00%	87,75%	88,88%	85,71%	80,00%	78,12%	88,88%	**86,87%**
N. trades	432	362	270	329	314	70	273	250	65	**263**
positive	301	264	194	243	230	50	180	163	39	**185**
negative	131	98	76	86	84	20	93	87	26	**78**

Table C.7: Test results for stochastic method (moving parameters, 252-day-window and 10 PC).

Bibliography

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

Evan Gatev, William N Goetzmann, and K Geert Rouwenhorst. Pairs trading: Performance of a relative-value arbitrage rule. *The Review of Financial Studies*, 19(3):797–827, 2006.

Andrew Pole. *Statistical arbitrage: algorithmic trading insights and techniques*. John Wiley & Sons, 2011.

Christopher Krauss. Statistical arbitrage pairs trading strategies: Review and outlook. *Journal of Economic Surveys*, 31(2):513–545, 2017.

Mark Cummins and Andrea Bucca. Quantitative spread trading on crude oil and refined products markets. *Quantitative Finance*, 12(12):1857–1875, 2012.

Zhengqin Zeng and Chi-Guhn Lee. Pairs trading: optimal thresholds and profitability. *Quantitative Finance*, 14(11):1881–1893, 2014.

Gangadharrao S Maddala and In-Moo Kim. Unit roots, cointegration, and structural change. 1998.

Helmut Lütkepohl. *New introduction to multiple time series analysis*. Springer Science & Business Media, 2005.

David A Dickey and Wayne A Fuller. Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American statistical association*, 74(366a):427–431, 1979.

Harold Edwin Hurst. Long-term storage capacity of reservoirs. *Transactions of the American society of civil engineers*, 116(1):770–799, 1951.

Ernest P. Chan. *Algorithmic trading: winning strategies and their rationale*. Wiley trading series. Wiley, Hoboken, New Jersey, 2013. ISBN 9781118460146.

Clive WJ Granger. *Co-integrated variables and error-correcting models*. PhD thesis, Discussion Paper 83-13. Department of Economics, University of California at ..., 1983.

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.

Hossein Rad, Rand Kwong Yew Low, and Robert Faff. The profitability of pairs trading strategies: distance, cointegration and copula methods. *Quantitative Finance*, 16(10):1541–1558, 2016.

Nicolas Huck and Komivi Afawubo. Pairs trading and selection methods: is cointegration superior? *Applied Economics*, 47(6):599–613, 2015.

Marco Avellaneda and Jeong-Hyun Lee. Statistical arbitrage in the us equities market. *Quantitative Finance*, 10(7):761–782, 2010.

Robert J Elliott, John Van Der Hoek*, and William P Malcolm. Pairs trading. *Quantitative Finance*, 5(3):271–276, 2005.

William K Bertram. Optimal trading strategies for itô diffusion processes. *Physica A: Statistical Mechanics and its Applications*, 388(14):2865–2873, 2009.

Binh Do, Robert Faff, and Kais Hamza. A new approach to modeling and estimation for pairs trading. In *Proceedings of 2006 financial management association European conference*, volume 1, pages 87–99, 2006.

Ian Jolliffe. Principal component analysis (pp. 1094-1096). *Springer Berlin Heidelberg. RESUME SELİN DEĞİRMECİ Marmara University, Goztepe Campus ProQuest Number: ProQuest). Copyright of the Dissertation is held by the Author. All Rights Reserved*, 28243034:28243034, 2011.

Richard Bellman. Dynamic programming, system identification, and suboptimization. *SIAM Journal on Control*, 4(1):1–5, 1966.

Pavel Berkhin. A survey of clustering data mining techniques. *Grouping multidimensional data: Recent advances in clustering*, pages 25–71, 2006.

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.

Jon Scott Armstrong. *Principles of forecasting: a handbook for researchers and practitioners*, volume 30. Springer, 2001.

José Pedro Ramos-Requena, JE Trinidad-Segovia, and MA Sánchez-Granero. Introducing hurst exponent in pair trading. *Physica A: statistical mechanics and its applications*, 488:39–45, 2017.

Enrico Bibbona, Gianna Panfilo, and Patrizia Tavella. The ornstein–uhlenbeck process as a model of a low pass filtered white noise. *Metrologia*, 45(6):S117, 2008.

John Stachel, Diana Kormos Buchwald, and Peter Gabriel Bergmann. *The collected papers of Albert Einstein*, volume 1. Princeton University Press Princeton, 1987.

Ludwig Arnold. Stochastic differential equations. *New York*, 2, 1974.

Ming Chen Wang and George Eugene Uhlenbeck. On the theory of the brownian motion ii. *Reviews of modern physics*, 17(2-3):323, 1945.

Binh Do and Robert Faff. Does simple pairs trading still work? *Financial Analysts Journal*, 66(4): 83–95, 2010.

William K Bertram. Analytic solutions for optimal statistical arbitrage trading. *Physica A: Statistical mechanics and its applications*, 389(11):2234–2243, 2010.

Donald A Darling and AJF58908 Siegert. The first passage problem for a continuous markov process. *The Annals of Mathematical Statistics*, pages 624–639, 1953.

Yaozhong Hu and Hongwei Long. Parameter estimation for ornstein-uhlenbeck processes driven by α-stable lévy motions. *Communications on Stochastic Analysis*, 1(2):1, 2007.

João Caldeira and Guilherme V Moura. Selection of a portfolio of pairs based on cointegration: A statistical arbitrage strategy. *Available at SSRN 2196391*, 2013.

Christian L Dunis, Gianluigi Giorgioni, Jason Laws, and Jozef Rudy. Statistical arbitrage and high-frequency data with an application to eurostoxx 50 equities. *Liverpool Business School, Working paper*, 2010.

Binh Do and Robert Faff. Are pairs trading profits robust to trading costs? *Journal of Financial Research*, 35(2):261–287, 2012.

William F. Sharpe. The sharpe ratio. *The Journal of Portfolio Management*, 21(1):49–58, 1994.

3-Month Treasury Bill: Secondary Market Rate. https://fred.stlouisfed.org/series/TB3MS/.

Andrew W Lo. The statistics of sharpe ratios. *Financial analysts journal*, 58(4):36–52, 2002.

List of Figures

List of Tables