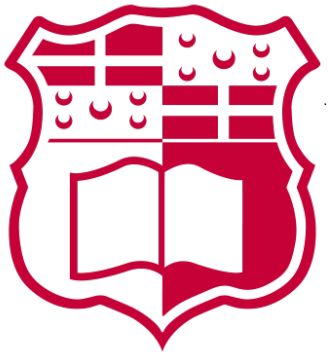

University of Malta

Statistical Arbitrage in Commodity Markets through PCA and OPTICS Clustering

Isaac Cuschieri

Supervisor: Dr. Christian Manicaro

A dissertation submitted in partial fulfilment of the
requirements for the Masters in Banking, Finance & Investments.
Faculty of Economics, Management and Accountancy,
University of Malta

October, 2024

Abstract

This thesis explored the application of statistical arbitrage strategies on commodity related assets. The asset universe consisted of a diversified basket of 55 assets spanning three asset classes: commodity futures, commodity-linked equities, and commodity currencies. Two strategies were employed: a traditional PCA-based approach and a method that additionally involved clustering the assets using OPTICS. Over the period from 2014 to 2024, both strategies generated slight yet consistent returns. Notably, the strategy incorporating OPTICS clustering outperformed, both in absolute returns and also risk adjusted performance, suggesting that the inclusion of a clustering step may provide additional benefits in such strategies. Moreover, when tested on a post COVID-19 period, the PCA approach failed to generate returns, while the OPTICS strategy remained slightly profitable. Additional results are presented on the characteristics of the residual parametrisation, as well as a insights into which asset clusters and sectors performed the best. Any returns attributable to both strategies proved to be uncorrelated both with a broad based commodity index and also the S&P500.

I would like to dedicate this to my family.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Preamble

In the early 20th century, biologist Thomas Hunt Morgan conducted a series of experiments on planarians. He undertook different mechanical and chemical interventions, such as cutting planarians in different orientations, and subjecting them to different environmental perturbations, which disrupted their physical form. He observed that, regardless of the nature of the cut or the specific perturbation, the pieces would regenerate into fully formed identical planarians (Morgan, 1901). This remarkable regenerative ability was for the most part due to inherent systematic factors within the organism that governed its structure and function, ensuring that, despite different external disruptions, planarians always reverted to their intrinsic form. This work laid the foundation for understanding how organisms maintain regularity in the face of disruptions.

Much like the planarian has a tendency to converge towards it's inherent form, even in the face of different external forces, in financial markets, the framework for sta-

tistical arbitrage is based on the fundamental principle, that, assets that experience idiosyncratic shocks, such as a sudden price spike or drop, are more likely to tend their long term mean with respect to similar assets. This reversion occurs because, despite these temporary deviations, assets are driven by shared systematic forces, economic conditions, as well as sector specific trends. These forces, combined with the inherent nature of the assets themselves, ensure that even after distinct disruptions, they tend return maintain their equilibrium relationship with respect to other similar assets. This convergence creates opportunities for statistical arbitrage, as temporary divergences signal potential trades where assets are expected to revert to their stable, long-term mean.

This analogy is of course mostly illustrative, as there are some fundamental differences between financial markets and biological systems. For example, while many assets tend to revert to their relative long-term means after idiosyncratic shocks, there are numerous examples of divergences between similar assets that do not result in convergence. Asset specific factors, structural market changes, shifts in investor sentiment, or even macroeconomic forces can lead to prolonged or even permanent deviations. However, just as how the planarian tends to it's intrinsic form in the presence of external forces, statistical arbitrage strategies seek to exploit the tendency of assets to hold a well founded equilibrium relationship, allowing traders to exploit opportunities resulting from temporary deviations that disrupt this balance.

1.2 Background and Context

Statistical Arbitrage is often credited as having been pioneered in the late 1980s by Nunzio Tartaglia (Gatev et al., 2006), who gathered a team of physicists and mathematicians to study statistical anomalies within equity markets. At its core, the concept is relatively straightforward. When two different assets have historically exhibited similar price movements, based on a particular measure, these assets become candidates for further consideration. If a sufficient divergence between their prices is determined to have occurred, the arbitrageur could then take a long position in the under performing asset while shorting the over performing one.

While traditional arbitrage guarantees deterministic profits, statistical arbitrage yields profits that are inherently stochastic. This implies that the strategy is only expected to be profitable over time, based on the expected returns across a sufficiently large series of trades (Hoel, 2013). In the long term, there should exist a point where the probability of gains outweighs that of losses, thus enabling a portfolio to accumulate consistent wealth through the implementation of such a strategy.

In the contemporary state of play, statistical arbitrage strategies have become increasingly complicated (Pole, 2011). This complexity is driven by the integration of ever-advancing quantitative methods that serve the necessity of exploiting any possible opportunities that arise in markets which are becoming increasingly competitive. This shift demands constant refinement, as only the most proficient and adaptable strategies can thrive in the modern environment.

1.3 Research Objectives

The analysis aims to investigate the viability and confirm the presence of exploitable statistical arbitrage opportunities for commodity-related assets over the considered time period. In the process, several ancillary but relevant questions will be addressed, particularly regarding the techniques employed to achieve this goal.

By conducting a comparative analysis on the same time-frame for strategies with and without the *Ordering Points To Identify the Clustering Structure* (OPTICS) clustering algorithm, the study will determine whether there is any added benefit to incorporating this clustering step in such a procedure. Such an advantage could provide support for the use of other unsupervised learning techniques, which are becoming increasingly popular in the context of strategies for statistical arbitrage. Additionally, the validity of performing *Principal Component Analysis* (PCA) as a pre-processing step will be also be scrutinized.

Much of the existing literature on statistical arbitrage focuses primarily on equity markets, where less attention is given towards whether these strategies can be effectively applied to commodity markets (Lazzarino et al., 2018). This paper aims to further contribute with regards to the viability of statistical arbitrage strategies for commodity-related assets. Furthermore, the inclusion of commodity assets spanning different asset classes will help identify whether meaningful long-term statistical arbitrage relationships can exist between assets which are usually separated by asset class boundaries before the formation stage.

1.4 Thesis Structure

The organisation of this thesis can be broken down as follows:

- **Chapter 2** contains the literature review which provides an overview of some of the main studies relevant to statistical arbitrage in general, as well as an overview of statistical arbitrage implementations which are of relevance to the one undertaken in this paper. A section on studies which consider statistical arbitrage applications in the context of commodities is also provided.

- **Chapter 3** provides the methodology undertaken in this paper. At each step of this methodology, reference is made to the theoretical basis behind any of the concepts which are being employed. There is also provided in this section an explanation behind each decision which had to be made along the way in order to arrive to the results which were obtained.

- **Chapter 4** provides a break down of the results obtained. These results come in 3 main forms; there are first results which pertain to the structure of the representation obtained through performing the PCA, secondly there are results which regards to the properties and nature of the clusters which were formed through the OPTICS, and finally the trading performance results are provided. This section further breaks down the results into their asset specific components and provides a discussion on the performance of the results.

- **Chapter 5** gives a brief retrospective overview of the results and discusses some of the limitations with regards to statistical arbitrage, both in general and also those which pertain to this specifications. Furthermore some recommendations are provided for similar future exercises.

Chapter 2

Literature Review

2.1 Background

Throughout the years, strategies for statistical arbitrage have garnered the attention of both academic researchers and industry practitioners alike. Numerous professional traders, institutional investors, and hedge fund managers continue to employ such strategies, (Pole, 2011). Around the mid-2000s and continuing into the 2010s, a renewed interest in statistical arbitrage sparked a resurgence of both analytical and empirical studies (Vidyamurthy, 2004; Elliott et al., 2005; Jurek and Yang, 2007; Huck, 2010; Avellaneda and Lee, 2010; Bertram, 2010; Do and Faff, 2012).

Pairs trading can be considered as an 'ancestor' to modern statistical arbitrage, relying on the price relationships between two correlated stocks to exploit temporary mainsprings. Gatev et al. (1999), recognized for their foundational work on the topic, authored both the 1999 and 2006 papers that conceptualized a pairs trading strategy which utilized a simple distance method which minimized the historic sum of squared distances between both the prices of two of *'coupled'* assets. Their com-

prehensive back-testing on U.S. equities from 1962 to 2002 employed a two-stage methodology, which consisted of a formation period spanning 1 year which was followed by a 6 month long trading phase. By applying a simple trading rule that initiated positions when price divergence exceeded two standard deviations, they achieved consistent annualized average excess returns of around 11 per cent, which held even when considering conservative transaction costs. The study attributed these profits to the temporary mispricing of closely related stocks, driven by a common return factor not explained by traditional risk models such as those of Fama and French. Do and Faff (2010) replicated the methodology which was undertaken by Gatev et al. (2006) over a longer time frame and found that almost one third of the pairs which were based on the distance method failed to converge. In this regard they argued that there might be limitations to the distance method, along with an increase in competitiveness for arbitrage opportunities due to technological developments.

In his comprehensive and widely regarded review of statistical arbitrage strategies, Krauss (2017) noted that excess returns such as those observed by Gatev et al. (2006), are among the few *'market phenomena'* that have been consistently validated over the years. In his review, Krauss also presented a taxonomy for statistical arbitrage strategies, where he classified them into the following broad categories; distance, co-integration, time-series, stochastic control and *'other approaches'*. One should note that these approaches are far from being mutually exclusive, and in practice elements from each of these categories are utilized in implementing strategies.

2.2 The Co-Integration Approach

Among proposing many fundamental ideas, Vidyamurthy (2004) sought to formalize the relationship between pairs of securities by employing the concept of co-integration. Recognizing that asset prices are often non-stationary and exhibit stochastic trends, Vidyamurthy decomposed the price series into their non-stationary and stationary components to facilitate meaningful statistical analysis. Considering two different series of order one, $I(1)$, which are co-integrated, series x_t and y_t could be decomposed as follows:

$$x_t = i_{x_t} + \varepsilon_{x_t} \tag{2.1}$$

$$y_t = i_{y_t} + \varepsilon_{y_t} \tag{2.2}$$

where i_{x_t} and i_{y_t} represent the idiosyncratic non-stationary components which capture the underlying stochastic trends of x_t and y_t and ε_{x_t} and ε_{y_t} denote the stationary components, or residuals, which fluctuate around a constant mean. Building upon this decomposition, Vidyamurthy here introduced the construction of a co-integrated series, z_t, which serves as the *spread* between both of the assets. This spread is defined using the co-integration coefficient γ, which quantifies the long-term equilibrium relationship between x_t and y_t such that $i_{x_t} = \gamma i_{y_t}$

$$z_t = x_t - \gamma y_t = (i_{x_t} - \gamma i_{y_t}) + (\varepsilon_{x_t} - \varepsilon_{y_t}) \tag{2.3}$$

Thus the non-stationary trends in x_t and y_t are perfectly offset by the co-integration coefficient γ, thereby eliminating any stochastic trends from the spread z_t. , This gives a spread z_t which is mean-reverting, and ideal candidate for pairs trading. Lin et al. (2006) further developed Vidyamurthy's framework by incorporating a

stop-loss mechanism into this pairs trading framework with respect to the minimum profit required per trade. In their comparative analysis, Huck and Afawubo (2015) found that over the S&P 500, the co-integration approach performed better than the distance method as defined per Gatev et al. (2006).

The framework proposed by Avellaneda and Lee (2010) was seminal in advancing statistical arbitrage strategies. They decomposed returns into systematic and idiosyncratic components using two distinct approaches. The first approach involved regressing the returns of a number of S&P 500 stocks onto pre-defined sector ETFs to isolate the systematic components. Inspired by Jolliffe (2002), the second approach employed a multi-factor model which considered a number of statistical PCA eigenvectors to represent the systematic components. This framework was foundational for the the one followed in the initial part of the methodology of this paper. In Avellaneda and Lee's framework the residuals which resulted from both of the aforementioned approaches were characterized by an Ornstein-Uhlenbeck (OU) process. This enabled the generation of trading signals based on deviations from equilibrium, quantified by a dimensionless s-score. Building on this model, Yeo and Papanicolaou (2017) extended the framework by focusing on risk management and presenting an optimization approach for tne investment allocation in response to the trading signals. Lettau and Pelger (2020) introduced a novel approach for deriving the factors for asset pricing, one which builds on traditional PCA by penalizing pricing errors in expected returns. This approach identified factors with high Sharpe ratios that were on occasions overlooked by PCA.

With regards to the application of the OU process for modeling spreads, a number of analytical results have been formulated. Bertram (2010) derived analytical formulas

for the trading phase, which followed on the assumption that the price of the spread between two assets follows an exponential OU process. He approached the problem by analyzing the first-passage time for the process, where he initially derived expressions for the variance and the mean of the trade duration. Subsequently, he provided formulas for the variance and the expected return per unit time. To conclude, Bertram proposed a solution for selecting optimal trading thresholds through maximization of both the Sharpe ratio and also the expected return. Zeng and Lee (2014) expanded upon Bertram's work by accounting for short positions as well, were they formulated a polynomial expression with regards to the expectation of the first-passage time in an OU process which had a two sided boundary. Endres and Stübinger (2019) formulated an optimal pairs trading strategy which was founded on the Lévy driven OU process. The strategy's objective function was explicitly represented, thus enabling optimization without the need for Monte Carlo methods. Through the maximisation of the expected return, which expressed with respect to the first passage time of the spread process, the model identified optimal entry and exit signals. By applying this model on high-frequency data from S&P 500 equities starting from 1998, through to 2015, they divided this data into 10 economic sectors and empirical back-testing demonstrated solid evidence for the strategy's profitability and value-added in considering a Lévy-driven model.

2.3 Statistical Arbitrage Pertaining to Commodity Markets

The majority of empirical literature for statistical arbitrage and pairs trading focuses on implementing these strategies within the context of equity markets, most often

in the U.S. In their expansive literature review, Lazzarino et al. (2018) found 165 papers regarding statistical arbitrage starting from 1995, through to 2016. From the 165 papers, 104 were on equities, 40 were on fixed income securities, whilst only 9 papers concerned commodities. There however still exist a number of studies which provide empirical implementations of these strategies within commodity markets beyond the ones surveyed in Lazzarino et al. (2018), especially in recent years. One of the first applications to commodity was the one of Girma and Paulson (1999), who focused on the spread between the price of petroleum futures and the prices for the futures representing corresponding refined products, such as heating oil and gasoline, over the period starting from 1983, through to 1994. They found that multiple variations of this spread could be considered as being stationary, and therefore, suitable for pairs trading. In their strategy, trades were initiated when the spread deviated from the mean by a factor of the standard deviation from the moving average, calculated over a specific number of days. The trades were closed when the spread reverted back to the moving average. Their results showed annual returns exceeding 15 per cent, even after taking into consideration accounting costs.

Bianchi et al. (2009) implemented a version of the strategy introduced by Gatev et al. (2006) in the context of commodity market futures over the period from 1990 to 2008. In their analysis, they reported both statistically and economically significant excess returns generated by the strategy. Additionally, they noted that these returns were achieved with relatively low exposure to systematic risk factors. Likewise, by analyzing the returns generated from a co-integration-based statistical arbitrage strategy across various European energy sectors, Hain et al. (2018) identified significant risk-adjusted excess returns. These results were found to be distinct from those generated by simple contrarian or momentum-based strategies.

Similarly, Nakajima (2019) explored arbitrage opportunities between wholesale futures for electricity and similar futures for natural gas on the New York Mercantile Exchange (NYMEX), under the assumption of a co-integrated relationship between the prices for power and natural gas. The results, based on data from 2014 to 2017, demonstrated the potential for significant profits, with yield rates reaching as high as 30 per cent, reinforcing the possibility of generating returns within the energy futures market.

Mikkelsen (2018) investigated the viability of performing statistical arbitrage with regards the stocks of 18 seafood companies which were listed under the Norwegian consumer goods sector of the Oslo Stock Exchange (OSE). Through the use of both high frequency and also daily data starting from January 2005, and all the way through to December 2014, two variations were applied: the traditional distance approach and a co-integration based method, such that both results could be compared. The findings showed a high rate of non-convergence for both methods, suggesting that after considering transaction costs, neither strategy produced significant profits, leaving the question of which approach is better suited for pairs selection in this case, unresolved. More recently, He et al. (2023) implemented pairs trading at high frequency intervals of 1 minute with regards to China's futures market. The framework incorporated co-integration testing, Kalman filtering, and the user of the Hurst index on data from 47 commodities which were sufficiently liquid. The performance of the strategy was bench-marked against the Wenhua Commodity Index. The findings revealed that, after accounting for transaction costs, the strategy achieved a cumulative return of 81 per cent within the sample and 21 per cent out-of-sample, with an impressive maximum draw down of less than 1 per cent out-of-sample.

2.4 Clustering

OPTICS, which was utilized in the second implementation considered in this paper, is a density-based clustering algorithm that creates an ordered representation of the dataset, capturing the clustering structure across a wide range of parameters. It works by calculating the reachability distance of points from core objects and organizing them in increasing order, allowing for efficient density-based cluster exploration (Bhattacharjee and Mitra, 2021).

In the context of pair selection, clustering via unsupervised learning could provide an objective way to group assets based on statistical similarities rather than subjective assumptions. For detecting suitable pairs, density-based clustering techniques are appropriate since they can identify natural groupings of assets without needing to pre-define the number of clusters. The most popular of these algorithms is DBSCAN *(Density-Based Spatial Clustering of Applications with Noise)* as introduced by Ester et al. (1996), however this is most appropriate for forming clusters with uniform density which is not necessarily implied for the case of pair formation. In this analysis, OPTICS is utilized in place of DBSCAN due to it's ability to work with clusters with varying densities. As pioneered by Ankerst et al. (1999), OPTICS identifies clusters at different density levels through allowing for a variable neighborhood radius ε, which makes it more suitable for pair formation, where asset clusters often have non-homogeneous densities.

Berkhin (2006) noted that when applying an unsupervised learning algorithm, it is crucial to limit the number of features as high dimensionality can lead to an increased chance of including irrelevant features and also exacerbates the curse of

dimensionality, where the volume of the feature space grows exponentially with each added dimension. According to Berkhin (2006), these issues become significant beyond 15 dimensions.

The most famous application of OPTICS for statistical arbitrage is that of Sarmento and Horta (2020), where they utilized PCA in order extract the systemic factors of risk within their security universe which consisted of 208 ETFs. OPTICS was applied to the PCA representation in order to identify potentially profitable pairs more efficiently. This approach proved beneficial, such that over the period between January 2009 and December 2018 they reported an annualized Sharpe Ratio of 3.79 with 86 per cent of the pairs being profitable when the clustering was implemented, which is greater than the annual SR of 3.58 and the 79 per cent of pairs being profitable for the case without clustering.

Wang et al. (2022) implemented a particular approach towards statistical arbitrage which incorporated a number of clustering techniques. These included OPTICS, association rule algorithms, and the bipartite graph partition algorithm which were all applied in order to derive many to many pairs. The results in this study demonstrated that this new approach effectively selected many-to-many pairs for trading, providing significantly a larger amount of trading opportunities over traditional pair trading methods. Han et al. (2023) applied a number of clustering methods which included k-means clustering, DBSCAN, and agglomerative clustering where they used both firm characteristics and past returns to identify trading pairs. Unlike traditional strategies for pairs trading that rely solely on the time series of returns, the inclusion of firm characteristics with price data to select pairs differentiated their study. They showed that incorporating firm characteristics significantly enhanced

pair identification and improved the strategy's performance. Applied to U.S. stock market data starting from January 1980, all the way through to December 2020, the market neutral portfolio created via agglomerative clustering generated a statistically significant mean annualised return of 24.8 per cent with a corresponding Sharpe ratio of 2.69. Even after taking into account transaction costs and also ignoring stocks at the bottom 20 per cent of NYSE stocks in terms of market cap, the strategy remained profitable. They also performed a number of robustness tests in order to confirm that data snooping did not influence the results.

2.4.1 Theoretical Background

Clustering involves organizing a dataset into meaningful groups or sub classes, where similar data points are grouped together based on shared characteristics (Alhamazani et al., 2014). This process helps in discovering patterns within datasets.

OPTICS was introduced to address limitations in traditional clustering methods, such as DBSCAN, which require a pre-set density threshold, ε and may struggle to handle clusters of varying densities. Instead of directly clustering the data, OPTICS organizes it based on the density-based structure of the dataset. This structure allows the detection of clusters over a wide range of parameters, making the algorithm versatile for various applications. It generates an ordering of the points such that points that are close in terms of density are located near each other in the output ordering.

As outlined by Bhattacharjee and Mitra (2021), the algorithm works by selecting a point, termed as a core object, and adding it to an ordered list. It then expands outward by considering points that are directly reachable from this core point—points

within a certain density radius ε. These reachable points are kept in a "seed list," ordered by increasing reachability distance from the closest core object. The reachability distance between points reflects how far they are from their nearest core point, capturing the density relationship in the data.

For each subsequent point in the seed list, its reachability distance is calculated, and it is written to the ordered file, continuing the cluster expansion. If a new point is found to be a core object, its directly reachable points are added to the seed list, maintaining the ordering by reachability distance. This process continues until all points are ordered, revealing the inherent cluster structure across multiple density scales.

In this case the time complexity of the algorithm is at worst, $O(N^2)$. However, through the application of tree based methods such as a k-d tree or R-tree to efficiently manage spatial relationships between points, the time complexity can be cut down to $O(N \log N)$, making it feasible to implement this for larger datasets.

Chapter 3

Methodology

3.1 Overview

This chapter gives an outline for the methodology undertaken for deriving and implementing the proposed strategies. The first step towards implementing these strategies was the process of collecting and processing the data. A number of key decisions were made in this regard, all of which will be discussed in the first part of this methodology. In total 55 commodity related assets were used. Following the collection of the data, a number of transformations were undertaken in order to derive the correlation matrix. This correlation matrix was then used in order to perform the PCA for this exercise. In this regard a number of procedures were carried out in order to ensure that the appropriate number of PCA components was used, which in this case was determined to be 10. Using the results from the PCA, OPTICS clustering was undertaken, were a number of clusters were determined. The results from both the PCA and the OPTICS clustering were then used in order to generate the residuals between the assets and the eigenportfolio formed, and also within the asset clusters derived through the OPTICS clustering. These

residuals were tested both in terms of their stationarity using the ADF test and also on their rate of mean reversion. For the latter, OU parametrisation using the maximum likelihood estimation approach was undertaken. This yielded a number of residuals which were modeled and used to generate the positions for opening and closing trades. Since residuals were generated from both the PCA directly and also through the OPTICS clustering two strategies were performed over the same trading period. These strategies were then both evaluated and compared using a number of metrics, which are defined in the last section of the chapter. Figure 3.1 provides a visualisation of the key steps underlying this methodology.

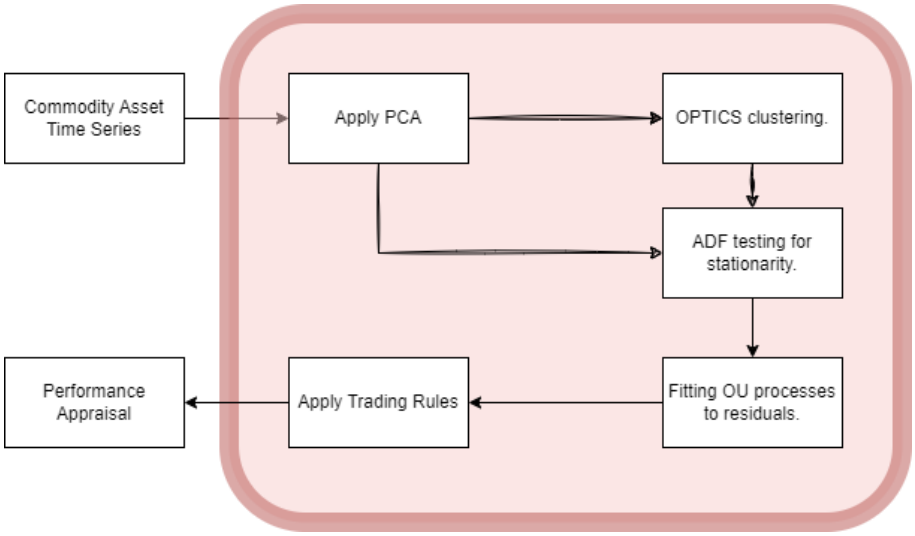

Figure 3.1: Overview of the methodology, where procedures in the red area are conducted on a rolling window basis.

As shown in the above diagram, following the data collection phase, the steps of this methodology were undertaken on a rolling window basis, which consisted both of a formation period and a trading period, similar to the approach taken by Gatev et al. (2006).

3.2 Data Methodology

3.2.1 Data Description & Retrieval

For this analysis, it was deemed important to choose a dataset which was both representative of commodity markets and also complete. In this regard a process of shortlisting was undergone, where data for different commodity assets was fetched through the a combination of financial Python packages. The data considered was at a daily interval from August of 2012, up until September of 2024, since 10 years was the most common time frame used in similar cases in literature. Prior data all the way back to the start of 2005 was also initially collected, but it was found to be of insufficient quality, thus this data was excluded from the analysis. Additionally, a number of assets were excluded from the dataset due to incomplete data.

The finally considered data comprised of three major asset classes within the commodity markets: commodity futures, equities, and currencies. The futures covered energy products, industrial metals, agricultural products, and soft commodities, as shown in Table 3.1. Equities were chosen from companies in the materials and energy sectors, primarily involved in mining and energy production, as detailed in Table 3.2 and the currencies chosen were tied to commodity-dependent economies that are influenced by the export of key resources, as listed in Table 3.3.

This consideration of the different asset classes could prove beneficial in the process of cluster formation, as it allows for clusters or pairs to be formed across different asset categories. For example, a pair could be constructed between the future contract of a commodity like gold, and Barrick Gold, which derives approximately 88

Ticker	Exchange	Description	Contract Size
XAU=	OTC	Gold Spot	100 troy ounces
XAG=	OTC	Silver Spot	5,000 troy ounces
XPT=	OTC	Platinum Spot	50 troy ounces
XPD=	OTC	Palladium Spot	100 troy ounces
SAFc1	LME	Aluminum C1	25 metric tons
HG=F	NYMEX	Copper Futures	25,000 pounds
LCOc1	ICE	Brent Crude Oil Front Month	1,000 barrels
CLc1	NYMEX	WTI Crude Oil Front Month	1,000 barrels
HO=F	NYMEX	Heating Oil	42,000 gallons
NGc1	NYMEX	NYMEX Henry Hub Natural Gas	10,000 mmBtu
RBc1	NYMEX	NYMEX RBOB Gasoline	42,000 gallons
NGLNMc1	ICE	ICE UK NBP Natural Gas	1,000 therms
Wc1	CBOT	CBoT Wheat Composite	5,000 bushels
BL2c1	Euronext	Euronext Paris Milling Wheat	50 metric tons
Cc1	CBOT	CBoT Corn	5,000 bushels
Sc1	CBOT	CBoT Soybeans	5,000 bushels
COMc1	Euronext	Euronext Paris Rapeseed	50 metric tons
RSc1	ICE	ICE-US Canola	20 metric tons
CTc1	ICE	ICE-US Cotton No. 2	50,000 pounds
LRCc1	LIFFE	LIFFE Robusta Coffee	10 metric tons
CCc1	ICE	ICE-US Cocoa Futures	10 metric tons
KCc1	ICE	ICE-US Coffee C Futures	37,500 pounds
OJC1	ICE	Orange Juice C1	15,000 pounds
LCCc1	ICE Europe	ICE Europe London Cocoa	10 metric tons
SBc1	ICE	ICE-US Sugar No. 11	112,000 pounds
JRUc1	Osaka Exchange	Osaka Exchange Rubber	5 metric tons
LSUc1	ICE	ICE White Sugar No. 5	50 metric tons
LHc1	CME	CME Lean Hogs	40,000 pounds
LCc1	CME	CME Live Cattle	40,000 pounds

Table 3.1: The commodity future contracts considered.

per cent of its total revenue from gold mining (Barrick Gold, 2023). This would possibly provide additional trading opportunities. All the equities chosen in this analysis were selected based on their involment in particular commodities. For instance ExxonMobil, Chevron, and Occidental were chosen because they are directly tied to energy commodities such as WTI Crude Oil, Brent Crude Oil, and Natural Gas, making them ideal clustering candidates for statistical arbitrage between these equities and the respective commodity futures.

Ticker	Name	GICS Sector	Market Cap (in billion $)
BHP	BHP	Materials	216.0
RIO	Rio Tinto	Materials	167.0
VALE	Vale	Materials	283.0
FCX	Freeport-McMoRan	Materials	45.9
NEM	Newmont	Materials	5.0
GOLD	Barrick Gold	Materials	35.8
AA	Alcoa	Materials	10.0
X	US Steel	Materials	8.8
MT	ArcelorMittal	Materials	19.0
MOS	Mosaic	Materials	27.0
CF	CF Industries	Materials	81.0
ADM	Archer Daniels Midland	Consumer Staples	26.0
BG	Bunge	Consumer Staples	13.0
CVX	Chevron	Energy	270.0
XOM	ExxonMobil	Energy	518.0
OXY	Occidental	Energy	46.0
SU	Suncor	Energy	67.0

Table 3.2: The commodity related equities considered.

Similarly to the case for the equities, commodity currencies, as detailed in Table 3.3, were selected due to their strong economic ties to key exports within their respective regions. All currencies were considered against the US dollar and the Euro and Pound were included to contextualize thee commodity-linked currencies These currencies represented economies with significant exposure to specific resources, which pre-disposes them to exhibit higher sensitivity to fluctuations in commodity prices.

Ticker	Currency	Continent
EUR=TRB	Euro	Europe
GBP=TRB	British Pound	Europe
AUD=TRB	Australian Dollar	Oceania
CAD=TRB	Canadian Dollar	North America
NZD=TRB	New Zealand Dollar	Oceania
CNY=TRB	Chinese Yuan	Asia
NOK=TRB	Norwegian Krone	Europe
BRL=TRB	Brazilian Real	South America
ZAR=TRB	South African Rand	Africa

Table 3.3: The selected currency pairs, all against the US dollar.

3.2.2 Data Processing

Exchange data over long periods often contains missing values, as was the case with the data used in this study. As such, assets series with more than 1 per cent missing values were entirely excluded. On the remaining data, the standard *Last Observation Carried Forward* method was applied, (Van Buuren, 2018). This method ensured that any values which were absent were substituted by the most recent available observation, which preserved the continuity of the data. Additionally, the data was also inspected for *'clusters'* of missing values or anomalies during key market events or holidays. This inspection ensured that no significant gaps were present in the used data and thus the forward fill method did not bias the data, particularly during high volatility periods. A number of non-consequential transformations were subsequently carried out in order streamline the data for later stages of this analysis. These transformations did not impact the integrity or the substance of the data but rather made it easier to manage for the rest of the analysis.

3.3 Construction of the Correlation Matrix

3.3.1 Standardizing Returns

Once the dataset was finalized, mid-price time series were generated for each asset. To construct the correlation matrix for Principal Component Analysis (PCA), the first step involved calculating returns by differencing the prices. Returns were used in this analysis instead of prices because return time series tend to exhibit stationarity more consistently than price series (Connor and Korajczyk, 1986). In line with this, the methodology outlined by Avellaneda and Lee (2010) was followed. For N assets observed over M days, the price of asset i at a time t_0, relative to the preceding $M+1$ days, was expressed as:

$$R_{ik} = \frac{P_i(t_0 - (k-1)\Delta t) - P_i(t_0 - k\Delta t)}{P_i(t_0 - k\Delta t)}, \quad k = 1, ..., M, \quad i = 1, ..., N, \tag{3.1}$$

where $\Delta t = \frac{1}{252}$ for daily data. These returns were then standardized to Y_{ik} as follows:

$$Y_{ik} = \frac{R_{ik} - \langle R_i \rangle}{\sigma_i}, \tag{3.2}$$

where $\langle R_i \rangle$ is the mean over time, calculated as:

$$\langle R_i \rangle = \frac{1}{M} \sum_{k=1}^{M} R_{ik}, \tag{3.3}$$

and σ_i is the standard deviation, given by:

$$\sigma_i^2 = \frac{1}{M-1} \sum_{k=1}^{M} (R_{ik} - \langle R_i \rangle)^2. \tag{3.4}$$

To construct the correlation matrix, it is important to ensure that the return series used are stationary (Granger and Newbold, 1974). One way to verify this is by examining the autocorrelation functions (ACFs) of the calculated returns. The autocorrelation α_i of the return series R_i at lag τ is defined as:

$$\alpha_i(\tau) = \frac{\sum_{k=1}^{M-\tau}(R_{ik} - \langle R_i \rangle)(R_{i(k+\tau)} - \langle R_i \rangle)}{\sum_{k=1}^{M}(R_{ik} - \langle R_i \rangle)^2}. \tag{3.5}$$

Literature suggests that for stationary time series, the ACF tends to diminish rapidly, with correlations between observations at larger time lags approaching zero. This behavior signifies the core property of stationarity, where a series maintains consistent statistical attributes over time, such as a constant mean and variance (Hyndman and Athanasopoulos, 2018). In contrast, non-stationary time series typi-

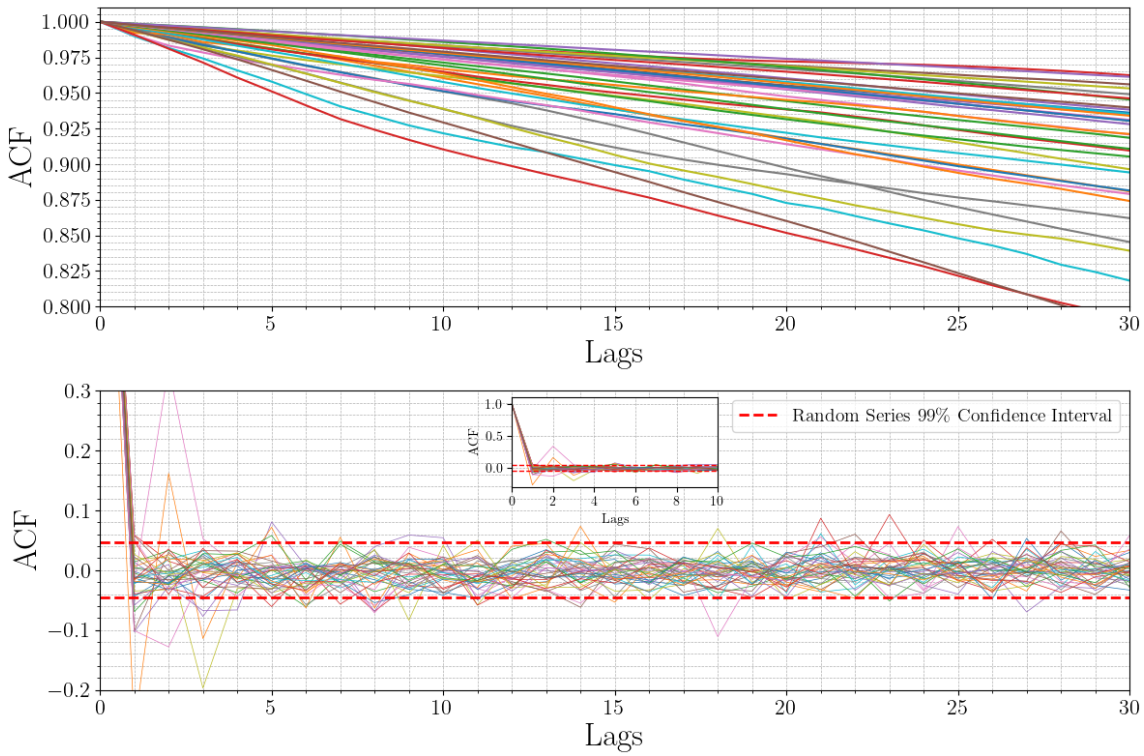

Figure 3.2: (Top) ACF of Prices vs. (Bottom) ACF of Returns.

cally exhibit a slower rate of decay in their ACF, reflecting long-lasting correlations. In certain cases, particularly when a trend is present, the ACF may remain significantly high even at larger lags, indicating a persistent, non-decaying correlation structure over time.

In Figure 3.2, the ACFs of both prices and returns for all the time series in the selected universe are displayed. The ACF for prices shows a slow and gradual decay over the first 30 lags, indicating persistent correlations and a non-stationary nature. In contrast, the ACF for returns quickly decays to zero, signifying that these series are stationary. For comparison, the 99 per cent confidence intervals for the ACF of Gaussian white noise, $X_{i,j} \sim \mathcal{N}(0, 1)$, are overlaid on the return ACF plot. Apart from a few exceptions, it is clear that the returns largely exhibit stationary behavior, staying within these confidence bounds. Given this satisfactory structure, the returns shown in Figure 3.3 were used to create the correlation matrix.

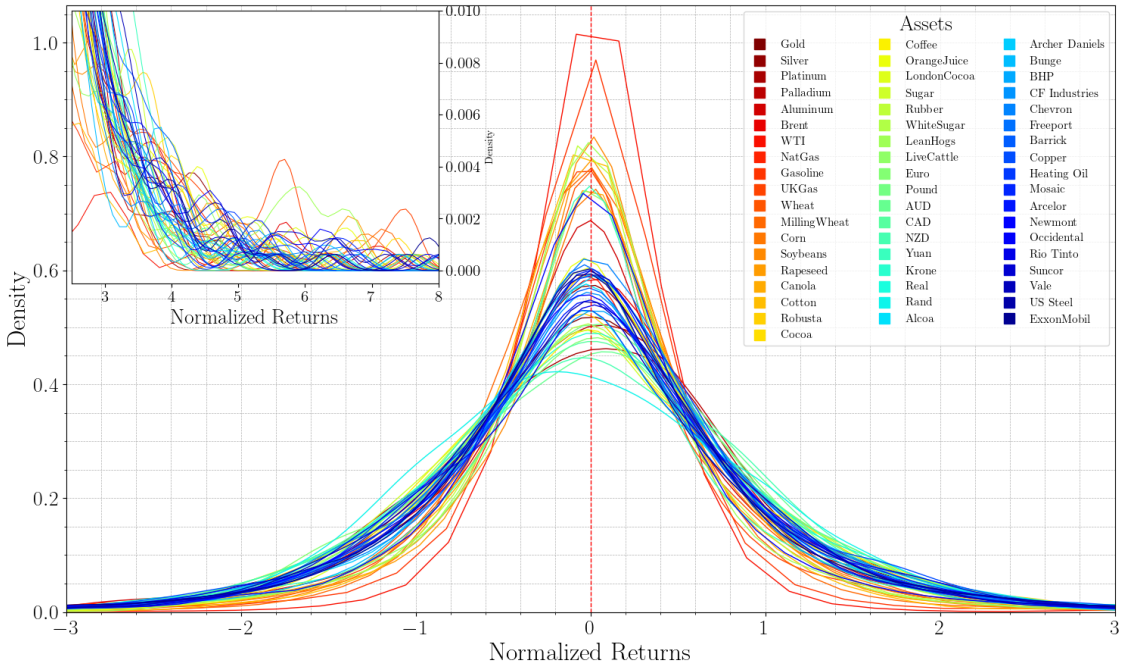

Figure 3.3: CDFs of the standardized returns for the dataset.

3.3.2 Calculating Correlations

The coefficient of correlation ρ_{ij} between two assets i and j can be calculated as follows:

$$\rho_{ij} = \frac{1}{M-1} \sum_{k=1}^{M} \frac{(R_{ik} - \langle R_i \rangle)(R_{jk} - \langle R_j \rangle)}{\sigma_i \sigma_j} = \frac{1}{M-1} \sum_{k=1}^{M} Y_{ik} Y_{jk}, \tag{3.6}$$

where R_{ik} and R_{jk} are the returns of assets i and j, respectively, over the time horizon $k = 1, \ldots, M$. Here, $\langle R_i \rangle$ and $\langle R_j \rangle$ represent the average returns of assets i and j, while σ_i and σ_j are their respective standard deviations. The $N \times N$ correlation matrix, $\mathbf{C}$, is then constructed from the individual correlation coefficients ρ_{ij}:

$$\mathbf{C} = \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1N} \\ \rho_{21} & 1 & \cdots & \rho_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{N1} & \rho_{N2} & \cdots & 1 \end{pmatrix}. \tag{3.7}$$

The resulting matrix, $\mathbf{C}$, is a symmetric, non-negative definite matrix which contains ones along the diagonal. This symmetry ensures that the matrix has real, non-negative eigenvalues, which is a key requirement for performing the PCA. In PCA, the correlation matrix is used to extract the common factors driving asset returns by analyzing the eigenvectors *(principal components)* and eigenvalues *(variance explained)*.

3.4 Principal Component Analysis (PCA)

The next step in the analysis was to perform a PCA, starting from the standardized returns. A PCA extracts meaningful information from complex datasets by identifying a linear transformation that projects a set of observed variables into a new set of uncorrelated variables, which are known as principal components. These principal components capture the underlying structure of the data by emphasizing the directions of maximum variance. To achieve this, PCA leverages the eigenvectors and eigenvalues pertaining to the empirical correlation matrix, which serves as the foundation for this transformation. The eigenvalues, denoted as $\{\lambda_j\}_{j=1}^{N}$, are ordered in descending magnitude as follows:

$$N \geq \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N \geq 0, \tag{3.8}$$

where N represents the number of dimensions or variables within the dataset. These eigenvalues reflect the amount of explainable variance by each corresponding principal component, with larger eigenvalues indicating greater importance in capturing the variance. Each eigenvector represents the direction of maximum variance in the data, and can be written as:

$$\nu^{(j)} = \left(\nu_1^{(j)}, \nu_2^{(j)}, \ldots, \nu_N^{(j)}\right)^T, \tag{3.9}$$

where $\nu^{(j)}$ corresponds to the j-th principal component, and the entries $\nu_1^{(j)}, \nu_2^{(j)}, \ldots, \nu_N^{(j)}$ represent the contributions of the original variables to this component. The percentage of variance explained by each eigenvalue λ_k is calculated using the following

formula:

$$\text{Variance Percentage of } \lambda_k = \frac{\lambda_k}{\sum_{j=1}^{N} \lambda_j}. \tag{3.10}$$

This ratio indicates how much of the total variability in the data is accounted for by each principal component, with the sum of all variance percentages adding up to 100 per cent. Furthermore, the empirical correlation matrix C_{ij}, an $N \times N$ diagonalizable matrix, can be decomposed into its eigenvectors and eigenvalues as $C = V\Lambda V^T$, where V is the eigenvectors matrix, and Λ is the diagonal matrix of eigenvalues. This decomposition allows for the expression of the data in terms of its principal components, with each eigenvector contributing to the overall structure based on its corresponding eigenvalue.

3.4.1 Choosing the number of PCA Factor

Due to the varying nature of the data over time, the number of principal components which have relevance in explaining a proportion of the total variance is not constant. For example, during periods of high market volatility, the number of principal components which is necessary to explain the same proportion of variance, decreases (Caneo and Kristjanpoller, 2021).

In order to get an empirical sense for the appropriate number of PCA factors in this case, as shown in Figure 3.4, PCA was applied to the dataset using rolling windows with a length of 1 year. The number of components needed to explain 40%, 50%, 60%, and 70% of the total variance was then calculated at each time step. Thus the figure represents how the amount of principal components required to capture different levels of variance has fluctuated over the years which were considered.

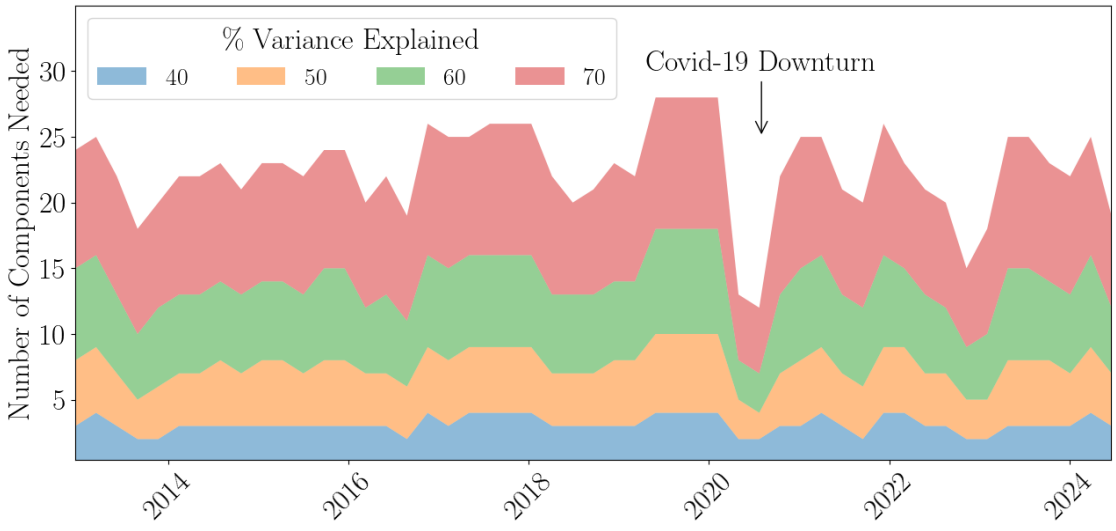

Figure 3.4: The number of Principal Components required to explain different thresholds of the total variance.

A significant observation that could be made in this case, is the period marked as the COVID-19 downturn around 2020. During this period, fewer principal components were necessary in order to explain the same percentage of variance. This behavior can be derived from heightened levels of market volatility and synchronized movement across the considered assets during the crisis. The extreme market conditions, such as sharp declines and rebounds, likely caused several key factors to dominate the variance, thus reducing the need for a large number of components in order to explain a large portion of the total variance in the data.

3.4.1.1 Parallel Analysis

A parallel analysis was also performed to determine the appropriate number of PCA factors for this case. In parallel analysis, the eigenvalues of the actual empirical data are compared with respect to the eigenvalues of randomly generated noise. In the event that the eigenvalues obtained for the actual data are smaller than those from the generated noise, it is considered that these eigenvectors do not reflect any true underlying structure but instead resemble the noise one would expect from purely random data (Iacobucci et al., 2022).

In a typical scree plot, the smaller eigenvalues (appearing towards the right of the plot) usually represent noise and random sampling error, while the larger eigenvalues (towards the left) indicate a genuine underlying structure in the data. When including random noise for parallel analysis, the process helps systematically compare the scree plot of actual data with a scree plot generated from random noise.

In this regard, random data was generated as $X_{i,j} \sim \mathcal{N}(0,1)$ for $i,j = 1,2,\ldots,N$ and the resulting scree plots for both the random data, and also the commoditites

data are shown in figure 3.5. The blue bars represent the principal components for the commodity data, where the first principal component (PC1) explains the highest amount of variance, approximately 22 per cent. From there, the explained variance steadily decreases across subsequent components. This pattern is common in high-dimensional data, where the first few principal components capture most of the variance, while later components explain progressively less. The green bars, on the other hand, represent the same decomposition applied to random Gaussian white noise.

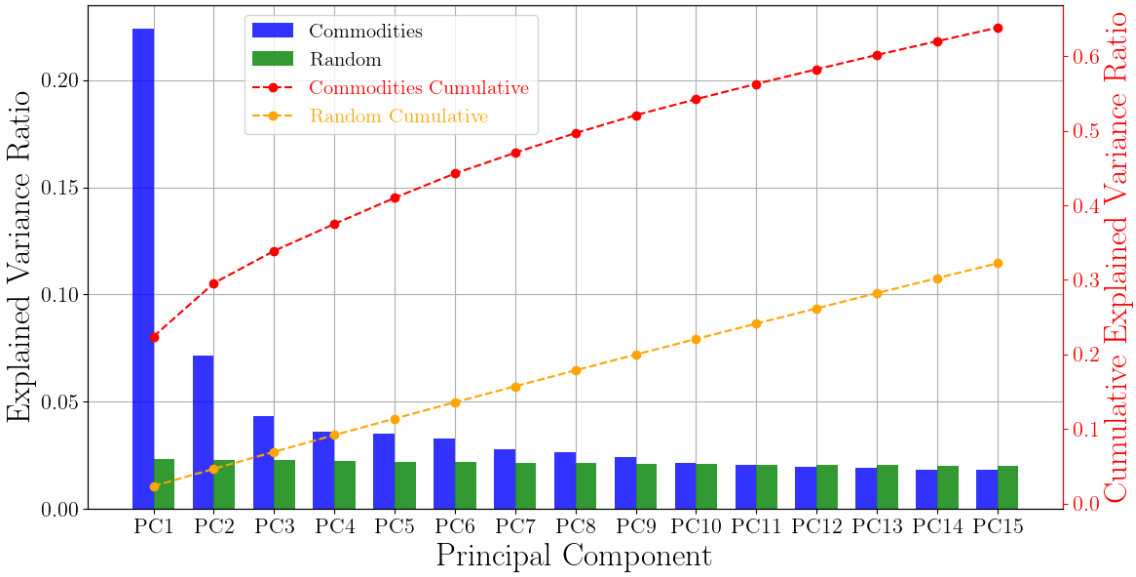

Figure 3.5: Comparison of the scree plot for the actual data against the scree plot generated from Gaussian noise, highlighting the distinction between meaningful components and random noise.

For the Gaussian noise the explained variance is more evenly distributed across the components compared to the commodity data, with no single component explaining a large portion of the variance. This is expected for white noise, which lacks the structured variability typically present in real-world data like commodity prices.

This analysis provides evidence that for the commodity data considered, 10 PCA components would provide meaningful insights beyond random noise. The number of components is also small enough to reduce the risk curse of dimensionality issues, such as those highlighted by Berkhin (2006).

3.5 Residual Estimation

3.5.1 Eigenportfolio Formation

For each of the principal components found in the PCA, continuing as per (Avellaneda and Lee, 2010), an *eigenportfolio* can be formed by dividing each of the eigenvectors derived through the PCA by the respective standard deviations:

$$Q_i^{(j)} = \frac{v_i^{(j)}}{\bar{\sigma}_i} \tag{3.11}$$

where $Q_i^{(j)}$ represents the weight of the asset i in the j-th eigenportfolio, $v_i^{(j)}$ is the i-th component of the eigenvector corresponding to the j-th principal component, and $\bar{\sigma}_i$ is the standard deviation of return pertaining to asset i. The returns of each eigenportfolio, F_{jk}, are then calculated as:

$$F_{jk} = \sum_{i=1}^{N} \frac{v_i^{(j)}}{\bar{\sigma}_i} R_{ik}; j = 1, 2, ..., m. \tag{3.12}$$

where R_{ik} denotes the return of asset i at time k, and m represents the number of principal components used.

Along the lines of Arbitrage Pricing Theory (APT) , the return of an asset can be decomposed into components of multi-factor model plus residuals. In this case, the factors correspond to the first m principal components from the PCA. The return for asset i is thus expressed as:

$$R_i = \sum_{j=1}^{m} \beta_{ij} F_j + \tilde{R}_i \quad \text{or} \quad \tilde{R}_i = R_i - \sum_{j=1}^{m} \beta_{ij} F_j. \tag{3.13}$$

3.5.2 Eigenportfolio Formation

Further in line with the approach of Avellaneda and Lee (2010), for each principal component obtained from the PCA, an *eigenportfolio* was constructed by scaling each component of the eigenvector by the corresponding asset's standard deviation:

$$Q_i^{(j)} = \frac{v_i^{(j)}}{\bar{\sigma}_i}, \tag{3.14}$$

where: $Q_i^{(j)}$ represents the weight of asset i in the j-th eigenportfolio, $v_i^{(j)}$ is the i-th component pertaining to the eigenvector associated with the j-th principal component, and $\bar{\sigma}_i$ corresponds to the the standard deviation in returns of asset i. The returns of each eigenportfolio, $F_{j,k}$, are then calculated as:

$$F_{j,k} = \sum_{i=1}^{N} Q_i^{(j)} R_{i,k}, \quad \text{for } j = 1, 2, \ldots, m, \tag{3.15}$$

where $R_{i,k}$ denotes the returns for asset i at time k, and m is the number of principal components utilized.

Along the lines of the APT of Ross (2013), the return of an asset can be decomposed into a multi-factor model plus residuals. In this context, the factors correspond to the first m principal components derived from the PCA. Thus, the return for asset i is expressed as:

$$R_{i,k} = \sum_{j=1}^{m} \beta_{ij} F_{j,k} + \tilde{R}_{i,k} \quad \text{or} \quad \tilde{R}_{i,k} = R_{i,k} - \sum_{j=1}^{m} \beta_{ij} F_{j,k}, \tag{3.16}$$

where: β_{ij} represents the sensitivity of asset i with respect to factor j, and $\tilde{R}_{i,k}$ is the residual for the return of asset i at a time k.

3.5.3 Linear Regression Between Assets & PCA Components

As per equation 3.13 a set of linear regressions were performed between the returns of each asset and the corresponding returns of the principal components. For each asset i, the returns R_i were regressed against the returns of the first $m = 10$ principal components , denoted by $F_1, F_2, \ldots, F_{10}$.

By estimating the beta coefficients and residuals, the asset's returns were decomposed into two parts: systematic returns driven by the principal components and idiosyncratic returns represented by the residuals. The residuals, obtained for each asset for one of the windows considered are in Figure 3.6.

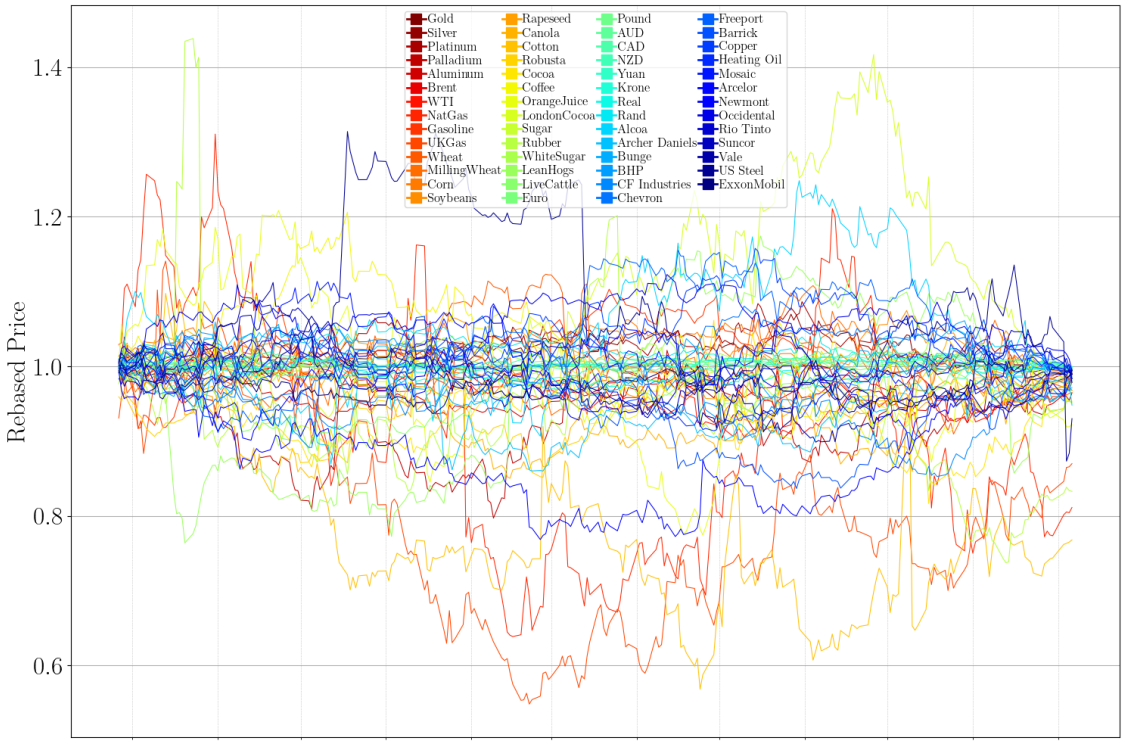

Figure 3.6: Rebased residuals from the regression of each asset against the first 10 principal components, generated on *25/09/2023*, showing varying degrees of stationarity across the residuals.

3.6 OPTICS-based Clustering

3.6.1 Cluster Formation

After performing the PCA and obtaining the asset loadings on the first 10 Principal components, the OPTICS algorithm was employed to identify clusters within the compacted representation.

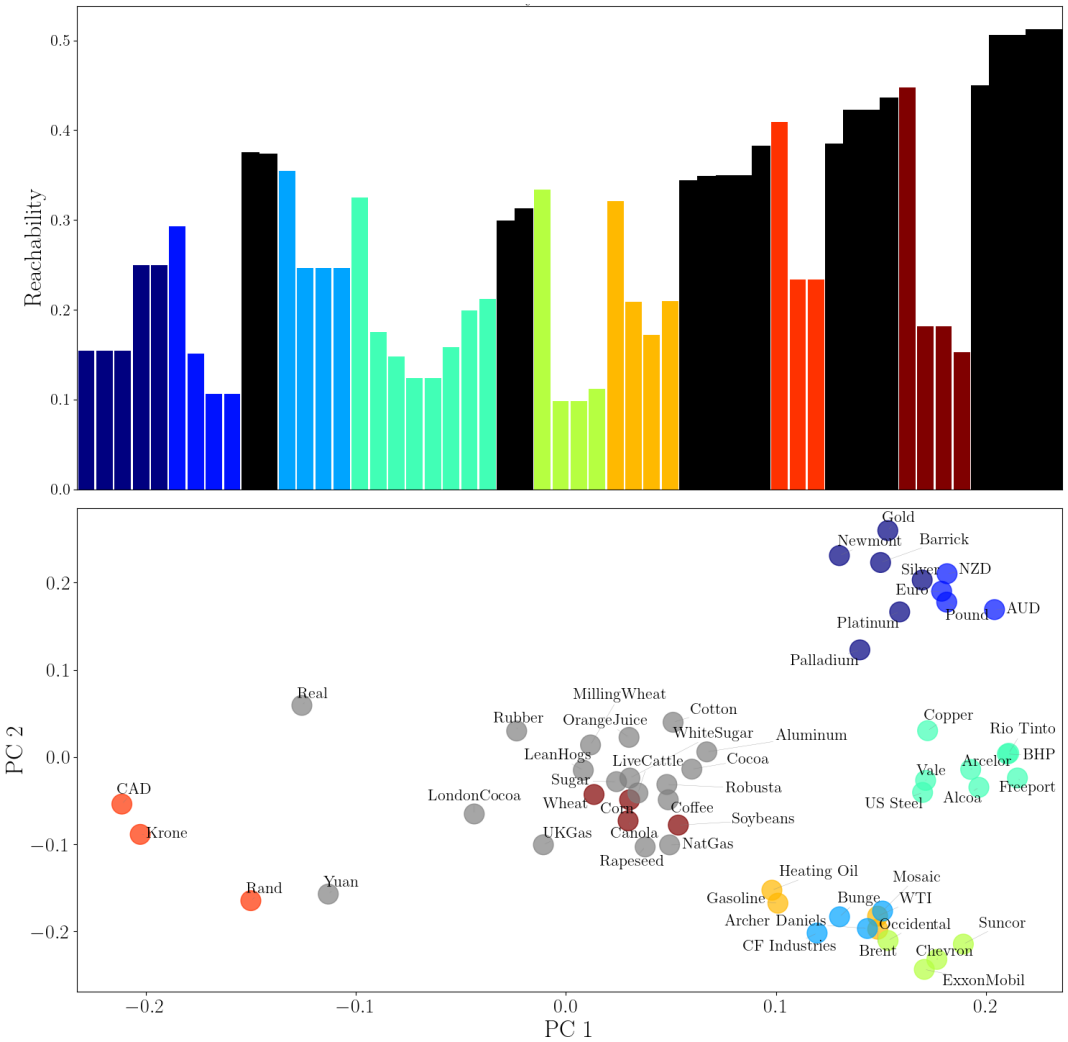

Figure 3.7: OPTICS Reachability Plot for Commodities and Related Assets. The reachability plot (Top) highlights clusters based on varying density, while the PCA scatter plot (Bottom) shows the relationships between assets with regards to the first two principal components.

This process was carried out on a 252 day rolling window basis. A minimum of 3 samples per cluster was specified, and an iterative process was conducted to determine the most suitable ε-distance threshold. A maximum size of 8 was also set for the clusters in order to avoid overhead in later steps of this analysis. Through a visual inspection of a number of reachability plots, an ε of 0.25 was identified as providing the most effective clustering results. This value gave balance between detecting meaningful clusters and minimizing the amount of noise. An example for the reachability plot with $\varepsilon = 0.25$ is shown in figure 3.7. A two-dimensional representation on the first 2 principal components was generated for each rolling iteration, as illustrated in the figure above. In addition to the PCA-based visualizations as shown in figure 3.7, two-dimensional plots were also created using 2 dimensional t-SNE. Over the entire simulation, using $\varepsilon = 0.25$, a range of approximately 3 to 7 simultaneous clusters were consistently identified at various points in time. An example of one of these clusters is given in figure 3.8

Figure 3.8: An example of a cluster containing Gold, Silver, Barrick, and Newmont, which was formed on 03/09/2023

3.7 Modeling Residuals and Spreads

Up until this stage of the analysis, residuals have been derived through two methods.

1. *PCA-factor based portfolio*: Residuals $\tilde{R}_i$ for each asset were obtained by excluding the influence of the first m principal components from the returns R_i, as shown in Equation (3.13). These residuals represented the idiosyncratic component of each asset's returns with respect to the principal components. This captured the movements not explained by the common factors identified through PCA.

2. *OPTICS-based clustering* was used to group assets based on similarities in their PCA loadings. Within each cluster, residuals were constructed through regressing each asset with the synthetic barycenter of the cluster. These spreads represent relative mispricings in the relationships among the closely related assets.

In both cases, as per (Avellaneda and Lee, 2010), for two assets with prices with time series P_t and Q_t, one can represent the co-integration between there two prices with:

$$\ln\left(\frac{P_t}{P_0}\right) = \alpha(t - t_0) + \beta \ln\left(\frac{Q_t}{Q_0}\right) + X_t \tag{3.17}$$

where α captures the time drift specific to asset P with respect to asset Q, whilst β gives the sensitivity of the relationship of asset P with asset Q. In it's differential form, this could be represented as

$$\frac{dP_t}{P_t} = \alpha dt + \beta \frac{dQ_t}{Q_t} + dX_t \tag{3.18}$$

such that X_t is the stationary residual process. This will be revisited in section 3.7.2

For case (1); P_t could be interpreted as the price of an asset with regards to the sum of a number of m PCA factors, $\sum_{j=1}^{m} \beta_j F_t^{(j)}$, where in case (2) for the clustering-based strategy, these factors $F_t^{(j)}$ could be interpreted as the projection of each asset which is not P, on the mean of the cluster. such that:

$$dX_t = \frac{dP_t}{P_t} - \alpha dt + \sum_{j=1}^{m} \beta_j F_t^{(j)} \tag{3.19}$$

where for this exercise it was assumed that $\alpha = 0$ since P and Q are co-integrated.

The next step of the process was to go through all the empirical residuals and determine their suitability for being used for generating trading signals based on their stationarity and mean-reverting behavior.

3.7.1 Augmented Dickey-Fuller (ADF) testing

To assess whether the residuals were stationary, the ADF test was employed. The ADF test determines the presence of a unit root within a time series and in this case it could be formulated as:

$$\Delta \tilde{R}_t = \alpha + \beta \tilde{R}_{t-1} + \sum_{k=1}^{p} \phi_k \Delta \tilde{R}_{t-k} + \epsilon_t \tag{3.20}$$

In this regression equation, $\Delta \tilde{R}_t = \tilde{R}_t - \tilde{R}_{t-1}$ represents the first difference derived from the residuals. The constant term α captures any deterministic trend, while β is the coefficient associated with the lagged level of the residual series $\tilde{R}_{t-1}$. The coefficients ϕ_k account for the autoregressive lagged differences $\Delta \tilde{R}_{t-k}$, and ϵ_t is the white noise error term representing any random deviations.

The hypotheses tested through the ADF test are:

- H_0: The time series contains a unit root, indicating non-stationarity ($\beta = 0$).
- H_1: The time series is stationary ($\beta < 0$).

The ADF test was applied to each residual series $\tilde{R}_i$ generated through both the PCA and also the OPTICS approach. Residuals and spreads with p-values less than a significance level $\alpha = 0.05$ were considered stationary and were shortlisted for the next stage.

3.7.2 Modeling Mean-Reverting Behavior: Ornstein-Uhlenbeck (OU) Process

For the residuals X_t identified as stationary via the Augmented Dickey-Fuller (ADF) test, their dynamics were modeled using the Ornstein-Uhlenbeck (OU) process, consistent with the approach taken by Avellaneda and Lee (2010). The OU process is a continuous-time stochastic process exhibiting mean-reverting behavior, thus it can be defined by the stochastic differential equation:

$$dX_t = \kappa(\mu - X_t)dt + \sigma dW_t, \tag{3.21}$$

where $\kappa > 0$ is the rate of mean reversion, μ corresponds to the long-term mean, $\sigma > 0$ pertains to the volatility, and dW_t represents the increment of a standard Wiener process. This stochastic differential equation could be discretized using the Euler Maruyama method (Kloeden et al., 1992), as follows:

$$X_{t+\Delta t} = X_t + \kappa(\mu - X_t)\Delta t + \sigma\sqrt{\Delta t}\epsilon_t, \tag{3.22}$$

where $\epsilon_t \sim \mathcal{N}(0, 1)$.

3.7.3 Selection Criteria

After estimating the OU process parameters for each residual series using MLE, selection criteria were applied to identify suitable trading strategies based on mean-reversion. The criteria used were:

- **Statistical Stationarity**: Only residuals that passed the Augmented Dickey-Fuller (ADF) test at a significance level of $\alpha = 0.05$ were considered, ensuring the series was stationary.
- **Mean-Reversion Rate**: Residuals with a high rate of mean reversion κ were preferred, indicating quicker reversion to the mean after deviations which would imply that trades could be entered and exit at faster periods.

3.8 Back-testing

3.8.1 Trading Signal Generation

Once the residuals and were identified as mean-reverting with a sufficient level of κ, the next step was to generate trading signals based on the deviations of these series from their equilibrium levels. Trading signals were generated by calculating the standardized scores of the residuals and spreads, known as S-scores, and setting thresholds to trigger long or short trades.

3.8.2 S-Score Calculation

The S-score was calculated to quantify how far the residuals or spreads deviate from their long-term mean, expressed in terms of standard deviations. For any residual

series $\tilde{R}_i$ or spread $S_{ij}(t)$, the S-score at time t is defined as:

$$S(t) = \frac{X_t - \mu}{\sigma},$$

where X_t is the value of the residual or spread at time t, μ is level of the long-term mean estimated through the OU process and σ is the volatility parameter of the OU process. This score allowed for a normalized comparison of the residual relative to its historical behavior.

3.8.3 Thresholds for Opening Long and Short Positions.

Trading rules were established based on the calculated S-scores. The core idea is to take a long position when the residual or spread is a defined level of distance below the mean, and to take a short position when the residual or spread is a defined level of distance above the mean.

The following thresholds were used to trigger trading signals: a long position was entered when $S(t) < -1.75$, indicating that the residual or spread is 1.75 standard deviations below its mean. Conversely, a short position was entered when $S(t) > 1.75$, indicating that the residual or spread is 1.75 SDs over it's mean. The expectation is that the series will revert to its mean, generating profit in both cases. Positions were closed when the S-score crosses its mean, μ. An example is shown below in Figure 4.4, where the standardized residuals for a cluster which was identified between September of 2023 and 2024 are shown. This cluster consisted of four assets: Gold, Silver, Barrick, and Newmont. Their standardized residuals are shown fluctuating around the mean. The red lines indicate the upper and lower thresholds of $S(t) = 1.75$ and $S(t) = -1.75$, which were used to trigger trading signals.

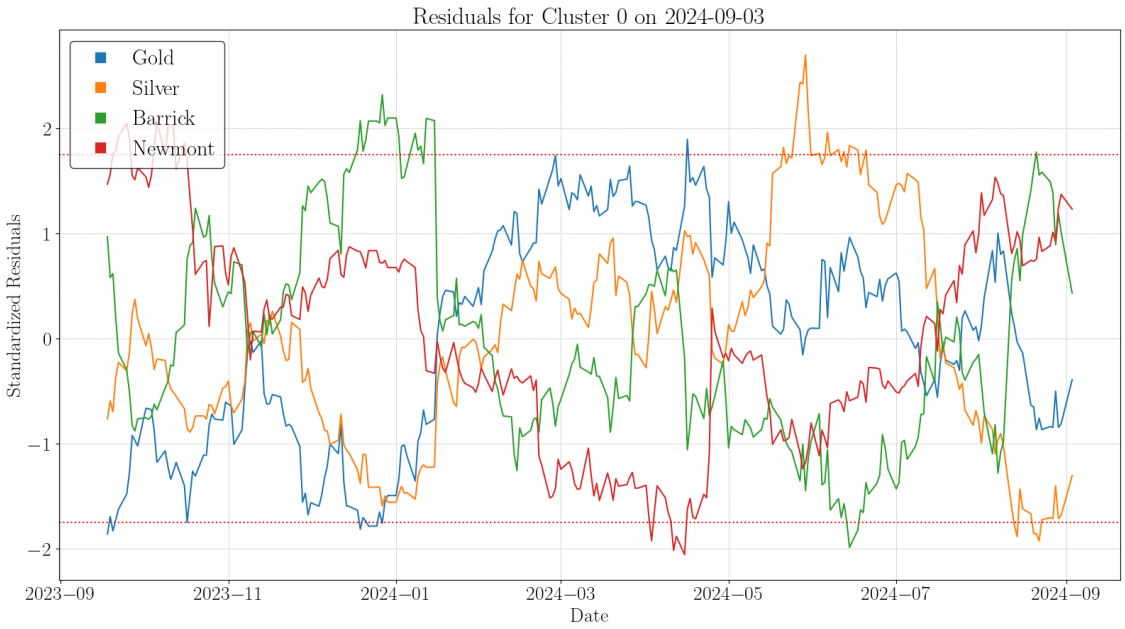

Figure 3.9: Standardized residuals for a cluster consisting of Gold, Silver, Barrick, and Newmont identified between September 2023 and 2024. The residuals fluctuate around the mean, with red lines marking the thresholds used to trigger long and short trading signals.

For example, when the residual of Gold *(blue line)* fell below the lower threshold of $S(t) = -1.75$ at the end of 2023, a long position was taken as explained in Section 3.8.4. Similarly, when the residual for Silver *(orange line)* rose above the upper threshold of $S(t) = 1.75$ in mid-2024, a short position was triggered. Positions were closed when the residuals crossed back to the mean, as seen in several cases where the lines reverted to the zero level.

The thresholds of ± 1.75 were pre-determined, and no optimization was conducted with regard to profits. This approach mirrors those found in similar studies, such as Gatev et al. (2006), where fixed thresholds were applied to avoid overfitting to historical data.

3.8.4 Implementing the Positions

To maintain market neutrality, the portfolio positions were constructed in such a way that no directional exposure systematic factors was present. This was achieved by adjusting the long and short positions as per the trading signals and corresponding factor loadings.

In the event of a long signal for an asset P, a long position of 1 unit was taken in P, while an offsetting short position of 1 unit was allocated across the factors F_j according to the relative β coefficients. For the PCA-based strategy, this involved going short on each principal component in proportion to the β_j values, ensuring that the sum of the short positions across the components equaled 1 unit. In contrast, for the OPTICS-based clustering strategy, the short position was distributed as 1 unit across the mean of the remaining assets in the cluster (excluding P), with the allocation based on the respective β_j values. In the case of a short signal, the inverse was implemented: a short position was taken in P, and the offsetting long position was distributed either across the principal components (in the PCA strategy) or across the mean of the cluster (in the OPTICS-based strategy) as per the same β allocations.

3.9 Performance Evaluation

Evaluating the performance of trading strategies is crucial for determining their viability and effectiveness. In this regard, the return series for an asset can be considered a complete and scale-free metric of investment performance. Furthermore, a useful property of returns is that their multiplication gives the return over a longer period, as shown below:

$$1 + R^{(t)} = \frac{P_t}{P_{t-1}} = \prod_{j=0}^{t-1} \frac{P_{t-j}}{P_{t-j-1}} = \prod_{j=0}^{t-1} (1 + R_{t-j}) \tag{3.23}$$

thus the cumulative return over a period t can be decomposed into the product of individual returns over smaller intervals. Building on this, the equity Curve for a trading strategy can be calculated by compounding the daily returns. The cumulative equity over time T is given by:

$$\text{Equity}_T = \prod_{t=1}^{T} (1 + \text{Position}_t \times R_t) \tag{3.24}$$

Several metrics were computed in order to evaluate and compare the performance of the OPTICS and PCA strategies. These metrics included the total return, annualized return, annualized Volatility, maximum drawdown, the Calmar ratio, and the cumulative excess return. To start off, the total return simply measures the overall growth of the investment over the entire period:

$$\text{Total Return} = (\text{Equity}_{Final} - 1) \times 100\% \tag{3.25}$$

Next the annualized return was computed in order to standardize the total return over a year, allowing comparison computed over time frames of different lengths:

$$\text{Annualized Return} = (\text{Equity}_{\text{Final}})^{\left(\frac{252}{N}\right)} - 1 \tag{3.26}$$

where N is the number of trading days. Similarly the annualized volatility was also computed as this represented an annualised metric for the volatility of the returns,

$$\text{Annualized Volatility} = \sigma_{\text{daily}} \times \sqrt{252} \tag{3.27}$$

which helped in quantifying the inherent risk by measuring the variability of the returns. Additionally the maximum draw down was worked out in order to quantify the largest peak to trough decline in the equity curve:

$$\text{Maximum Drawdown} = \max\left(\frac{\text{Peak} - \text{Trough}}{\text{Peak}}\right) \tag{3.28}$$

This metric highlighted the worst-case scenario in terms of capital loss over the period under consideration. The Calmar Ratio was also worked out on order to quantify the risk-adjusted return as a function of both the annualised return and also the maximum draw-down, as shown below;

$$\text{Calmar Ratio} = \frac{\text{Annualized Return}}{\text{Maximum Drawdown}} \tag{3.29}$$

This gave an indication as to how well the strategy compensated for the the degree of risk taken, with higher values corresponding to better risk-adjusted performance. The cumulative excess return was also calculated in order to aggregate return in excess of a benchmark index as follows;

$$\text{Cumulative Excess Return} = \prod_{t=1}^{T}(1 + R_{\text{Strategy},t} - R_{\text{Benchmark},t}) - 1 \tag{3.30}$$

Chapter 4

Results & Discussion

4.1 Overview

This chapter presents the outcomes and evaluation of the proposed statistical arbitrage strategies based on both the PCA and OPTICS approaches. To maintain clarity and coherence, the results and discussions are interwoven throughout. The first part of this chapter will explore the resulting structure derived through the PCA for the asset universe considered. This gives an insight into the relationships between the asset classes and assets which were considered. Following this, the characteristics of the nature and the frequency of the resulting clusters formed during the formation phase are presented and discussed. The parameterisation of the residuals is then examined, where a number of salient insights arise. Following this the performance of both of the trading strategies, driven by the residuals from these methods, is evaluated using a set of metrics. The discussion explores the implications of the empirical findings on the performance of the both of the strategies, highlighting key points of comparison between the PCA and OPTICS driven approaches.

4.2 PCA Results

As mentioned in the methodology, through a parallel analysis over the entire time span of the dataset, it was determined that considering 10 components for the PCA would provide the right balance between capturing as much of the variance of the dataset while having eigenvectors that provide information beyond random noise.

4.2.1 Eigenvectors

Figure 4.1 displays the coefficients for the assets within the studied universe on the first three principal components (PCs), represented as a bar each principal components. The coefficients indicate the contribution of each asset to the respective eigenvector.

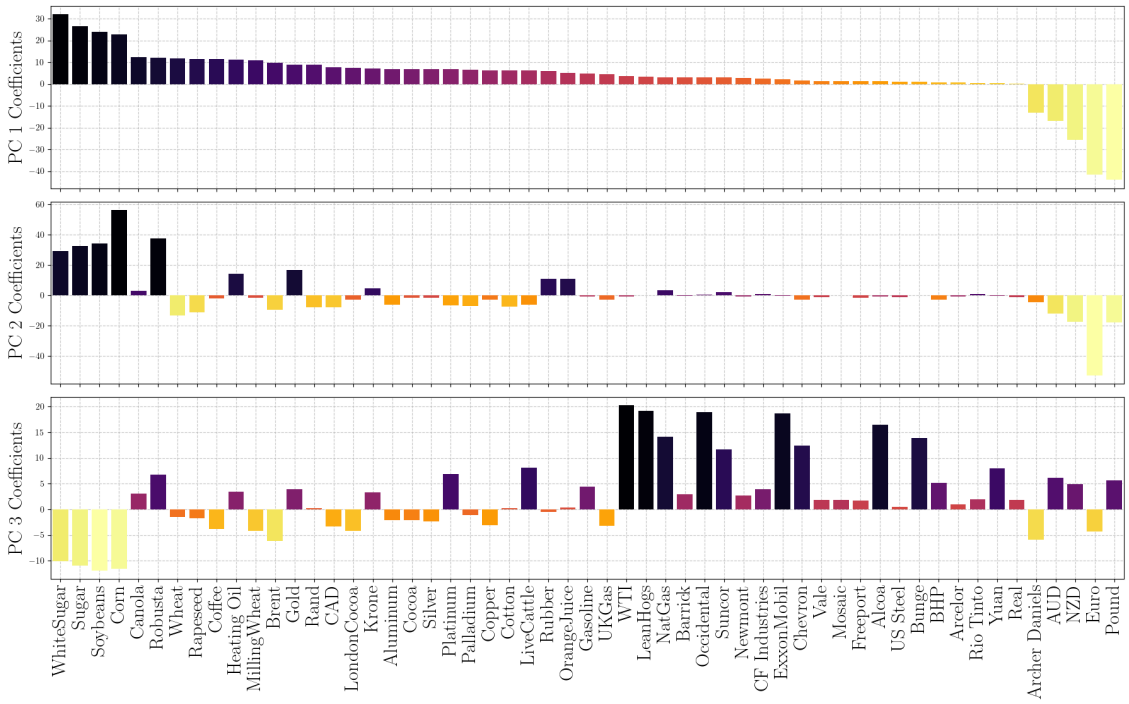

Figure 4.1: The coefficient for each asset on the first 3 eigenvectors sorted by the magnitude of the coefficient for the first eigenvector.

The first principal component (PC1) captured 22.54 per cent of the variance in the dataset as shown in table 4.1. The majority of the assets have a positive contribution towards the first eigenvector. This fits well with the notion that the first eigenvector represents a *'market'* influence that is common throughout most of the assets (Plerou et al., 2002). While this common influence is well documented for stock markets, this result shows that even for commodities spanning multiple asset categories, such a common factor could be deduced. Beyond the first principal component, the second and third components, which capture variances which are orthagonal the first and second components respectively, show a distribution of coefficients which has a more balanced mix of positive and negative contributions compared to PC1. This could indicate that these eigenvectors are capturing variances which are asset group, or industry specific. The coefficients for the remaining seven eigenvectors are provided in the appendix.

PC1	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**	**PC7**	**PC8**	**PC9**	**PC10**
0.2254	0.0715	0.0439	0.0358	0.0350	0.0323	0.0279	0.0264	0.0236	0.0216

Table 4.1: Explained Variance by each Principal Component

The concept of 'coherence' described in (Avellaneda and Lee, 2010) is also evident in these results. This is the notion that assets in similar industries have coefficients which are similar to those assets within the same categories and industries. For instance, large positive coefficients are observed for a set of soft commodities in the first eigenvector, while negative coefficients are found for a group of currencies *(the Pound, Euro, & Antipodean Dollars)*, indicating distinct group-specific movements.

4.2.2 Insights into the Asset Universe Structure

Focusing on the first two dimensions, which capture the majority of the dataset's variance, allows for a representative visualization of the relationships and proximities between different assets and their respective groups. The left panel of 4.2 showcases the projection onto the first two principal components derived from PCA, while the right panel presents the same data transformed using t-SNE.

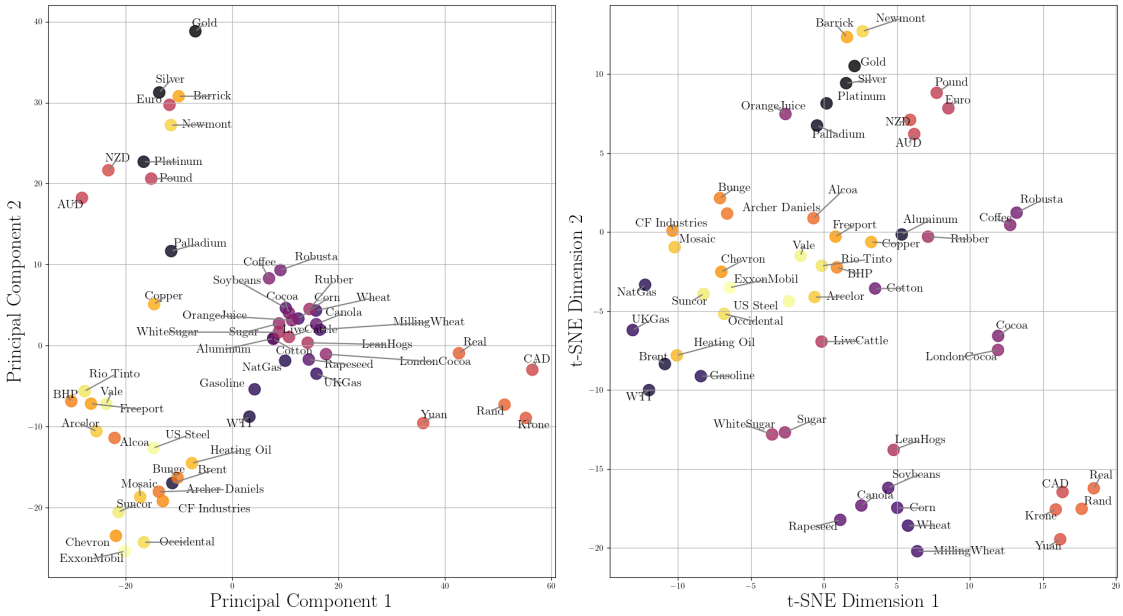

Figure 4.2: 2D representation of the universe, by the first 2 principal components, (Left), and first 2 t-SNE dimensions, (Right).

The PCA representation reveals broad clustering, with the grouping of assets indicating that precious metals and certain currencies display some relationship, whilst energy-related futures and stocks also forming cluster on the opposite side of the second component. A broad clustering of various soft commodities is also evident. It is important to note that further differentiation between the assets would be represented by other principal components, which are not captured in this plot.

In this regard, t-SNE captured with more granularity the relationship between the assets. This is due to the fact the the t-SNE is not constrained to linear projections, thus this allowed for non-linear transformations that captured further relationships between the assets. This is evident for example with similar behavior shown between white sugar no.5 and sugar, coffee and robusta, and US Cocoa and it's London equivalent.

Other distinctions are also made by the t-SNE representation where for example a pattern is evident in the case of precious metals, where gold, silver, palladium, and platinum are tightly clustered alongside precious metal mining stocks like Newmont and Barrick. The slight offset of mining stocks, however, indicates that while correlated, these assets are subject to additional company-specific influences. A similar dynamic is observed within the industrial metals cluster, where Aluminum and Copper group closely with companies like Freeport, Rio Tinto, and BHP, showcasing the interconnectedness between base metal prices and the performance of firms that are involved in mining these metals. A distinct grouping of grains and soft commodities is also apparent, along with a grouping for currencies, where it is notable that two separate clusters emerge, reflecting differing relationships and behaviors within the currency group.

4.3 Clustering Results

The clustering results obtained using the OPTICS algorithm on the PCA factor loadings are presented in this section. This clustering was done using $m = 10$ factors for the PCA, therefore one can interpret this as a segmentation on this more compact representation of the chosen asset universe. These results were obtained by applying the methodology shown in Figure 3.7 on a rolling window basis using the parameters $\epsilon = 0.25$ and min_samples $= 2$.

4.3.1 Cluster Sizing

Given that the OPTICS algorithm does note explicitly define the number of members in a cluster, varying cluster sizes were obtained throughout the analysis. Clusters with less then 3 members were excluded from this analysis since these would either represent the case of pairs trading for a size of 2, or not exist for any size < 2.

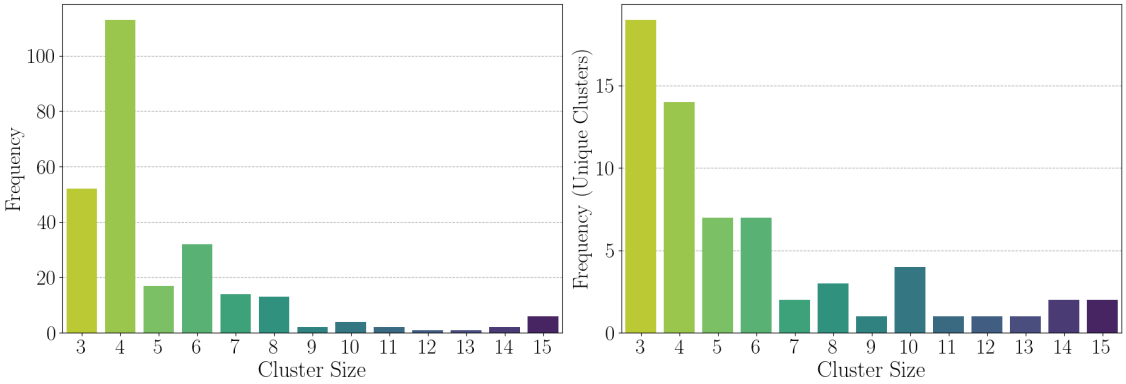

Figure 4.3: *Left*: Frequency Distribution of Cluster Size Obtained. *Right*: Distribution of Unique Clusters by Size.

As shown in Figure 4.3, cluster sizes ranged from 3 to 15. The most frequent cluster size was 4, appearing 113 times. Approximately 93.5 per cent of the clusters had sizes of 8 or less indicating a tendency towards smaller groupings, which is expected.

In rare cases, clusters as large as 15 members were observed, but these were outliers, appearing only 6 times and accounting for less than 2.3 per cent of the total cluster instances. Such large clusters were generally not persistent and as such, in order to ensure robustness, these large clusters were not included in the main strategy.

4.3.2 Observed Clusters

A total of 259 cluster instances were observed. A number of these cluster instances would contain a similar cluster with 1 or 2 asset substituted. In order to get a representative indication of the resulting groupings, a sample of the most frequent cluster from each broad grouping was enumerated as shown in figure 4.4.

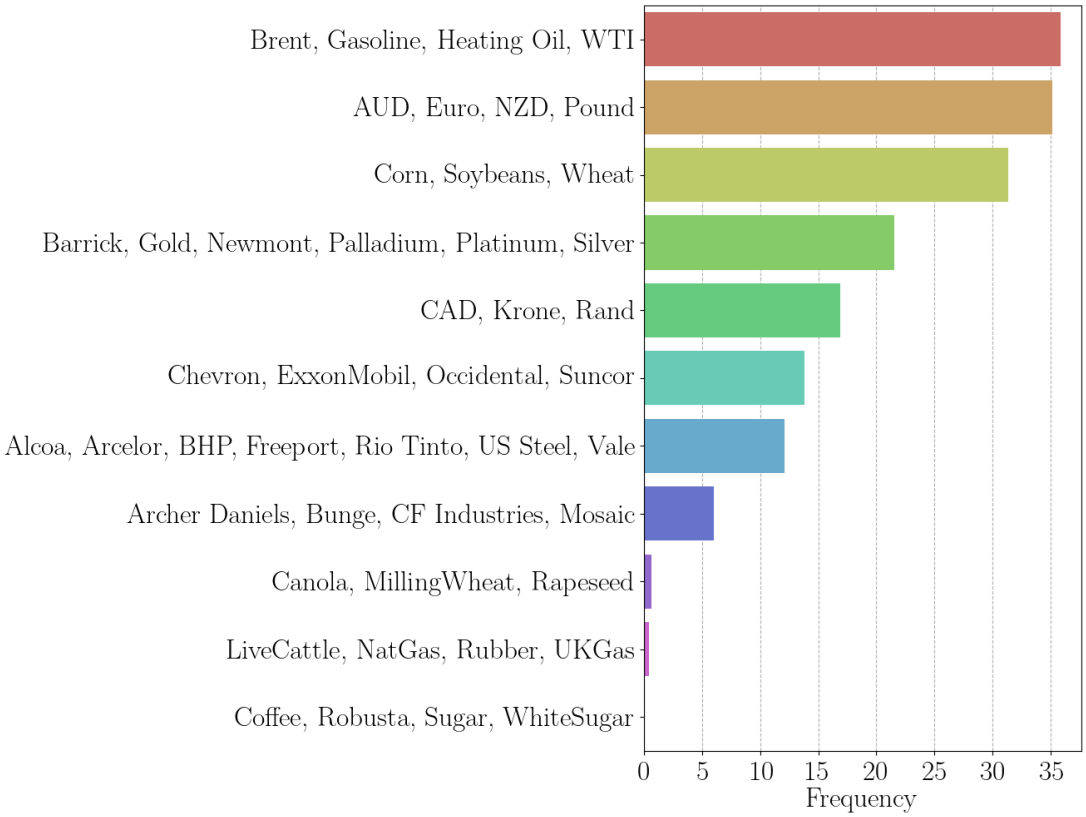

Figure 4.4: Unique cluster groupings for different assets sorted by frequency.

It is evident that the most frequently observed clusters contained assets which are

within the same expected fundamental sector, which is indicative that the PCA compact representation did not loose any important information about the assets. An interesting observation from figure 4.4 is that soft commodities in general tended to cluster less frequently than hard commodities, energy commodities and currencies. This might suggest that soft commodities may exhibit more idiosyncratic price movements due to localized supply and demand factors *and/or* seasonal variations, which might make them less susceptible to general macro-economic market drivers.

The most frequent cluster, comprising Brent, Gasoline, Heating Oil, and WTI, captured a strong and consistent correlation between these oil based commodity futures, which is expected and can be attributed to shared supply and demand factors, as well as geopolitical influences affecting the global oil market,(Nakajima, 2019), (Fanelli, 2024). The second most frequent cluster included the AUD, Euro, NZD, and Pound, forming a basket of G-10 currencies. The close relationship between the AUD and NZD is well-established (Smyth, 2009) (Stephens, 2007), whilst the inclusion of the Euro and Pound may reflect the impact of a common underlying dollar factor, as all the currency pairs in this analysis are measured relative to the US dollar.

The cluster comprising Barrick, Gold, Newmont, Palladium, Platinum, and Silver highlights a distinct grouping of precious metals alongside two major mining companies. Barrick and Newmont are primarily gold miners but also extract other metals such as copper, silver, and zinc. The relationship between these mining companies and precious metals is expected and has been well-documented in the literature (Parrey et al., 2024), reflecting shared exposure to precious metal price movements and similar sensitivities to macroeconomic factors.

The largest cluster comprising soft commodities included Corn, Soybeans, and Wheat. The relationship between corn and soybeans is well expected given their shared seasonal harvesting cycles (Sørensen, 2002). Additionally, both crops often compete for the same agricultural land, which strengthens their correlation. Evidence of factor spillover between wheat and corn has also been demonstrated, as these grains are not only influenced by similar supply and demand shocks but also respond similarly to global agricultural policies and weather patterns (Tonin et al., 2020).

Overall, the results obtained from this clustering make sense from a fundamental perspective. This clustering analysis provided a distinct and appropriate compacted representation of the assets, which could be used in order to create clusters which could be modeled to produce residuals from which trading opportunities could be derived.

4.4 Back-Testing Results

The back testing for both the PCA and the OPTICS based strategies was carried out on a rolling window basis such that for each date considered, a window which was set to 1 year was considered for generating the trading positions of that day. The PCA and OPTICS clustering were recalculated every 60 days as per the methodology. This procedure was carried out over 2897 iterations and yielded the results which will be discussed below.

4.4.1 Empirical Characteristics of the OU Parameterization of Residuals

A number of residuals were obtained throughout the back testing. For example, in the case of the PCA, a residual was generated between each asset and the risk factors over the iterations. This yielded a total of 159,335 residuals, each parameterized as an Ornstein Uhlenbeck Process as described in the methodology. For each residual, this gave respective θ, μ, and σ values. An ADF test was also carried out for each residual, providing an associated p-value.

The p-value and the characteristic time of mean reversion, $\tau = 1/\theta$, are of particular interest since they were used to determine which residuals are used for trading. Their resulting distributions are provided in Figure 4.5, where the cut-offs for selecting the residuals are indicated with the red dotted lines.

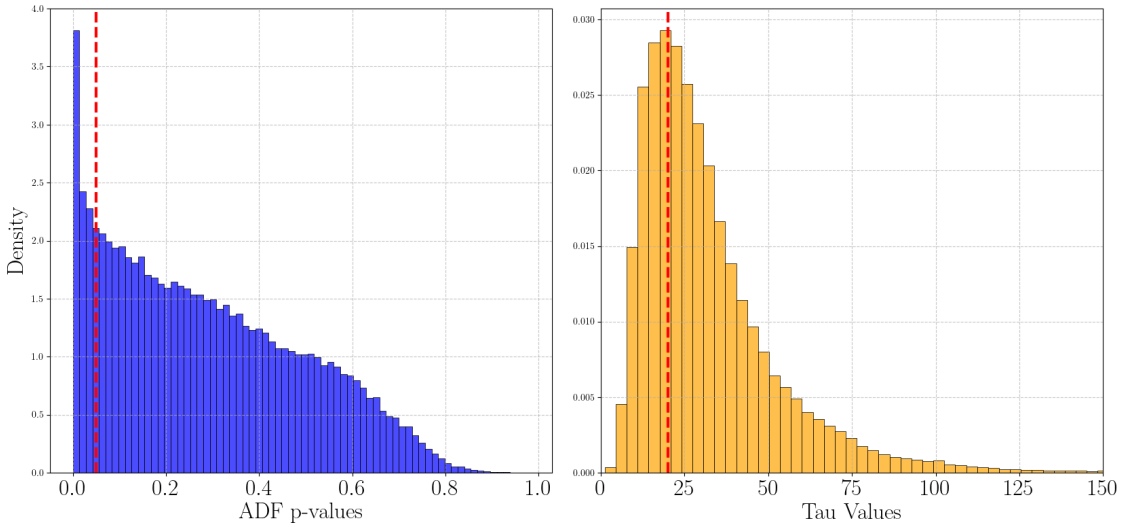

Figure 4.5: Left: Distribution of the p-values of the ADF test. Right: Distribution of the observed OU-τ values.

The distribution of the characteristic time of mean reversion, τ, shows a concentration of values around the lower end of the scale. There is a significant right skew, with most of the residuals having a characteristic time of less than 50 days. There is also a tail of values that extends up to a maximum of 252 days, which is the window size considered. The chosen cutoff of 20 days includes 31.35 per cent of the residuals and happens to be very close to the mode of the distribution.

It is also evident that the level of significance chosen, 95 per cent, shortlisted only a small portion of the total residuals. Specifically, only 13.63 per cent of the residuals had p-values below the 0.05 threshold, indicating stationarity at this level of confidence. There is a gradual increase in the proportion of residuals with lower p-values, with a distinctly higher likelihood of residuals having p-values at small levels close to zero.

In total, when applying the cutoffs for both τ and the ADF, 12.87 per cent of the

total residuals were used for trading. This is quite close to the number of residuals with a valid degree of stationarity, since as evident in Figure 4.6, there is a direct relationship between the degree of stationarity and the time of mean reversion of the residual.

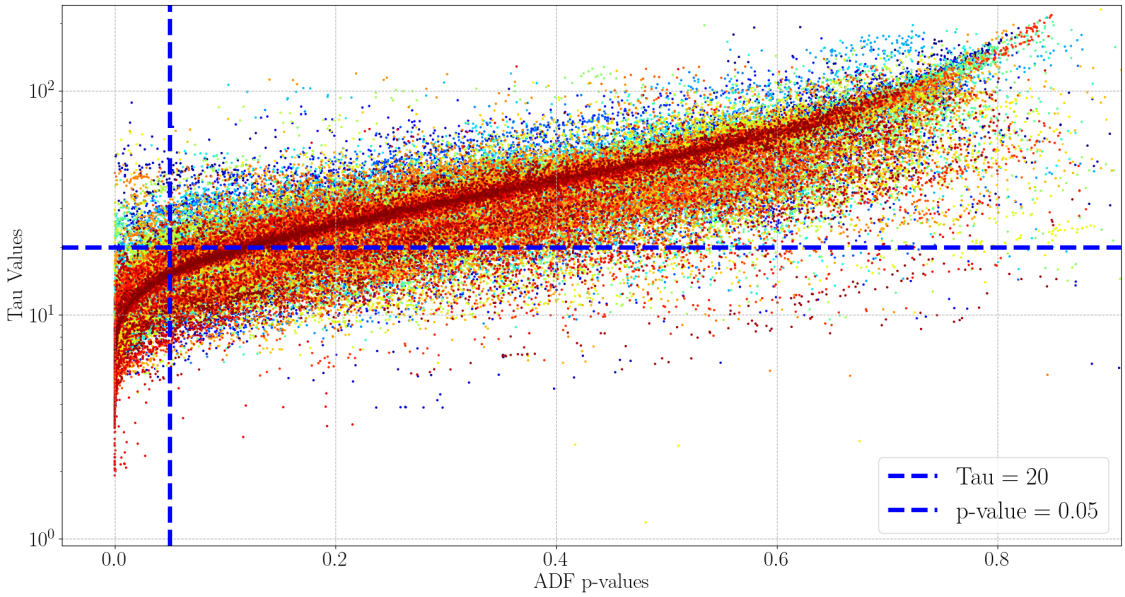

Figure 4.6: Scatter plot showing the distribution of the mean-reversion parameter, τ, versus the ADF p-value for each residual. The color gradient ranges from blue (earlier in time) to red (later in time).

In fact, it is evident that there is a clear trend where the characteristic time is lower for residuals with lower p-values. Therefore, by selecting series based on satisfactory p-values, one is also indirectly filtering for residuals that exhibit faster mean reversion. Nevertheless, there are still a small number of residuals *(which fall in the top left quadrant of Figure 4.6)* that, despite being stationary, do not revert to the mean quickly enough. This highlights the importance of filtering based on both criteria to ensure the residuals used for trading are both stationary and exhibit fast mean reversion.

The OU parameterisation also resulted in values for σ, the diffusion coefficient as per Equation 3.21. Figure 4.7 shows the volatility, as per σ, as it varied throughout time.

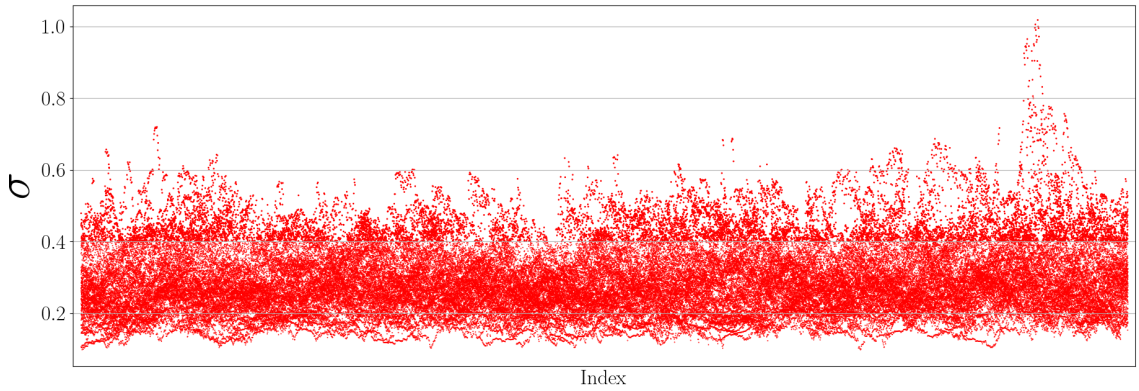

Figure 4.7: Ornstein Uhlenbeck σ throughout time.

It is clear from Figure 4.7 that the majority of the values for σ fell between 0.1 and 0.6. This is a reasonable level of volatility. There are also hints of clustering of the σ term, with a relatively large concentration of high σ residuals occurring towards the end of the simulation. It is noteworthy that this period of high σ value.

4.4.2 Strategy Performance

The summary of the results for each strategy, both with and without transaction costs, is presented in Table 4.2. For reference, the performance of the Bloomberg Commodity Index (BCOM) over the same period is also included. The BCOM is a broad-based index that tracks the performance of 23 commodities, spanning the energy, metals, and agricultural sectors, with weightings based on their economic significance (Shahzad et al., 2022). It is important to note that while the BCOM is provided for comparison, caution should be exercised in making direct comparisons between this basket and the performance of the strategies. This is because the BCOM represents a passive investment in a broad set of commodities, while the

strategies in question are based on active trading, which have different risk-return characteristics and objectives.

From these results, it is evident that both strategies yield positive yet modest returns throughout the backtesting period. In terms of the magnitude of returns, the OPTICS-based strategy seems to perform better than the solely PCA-based strategy. Additionally, the OPTICS-based strategy exhibits higher volatility, as reflected in the annualized volatility which is 0.47 per cent for OPTICS and 0.21 per cent for the PCA strategy. This increased volatility is also accompanied by a higher maximum drawdown, with OPTICS experiencing a drawdown of 0.68 per cent versus PCA's 0.44 per cent.

Despite the higher volatility, the OPTICS strategy maintains a superior Calmar Ratio of 0.95 compared to PCA's 0.24, indicating a more favorable risk-adjusted return profile. When transaction costs are factored in, all metrics retain a similar profile, and the OPTICS strategy still outperforms the PCA strategy.

	PCA	OPTICS	PCA w/ Cost	OPTICS w/ Cost	PCA w/ Cost post C-19	OPTICS w/ Cost post C-19	Bloomberg Commodity Index
Total Return	1.21%	7.71%	0.66%	7.01%	-0.11%	2.71%	-24.57%
Max Drawdown	0.44%	0.68%	0.48%	0.70%	0.57%	0.40%	57.11%
Annualised Return	0.10%	0.65%	0.06%	0.59%	-0.02%	0.62%	-2.42%
Annualised Volatility	0.21%	0.47%	0.21%	0.47%	0.21%	0.46%	13.92%
Calmar Ratio	0.24	0.95	0.12	0.84	-0.04	1.54	-0.04

Table 4.2: Performance comparison of PCA, OPTICS strategies with costs and post-Covid adjustments against Bloomberg Commodity Index

Results for the strategy performances after the COVID-19 pandemic are also considered. This distinction is made to highlight any possible changes in strategy performance due to shifts in market structure post-COVID-19. The cutoff date for this analysis was set as the second half of 2020 (*starting from July 1, 2020*), which marks the point where the VIX recovered from the COVID-19 shock and returned to its new mean. Since this cutoff date, there is a divergence between the results of both strategies: the OPTICS-based strategy has achieved a total return of 2.71 per cent, while the PCA-based strategy has failed to generate any returns during the same period. This divergence in performance could be observed in Figure 4.8.

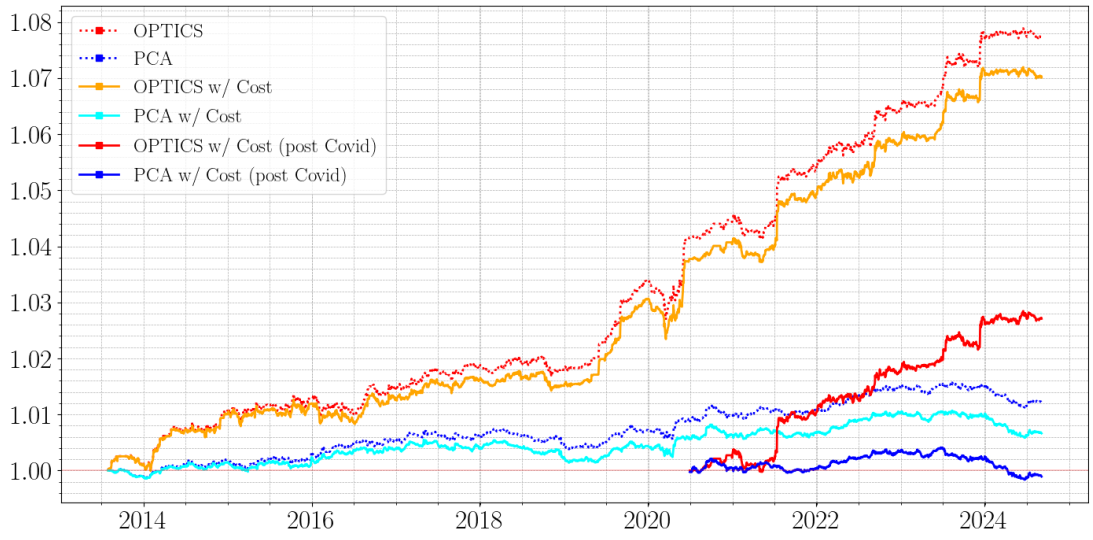

Figure 4.8: The performance of each strategy, with and without costs, over time.

In Figure 4.8, the cumulative performance of the strategies over time is depicted. Both OPTICS and PCA-based strategies exhibit a steady increase in equity throughout the back-testing period, albeit with varying magnitudes. Notably, the OPTICS based strategy consistently outperforms the PCA strategy, especially after 2018, where the gap between the returns of both strategies widens significantly.

In terms of volatility, the OPTICS based strategy exhibits more fluctuations in it's returns than the PCA based strategy, which is consistent with the higher annualized volatility observed. Despite this, the OPTICS strategy delivers a better risk-adjusted return profile when compared to the PCA-based strategy.

It is noteworthy that both strategies maintained stable performance without experiencing significant drawdowns throughout the trading period, including during the COVID-19 pandemic and downturns in specific commodity sectors.

4.4.3 Asset-Wise Attribution of Returns

The modest yet consistent cumulative performance performance of both strategies, could be attributed to the nature of the implemented strategies, were at any point in time there was a number of exposures open to different assets due to the fact that the investment budget was proportionally divided between the signals generated from the different residuals for both that case of the PCA based strategy and also the OPTICS based strategy.

In this regard, the cumulative performance could be broken down on an asset by asset basis in order to interpret better the performance of both the strategies. For the case of the PCA based strategy, at each point in time, the residual between each asset the eigenvectors could be constructed. As such, throughout the trading period, the performance attributable to each asset could be considered as returns generated between each asset and all the eigenvectors through time.

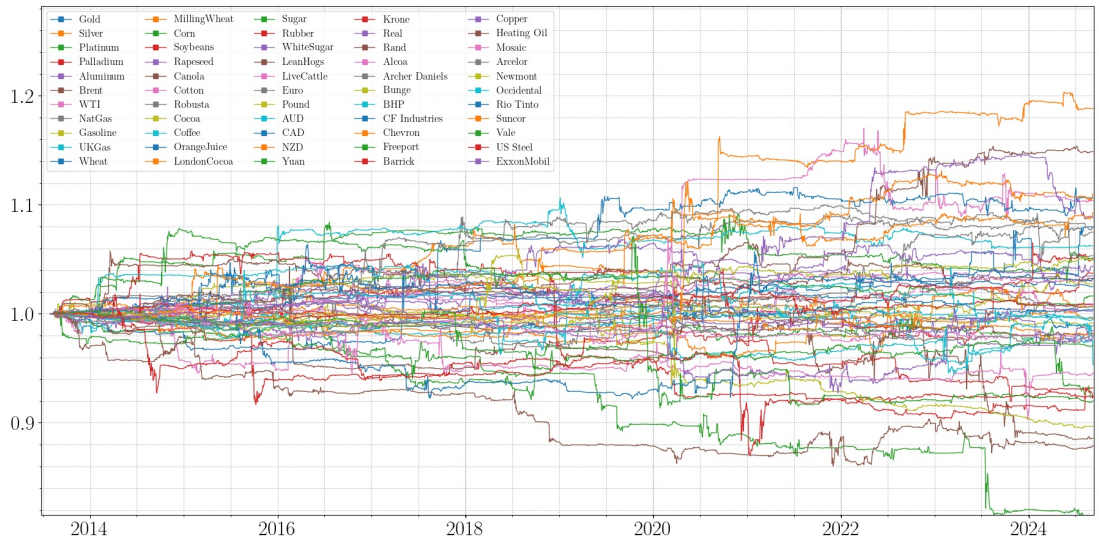

Figure 4.9: PCA based strategy: Returns Attribution by Asset.

Figure 4.9 shows this breakdown of of the returns. It is evident that the outcomes vary across assets, with some generating a total profit through their residuals, while others contributed to a total loss. On average the combination of all these returns yielded the slightly positive total returns which resulted for this strategy.

	Milling Wheat	**Canola**	**Chevron**	**Wheat**	**WTI**	**Suncor**	**Copper**	**Arcelor**	**Orange Juice**	**Robusta**
Total Returns	1.180%	1.141%	1.104%	1.099%	1.095%	1.087%	1.081%	1.073%	1.071%	1.069%

Table 4.3: Top 10 Assets with the highest total return for the PCA strategy.

Half of these assets are agricultural commodities, with two energy companies and WTI Crude Oil, as well as Copper and the Luxembourgish steel manufacturer, Arcelor. This composition suggests that these assets and/or asset groups might exhibit well-behaved mean-reverting residuals over the considered trading period.

Similarly, Figure 4.11 shows the contribution respective to each asset for the OP-

TICS based strategy. Here, the return for each asset represents returns generated through signals from residuals between that asset and any potential cluster mean. There is once again a range of outcomes for each asset, however this time there is noticeably longer periods of flat returns. These periods would corresponds to ranges in time where that particular asset was not part of any cluster.

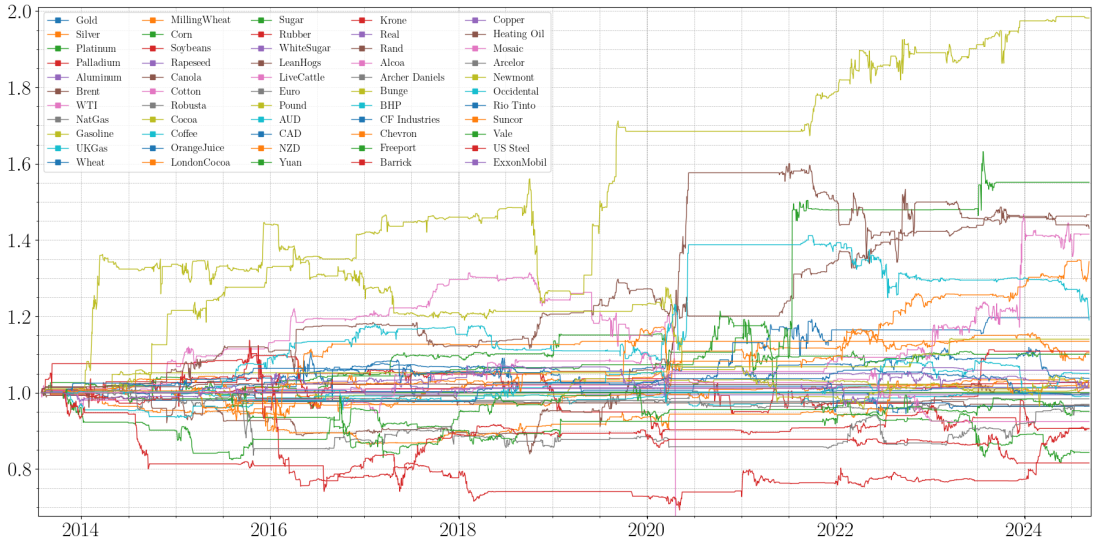

Figure 4.10: Optics based strategy: Returns Attribution by Asset.

In total, there were 15 assets, *Yuan, Aluminum, NatGas, UKGas, Cotton, Robusta, Cocoa, Coffee, OrangeJuice, LondonCocoa, Sugar, Rubber, WhiteSugar, LeanHogs, LiveCattle*, which over the entire trading period were not part of any cluster, and as such had no contribution towards the total final return of the strategy. This time

	Gasoline	**Corn**	**Brent**	**Heating Oil**	**Alcoa**	**Suncor**	**Wheat**	**Occidental**	**Bunge**	**Milling Wheat**
Total Returns	1.95%	1.53%	1.45%	1.40%	1.39%	1.32%	1.18%	1.17%	1.14%	1.13%

Table 4.4: Top 10 Assets with the highest total return for the OPTICS strategy.

Once again, almost half the assets in this list are agricultural commodities, with

Wheat and Milling Wheat notably appearing in the top performers for both the PCA and OPTICS approaches. The other half of this list is made up of energy commodities and companies, as well as the aluminum manufacturer Alcoa.

4.4.4 Discussion

Both of the strategies considered displayed a consistent performance which was resistant to shocks and major downturns. This stability could be attributed to the fact that in both cases, at most points in time, multiple positions were opened in different assets. This made it unlikely for the total holding to experience a net loss. Furthermore. The natural diversification within commodities due to their lower inter-correlation compared to equities, (International Monetary Fund., 2015) further contributes to the total stability in the performance of the strategies.

However, this diversification has also potentially limited the returns from the strategies. The total annualized returns are generally low for both strategies, with the PCA-based strategy achieving an annualized return of 0.10% and the OPTICS-based strategy achieving 0.65%. While diversification reduces the likelihood that all open positions will under-perform simultaneously, it also means that it is also unlikely for all exposures to generate returns at the same time. Consequently, in the long term, especially for the case of the OPTICS-based strategy, the positive returns slightly outweigh the negative ones, leading to small but consistent profits.

According to literature, returns generated through statistical arbitrage should generally exhibit low volatility, be positive, and demonstrate low correlation with broader market returns (Focardi et al., 2016). In the present analysis, both the PCA-based and OPTICS-based strategies align with these characteristics to a reasonable extent.

4.4.4.1 Comparisons with Benchmarks

In the context of investing in commodities, evaluating performance relative to a benchmark is more challenging than for other asset classes such as equities or fixed-income securities.

One of the most recognized broad commodity indices is the Bloomberg Commodity Index (BCOM), which tracks the performance of a variety of globally traded commodity futures, offering a broad measure of the health and trends within commodity markets. It aims to prevent over-concentration in any single commodity or sector. The index includes 23 commodity futures, organized into six sectors (Bloomberg, 2014). To preserve balance, no individual commodity can account for more than 15% of the index, and no single commodity combined with its derivatives can exceed 25%. Furthermore, no sector can represent more than 33% of the index, with these thresholds adjusted during the annual re-weighting process.

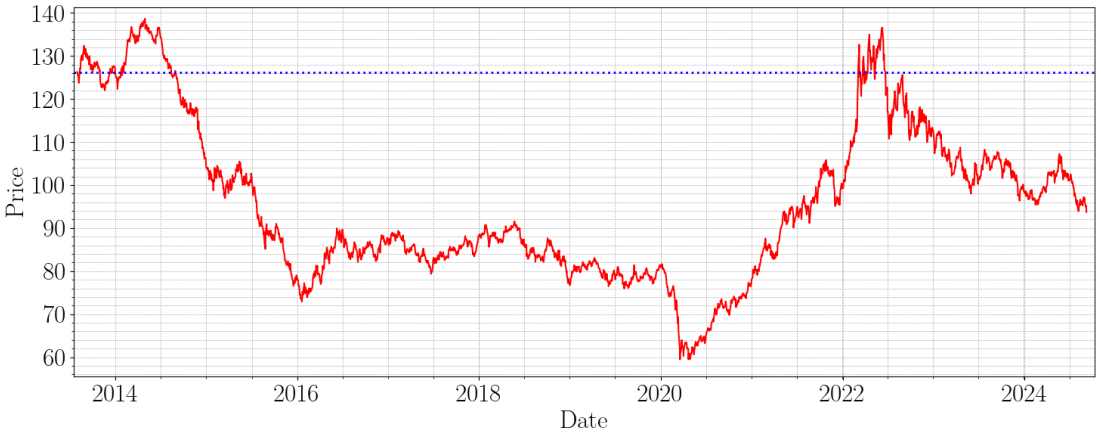

Figure 4.11: The performance of the Bloomberg Commodity Index over the same time horizon as the strategies considered in this paper.

When considering the performance of this index over the same time frame on which the PCA and OPTICS strategies were tested, it is evident that investing in this

index as a passive strategy would have yielded negative returns. In this regard, comparing the performance of the PCA and OPTICS strategies with this benchmark shows that both strategies outperformed. However, such a comparison may not be entirely meaningful due to the nature and scope of broad commodity indices.

This under performance is well documented in the literature. For example, Blocher et al. (2018) notes that indexing commodity futures, particularly through equally weighted indexing, is a passive strategy that is relatively simple to implement. While there are several such indices, passive investing in them has produced negative or flat returns over much of their historical performance, leading many practitioners to gradually move away from this approach.

One potentially useful takeaway when comparing broad commodity indices like the BCOM is the correlation of returns. Statistical arbitrage strategies typically provide returns uncorrelated with market-wide returns. In this case, the returns of both strategies tested satisfy this condition. A linear regression of the strategies' returns, the results of which can be found in the appendix, revealed a low beta relative to the BCOM's returns. A similar regression with the S&P 500 also showed that the returns generated by both strategies exhibited negligible correlation with market-wide returns.

Chapter 5

Conclusion

5.1 Summary of Findings

This analysis evaluated the effectiveness of the strategies both the PCA and OPTICS strategies amongst the considered commodity asset universe. The PCA was employed 10 principal components, were the resulting distribution of eigenvalues was quite standard, with a single dominant eigenvalue which captured the dominant market component. The OPTICS clustering generated a sufficient number of clusters throughout the period considered. Clusters typically ranged from three to seven assets in size, with smaller clusters of around four members being the most prevalent. These clusters often corresponded to the fundamental sectors, including oil-based commodities, precious metals, currencies and various grains. Notably, other soft commodities were less frequently grouped, suggesting that their price movements might be more influenced by localized factors and seasonal variations, making them less susceptible to broad drivers.

Testing of the trading strategies revealed that residuals derived from both PCA and

OPTICS-driven models exhibited characteristics suitable for statistical arbitrage. These residuals were parameterized using an OU process, with a good proportion of the residuals being sufficiently stationary and with a low enough characteristic time, which qualified them for trading strategies. Both strategies yielded small yet consistent returns. Performance comparisons indicated that the OPTICS-based strategy outperformed the PCA-based approach, achieving a total (0.47% annualized compared to PCA's 0.21%), it maintained a superior Calmar Ratio of 0.95 compared to PCA's 0.24, indicating more favorable risk-adjusted returns. Post-July 2020, the OPTICS strategy continued to generate positive returns (2.71%), while the PCA strategy's performance was negligible, underscoring the possible benefits of employing the OPTICS clustering

Asset-wise return attribution further highlighted the difference that there are from asset to asset. There was a degree of sectoral coherence in this regard, were assets from the same sectors had similar return profiles. The most notable returns came from energy-related futures and companies, metal producers, and a variety of agricultural commodities, with grains slightly outperforming the others. The number of positions which were spread across multiple assets and sectors, was a double edged sword, where this both acted as a stabiliser towards significant losses, but also a potential limiter towards the potential for large profits. Furthermore, whilst the OPTICS based clustering benefited from consistent performance, certain assets, especially soft commodities, consistently did not take part in any of the clusters, resulting in no contribution to the overall strategy returns.

5.2 General Limitations to Statistical Arbitrage

The limitations of statistical arbitrage, particularly in pairs trading, are driven by several key factors. As noted by Do and Faff (2010) one of the primary causes of declining profitability in statistical arbitrage implementations is the increasing proportion of non-convergent trades. These non-convergent trades prevent arbitrageurs from closing positions profitably due to the failure of the price spread to revert. Additionally, arbitrage risk plays a significant role in trade failure, where this can encompass a number of risks such as fundamental risk, synchronization risk and noise trader risk. Do and Faff (2010) describe fundamental risk as the possibility of unexpected disruptions to the relative pricing relationship between paired securities, such as company-specific events like the discovery of new assets or strategic changes that can cause lasting divergence. Noise-trader risk refers to irrational market activity by uninformed traders exacerbating the divergence, while synchronization risk arises from uncertainty regarding when arbitrageurs will exploit the mispricing.

Market structure and the flow of information also contribute to limitations in statistical arbitrage. Andrade et al. (2005) highlight how uninformed trading shocks and the slow diffusion of information across asset pairs can delay price convergence, making it difficult for arbitrageurs to exploit the divergence profitably. Idiosyncratic shocks, such as firm-specific events like earnings announcements or strategic investments, can change the factor loadings that drive the returns of a pair, potentially rendering the pair unsuitable for arbitrage (Andrade et al., 2005).

In emerging markets, the limits to arbitrage are particularly significant. Jacobs (2015) found that market structure issues, such as lack of liquidity and trading bar-

riers, can increase risk and limit the effectiveness of statistical arbitrage strategies. Over time, the profitability of pairs trading has been steadily declining, as observed by Rad et al. (2016), especially after 1985, due to increasing market efficiency, competition, and the impact of transaction costs.

Finally, despite the introduction of advanced statistical techniques, such as co-integration and copula based approaches, these methods still suffer from the same fundamental issues, including non-convergence and idiosyncratic shocks (Rad et al., 2016). These factors collectively explain why statistical arbitrage, especially in the case of pairs trading, has faced growing limitations and declining profitability.

5.3 Implementation Specific Limitations & Future Recommendations

There are several limitations associated with this specific implementation. One of these limitations is the relatively small size of the asset universe. While the inclusion of 55 assets was sufficient for generating a meaningful number of positions, with nearly all periods having 3-7 clusters contributing to signals in the OPTICS-based strategy, and multiple exposures open simultaneously in the PCA-based strategy, it is likely that a larger universe of assets could have provided even more robust and diversified opportunities. In studies conducted on equity markets, it is common for authors to include over 100 assets, with some including thousands. These numbers might be harder to achieve in this case due to the more limited range of available commodities, but expanding the universe could possibly improve the results.

Another limitation lies in the absence of optimization regarding both the assets used for trading, and the thresholds used for signal generation. While the lack of optimization reduces the risk of over fitting or data snooping, it may also result in under performance due to the absence of generalizable improvements that could be gained from refining these parameters. Implementing a limited degree of optimization which is supported with sufficient validation could potentially enhance the overall profitability of the strategy.

Additionally in this implementation a constant number of PCA factors was used throughout this exercise. This can be improved upon by considering a dynamic number of factors, similar to the implementation applied by Yeo (2016). This could allow the model to dynamically adapt to shifts in the market structure, leading to potentially better performing generated statistical PCA factors. More advanced trading rules could have also been implemented, such as for example using more than 2 thresholds, where separate thresholds for opening and closing the trades could be defined. Another possibility is that of including stop loss mechanisms in order to improve the risk profile of the strategy. The inclusion of take profits and stop losses could have further refined the strategy.

5.4 Final Remarks

In conclusion, both strategies yielded slight returns which had characteristics of returns typically derived through statistical arbitrage strategies. This means that they were positive, with low volatility, and they had a lack of correlation to broad market returns. The OPTICS-based statistical arbitrage strategy demonstrated superior performance and risk-adjusted returns compared to the approach which

only included the PCA. These returns are attributable to the diversification offered through the number of simultaneously opened positions, which on one side enhanced stability, but also may have limited the potential for higher gains. The asset specific results revealed sectoral patterns, which broadly aligned with the fundamental groups of the assets. Notable contributions to returns were attributable to energy-related assets, metals, and certain agricultural commodities. Future research could explore further refinements to these strategies, both through some degree of optimisation based on robust validation, and also in terms of having a larger asset universe with more data. Additionally, exploring adaptive trading mechanisms, such as dynamic factors and also employing sophisticated trading criteria, could improve upon this approach to statistical arbitrage. Ultimately, this exercise highlights the potential of methodological approaches towards investing, and invites further exploration into the mechanisms underlying relationships between assets.

Bibliography

Alhamazani, K., Ranjan, R., Jayaraman, P. P., Mitra, K., Wang, M., Huang, Z. G., Wang, L. and Rabhi, F. (2014), Real-time qos monitoring for cloud-based big data analytics applications in mobile environments, *in* '2014 IEEE 15th International Conference on Mobile Data Management', Vol. 1, IEEE, pp. 337–340.

Andrade, S., Di Pietro, V. and Seasholes, M. (2005), 'Understanding the profitability of pairs trading', *Unpublished working paper, UC Berkeley, Northwestern University* .

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

Avellaneda, M. and Lee, J. H. (2010), 'Statistical arbitrage in the us equities market', *Quantitative Finance* **10**(7), 761–782.

Barrick Gold (2023), 'Annual report 2023: Driving value, building growth'. Accessed: 2024-08-04.
URL: *https://www.barrick.com/English/investors/annual-report/default.aspx*

Berkhin, P. (2006), A survey of clustering data mining techniques, *in* 'Grouping multidimensional data: Recent advances in clustering', Springer, pp. 25–71.

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

Bhattacharjee, P. and Mitra, P. (2021), 'A survey of density based clustering algorithms', *Frontiers of Computer Science* **15**, 1–27.

Bianchi, R., Drew, M. and Zhu, R. (2009), Pairs trading profits in commodity futures markets, *in* 'Proceedings of Asian Finance Association 2009 International Conference', pp. 1–26.

Blocher, J., Cooper, R. and Molyboga, M. (2018), 'Benchmarking commodity investments', *Journal of Futures Markets* **38**(3), 340–358.

Bloomberg (2014), 'The bloomberg commodity index family - index methodology', PDF. Archived from the original (PDF) on 2015-03-19. Retrieved 2024-08-02. **URL:** *https://www.bloomberg.com*

Caneo, F. and Kristjanpoller, W. (2021), 'Improving statistical arbitrage investment strategy: Evidence from latin american stock markets', *International Journal of Finance & Economics* **26**(3), 4424–4440.

Connor, G. and Korajczyk, R. A. (1986), 'Performance measurement with the arbitrage pricing theory: A new framework for analysis', *Journal of Financial Economics* **15**(3), 373–394.

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

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

Elliott, R. J., Van Der Hoek*, J. and Malcolm, W. P. (2005), 'Pairs trading', *Quantitative Finance* **5**(3), 271–276.

Endres, S. and Stübinger, J. (2019), 'Optimal trading strategies for lévy-driven ornstein–uhlenbeck processes', *Applied Economics* **51**(29), 3153–3169.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X. et al. (1996), A density-based algorithm for discovering clusters in large spatial databases with noise, *in* 'kdd', Vol. 96, pp. 226–231.

Fanelli, V. (2024), 'Mean-reverting statistical arbitrage strategies in crude oil markets', *Risks* **12**(7), 106.

Focardi, S. M., Fabozzi, F. J. and Mitov, I. K. (2016), 'A new approach to statistical arbitrage: Strategies based on dynamic factor models of prices and their performance', *Journal of Banking & Finance* **65**, 134–155.

Gatev, E., Goetzmann, W. N. and Rouwenhorst, K. G. (1999), Pairs trading: Performance of a relative value arbitrage rule, Working paper, Yale School of Management's International Center for Finance.

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

Girma, P. B. and Paulson, A. S. (1999), 'Risk arbitrage opportunities in petroleum futures spreads', *Journal of Futures Markets* **19**(8), 931–955.

Granger, C. W. and Newbold, P. (1974), 'Spurious regressions in econometrics', *Journal of Econometrics* **2**(2), 111–120.

Hain, M., Hess, J. and Uhrig-Homburg, M. (2018), 'Relative value arbitrage in european commodity markets', *Energy Economics* **69**, 140–154.

Han, C., He, Z. and Toh, A. J. W. (2023), 'Pairs trading via unsupervised learning', *European Journal of Operational Research* **307**(2), 929–947.

He, C., Wang, T., Liu, X. and Huang, K. (2023), 'An innovative high-frequency statistical arbitrage in chinese futures market', *Journal of Innovation & Knowledge* **8**(4), 100429.

Hoel, C. H. (2013), Statistical arbitrage pairs: can cointegration capture market neutral profits?, Master's thesis.

Huck, N. (2010), 'Pairs trading and outranking: The multi-step-ahead forecasting case', *European Journal of Operational Research* **207**(3), 1702–1716.

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

Hyndman, R. J. and Athanasopoulos, G. (2018), *Forecasting: principles and practice*, OTexts, Melbourne, Australia.

Iacobucci, D., Ruvio, A., Román, S., Moon, S. and Herr, P. M. (2022), 'How many factors in factor analysis? new insights about parallel analysis with confidence intervals', *Journal of Business Research* **139**, 1026–1043.

International Monetary Fund. (2015), *Global Financial Stability Report, April 2015: Navigating Monetary Policy Challenges and Managing Risks*, International Monetary Fund.

Jacobs, H. (2015), 'What explains the dynamics of 100 anomalies?', *Journal of Banking & Finance* **57**, 65–85.

Jolliffe, I. T. (2002), *Principal component analysis for special types of data*, Springer.

Jurek, J. W. and Yang, H. (2007), Dynamic portfolio selection in arbitrage, *in* 'EFA 2006 meetings paper'.

Kloeden, P. E., Platen, E., Kloeden, P. E. and Platen, E. (1992), *Stochastic differential equations*, Springer.

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

Lazzarino, M., Berrill, J., Šević, A. et al. (2018), 'What is statistical arbitrage?', *Theoretical Economics Letters* **8**(05), 888.

Lettau, M. and Pelger, M. (2020), 'Factors that fit the time series and cross-section of stock returns', *The Review of Financial Studies* **33**(5), 2274–2325.

Lin, Y.-X., McCrae, M. and Gulati, C. (2006), 'Loss protection in pairs trading through minimum profit bounds: A cointegration approach', *Advances in Decision Sciences* **2006**.

Mikkelsen, A. (2018), 'Pairs trading: the case of norwegian seafood companies', *Applied Economics* **50**(3), 303–318.

Morgan, T. (1901), *Regeneration*, Columbia biological series, Macmillan.
URL: *https://books.google.com.mt/books?id=MtYKAAAAIAAJ*

Nakajima, T. (2019), 'Expectations for statistical arbitrage in energy futures markets', *Journal of Risk and Financial Management* **12**(1), 14.

Parrey, Z. A., Dar, A. B. and Paul, M. (2024), 'Revisiting precious metal mining stocks and precious metals as hedge, diversifiers and safe-havens: a multidimen-

sional scaling and wavelet quantile correlation perspective', *Empirical Economics* pp. 1–23.

Plerou, V., Gopikrishnan, P., Rosenow, B., Amaral, L. A. N., Guhr, T. and Stanley, H. E. (2002), 'Random matrix approach to cross correlations in financial data', *Physical Review E* **65**(6), 066126.

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

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

Ross, S. A. (2013), The arbitrage theory of capital asset pricing, *in* 'Handbook of the fundamentals of financial decision making: Part I', World Scientific, pp. 11–30.

Sarmento, S. M. and Horta, N. (2020), 'Enhancing a pairs trading strategy with the application of machine learning', *Expert Systems with Applications* **158**, 113490.

Shahzad, U., Jena, S. K., Tiwari, A. K., Doğan, B. and Magazzino, C. (2022), 'Time-frequency analysis between bloomberg commodity index (bcom) and wti crude oil prices', *Resources Policy* **78**, 102823.

Smyth, N. (2009), 'Dp2009/03 order flow and exchange rate changes: A look at the nzd/usd and aud/usd'.

Sørensen, C. (2002), 'Modeling seasonality in agricultural commodity futures', *Journal of Futures Markets: Futures, Options, and Other Derivative Products* **22**(5), 393–426.

Stephens, D. (2007), *A forecasting model for the NZD/AUD exchange rate*, Westpac Institutional Bank.

Tonin, J. M., Vieira, C. M., de Sousa Fragoso, R. M. and Martines Filho, J. G. (2020), 'Conditional correlation and volatility between spot and futures markets for soybean and corn', *Agribusiness* **36**(4), 707–724.

Van Buuren, S. (2018), *Flexible imputation of missing data*, CRC press.

Vidyamurthy, G. (2004), *Pairs trading: Quantitative methods and analysis*, Vol. 217, John Wiley & Sons.

Wang, Y., Li, X., Wu, P. and Xie, H. (2022), Many-to-many pair trading, *in* 'Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint International Conference on Web and Big Data', Springer, pp. 399–407.

Yeo, J. (2016), Factor Models, Mean-Reversion Time, and Statistical Arbitrage, PhD thesis, Stanford University.

Yeo, J. and Papanicolaou, G. (2017), 'Risk control of mean-reversion time in statistical arbitrage', *Risk and Decision Analysis* **6**(4), 263–290.

Zeng, Z. and Lee, C.-G. (2014), 'Pairs trading: optimal thresholds and profitability', *Quantitative Finance* **14**(11), 1881–1893.

Chapter 6

Appendix

6.1 Linear Regressions vs Benchmarks

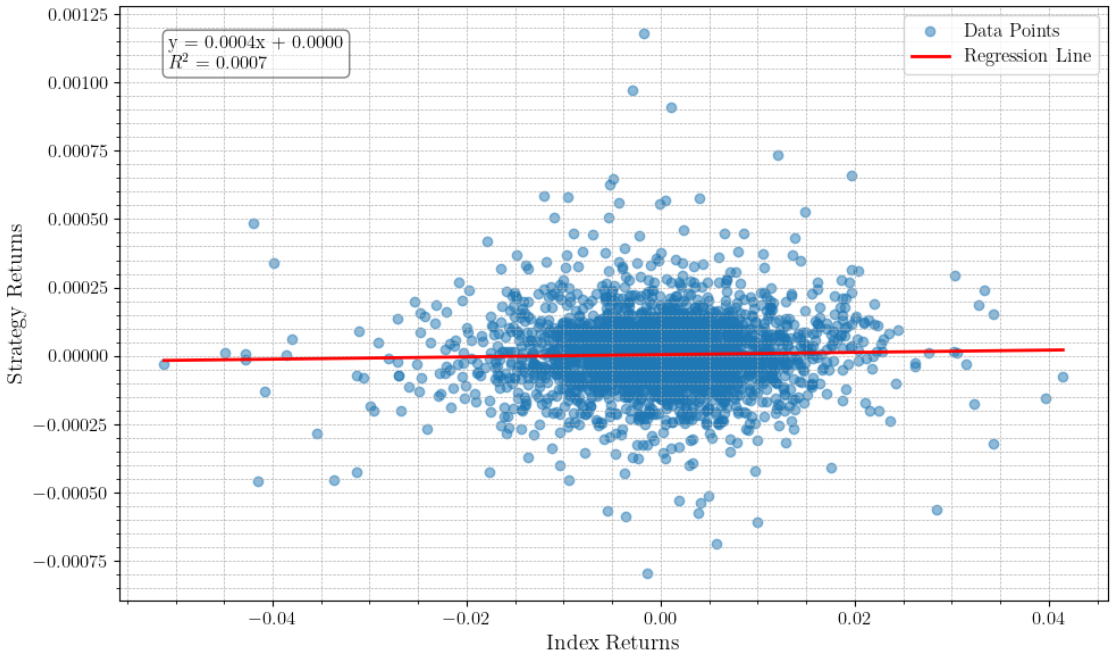

Figure 6.1: Linear regression between the PCA returns (y) & the Bloomberg Commodity Index (x).

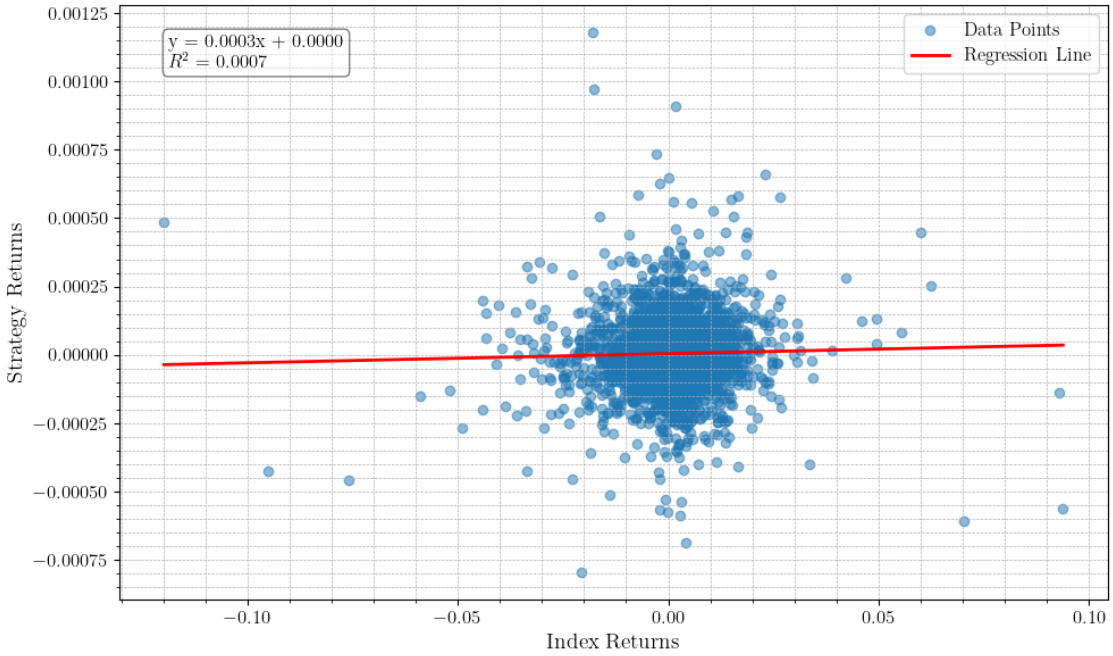

Figure 6.2: Linear regression between the PCA returns (y) & the S&P 500 (x).

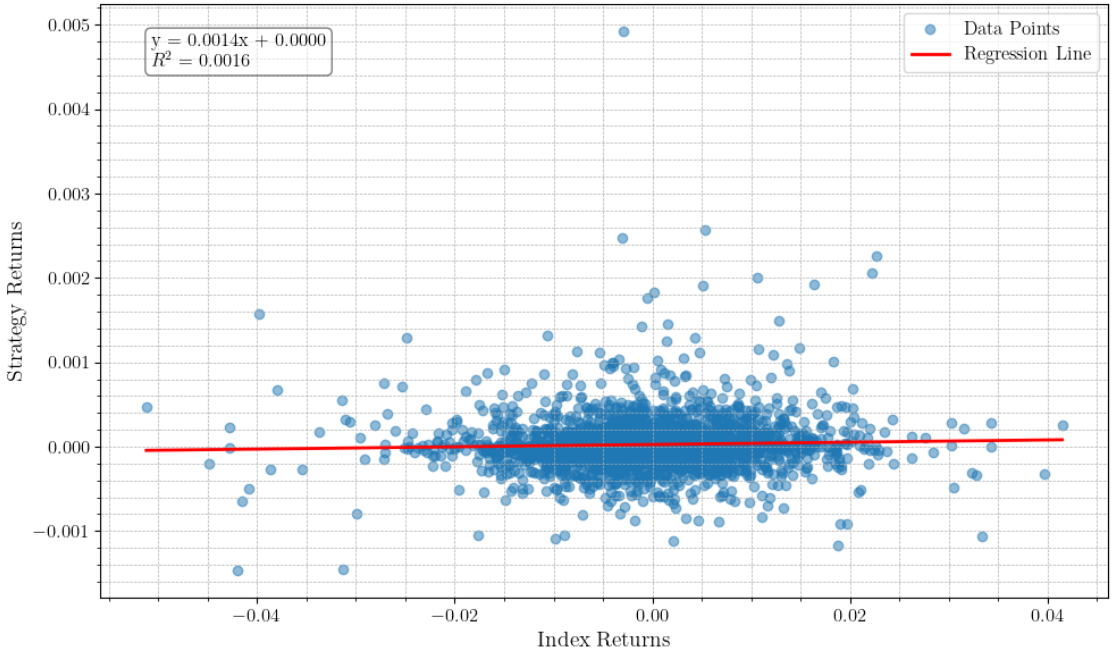

Figure 6.3: Linear regression between the OPTICS returns (y) & the Bloomberg Commodity Index (x).

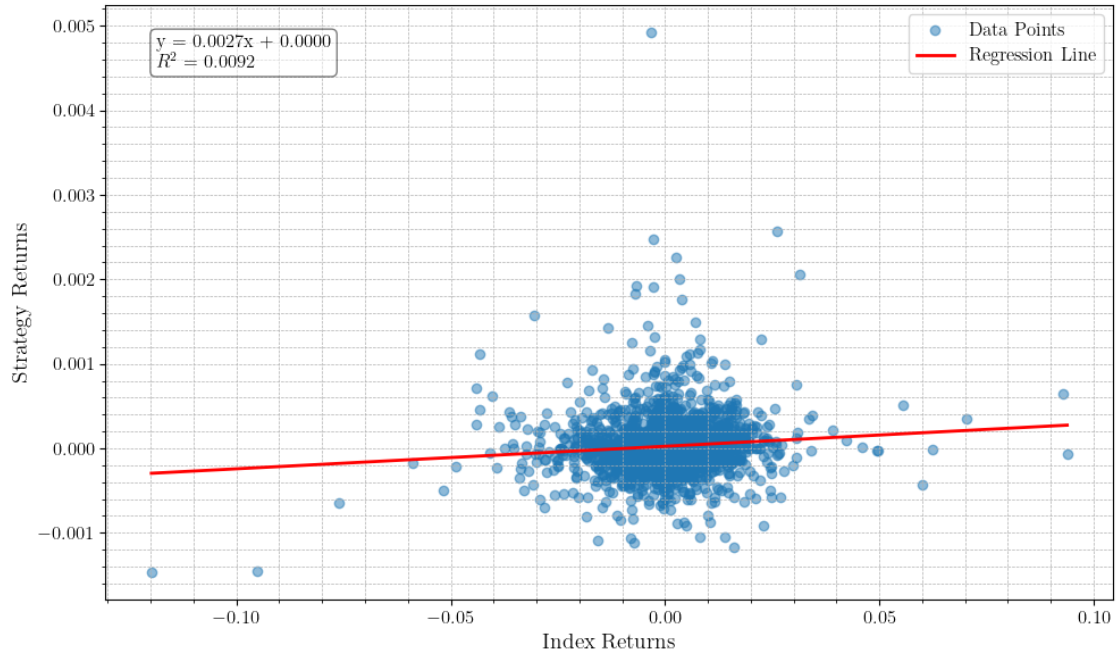

Figure 6.4: Linear regression between the OPTICS returns (y) & the S&P 500 (x).

6.2 Miscellaneous

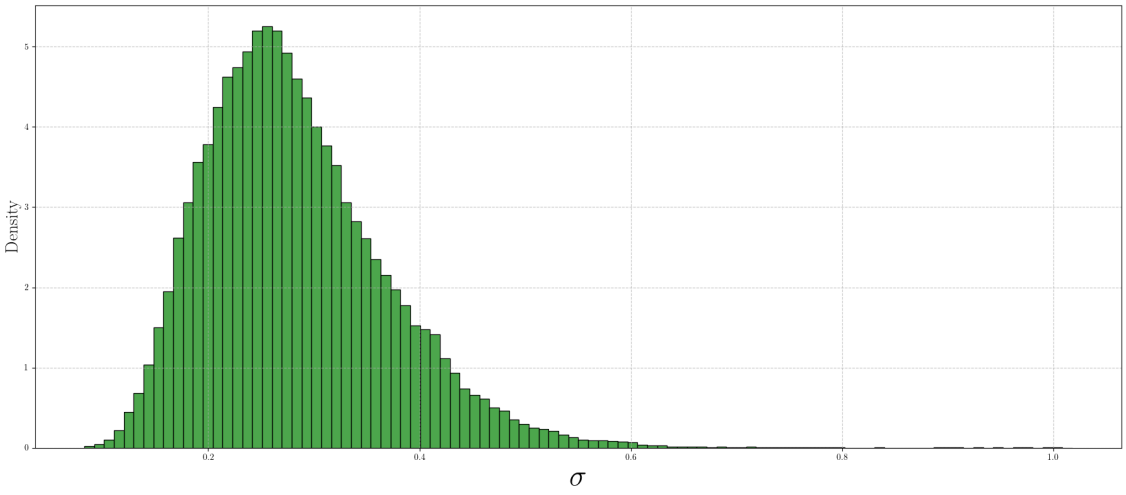

Figure 6.5: Empirical distribution of σ for the residuals.

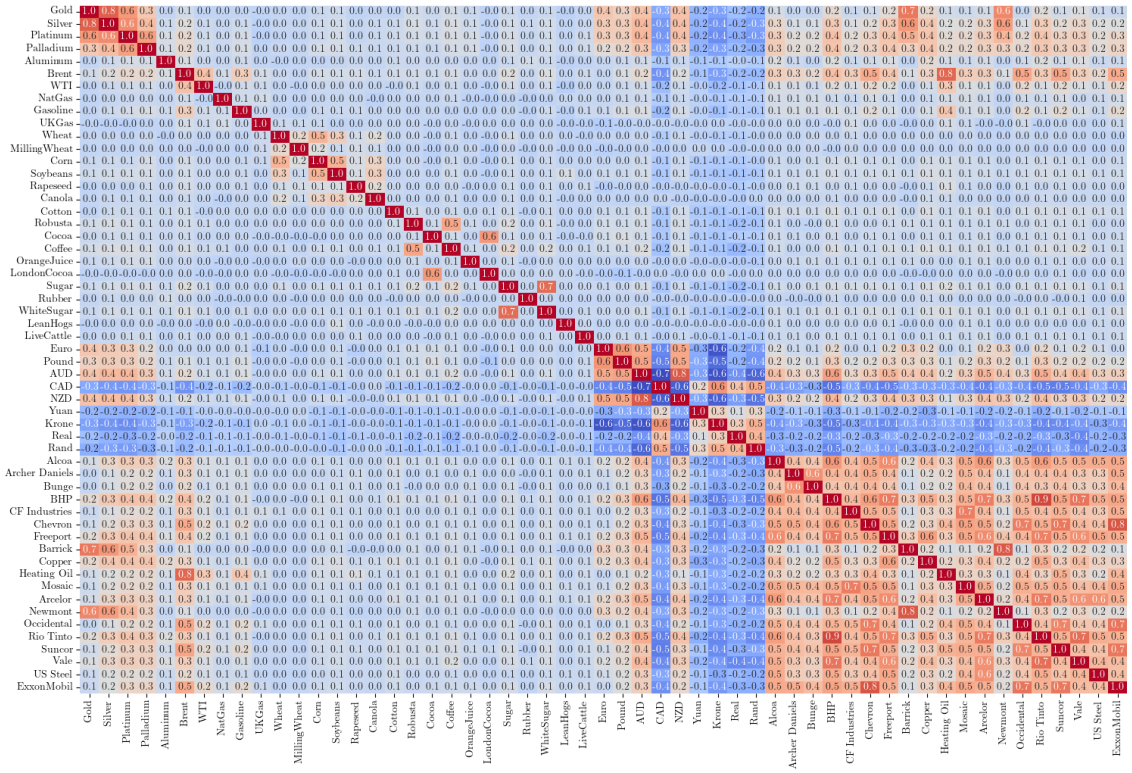

Figure 6.6: Correlation matrix for the asset universe.

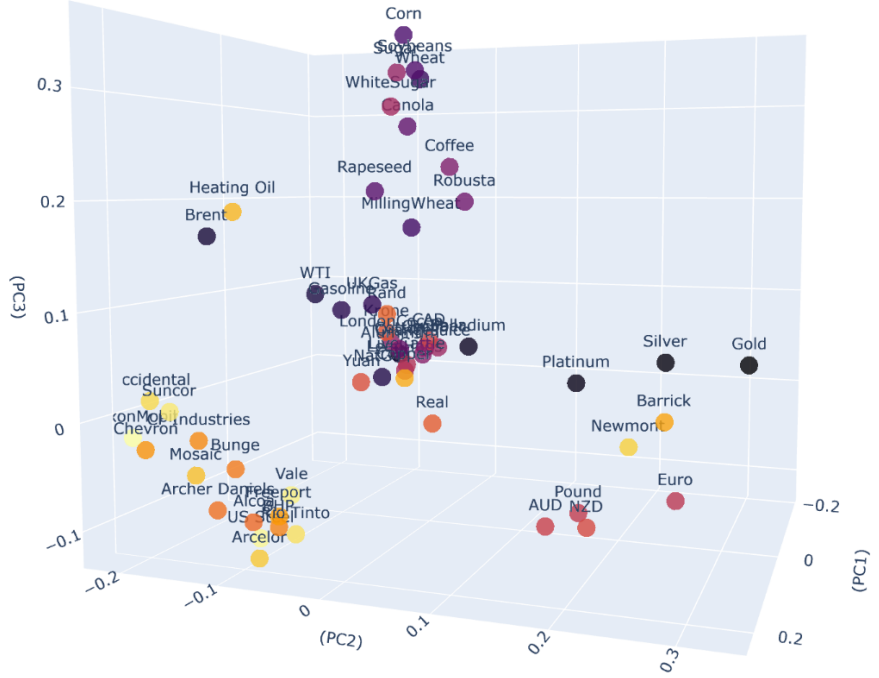

Figure 6.7: Representation of the asset universe over the first 3 PCs.

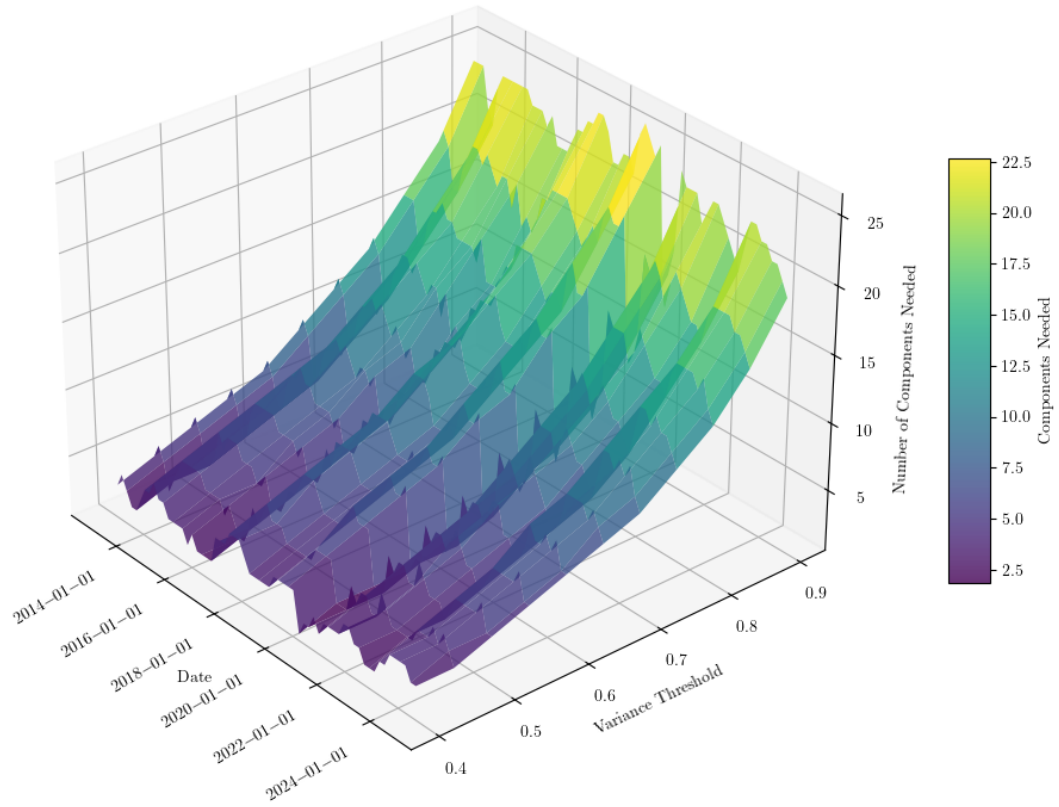

Figure 6.8: Number of PCA components needed over time for different thresholds.