DEPARTMENT OF ECONOMICS AND FINANCE

ASSET PRICING

Multidimensional cointegration for stocks picking: empirical evidence

Advisor: Prof. Nicola Borri

Co-advisor: Prof. Valerio Marchisio

747221

Anakin Gaffi

Academic Year: 2022/2023

Contents

1 Introduction

The advent of increased computational power has ushered in an era of progressively faster finance. This evolution has engendered a more mathematical and complex approach to trading, leading to the development of various pair trading and statistical arbitrage strategies, including those grounded in cointegration. This thesis endeavors to demonstrate the breadth and applicability of this concept, highlighting how it can be expanded dimensionally—specifically, by analyzing more than two stocks. The objective is to showcase the potential for achieving superior risk-return ratios through a multidimensional approach to cointegration. Through the simulation and analysis of the performances of portfolios constructed using this approach, this study aims not merely to ascertain the profitability of the strategy but also to affirm its capacity for risk neutrality, i.e., the ability to capitalize on both the bullish and bearish phases of the market, a trait highly esteemed by hedge funds.

Given its highly data-driven nature, this strategy necessitates the use of algorithms for its implementation, encompassing the selection of assets and the management of positions. However, this thesis will not extensively delve into the practical aspects of portfolio management or operational intricacies, as these are the domains of those who implement such strategies in practice. The principal aim is to broaden the cointegration approach within statistical trading, transitioning from the search for cointegrated pairs to the identification of cointegrated groups comprising multiple stocks. This necessitates devising a strategy for portfolio creation and subsequent performance analysis. This process was realized through the development of various Python functions, enabling the processing of a considerable number of portfolios across a dataset spanning over 15 years.

The inaugural chapter delineates the theoretical underpinnings of the strategy, interspersed with a succinct discussion on pair trading and the long/short strategy. Central to this discourse is cointegration, particularly its identification via the Augmented Dickey-Fuller test and the Johansen test. The ensuing chapter elucidates the structure of the trading strategy, commencing with dataset creation and the selection of cointegrated stocks, followed by portfolio construction upon determining the appropriate weights. The final chapter is dedicated to the analysis of portfolio performances, encompassing a comparative evaluation of the returns and volatility of various portfolios against the benchmark, culminating in the verification of the strategy's market neutrality. Additional analyses include assessing skewness and kurtosis to gain deeper insights into the simulated returns of the portfolios, and the computation of the Sharpe ratio to juxtapose the risk-return profiles of the diverse portfolios.

2 Theoretical basis of strategy: pair trading and cointegration

The first chapter outlines the theoretical foundations of the strategy. Starting with the economic fundamentals of trading such as long/short trading and market neutrality that characterize the strategies of multiple hedge funds. Next, is introduced the concept of pair trading and the underlying mathematical and empirical foundations, like the econometric roots of cointegration. After that there is a discression on tests that help identify cointegrated pairs: the Augmented Dickey Fuller test and the Johansen test.

2.1 Long/short strategy and market neutrality

Underlying any trading strategy implemented by hedge funds is the goal of obtaining a return independent of what is happening in the market. As much as this observation defies the Markowitz frontier and efficient market theory that is what hedge funds try to do. It was within one of the very first funds that Alfred Winslow Jones introduced the concept of long/short stategies as we know it today. This strategy is the foundation of many more complex investment methodologies. Basically, a long/short strategy consists of opening a long position in a security that is expected to increase in value while opening a short position in a security where the value is expected to decrease. The goal is to profit from both increases and decreases in stock prices.

Consider two stocks, X and Y, where X is the stock of a company with continuously increasing revenues, and Y is the stock of a declining company. Assuming that the price of X and Y at time t_0 is \$90 and \$70, respectively, a trading strategy is implemented where a long position is opened on X and a short position on Y. At time t_1, the price of X drops to \$80, resulting in a loss for the long position. However, the price of Y falls to \$40, which, due to the short position, leads to a profit from the decrease in price.

Stock	t_0	t_1	Δ	$\Delta\%$
X	90	80	-10	-11,1%
Y	-70	-40	30	42,8%
Portfolio	20	40	20	12,5%

Despite the loss in the long position on X and the profit in the short position on Y, the overall strategy yields a positive return of \$20, which translates to a return percentage of 12.5%. The objective of a long/short exposure is to increase the potential return of a trade and reduce the exposure to market risk. Accordingly, the long/short strategy can be said to be a market neutral strategy.

For instance, consider the betas of stocks X and Y are 1,5 and 1,2, respectively. We

compare the betas of two different portfolios: one in which we buy equal amounts of X and Y, and another in which we buy as much of X as we short Y. The beta of a portfolio is the weighted average of the betas of its constituents, leading to the following results:

$$\beta_{\text{portfolio}} = w_X \beta_X + w_Y \beta_Y \text{ where } w_X + w_Y = 1 \tag{1}$$

- For the equally weighted portfolio, the beta is calculated as:

$$\beta_{\text{long}} = \frac{1}{2} \times 1,5 + \frac{1}{2} \times 1,2 = 1,35$$

- For the long/short portfolio, the beta is calculated as:

$$\beta_{\text{long/short}} = \frac{1}{2} \times 1,5 - \frac{1}{2} \times 1,2 = 0,15$$

  Assuming equal dollar amounts invested in and shorted, the weights are equal and opposite, leading to a reduced net beta.

Through market neutrality, one aims to achieve returns that are uncorrelated with the market. This approach seeks to mitigate systemic market risks and relies on the skill of the investor or trader to identify undervalued or overvalued assets. In addition, while market neutral strategies do not necessarily reach a zero dollar cost condition - that is, a scenario where the long positions are fully funded by the short positions - they tend to be more capital-use efficient than traditional buy and hold strategies. This efficiency is due to the simultaneous utilization of long and short positions, effectively leveraging the capital in a more balanced and risk-adjusted manner. Compared to the latter, a long/short strategy is more complex and requires the implementation of more advanced risk management. In fact, the presence of short positions necessitates margin management to keep the position open, as well as incurring higher fees. This demands a more in-depth knowledge of market dynamics than is typically required for a long trade. On the other hand, a long/short strategy tends to expose a portfolio less to the risk of market collapse and overall volatility.

There are numerous methods for selecting stocks on which to apply the long/short strategy. Certainly, the most popular one is based on fundamental analysis, in which one tries to identify the intrinsic value of a company around which the price fluctuates in the medium term. Discounted Cash Flow (DCF) models are often used for this purpose. Therefore, to estimate intrinsic value, future cash flows are estimated and discounted at a rate of return, which represents the capital structure of the company.

To make the values of different companies comparable, different metrics such as the price to earnings ratio (P/E) and price to book value ratio (P/B) are used. In each case in fundamental analysis, accounting records such as the income statement and balance sheet

are used to assess a range within which, with some margin of safety, a price reflecting the intrinsic value of the company can be found.

The implementation of the long/short strategy is based on this by setting margins to open and close positions, as well as all portfolio management.

A more quantitative approach to the long/short strategy is pair trading, which employs mathematical and econometric models rather than fundamental assessments to identify stocks to trade on. This strategy involves identifying pairs of stocks whose prices are historically correlated and are expected to continue to move in tandem. When these pairs diverge in terms of their price relationship, a trade is executed: one stock is bought long, and the other is sold short. The expectation is that the prices will converge again, allowing the trader to profit from the relative movement of the two stocks.

2.2 Pair trading

Pair trading is a form of statistical arbitrage that took hold in the late 20th century as a complex investment strategy implemented by hedge funds. The model was first applied by Morgan Stanley in 1980 through Nunzio Tartaglia's quantitative group. Later thanks to the likes of Engle and Granger, econometric studies led to statistical arbitrage and the concept of cointegration.

The model exploits market inefficiencies, that is, when the price movements of two stocks, which are historically correlated, diverge it means that an arbitrage possibility arises. The underlying principle is that the two stocks are economically related, in fact they are often part of the same industry, consequently their price should move similarly. Consequently should the two prices diverge significantly without economic foundation the possibility of pair trade would arise based on the expectation that prices will converge again.

The models used by Morgan Stanley therefore used models that identified relationships between stocks based on common economic factors and exploited short-term fluctuations to open positions. The simplest approach to pair trading is based on tracking variance, which is the average distance between the prices of two stocks.

Mathematically it is defined as:

$$TV = \frac{1}{T}\sum_{t=1}^{T}(Q_{A,t} - Q_{B,t})^2 \tag{2}$$

Where $Q_{A,t}$ and $Q_{B,t}$ are rate between the log prices of A and B in t and in 1. $Q_t = P_t/P_1$. T is the number of periods considered. So, the tracking variance is the average of the squared diviation of normalized prices. Therefore, one identifies the spread between the two stocks as Δ and identifies the thresholds as a function of σ_Δ, that is, the standard

deviation. When thresholds are exceeded, positions are opened by buying the undervalued stock and shorting the overvalued stock. The position is usually closed when the spread converges to zero.

To give a brief example by considering Nasdaq stocks. Daily data for the past 20 years of Nasdaq stocks were downloaded from Bloomberg. Then a function was written in python to identify the pairs of stocks with a tracking variance less than 0.01, which is more than sufficient for the illustrative purpose of this example.

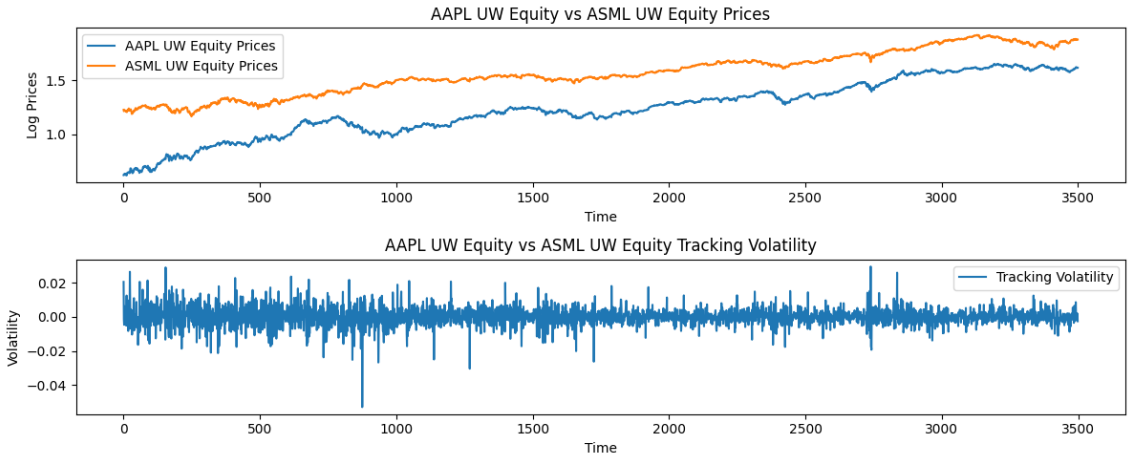

Figure 1: Normalized prices and tracking variance of two stocks.

Figure 1 shows in the first graph shows the log prices of AAPL-ASML, i.e., the first pair identified by the Python function with tracking variance less than 0.01. In fact, it can be seen that the two time series have a very similar trend. The graph below, on the other hand, shows the tracking variance; in fact, it can be seen that when the trend of the two time series diverges, the tracking variance increases, creating opportune profit opportunities.

From an empirical point of view, it is important to cite *Pairs Trading: Performance of a Relative-Value Arbitrage Rule* by Evan Gatev, William Goetzmann, and K. Geert Rouwenhorst, a study initially conducted in 1998 and later updated in 2006. This seminal work offers an in-depth analysis of the pair trading strategy, in which stocks are matched based on the minimum deviation between their normalized historical prices. The data considered in the study span daily records from 1962 to 1997. One of the key findings is that portfolios composed of the top 120 pairs can realize over 12 percent average annualized excess returns, which are higher than the estimated transaction costs. This study provides a detailed examination of the dynamics of mispricing among correlated stocks. Furthermore, it confirms that the returns from pair trading predominantly come from this mispricing rather than from conventional measures of risk, thus validating pair trading as a market-neutral strategy.

2.3 Cointegration

Cointegration is the basis of several modern pair trading strategies. This is a statistical property according to which nonstationary time series maintain an equilibrium relationship according to which the spread between these series remains stationary. In other words, it means that two time series are said to be cointegrated when although they are individually nonstationary (in trend), there is a linear combination of them that is stationary. Arbitrage opportunities are found in short-run fluctuations, since if a series is stationary it will return toward long-run equilibrium.
The fundamental concept of cointegration and the econometric study of time series is stationarity. Basically, a time series is said to be stationary if the stochastic process that generated it has the parameters that are independent of time.
Define a stochastic process like a collection of random variables $\{X_t(\omega); t \in T\}$ defined on the same probability space $(\Omega, \mathcal{F}, P)$, where T is the index set of the process. The process can be *discrete* if T is a subset of $\mathbb{Z}$, can be *continuous* if T is a subset of the real numbers $\mathbb{R}$.

So if T is a subset of $\mathbb{Z}$ set of time points is finite $T = \{t_1, \ldots, t_n\}$ and the cumulative distribution function (CDF) of $X = \{X_i(\omega); i \in T\}$ is defined as:

$$F_{t_1,\ldots,t_n}(x_{t_1}, \ldots, x_{t_n}) = P(X_{t_1}(\omega) \leq x_{t_1}(\omega), \ldots, X_{t_n}(\omega) \leq x_{t_n}(\omega)) \tag{3}$$

which, for the generic stochastic process X, is denoted as:

$$F_X(x_{t_1}, \ldots, x_{t_n}) \tag{4}$$

Given a definition of a stochastic process we now need to define stationarity, a fundamental characteristic for the study and prediction of time series. Mathematically, a strict stationary series requires that the joint distribution of any set of observations remains the same when shifted in time. However, this definition is quite stringent. In practice, often a weaker form, known as "second-order" or "weak-form" stationarity, is considered. For a time series X_t to be considered second-order stationary, it must satisfy the following criteria:

1. **Constant Mean:** The first moment of the series remains constant over time. Mathematically:
$$\mathbb{E}[X_i] = \mu, \quad \forall t \tag{5}$$

2. **Finite Variance:** The second moment the series is finite over time. Mathematically:
$$\mathbb{E}[X_i^2] < \infty, \quad \forall t \tag{6}$$

which implies that:

$$\mathbb{E}[(X_i - \mu)^2] < \infty, \quad \forall t \tag{7}$$

3. **Constant Covariance:** The covariance between the i^{th} term and the j^{th} term should not be a function of time, implying that every lag λ has constant variance. It is same for autocovariance function $\gamma_X(i, i+m)$ that depends only on m (the difference between i and $i+m$), not on time t. Mathematically:

$$\text{cov}(X_i, X_j) = \text{cov}(X_{i+k}, X_{j+k}), \quad \forall i, j, k \tag{8}$$

The above properties ensure that the series does not exhibit any trend, its fluctuations around the mean have a consistent spread, and its short-term movements are entirely represented by its autocovariances, which remain consistent across time.
Practically the characteristic of the stationary series that allows arbitrage is mean-reverting, that is, the fact that after deviating from the mean for a short period they tend to converge back to it. Empirically, one can consider the time series of a stock price as a random walk.

Let X_1, X_2, X_3,... be a sequence of independent and identically distributed (i.i.d.) random variables, where each X_i represents a single step of the walk. The position of the random walk at time n is then given by the sum of the first n steps: $S_n = S_0 + X_1 + X_2 + \ldots + X_n$.

$$S_n = S_0 + \sum_{i=1}^{n} X_i \tag{9}$$

If we consider that X_i can take value 1 or -1 with a chance of 50% each we will have that the expected value of S_n will be 0 while its variance σ_n^2, so since the variance depends on time a random walk is not stationary and not predictable. Consequently it is necessary to remove the time component and make the series stationary by taking their first difference as an example. Figure 2 shows AAPL's log price graph at the top, you can see that the time series is not stationary. While the bottom graph was generated by a function written in Python that takes a time series as input and applies a difference transformation iteratively until the time series is stationary. The differentiation process functions in this way, first the first difference is calculated, which is defined as:

$$\Delta Y_t = Y_t - Y_{t-1} \tag{10}$$

Sometimes, first differencing is insufficient to achieve stationarity. Higher-order differencing may then be employed:

$$\Delta^2 Y_t = \Delta(\Delta Y_t) = \Delta Y_t - \Delta Y_{t-1} = Y_t - 2Y_{t-1} + Y_{t-2} \tag{11}$$

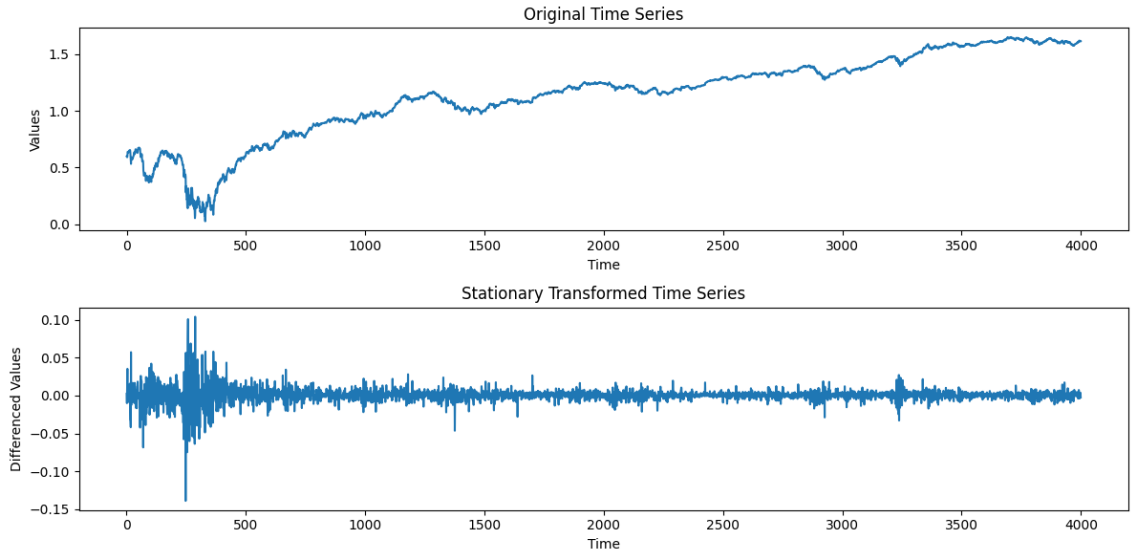

Figure 2: Normalized prices a first difference trasformation of a time series.

Generally, for d^{th} order differencing:

$$\Delta^d Y_t = \Delta(\Delta^{d-1} Y_t) \tag{12}$$

Therefore, the fundamental insight of cointegration is that it is possible to combine and transform nonstationary processes into stationary processes in order to study and predict them. In fact, two processes of can be called cointegrated if there is a linear combination of them that is stationary. If the prices of two stocks are cointegrated, it means that they comove. If y_t and x_t are historical stock prices there are two coefficients β and α such that:

$$\beta x_t + \alpha y_t = z_t \text{ where } z_t \text{ is a stationary process} \tag{13}$$

Given the long-run equilibrium relationship above, deviations from the short-run equilibrium can be exploited. Assuming that α is 1 and β is -1 if z_t is positive it means that x_t is overvalued relative to y_t so one must buy y_t and sell x_t, since over the long run z_t should return to 0. Conversely if z_t is negative one must sell y_t and buy x_t since the latter is undervalued compared to y_t which is overvalued. The same concept can be expanded by looking for a stationary series consisting of more than 2 stocks. Consider historical stock prices x_t, y_t, z_t, and w_t. There exist coefficients α, β, γ, and δ such that the linear combination:

$$\alpha x_t + \beta y_t + \gamma z_t + \delta w_t = u_t \tag{14}$$

where u_t is a stationary process.

2.4 Testing for cointegration: ADF test

The first tool for detecting stationarity, a prerequisite for cointegration, in the framework of time series analysis is the augmented dickey-fuller (ADF) test. In nonstationary processes the effects of short-term shocks persist in the long run, while in stationary processes such shocks are temporary. Mathematically it means that nonstationary series often have a unit root. The ADF test is precisely an expanded version of the Dickey-Fuller tests that are based on time series with an AR(1) process.Taking the simplest model, i.e., a random walk, as an example

$$y_t = \phi y_{t-1} + \epsilon_t \tag{15}$$

where ϕ is the coefficient of the lagged term at t-1 and ϵ_t is the error term in t. We can transform this AR(1) model into an MA(infinity) model by iteratively substituting lagged values of Y. The first substitution gives:

$$y_t = \phi(\phi y_{t-2} + \varepsilon_{t-1}) + \varepsilon_t$$

$$y_t = \phi^2 y_{t-2} + \phi\varepsilon_{t-1} + \varepsilon_t$$

Continuing this process, we obtain :

$$y_t = y_0^t + \sum_k^{t-1} \phi^k \varepsilon_{t-k}$$

$$y_t = \phi y_{t-1} + \epsilon_t = y_0^t + \sum_k^{t-1} \phi^k \varepsilon_{t-k} \tag{16}$$

Assuming that the expected value of ϵ_t is 0 and that the variance of each ϵ_t is the same we get that the expected value and variance of the process are:

$$E(y_t) = \phi^t y_0 \tag{17}$$

$$Var(y_t) = \sigma^2(\phi^0 + \phi^2 + \phi^4 + \ldots + \phi^2(t-1)) \tag{18}$$

Thus it can be seen that expected value and variance depend solely on the modulus of ϕ. Three possible scenarios arise, depending on whether the absolute value of ϕ is larger, smaller, or equal to 1. If $|\phi|$ is larger than one, it means that the series is nonstationary and will explode with time. If $|\phi|$ is smaller than 1, the expected value with time will converge to zero while the variance will converge to $\frac{\sigma^2}{1-\phi^2}$. Therefore, if the absolute value of the scaling coefficients is less than one, we are in a stationary condition since the expected value and variance are constant in time. In contrast, in the unit root case,

which means that the absolute value of ϕ is 1, the variance becomes time-dependent, i.e., $t\sigma^2$, and consequently nonstationary. Now that we can expand the model, moving to the Augmented Dickey Fuller, which considers higher-order correlation, including lagged difference terms of time series, so we move from an AR(1) process to an AR(p):

$$\Delta y_t = \mu + \delta y_{t-1} + \sum_{i=1}^{p} \gamma_i \Delta y_{t-i} + \epsilon_t \tag{19}$$

where μ is the drift term and γ_i is the coefficient of the lagged difference term. The null hypothesis of the ADF test is that:

$$H_0 : \delta = 0$$

$$H_1 : \delta < 0$$

To test the series, it is necessary to calculate the t-statistic, $t_{\hat{\delta}}$, of δ and compare it with the Dickey-Fuller distribution. If the t-statistic is less than the critical value of the Dickey-Fuller distribution, t_{critical}, the null hypothesis of a unit root is rejected. Therefore, the stationarity hypothesis is not rejected.
The ADF is used in the Engle-Granger approach to test the cointegration. The approach has 2 steps:

1. Estimate the regression between the two time series y_t and x_t to obtain the residuals $\hat{\epsilon}_t$.

2. Test the residuals for a unit root using the ADF test.

2.5 Testing for cointegration: Johansen test

As an alternative to the Engle-Granger approach using the ADF, one can use The Johansen Test. The latter allows one to analyze cointegration in multivariate time series. In fact, one can consider this test as a multivariate extension of the ADF. Although derived from the same econometric framework there are 2 different statistics for determining the cointegration relationship in a set of nonstationary time series: eigenvalue statistic and trace statistic. These differ on how the null hypothesis is tested.

- **Trace Statistics:** The null hypothesis is that the number of cointegrated vectors is less than or equal to r versus the hypothesis that it is larger.

$$\text{Trace} = -T \sum_{i=r+1}^{n} \ln(1 - \lambda_i),$$

where T is the sample size and λ_i is the eigenvalue. The test is done sequentially starting from $r = 1$ and the first r for which the null hypothesis is not rejected is taken as the estimator of r. Therefore, a relevant trace statistic means that there are more than r cointegrated vectors.

- **Eigenvalue Statistics:** It shares the null hypothesis with trace statistics, but as an alternative it has $r + 1$ cointegrating vectors.

$$\text{Max-eigen} = -T \ln(1 - \lambda_{r+1}),$$

As with the trace, the test is performed sequentially starting from $r = 1$ and the first r for which the null hypothesis is not rejected is taken as the estimator. The two tests are very similar so they can be used together to increase the consistency of the results.

From a mathematical point of view, a vector autoregressive model (VAR), i.e., the multivariate version of an AR process, is applied. In detail, a Vector Error Correction Model (VECM). Given a vector X_t of n nonstationary time series, a VAR model of order p can be written as:

$$X_t = \Phi_1 X_{t-1} + \Phi_2 X_{t-2} + \cdots + \Phi_p X_{t-p} + \mu + \varepsilon_t$$

In matrix form:

$$\begin{bmatrix} X_{1t} \\ X_{2t} \\ \vdots \\ X_{nt} \end{bmatrix} = \Phi_1 \begin{bmatrix} X_{1,t-1} \\ X_{2,t-1} \\ \vdots \\ X_{n,t-1} \end{bmatrix} + \Phi_2 \begin{bmatrix} X_{1,t-2} \\ X_{2,t-2} \\ \vdots \\ X_{n,t-2} \end{bmatrix} + \cdots + \Phi_p \begin{bmatrix} X_{1,t-p} \\ X_{2,t-p} \\ \vdots \\ X_{n,t-p} \end{bmatrix} + \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix} + \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \\ \vdots \\ \varepsilon_{nt} \end{bmatrix}$$

Where each Φ_i (where $i = 1, 2, \ldots, p$) is represented as:

$$\Phi_i = \begin{bmatrix} \phi_{11}^{(i)} & \phi_{12}^{(i)} & \cdots & \phi_{1n}^{(i)} \\ \phi_{21}^{(i)} & \phi_{22}^{(i)} & \cdots & \phi_{2n}^{(i)} \\ \vdots & \vdots & \ddots & \vdots \\ \phi_{n1}^{(i)} & \phi_{n2}^{(i)} & \cdots & \phi_{nn}^{(i)} \end{bmatrix}$$

Now taking the first difference of X_t we get a Vector Error Correction Model (VECM):

$$\Delta X_t = \Pi X_{t-1} + \sum_{i=1}^{p-1} \Gamma_i \Delta X_{t-i} + \mu + \varepsilon_t$$

Where:

- $\Pi = -(\Phi_1 + \Phi_2 + \cdots + \Phi_p - I)$ is an $n \times n$ matrix.
- $\Gamma_i = -(\Phi_{i+1} + \Phi_{i+2} + \cdots + \Phi_p)$ for $i = 1, 2, \ldots, p-1$ are $n \times n$ matrices.
- $\Delta X_t = X_t - X_{t-1}$ is the first difference of X_t.
- μ is an $n \times 1$ vector of constants.
- ε_t is an $n \times 1$ vector of error terms.

The VECM makes it possible to capture the relationships between different variables by also allowing the identification of cointegration between multiple non-stationary time series. Looking at the formula of the VECM, it can be seen that Π represents the long-term relationships among the various time series, while Γ_i encloses the coefficients showing the short-term dimanics. The Johansen test is based on the decomposition and analysis of the eigenvalues and eigenvectors of the Π matrix. In fact, eigenvalues are used to calculate statistics and thus measure the degree of cointegration of the time series, while eigenvectors are the cointegrating vectors. Considering the above, the rank ofΠ indicates the number of cointegrated vectors, so if the rank is zero, it means that there is no cointegrated time series; on the other hand, if the matrix is full rank, it means that all variables are cointegrated. Unlike the Engle-Granger approach the Johansen test allows for more agile work in order to individual multiple cointegrated time series.

3 Strategy implementation

This section illustrates the structure of the trading strategy. It delves into the rationale behind the strategy and the details of its practical implementation. The steps are as follows:

1. Choose the data to be analyzed and create the datasets, aiming to optimize the utilization of available data. Then,identify the size of the training set, validation set, and test set.

2. Select cointegrated sets of stocks and categorize them.

3. Analyze the group of the cointegrated sets of stock to establish operational parameters and rank the groups.

4. Simulate the application of the trading parameters obtained from the training set on the test set and calculate the returns of the stock groups.

5. Apply the weights to the stock groups and create the portfolios.

The entire strategy, from initial data preparation to the analysis and comparison of returns, was meticulously crafted using Visual Studio Code with Python.

3.1 Dataset preparation

The dataset selection is pivotal to the trading strategy. The chosen dates consist of the daily data from the constituent stocks of prominents market indices, particularly the S&P and Euronext. The rationale behind selecting these datasets is to analyze a broad spectrum of stocks while mitigating the risk associated with illiquid stocks, minimizing the liquidity risk inherent in this strategy. By taking other indexes it would probably be possible to identify many more cointegrated stocks, but from an operational point of view the low liquidity of the traded stocks would have a negative impact on the performance of the strategy.
The daily data considered are those from 08/03/2007 to 13/02/2023 downloaded from bloomberg, with a total of 4158 observations per share. Such a large dataset was chosen in order to have a more consistent analysis from a time perspective as well as to consider only stocks that have been in the index for a long time. For these reasons, the first dataframe created containing analyzable data contains 455 stocks.
This study adopts a rigorous empirical approach to validate the proposed trading strategy, characterized by a meticulous division of the dataset into distinct subsets: training, validation, and testing. Each subset plays a pivotal role in the comprehensive evaluation of the

strategy's efficacy and robustness. The initial phase involves the exploitation of the training set to discern and identify groups of cointegrated stocks. This phase is foundational, establishing the underpinnings of the trading strategy by leveraging statistical techniques to reveal potential cointegration relationships among various stock entities. Subsequent to the training phase, the validation set is employed to rigorously scrutinize the persistence of the cointegration assumptions. This critical evaluation serves to confirm or refute the initial findings, providing an essential checkpoint before further deployment of the strategy. Upon successful validation, the strategy progresses to a crucial stage of parameter analysis and optimization, informed by the data and insights accrued from both training and validation phases. The culmination of this process is marked by a simulated implementation of the trading strategy on the validated pairs in the test set. This simulation is instrumental in gauging the practical applicability and potential profitability of the strategy under real-market conditions.

This tripartite methodological framework is imperative for approximating real-life trading scenarios, thereby enabling a rigorous and objective assessment of the model's predictive prowess. Absent this structured demarcation, the analysis risks being predicated on prospective data, thus yielding a potentially erroneous representation of the strategy's true performance capabilities.

Cointegration, a statistical relationship between time series, is inherently transient in nature. Consequently, the search for cointegrated stocks across an excessively expansive dataset is not practically advantageous. It is imperative to identify a temporal window within which groups of cointegrated stocks manifest and sustain their interdependencies, thus facilitating the strategic opening and closing of trading positions. In the realm of empirical research, there exists no universally prescribed guideline for the optimal duration of these windows. The choice is largely influenced by the specificities of the study at hand. For the purposes of this analysis, a one-year period has been designated for the formation of groups, followed by a subsequent six-month window for the execution of trading operations. These particular time frames were determined through extensive empirical testing.

To maximize the utility of the available dataset, the data was segmented into 31 subdatasets. Each of these subdatasets comprises approximately 378 observations, roughly equating to one and a half years of financial data. Furthermore, each dataset was equitably divided into three distinct sets: training, validation, and test, each encompassing 126 observations. The methodology employed involves using the first six months of data to identify cointegrated groups. The subsequent six months are dedicated to the validation of these groups. Finally, the last six months of each subdataset are utilized to simulate the trading strategy. This approach employs a dynamic rolling window, wherein the valida-

tion and test sets of one subdataset seamlessly become the training and validation sets of the subsequent subdataset, respectively.

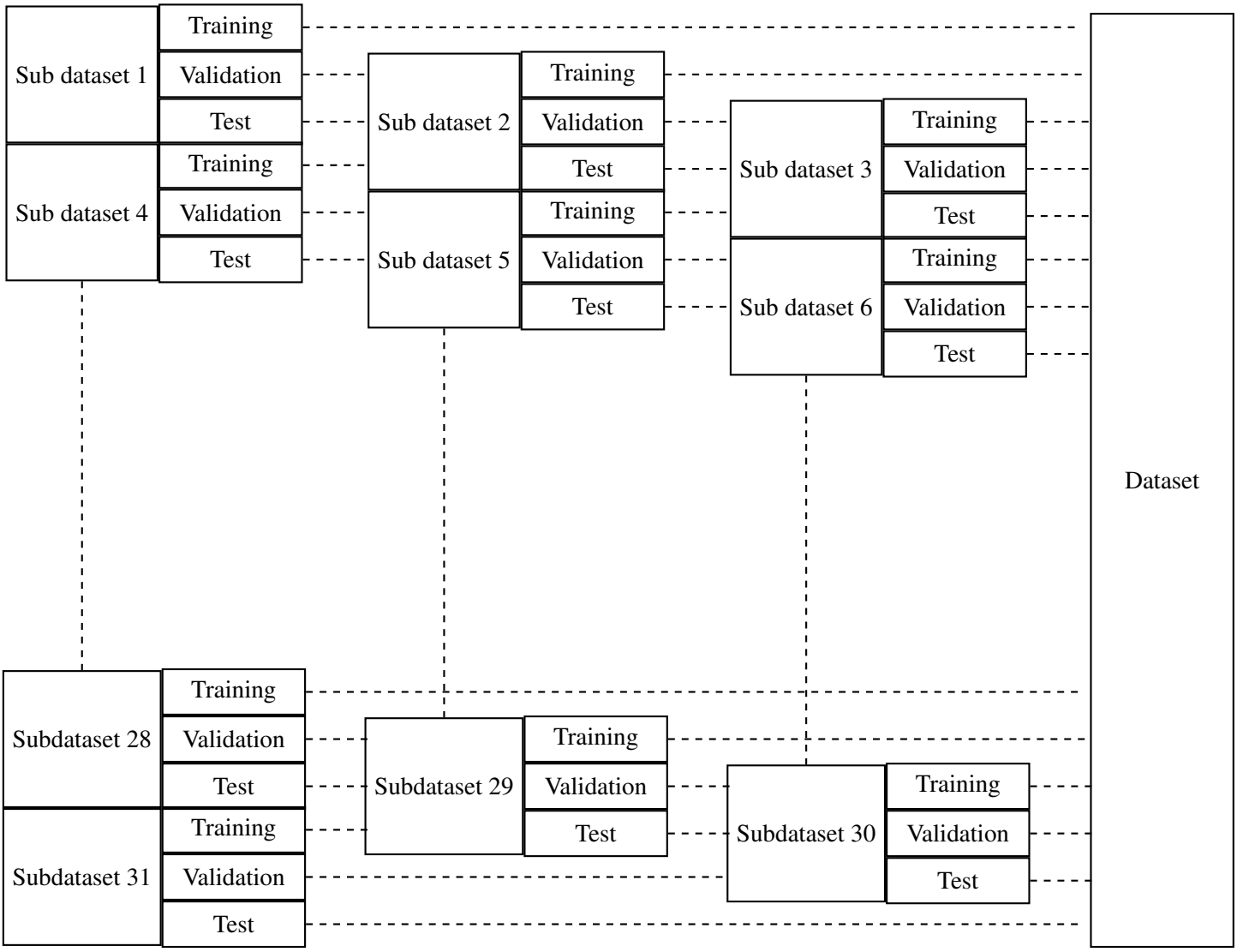

Figure 3: Dynamic rolling window dataset structure

3.2 Sets selection

Having prepared the datasets, one can proceed to search for cointegrated sets of stocks. Beyond the conventional approach of pairing stocks, this study expands its scope to encompass sets comprising multiple stocks, specifically targeting groupings of 3 to 10 stocks. This methodology aims to provide a more comprehensive understanding of the cointegration phenomenon within the stock market. The primary challenge in this en-

deavor lies in the combinatorial nature of the problem. The task involves assessing the potential cointegration among numerous combinations of stocks, quantified by the binomial coefficient:

$$C(n,k) = \frac{n!}{k!(n-k)!} \tag{20}$$

where n represents the total number of stocks under consideration and k is the number of stocks in a given set. While the analysis of pairs ($k = 2$)—yielding 103,285 combinations—remains computationally feasible, the expansion to larger sets introduces a significant computational burden. This is attributable to the exponential growth in the number of combinations, as depicted in the ensuing table: The computational challenge in identi-

k	Set	Time
2	103,285	2 min 9 sec
3	15,596,035	5 hours 24 min 55 sec
4	1,762,351,955	25 days
5	158,964,146,341	6 years
6	11,922,310,975,575	472 years
7	764,731,089,719,025	30.000 years
8	4.28×10^{16}	1.696.959 years
9	2.13×10^{18}	84.290.219 years
10	9.49×10^{19}	3.760.278.779 years

Table 1: Number of sets n=455 for k

fying cointegrated stock sets extends beyond the mere counting of possible combinations. The last column of the provided table estimates the time required for the available hardware to analyze all potential combinations. Given the enormity of this computational task, it became impractical to exhaustively analyze every possible set.

As a result of these computational constraints, a random sampling methodology was employed. For each value of k, representing the number of stocks in a set, 200,000 sets were randomly selected. The level of cointegration within these randomly chosen sets was then rigorously tested. The decision to limit the number of sets to 200,000 for each k was also dictated by hardware limitations. To test all chosen sets across all subdatasets required approximately 12 hours. This approach provides a viable method for exploring cointegration in larger sets of stocks. The adoption of a random sampling strategy offers a practical solution to the otherwise prohibitive task of exhaustive analysis.

To assess the presence of cointegration in the sample sets of stocks, we utilized the Johansen test methodology, as discussed in the previous chapter. Specifically, the Python function `coint_johansen` was employed for this purpose. This function, which operates on time series data as input, returns several critical pieces of information crucial to our analysis:

- Trace statistic and Maximum eigenvalue statistic, along with their corresponding critical values.
- Cointegration vectors.

The lag parameter in the Vector Error Correction Model (VECM) was setted equal 1, the function consequently computes the Π matrix and a single Γ matrix. However, our analysis primarily focuses on the eigenvalues of the Π matrix, which are central to the Johansen test. The `coint_johansen` function provides critical values at the 10%, 5%, and 1% significance levels. Therefore, all sets for which both the Trace statistic and the Maximum eigenvalue statistic exceed these thresholds, and thus cannot be rejected, are categorized into dictionaries according to their significance level. A non-rejection scenario implies that the cointegration relationship encompasses all the time series in the set, indicating that all betas in the cointegrated vectors are non-zero. Initially, the test was conducted on the training dataset. Subsequently, the sets identified as cointegrated were subjected to another Johansen test, this time including the validation dataset. If the statistical values continued to surpass the critical thresholds, the sets were retained; otherwise, they were discarded. The cointegrated sets identified thus far demonstrate resilience in terms of cointegration. However, our requirement extends beyond mere cointegration; we necessitate that the linear combination of prices, as indicated by the Johansen test, be genuinely stationary. The Johansen test, while adept at capturing relationships among multiple time series, exhibits limitations in verifying the stationarity of these relationships, especially when juxtaposed with the Augmented Dickey-Fuller (ADF) test. To ascertain the stationarity of the time series constructed using the cointegrating vectors, we apply the ADF test to the entire dataset, encompassing both training and validation periods. Sets whose cointegrated vector yield a p-value exceeding 0.05 in the ADF test are subsequently discarded. This process results in a refined collection of cointegrated sets, each exhibiting stationary characteristics. A notable observation from this secondary selection phase is the elimination of sets across all confidence levels of cointegration identified by the Johansen test. This underscores a disparity in the information captured by the Johansen and ADF tests from the time series relationships. Additionally, a higher proportion of sets with fewer stocks (smaller size) are discarded, suggesting complexities that cannot be readily explained due to observable data asymmetry. After this rigorous selection procedure, the number of cointegrated sets diminishes significantly. The table below illustrates the reduced number of sets identified in each period for each set dimension after the application of the ADF test:

Subdataset/Set	2	3	4	5	6	7	8	9	10
1	290	571	249	128	58	51	21	22	16
2	337	898	257	98	29	12	3	4	1
3	298	1346	858	418	214	193	186	82	132
4	406	788	240	104	71	61	69	153	150
5	241	565	285	127	79	57	35	22	22
6	125	189	72	37	22	6	22	7	17
7	383	800	179	56	28	16	29	46	89
8	821	1533	510	188	76	25	22	16	15
9	378	888	396	171	90	72	35	21	13
10	273	592	177	59	41	11	11	10	7
11	222	416	145	88	51	26	21	6	8
12	254	524	199	69	30	22	15	18	12
13	381	781	205	81	30	17	8	8	4
14	72	245	146	49	31	15	12	9	6
15	410	812	285	127	70	34	23	15	11
16	177	507	149	49	23	20	10	5	7
17	183	585	447	321	200	138	72	42	38
18	356	866	448	223	127	67	58	34	28
19	1540	2386	748	223	79	31	19	14	10
20	194	233	50	19	2	4	0	1	2
21	335	702	287	170	94	51	38	29	33
22	332	834	400	226	125	71	44	37	35
23	229	543	186	80	34	21	14	18	2
24	279	591	229	70	51	25	14	17	16
25	119	1101	801	552	441	336	239	194	163
26	2089	6209	2758	1066	362	150	60	44	15
27	278	433	120	36	15	9	2	1	3
28	414	641	164	62	26	13	7	5	11
29	253	425	131	54	37	16	10	9	7
30	392	887	328	182	46	21	17	22	6
31	197	641	346	168	78	35	21	18	10

Table 2: Number of cointegrated set for each subdataset

It is observable that the number of cointegrated sets diminishes as the size of the stocks involved in the cointegration increases. This phenomenon can be attributed to the computational constraints imposed by the sheer volume of potential sets that can be analyzed. While one might initially assume that the absolute number of cointegrated sets would escalate with an increase in set size, the reality is nuanced.

As the number of analyzable sets grows exponentially with the size, the proportion of cointegrated sets within this expanding universe represents an increasingly smaller fraction. This leads us to the concept of the *curse of dimensionality*, a term originally coined in the context of high-dimensional statistical analysis. In essence, as the dimensions of an

analyzable space expand, the data sought within that space become increasingly sparse. This sparsity renders the observation of such data more challenging, as they are dispersed over a progressively larger volume.

Thus, the curse of dimensionality aptly describes the situation encountered in the search for cointegrated sets: the larger the set size, the more rarefied the cointegrated sets become, making them increasingly elusive in a vast dimensional space.

3.3 Sets Ranking and cointegration analysis

Despite the extensive search process, the number of identified sets remains excessively large. Given the high number of iterations involved, it becomes imperative to mitigate the occurrence of false positives. Consequently, the strategy necessitates a further reduction in the number of sets, focusing on those that are more likely to yield favorable trading outcomes.

A crucial criterion for sorting the sets is their mean reversion characteristic. Mean reversion, in this context, refers to the rate at which a stationary time series, after deviating from its mean, reverts back to it. The time series of interest here is the cointegrated series derived from the cointegrated sets. Sets exhibiting higher mean reversion rates offer more frequent opportunities for profit. This is due to the fact that a set with faster mean reversion diminishes trading uncertainty and reduces market exposure. To quantitatively assess and rank the mean reversion of the sets, the concept of half-life is employed. Half-life is defined as the time required for a financial instrument to revert to half of its deviation from the mean.

From a formal perspective, half-life is deduced from a mean reversion model. In such a model, where the current value of a process is a function of its deviation from the mean in the preceding period, the rate of autocorrelation decay is calculated. As such, sets with a lower half-life are deemed more attractive for implementation in a cointegration-based trading strategy.

Formally, a mean-reverting process is modeled as a first-order autoregressive process (AR(1)). This can be represented by the equation:

$$X_t = \phi X_{t-1} + e_t$$

where X_t is the cointegrated series or spread series.

Once the parameter ϕ is estimated through Ordinary Least Squares (OLS) regression, the half-life (denoted as h) of the process is calculated using the formula:

$$h = -\frac{\ln(2)}{\ln(\phi)} \tag{21}$$

This calculation indicates the time required for the process to revert to half of its current deviation from the mean. Subsequently, the sets are sorted based on their calculated half-life, starting with the lowest. From this sorting, the top 100 sets with the shortest half-life are retained for further analysis or trading strategies.

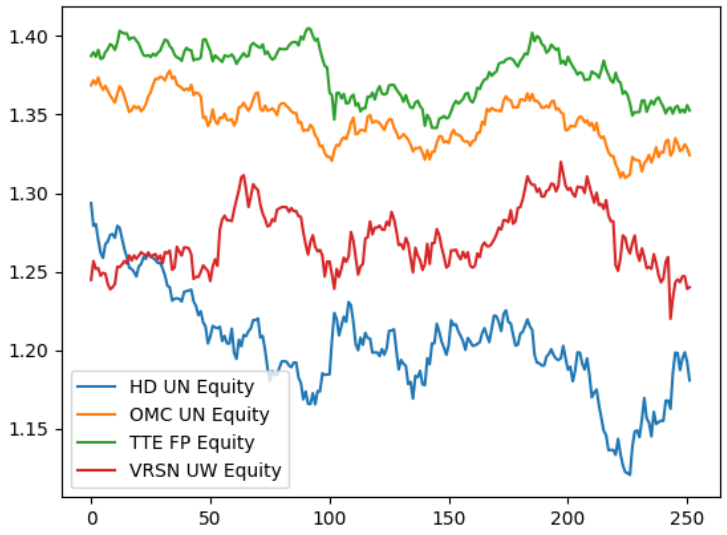

Figure 4: Normalized prices of HD, OMT, TTE and VRSN.

Figure 4 illustrates the normalized prices of four cointegrated stocks identified as viable for the trading strategy. These include HD (The Home Depot), a major U.S. company in the DIY retail sector; OMC (Omnicom Group), operating in the advertising industry; TotalEnergies (TTE), in the fossil fuels sector; and Verisign (VRSN), which specializes in grid infrastructure.

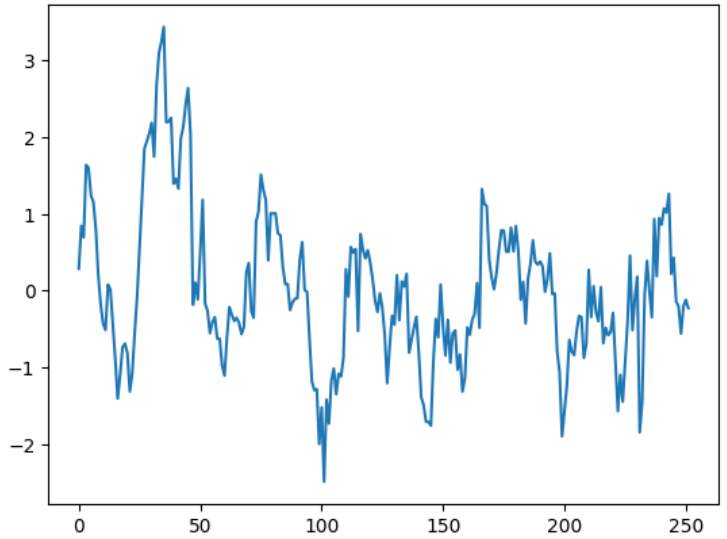

Figure 5: Normalized cointegrated series of HD, OMT, TTE and VRSN.

Figure 5 depicts the cointegrated vector of these four stocks, expressed formally as:

$$Cv_t = X \cdot \beta$$

where Cv_t is the cointegrated vector, X is the matrix of normalized prices,

$$X = [X_{\text{HD}}, X_{\text{OMC}}, X_{\text{TTE}}, X_{\text{VRSN}}]$$

and β is the vector of standardized cointegration coefficients,

$$\beta = [\beta_{\text{HD}}, \beta_{\text{OMC}}, \beta_{\text{TTE}}, \beta_{\text{VRSN}}] = [-1.199, 4.098, -0.818, -1.080]$$

Here, X is an $n \times k$ matrix, and β is a $k \times 1$ vector, where n is the number of observations and k is the number of stocks in the set. The obtained cointegrated vector Cv_t is then standardized:

$$\text{scv}_t = \frac{(Cv_t - \mu_{Cv})}{\sigma_{Cv}}$$

where μ is the mean of the vector Cv_t and σ is its standard deviation. To establish trading parameters based on the standardized cointegrated vector SCv_t, positions are opened when Scv_t exceeds the thresholds of 1 or -1. Specifically, the conditions for opening positions are as follows:

If SCv_t exceeds 1, a short position is opened. In the above example, this means opening a long position on HD, TTE, and VRSN for 1.199, 0.818, and 0.451, respectively, and opening a short position of 4.09 on OMC.he short position is closed when SCv_t returns to the mean, which is 0. So, if there is a short position, you will add to open long positions on assets with negative beta, while opening shorts on those with positive beta.

	HD	OMC	TTE	VRSN	Total
Opening price	\$3.562	\$3.953	\$3.997	\$3.389	
Closing price	\$3.556	\$3.905	\$3.961	\$3.436	
Starting position	\$1.199	\$-4.098	\$0.818	\$1.080	\$-1
Closing position	\$1.197	\$-4.048	\$0.811	\$1.095	\$-0.944
Return	\$-0.002	\$0.050	\$-0.007	\$0.015	\$0.056

The table provides an example of a short position identified in the training set. The weights are already standardized, and their sum is equal to either 1 or -1. It's worth noting that not all positions need to be in the positive to make the trade as a whole profitable.

If SCv_t falls below -1, a long position is opened. As illustrated in the following table, this entails opening long positions on assets with positive beta and opening short positions on assets with negative beta.

	HD	OMC	TTE	VRSN	Total
Opening price	\$3.588	\$3.959	\$4.002	\$3.375	
Closing price	\$3.594	\$4.004	\$3.996	\$3.388	
Starting position	\$-1.199	\$4.098	\$-0.818	\$-1.080	\$1
Closing position	\$-1.201	\$4.144	\$-0.817	\$-1.084	\$1.041
Return	\$-0.002	\$0.046	\$0.001	\$-0.004	\$0.041

Operationally speaking, this strategy involves a long/short approach in which it is also possible to create zero-dollar portfolios, where the long positions finance the short positions. In terms of parameters, it is also possible to adjust the strategy thresholds by multiplying them by a scalar. Specifically , the thresholds for opening positions can be modified as follows: $+1/-1 \times$ std $\times$ scalar. If the resulting number is less than 1, it will trigger more trades with lower margins. Conversely, if the number is greater than 1, it will reduce the number of trade opportunities but potentially result in more profitable trades.

3.4 Trading results

In each designated period of our dataset analysis, the initial year is dedicated to identifying cointegrated sets and calculating their respective weights. This entails applying the weights identified within the training set to the test set for each group of stocks. For the subsequent six-month period, a simulation of the trading strategy is undertaken. The process involves standardizing the cointegrated series for each stock set, monitoring the standardized series to identify when it exceeds predefined thresholds (thus triggering the opening and closing of positions), and forcefully closing any open position at the end of the period, irrespective of profit or loss. Future decisions regarding the allocation of investable capital among various sets necessitate a clear methodology for calculating returns on individual sets. Returns are fully reinvested throughout the trading period, spanning six months. This implies that profits derived from a trade within a cointegrated set are reinvested in the same stock set until the end of the trading period. As a result, returns for each set in each period are compounded continuously, causing the relative weight of each set in the portfolio to adjust dynamically, becoming either heavier or lighter by the end of the six-month interval.

Given the inherently quantitative nature of the proposed trading strategy, which encompasses both long and short positions, the segmentation of capital allocation emerges as a critical risk management measure. This strategy, while analytically rigorous, carries the potential for substantial loss, including the complete erosion of invested capital under certain circumstances. A primary concern is the possibility of incurring significant losses due to large and negative beta values on a particular stock within a cointegrated set. Should

the cointegration framework fail to hold, or if the stock price deviates excessively, there is a heightened risk of losing the entire capital allocated to that set. In scenarios where multiple sets simultaneously perform poorly, an evenly redistributed return mechanism could potentially deplete the entire capital base. To mitigate such risks, it is imperative to discretely parcel out the capital to each set. This approach ensures that underperformance in one specific set does not disproportionately impact the overall portfolio. Incorporating short trades into the strategy necessitates the consideration of a finite budget allocation for each set. This budget constraint implies that in instances where a set's performance is so dismal that its value diminishes to zero, any subsequent negative returns will not be considered. Under these circumstances, the capital allocated to the underperforming set would be deemed as lost. This measure serves to contain the impact of negative returns, preventing them from cascading across the entire portfolio.
In the realm of financial instruments suitable for the application of our trading strategy, Contracts for Differences (CFDs) emerge as the instrument of choice. CFDs represent an agreement between an investor and a financial institution to trade on the future value of an underlying asset. The key aspect of a CFD is that the profit or loss is determined by the difference between the asset's value at the opening and closing of the contract, with the net difference being settled in cash. CFDs offer several distinct advantages that align well with the requirements of our trading strategy:

1. **Leverage:** CFDs enable investors to manage trades that are significantly larger than the actual capital invested. This leverage aspect is particularly beneficial in amplifying the potential returns from small price movements of the underlying assets.

2. **Fractional Ownership:** The flexibility to purchase any fraction of a stock through CFDs enhances the adaptability of the strategy, especially when dealing with high-value stocks.

3. **Liquidity and Lower Costs:** CFDs are known for their high liquidity and relatively lower costs and fees. This feature makes them ideal for strategies that entail a high frequency of daily trades.

The characteristics of CFDs, including leverage, fractional ownership, and cost-effectiveness, synergize well with the proposed trading strategy. They facilitate a more dynamic and responsive approach to market movements, thereby enhancing the potential for profit realization in a highly fluid trading environment.
Having established the foundational assumptions and provided necessary clarifications, we now proceed to elucidate the methodology employed for calculating the returns of individual sets for each designated period.

The returns of each stock within a set are calculated using the formula:

$$r_{i,t} = \frac{P_{i,t+1} - P_{i,t}}{P_{i,t}}$$

where i denotes the stock, t represents the period, and $P_{t+1,i}$ is the price of the stock in the subsequent period $t + 1$.

To account for the nature of the trade, whether long or short, the returns are adjusted as follows:

$$r_{i,t}^{(I)} = r_{i,t} \times I_t$$

Here, I_t assumes the values 1, -1, or 0, corresponding to a long position, a short position, or no position, respectively.

The return for the entire period T for stock i in the set is calculated using the formula:

$$r_{i,T}^{(I)} = \left(\prod_{t=0}^{T=t} (1 + r_{i,t}^{(I)}) \right) - 1$$

This formula represents the product of the day-by-day, position-adjusted returns.

Given the above, the return of a set over the period T is:

$$r_{s,T} = \sum_{i \in S} W_i \times r_{i,T}^{(I)}$$

where $W_i = \frac{\beta_i}{\sum_{i \in S} \beta_i}$ and β_i is the cointegration coefficient of the time series in the set.

In complete form, the return of a set for one period is calculated as:

$$r_{s,T} = \sum_{i \in S} W_i \times \left(\prod_{t=0}^{T=t} (1 + r_{i,t} \times I_t) \right) - 1 \tag{22}$$

As deduced from the formula, the returns from each individual set are not only reinvested in the same set for the entire period, but the allocation within the sets also remains constant from the onset of the period. This implies that the returns generated by individual stocks within a set are compulsorily reinvested.

In the context of trading with Contracts for Difference (CFDs), the estimation of commissions takes into account both the bid/ask spread and overnight commissions:

- The bid/ask spread, representative of the difference between the buying and selling prices, functions akin to a commission applied to each trade. Given that the underlying assets are shares of the S&P 500 and Euronext, the spread applied can be considered relatively low, typically around 0.05%. The spread cost for a trading

period is calculated by multiplying this percentage with the trading volumes of the period.

- Overnight commissions refer to fees levied by CFD brokers for maintaining an open position overnight. This fee is applicable for each day the position remains open. An annual rate of 5% can be considered standard for such commissions. The total cost of these overnight commissions is calculated by taking the average transaction amount, multiplying it by the average duration of the transaction, and then multiplying this by the total number of transactions. This product is further multiplied by the annual overnight rate, which is then divided by 365, the number of days in a year. This calculation method considers the daily accrual of the overnight rate over the entire period of the transaction.

Set	Ave Return	Duration	Ave Spread	Ave Overnight	Net Ave Return
2	0.0099	13.5541	0.0009	0.0014	0.0076
3	0.0321	11.4186	0.0056	0.0070	0.0195
4	0.0323	10.3379	0.0077	0.0087	0.0159
5	0.0241	9.6522	0.0064	0.0068	0.0108
6	0.0279	9.0653	0.0080	0.0080	0.0120
7	0.0251	8.4735	0.0086	0.0080	0.0085
8	0.0472	8.0770	0.0158	0.0140	0.0174
9	0.0369	7.7262	0.0146	0.0124	0.0099
10	0.0428	7.3202	0.0129	0.0104	0.0195

Table 3: Average returns and costs for trade for each set

A notable trend observed from the data is the inverse relationship between the number of stocks in a set and the duration of trades. As the set size increases, the duration of each trade tends to shorten. This observation would generally lead to the expectation of reduced overnight commission costs, due to the shorter holding period of each trade. However, an increase in the number of stocks within a set also leads to an increase in leverage. This heightened leverage results in higher commissions related to the traded quantity. Furthermore, the data indicates a positive correlation between set size and both overnight commission and spread costs. As the number of stocks in a set escalates, there is a discernible increase in both overnight commissions and the spread. This pattern highlights the intricate balance between set size, trade duration, and the cumulative cost impact on the overall trading strategy. The table also provides insight into the average return per trade, both before and after accounting for commission costs. This data facilitates a comprehensive understanding of the net profitability of trades across various set sizes, considering the direct and indirect costs associated with each trade.

The graph below illustrates the performance of individual sets during a specific period. In this analysis, we focus on sets composed of 7 stocks identified in period 6. The choice of this period is primarily for the sake of graph comprehensibility, as it allows a clear presentation of the relevant information.

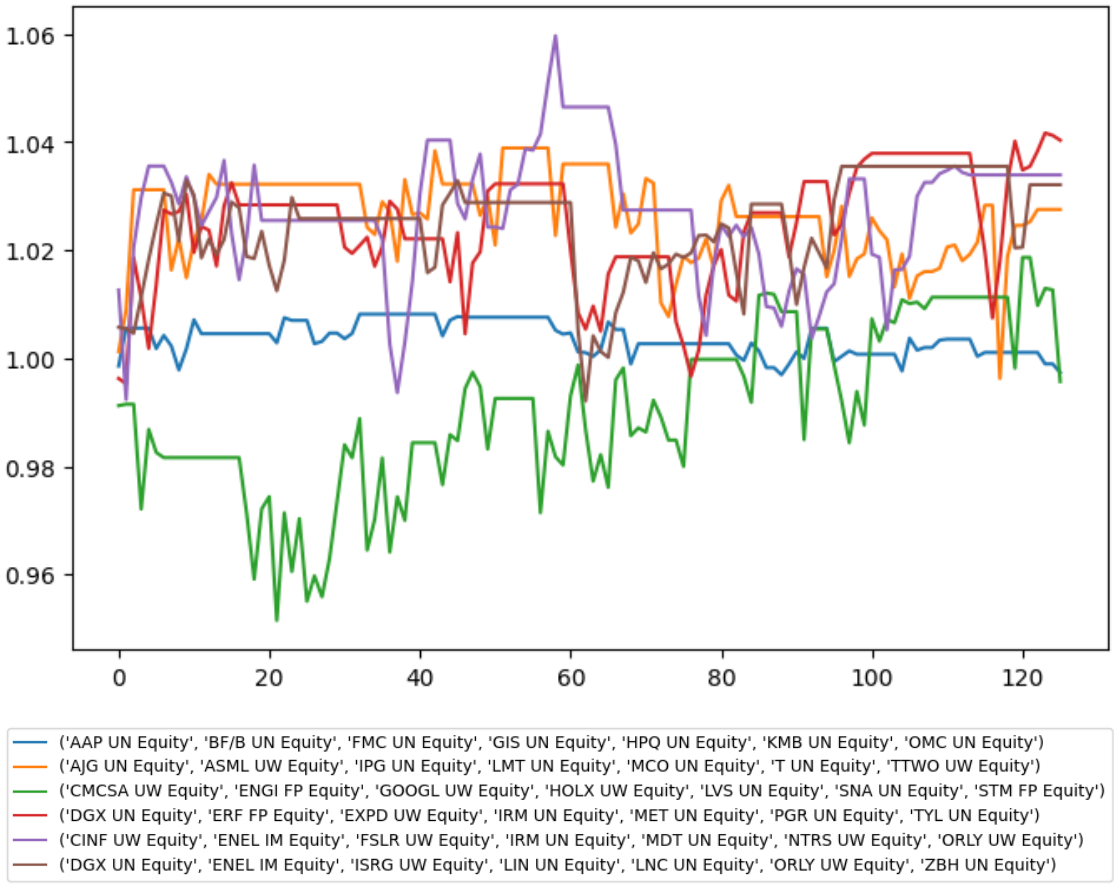

Figure 6: P&L of each set.

The sets displayed in the graph are carefully selected based on their composition of 7 stocks during the specified period. It is noteworthy that certain stocks, such as ENEL, are present in multiple sets. The potential impact of such overlap on exposure is contingent upon the cointegration coefficients of the stocks within the sets and the specific nature of open trades.

Figure 7 provides insights into the average performance of the sets previously depicted in Figure 6, considering both gross and net returns. The analysis distinguishes between returns before and after accounting for various fees, namely overnight fees and the spread. In the absence of commissions, the average return for the period stands at 4%, showcasing the strategy's raw potential. However, when accounting for fees, the net return is approx-

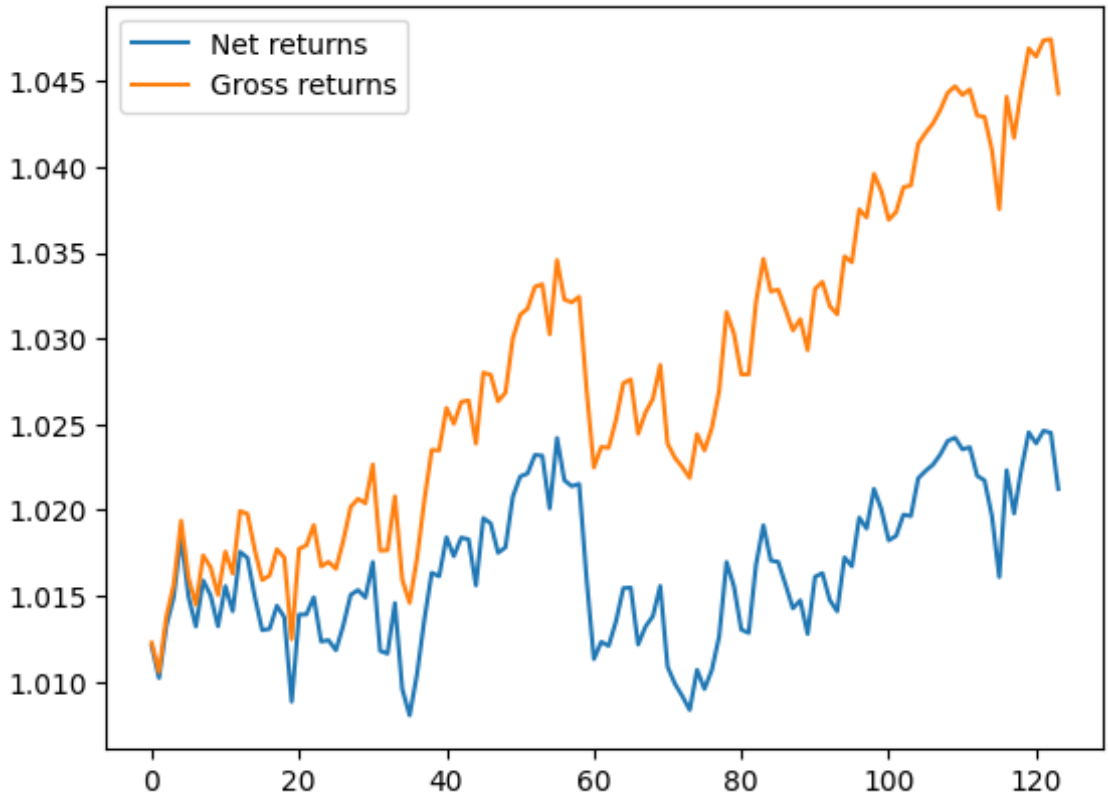

Figure 7: Average gross P&L vs net P&L.

imately 2%. This stark contrast underscores the substantial impact that fees have on the overall results of the trading strategy. The following table shows the average cumulative returns net of fees per set each size in each period

Table 4:

Index	2	3	4	5	6	7	8	9	10
1	0.0465	-0.4442	0.0634	0.0585	0.0672	0.0667	-0.0577	0.1201	-0.0592
2	0.0907	0.1020	0.2699	-0.0034	-0.0371	-0.0026	-0.0209	0	0
3	0.1331	0.0744	0.0392	0.0640	0.0926	-0.0389	0.5545	0.1056	0.0684
4	0.0334	0.0947	0.0793	0.0585	0.0318	0.0468	-0.0042	0.0678	-0.0174
5	0.0669	0.1721	0.0545	0.0540	0.0628	0.0054	142.3087	0.0673	0.0408
6	0.1894	0.0630	0.0324	0.1049	-0.0028	0.0212	0.1148	0.0324	-0.0112
7	0.0304	0.0626	0.0549	0.0349	-0.0174	0.0334	0.0108	0.0335	-0.0074
8	0.0584	0.1055	0.0322	0.0387	-0.0195	0.0703	0.0109	0.0738	-0.0016
9	0.0720	0.0769	0.0489	0.0385	0.0191	0.0203	0.1623	0.0377	0.0059
10	0.0123	0.0588	0.0401	0.0040	0.0747	0.0195	-0.0227	0.0174	0.2053
11	0.0472	0.1829	0.0612	0.0330	0.0779	0.1082	0.0336	0.0081	0.0266
12	0.0362	0.0537	0.0472	0.0303	-0.0654	0.0061	0.0824	0.0549	0.0528
13	0.0216	0.0480	0.0522	0.0683	0.2562	0.0845	0.0030	0.2297	-0.0083
14	0.0347	0.0462	0.2844	0.0507	0.0286	0.0521	-0.0157	0.0145	-0.0087
15	0.0071	0.0467	0.0228	0.0375	0.0395	0.0163	0.0068	0.0103	-0.0108
16	0.0295	0.0478	0.1261	0.0204	0.0082	0.1500	0.0161	0.0246	0.0046
17	0.0346	0.0746	0.0827	0.0404	0.0106	0.0147	-0.1275	0.0469	-0.0419
18	0.0169	0.0324	-0.0228	0.0486	0.0420	0.0082	0.0379	0.0179	0.0260
19	0.0334	-0.0410	0.0508	0.1049	0.0256	0.0242	0.0223	0.0359	-0.0464
20	0.0236	0.0578	0.0217	0.1297	0.5011	-0.0024	0	-0.0045	-0.0140
21	0.0202	0.1233	0.1434	0.0346	0.0076	0.0428	0.0001	0.0354	-0.0015
22	0.1086	-0.0685	0.0515	0.0283	0.0995	0.0232	0.0519	0.0617	0.0197
23	0.0278	0.0277	0.0995	0.0256	-0.0772	0.1699	0.0035	0.0345	0.1073
24	0.0115	0.0835	0.0430	0.0373	0.2760	0.0115	0.0061	0.0164	-0.0212
25	-0.0341	0.0037	-0.0057	-0.0138	-0.0368	-0.0236	-0.2117	0.0522	-0.0672
26	0.0171	0.1639	0.2092	0.0476	0.0254	0.0307	-0.0264	0.1968	-0.0122
27	0.0099	0.0367	0.0462	0.0200	-0.0127	0.0301	-0.0073	0.0066	-0.0037
28	0.0159	0.0364	0.0645	0.0271	0.0130	0.0641	-0.0117	0.0290	-0.0392
29	0.0196	0.0367	0.0419	0.0249	0.0084	0.0008	0.0091	0.0578	0.0007
30	0.0247	0.1034	0.0515	-0.0986	0.0494	0.0924	0.0391	0.0112	0.0744
31	0.0639	0.0834	0.0947	0.0720	0.0472	0.0394	-0.0057	0.0017	-0.0391

3.5 Creation of Portofolios

Given a set of N assets, the objective of portfolio optimization is to determine the optimal allocation of weights to these assets, maximizing return and minimizing risk. This problem can be formalized using the framework of Modern Portfolio Theory, particularly the Markowitz optimization model. The optimization problem is posed as follows:

$$\begin{aligned} \text{Minimize} \quad & \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} w_i w_j \sigma_{ij} \\ \text{subject to} \quad & \sum_{i=1}^{N} w_i = 1 \\ & \sum_{i=1}^{N} w_i R_i = \text{target return} \\ & w_i \geq 0 \text{ per } i = 1, 2, \ldots, N \end{aligned} \tag{23}$$

Here, w_i represents the weight of the i-th asset in the portfolio. The term σ_{ij} denotes the covariance between the returns of assets i and j, and R_i represents the expected return of the i-th asset. The first constraint ensures that the sum of the weights is equal to 1 (i.e., the entire budget is allocated), the second constraint sets the portfolio's expected return to a pre-specified target value, and the final constraint ensures that the weights are non-negative, reflecting a no-short-selling condition.

Markowitz's Portfolio Theory, a cornerstone of modern portfolio management, is predicated on several assumptions that pose significant challenges from an implementation standpoint. Key among these is the assumption of a normal distribution of returns. This assumption becomes particularly tenuous during periods of extreme market volatility or financial crises, as well as in scenarios involving assets that are cointegrated, as is the case here.

Furthermore, Markowitz's model assumes the stability of parameters such as returns over time. Given that portfolio optimization is inherently sensitive to parameter inputs, even minor variations in these parameters can lead to significant alterations in the portfolio's composition. This sensitivity underscores the practical difficulties of implementing Markowitz's model in dynamically changing markets.

In light of these challenges, a simpler and more pragmatic approach to capital allocation may be adopted: the equal weight portfolio strategy. Under this strategy, if there are n assets in the portfolio, each asset is assigned a weight:

$$w_i = \frac{1}{n} \quad \text{for } i = 1, 2, \ldots, n. \tag{24}$$

Here, w_i denotes the weight assigned to the i-th asset. This strategy ensures an egalitarian distribution of capital across all assets, circumventing the need for complex parameter estimation. The focus thus shifts solely to the number of assets in the portfolio, simplifying the portfolio construction process.

The efficiency of an equal weight portfolio is particularly notable when the underlying assets have similar returns and volatility profiles. An equal weight strategy, offers a compelling alternative to more complex weighting schemes, especially in contexts with significant uncertainty in parameter estimation.

DeMiguel's 2006 study underscores this point, revealing that equal weight portfolios often outperform more intricately constructed portfolios over the long term. This enhanced performance is attributed to the inherent uncertainty associated with forecasting future returns. Moreover, as highlighted in the studies conducted by Alexander Swade, one of the key advantages of equal weight portfolios in the long run can be traced to their increased exposure to small-cap companies, approximately three times higher than that of capital-weighted portfolios.

Swade's research contrasts equal weight portfolios with capital-weighted portfolios, the latter being the most prevalently utilized model in portfolio construction. In the context of this analysis, it becomes evident that the equal weight approach offers a robust alternative, especially when considering portfolios composed of cointegrated sets of assets.

Given the mathematical and statistical nature of cointegration, the decision was made to eschew market capitalization-based weights in favor of focusing on the characteristics of time series formed through cointegration coefficients. This approach stems from the belief that, for such sets, the market capitalization of individual stocks is less relevant than the dynamics captured by the cointegrated relationships.

An alternative approach to weight allocation in portfolio management, grounded in the principles of Modern Portfolio Theory, focuses on minimizing portfolio risk independent of returns. This method seeks to identify a combination of assets that achieves the lowest possible portfolio variance. The optimization problem can be formulated as follows:

$$\begin{aligned} \text{Minimize:} \quad & \sigma_p^2 = \mathbf{w}^\top \Sigma \mathbf{w}, \\ \text{subject to:} \quad & \sum_i w_i = 1, \end{aligned}$$

where:

- σ_p^2 represents the variance of the portfolio.

- $\mathbf{w}$ is the vector of portfolio weights.
- Σ is the covariance matrix of asset returns.
- $\mathbf{w}^\top$ denotes the transpose of the vector of weights.

However, this allocation scheme is not without its drawbacks. Primarily, it may lead to an overexposure to assets with low volatility. This is because the optimization process inherently favors assets with smaller variances in the pursuit of minimizing the overall portfolio variance. Such overexposure could potentially limit the portfolio's diversification benefits and may lead to suboptimal performance under certain market conditions.

In the realm of portfolio management, advancing beyond traditional capital allocation ratios, one can employ risk budgeting measures. This approach focuses not on the capital ratios but on the risk ratios one wishes to allocate to various assets. Unlike the traditional risk-return optimization approach, risk budgeting does not necessitate assumptions regarding various parameters. Instead, weights are determined in advance based on the level of risk one is willing to expose to each asset.

In the context of risk budgeting, risk parity is a concept analogous to equal weighting in capital allocation. Equal weighting implies allocating capital equally across assets, whereas risk parity involves distributing risk equally across assets. This is based on the understanding of how portfolio risk is defined and the individual risk contribution of each asset.

The formula for the risk contribution (RC_i) of a single asset in a portfolio is given by:

$$RC_i = \frac{w_i^2 \cdot \sigma_i^2 + \sum_{j \neq i} w_i w_j \cdot \sigma_{ij}}{\sigma_p^2} \tag{25}$$

where:

- RC_i is the risk contribution of asset i.
- w_i and w_j are the weights of assets i and j in the portfolio.
- σ_i^2 is the variance of asset i.
- σ_{ij} is the covariance between the returns of assets i and j.
- The summation $\sum_{j \neq i}$ denotes the sum over all assets j, excluding asset i.

To construct a risk parity portfolio, the goal is to equalize the risk contributions of all assets. This objective can be achieved by minimizing the difference between the risk

contributions of each asset in the portfolio:

$$\min_{\mathbf{w}} \sum_{i=1}^{N} \left(\frac{w_i^2 \cdot \sigma_i^2 + \sum_{j \neq i} w_i w_j \cdot \sigma_{ij}}{\sigma_p^2} - RC \right)^2 \tag{26}$$

In order to illustrate the differences in allocation strategies between equal-weighted and risk parity portfolios, we present two graphical representations. These graphs compare the allocation strategies using a set of seven stocks as an example. The process of

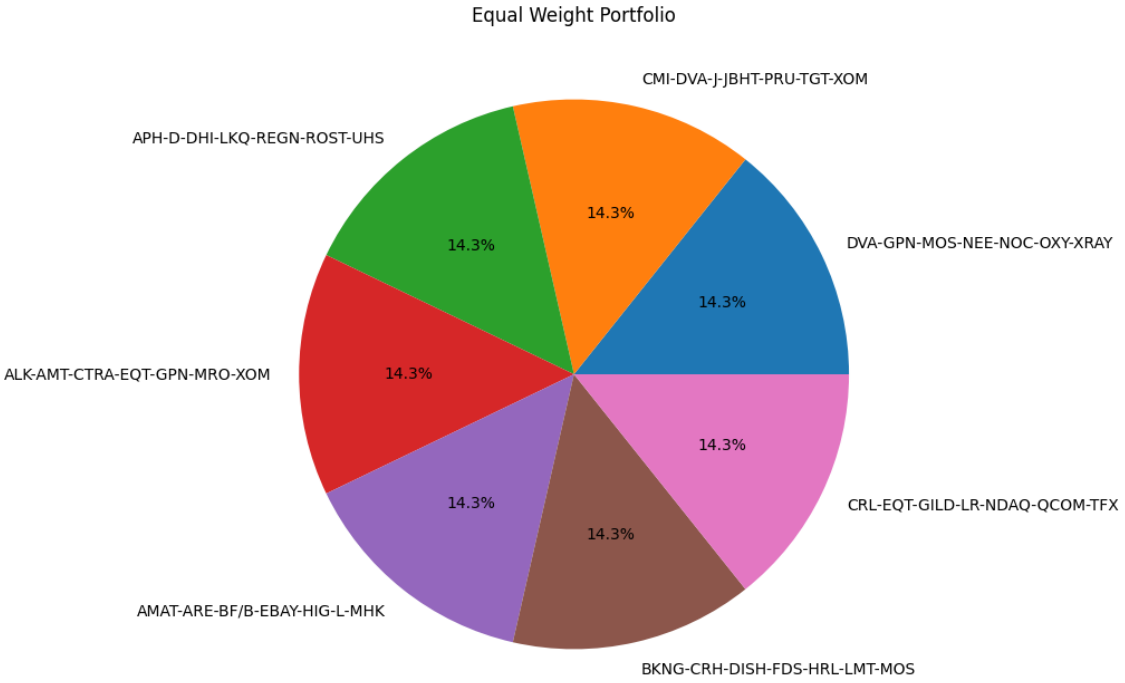

Figure 8: Equal weighted Portfolio.

portfolio construction, particularly when dealing with cointegrated sets of assets, involves a practical approach to determining the appropriate weights for each asset in the portfolio. This process can be described as follows:

1. **Identification of Cointegrated Sets:** In each period, the first step involves identifying sets of assets that exhibit cointegration. This identification is crucial as it dictates the composition of the portfolio for the upcoming period.

2. **Estimation of the Covariance Matrix:** Once the cointegrated sets have been identified, the next step is to estimate the covariance matrix of these sets. This estimation is typically performed using historical data, split into training and validation sets, to ensure robustness and validity of the covariance estimates.

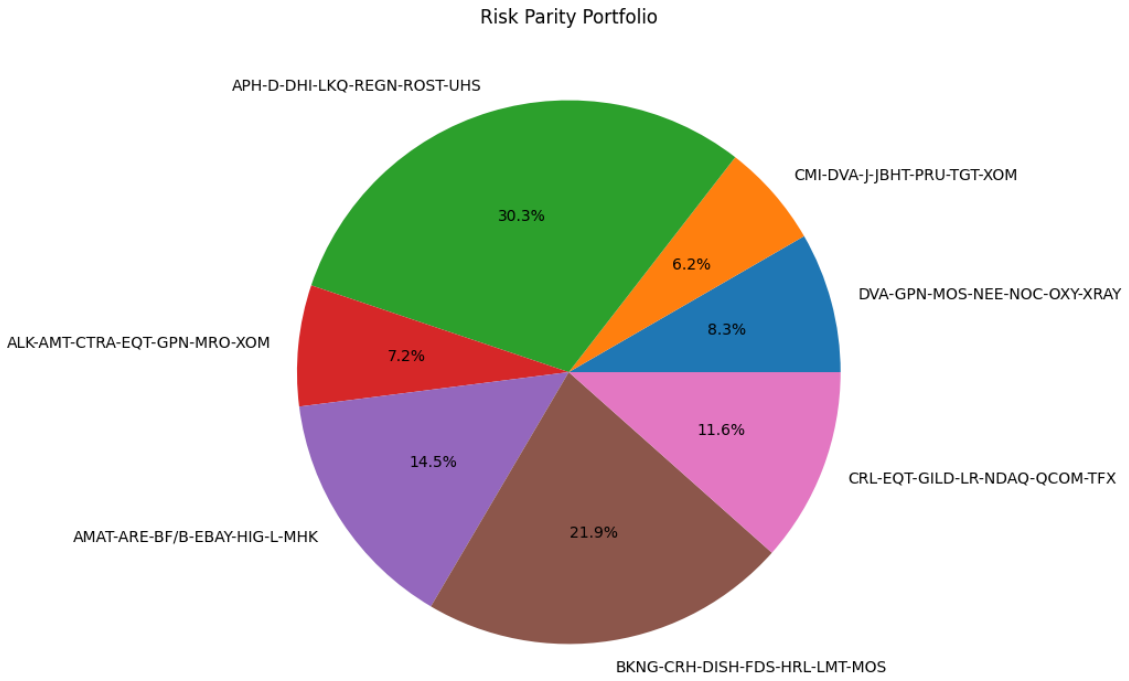

Figure 9: Risk Parity Portfolio.

3. **Calculation of Weights:** Based on the estimated covariance matrix, weights for each asset in the cointegrated sets are calculated. These weights are crucial as they determine the proportion of capital allocated to each asset in the portfolio.

4. **Application of Weights:** At the beginning of each period, the calculated weights are applied to the portfolio. This step marks the actual implementation of the portfolio strategy, where the theoretical weights are put into practice.

This approach ensures that the portfolio is constructed in a systematic and data-driven manner, taking into account the dynamic nature of asset relationships and market conditions. By updating the weights at the beginning of each period, the strategy remains responsive to the changing market environment.

4 Portfolio results

In this chapter, we undertake a comprehensive evaluation of the performance of 18 distinct portfolios, each constructed from cointegrated sets of stocks. For each set of cointegrated stocks, two distinct portfolios were constructed: one with equal capital allocation across each set (termed as "Equal Weighted"), and the other with risk distributed equally ("Risk Parity").

A crucial aspect of our analysis is to assess whether the strategy can be considered market neutral. Market neutrality in this context implies that the portfolio's performance is independent of the movements in the broader market, represented here by the S&P 500 index. To analyze the relationship between the S&P 500 and the returns of the portfolios, a regression analysis was conducted. This analysis involved regressing the returns of the portfolios against the returns of the S&P 500. The primary objective of this regression is to determine the beta of each portfolio. The beta, a measure of the portfolio's volatility relative to the market, is a critical indicator in this assessment. A beta close to zero would suggest that the portfolio is market neutral, meaning its returns are largely unaffected by market swings.

$$\begin{aligned} \text{Let:} \quad & R_p \text{ represent the portfolio returns,} \\ & R_m \text{ represent the market (S\&P 500) returns.} \\ \text{Then, the regression model is:} \quad & R_p = \alpha + \beta R_m + \epsilon, \end{aligned}$$

where α is the intercept, β is the portfolio's beta, and ϵ is the error term

	Beta	Alpha	r-value	p-value	Std Err
2	0.0027201	0.0003708	0.004075	0.7998	0.01073
3	0.0079362	0.0004207	0.009716	0.5455	0.01313
4	0.01700	0.0006066	0.02085	0.1946	0.01310
5	0.0064252	0.0003320	0.01357	0.3985	0.007609
6	-0.0067195	0.0004869	-0.006591	0.6817	0.01638
7	-0.01605	0.0004365	-0.01437	0.3711	0.01794
8	-0.42013	0.02191	-0.004244	0.7917	1.5909
9	0.0021105	0.0004102	0.003222	0.8411	0.01053
10	0.01004	0.0001524	0.01085	0.4997	0.01488

Table 5: Equal Weighted Portfolios regression results

Table 5 demonstrates that, with the exception of the portfolio composed of sets of 8 stocks, all other portfolios exhibit betas very close to 0. This finding confirms the strategy's risk neutrality with respect to the market. In contrast, the portfolio consisting of

	Beta	Alpha	r-value	p-value	Std Err
2	-2.137e-06	0.000147	-0.00001446	0.9993	0.002375
3	0.0005143	0.0001540	0.005809	0.7178	0.001423
4	0.0003613	0.0001241	0.004405	0.7840	0.001318
5	0.0018501	0.0001142	0.02122	0.1867	0.001401
6	0.0003638	0.00009314	0.002780	0.8626	0.002103
7	0.0022097	0.00006210	0.01834	0.2538	0.001936
8	0.0015805	-0.00003640	0.004695	0.7702	0.005409
9	-0.0002431	0.0001447	-0.001429	0.9291	0.002734
10	0.0101065	0.0001606	0.01093	0.4963	0.01485

Table 6: Risk Parity Portfolios regression results

8-stock sets has a notably negative beta relative to the market, specifically at -0.42. Similar observations can be made from Table 6, which pertains to the risk parity portfolios. Here again, the betas are consistently low across the board. A common feature in both tables is the r-value being very close to 0. It is important to note that this value represents the Pearson correlation coefficient, which ranges from -1 to 1. A value of 1 indicates perfect positive correlation, while -1 indicates perfect negative correlation. Furthermore, it is observed that in both tables, all p-values are significantly high, suggesting that the betas are not statistically significant. This implies that the beta values may not reliably predict future performance with respect to market movements. Regarding Jensen's alpha, the data show that they are positive for all portfolios except for the one consisting of 8 cointegrated stocks. This exception suggests a different performance characteristic for this specific portfolio configuration.

4.1 Portfolio Returns

In the following section, we present a series of graphs that illustrate the performance of various portfolios. For the sake of interpretability, these portfolios are paired based on the number of cointegrated stocks within their respective sets. This approach allows for a more nuanced comparison of portfolio behaviors under similar cointegration conditions. Each portfolio's performance is also compared against the S&P 500 index, serving as a benchmark. This comparison is crucial to understand how each portfolio stands relative to a broad market indicator and provides insights into their relative performance in different market conditions.

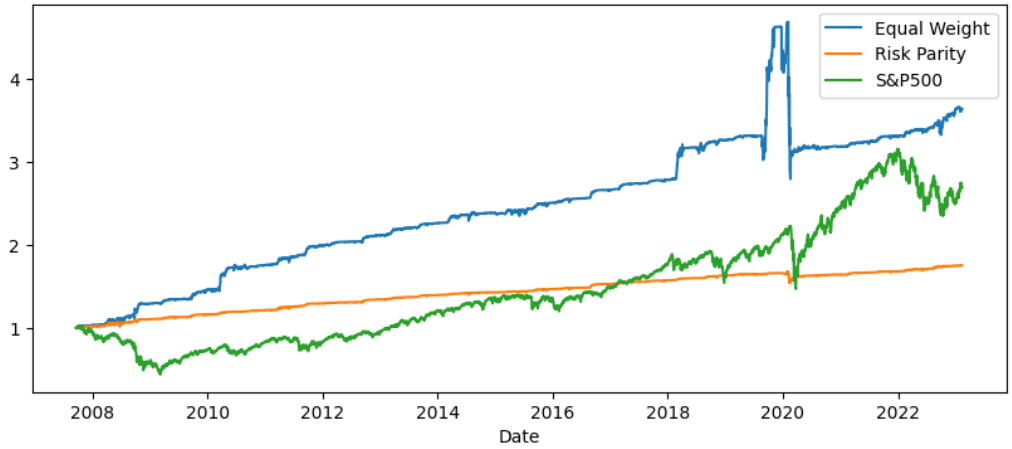

Figure 10: Cumulative returns of Portfolios of sets of 2 stocks.

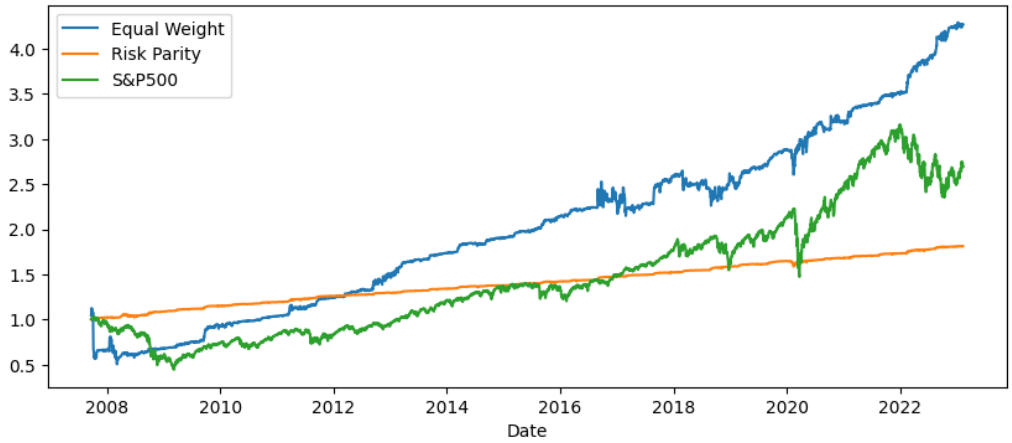

Figure 11: Cumulative returns of Portfolios of sets of 3 stocks.

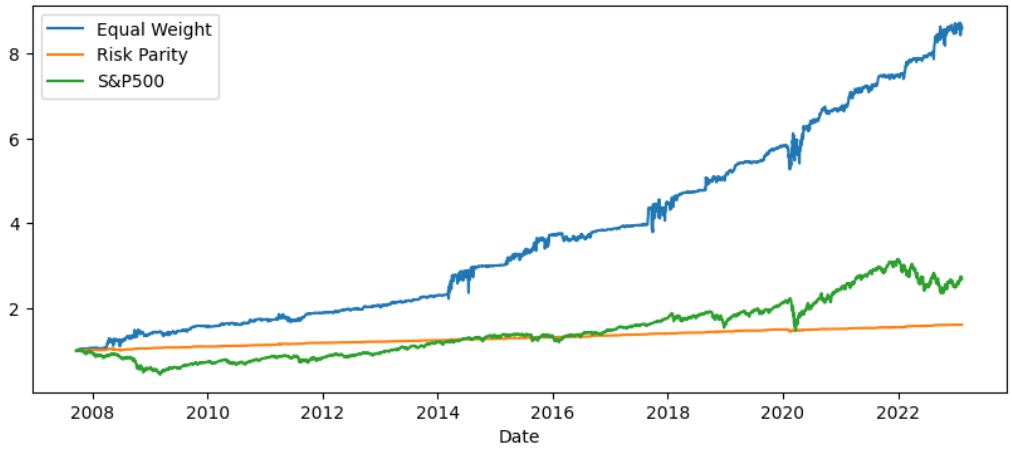

Figure 12: Cumulative returns of Portfolios of sets of 4 stocks.

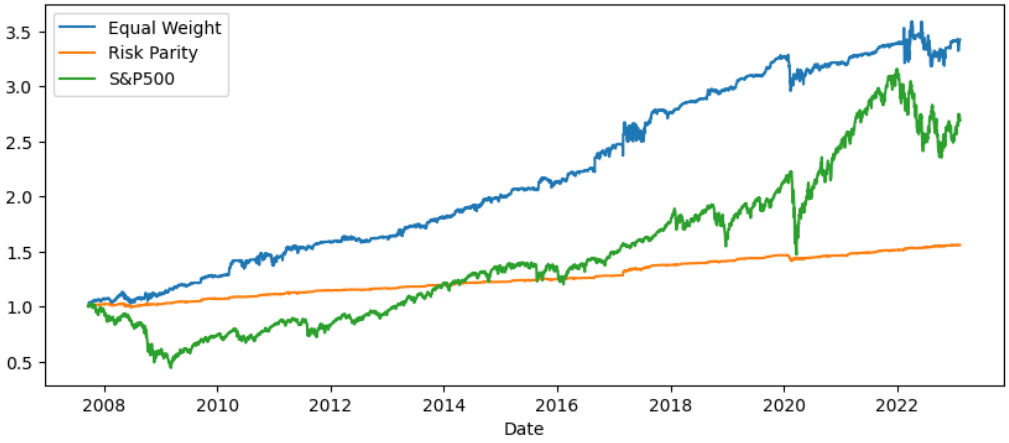

Figure 13: Cumulative returns of Portfolios of sets of 5 stocks.

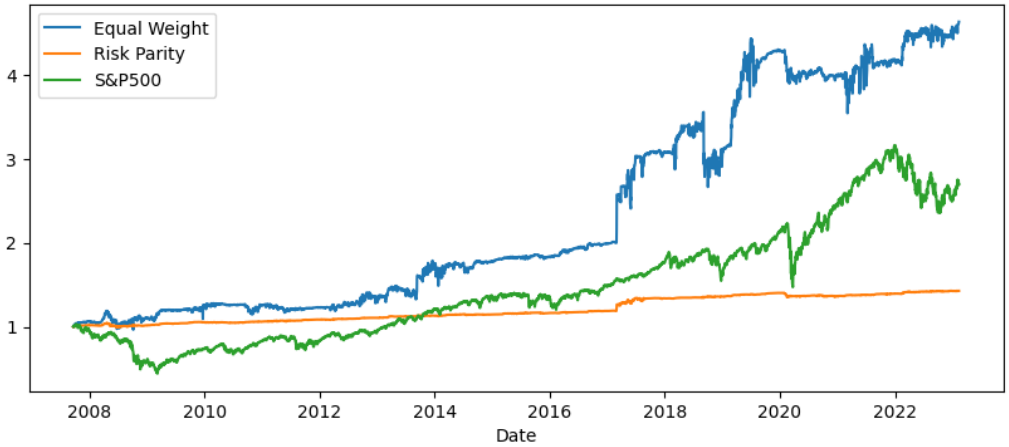

Figure 14: Cumulative returns of Portfolios of sets of 6 stocks.

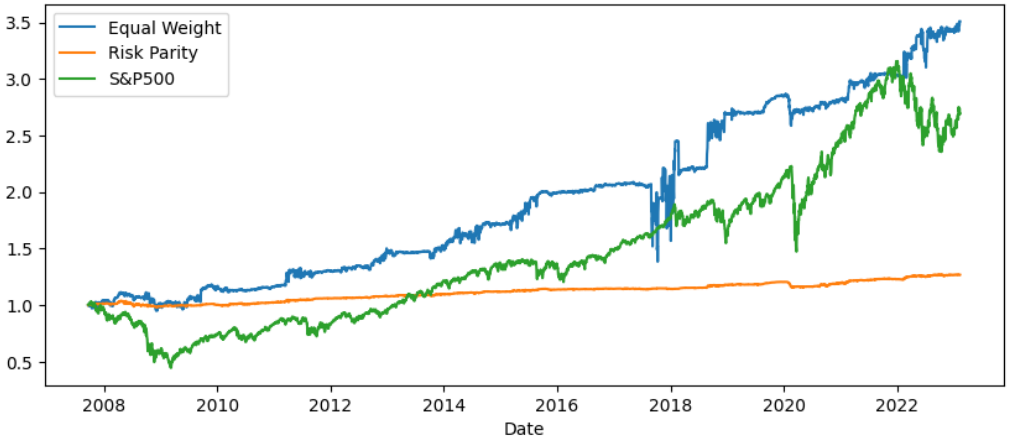

Figure 15: Cumulative returns of Portfolios of sets of 7 stocks.

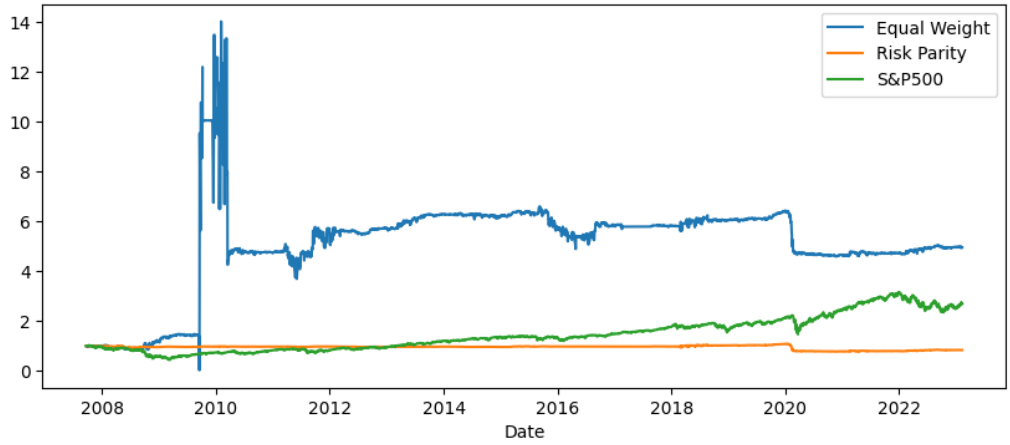

Figure 16: Cumulative returns of Portfolios of sets of 8 stocks.

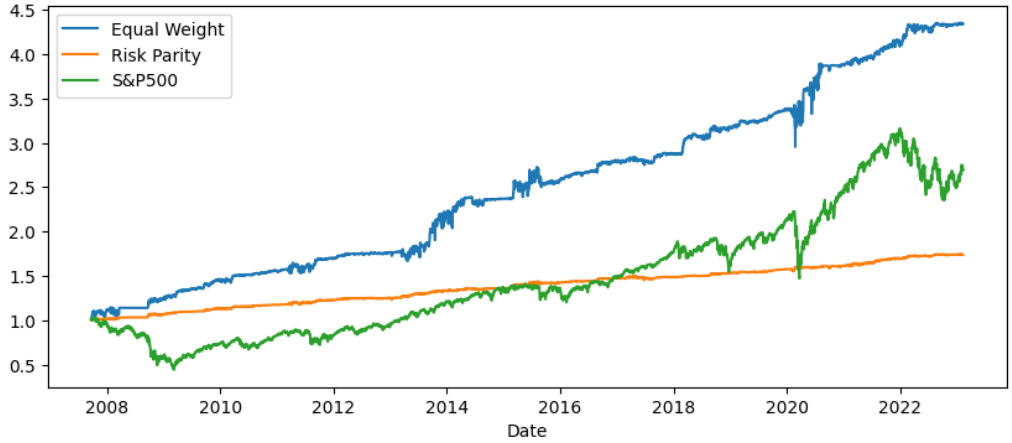

Figure 17: Cumulative returns of Portfolios of sets of 9 stocks.

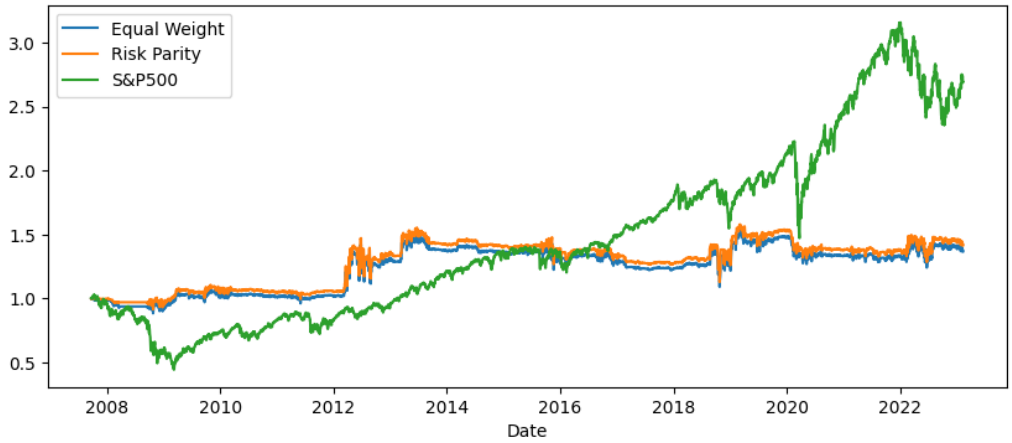

Figure 18: Cumulative returns of Portfolios of sets of 10 stocks.

As indicated in Table 7, among the equal-weighted portfolios, the one composed of 4-stock cointegrated sets emerges as the top performer with an impressive return of 761.14%. Observing its graphical representation in Figure 12, it is evident that the portfolio's value increases exponentially with rising market volatility. Portfolios constituted of 3, 4, 6, 8, and 9-stock cointegrated sets outperformed the classic pair-based cointegrated portfolio. The performance of portfolios with 5 and 7-stock sets is also commendable, closely matching the net return level. However, it is crucial to consider the long-term impact of fees on these portfolios, which are likely to be significantly higher than those for the pairs portfolio. Remarkably, all portfolios, except for the one composed of 10-stock sets, surpassed the benchmark. The portfolio with 1 0-stock sets yielded a mere return of 36.89%, marking it as the least successful. Moreover, the instability of the portfolio comprising 8-stock sets is notable. As previously discussed, the increase in the number of cointegrated stocks results in a reduction in the number of identifiable sets. This diminishing data pool renders the comparison of sets with more than 6 cointegrated stocks increasingly challenging. A particular issue arises in the portfolios composed of 10-stock sets, where at different times the portfolio consisted of only 1 or 2 sets. This limitation led to a scenario where the application of various weighting strategies resulted in similar portfolio performances. Notably, this scenario represents the only case where the risk parity portfolio outperforms the equal weighted portfolio. However, as depicted in Table 8 and corroborated by the respective graphical representations, the risk parity portfolios generally underperform, not only in comparison to their equal-weighted counterparts but also relative to the benchmark. A striking example is the portfolio with 8-stock sets, which even demonstrates a negative return of -16.12%.

Tables 10 and 11 detail the annualized returns of these portfolios. It is evident that, excluding the portfolio with 10-stock sets, the equal-weighted portfolios consistently yield higher returns than the benchmark. This pattern underscores the relative performance advantage of the equal-weighted strategy over the risk parity approach in most cases.

Set of 2	Set of 3	Set of 4	Set of 5	Set of 6	Set of 7	Set of 8	Set of 9	set of 10
263.85%	327.02%	761.14%	242.82%	362.97%	251.02%	396.07%	334.11%	36.89%

Table 7: Cumulative returns of Equal Weighted Portfolios

Set of 2	Set of 3	Set of 4	Set of 5	Set of 6	Set of 7	Set of 8	Set of 9	set of 10
75.60%	81.39%	61.72%	55.97%	42.85%	26.85%	-16.12%	74.22%	41.67%

Table 8: Cumulative returns of Risk Parity Portfolios

Cumulative	Annualized
169.33%	6.65%

Table 9: Returns of S&P 500

Set of 2	Set of 3	Set of 4	Set of 5	Set of 6	Set of 7	Set of 8	Set of 9	set of 10
8.75%	9.89%	15.02%	8.33%	10.47%	8.5%	10.97%	10.01%	2.06%

Table 10: Annualized returns of Equal Weighted Portfolios

Set of 2	Set of 3	Set of 4	Set of 5	Set of 6	Set of 7	Set of 8	Set of 9	set of 10
3.73%	3.95%	3.17%	2.93%	2.34%	1.56%	-1.14%	3.69%	2.29%

Table 11: Annualized returns of Risk Parity Portfolios

4.2 Volatility of Portfolios

The first measure for calculating portfolio risk is annualized volatility. Annualized volatility is a commonly used statistical measure of the variation in a portfolio's returns over time and is often employed as an indicator of the investment's risk. It signifies the extent to which the investment's returns can vary over a year. The higher the volatility, the greater the risk associated with the investment, as it indicates a higher uncertainty in future returns.

$$\text{Annualized Volatility} = \text{Daily Standard Deviation} \times \sqrt{252}$$

Where:

- Daily Standard Deviation is the standard deviation of daily returns.
- Number of Periods per Year depends on the frequency of the data: for daily data, it is typically 252 (the average number of trading days in a year).

This measure provides a clear idea of how much the portfolio's return can fluctuate on an annual basis, allowing investors to better assess the associated risk. In this analysis, we

Set of 2	Set of 3	Set of 4	Set of 5	Set of 6	Set of 7	Set of 8	Set of 9	Set of 10
0.1391	0.1702	0.1670	0.0986	0.2125	0.2327	20.6325	0.1365	0.1929

Table 12: Annualized volatility of Equal Weighted Portfolios

Set of 2	Set of 3	Set of 4	Set of 5	Set of 6	Set of 7	Set of 8	Set of 9	Set of 10
0.0308	0.0184	0.0171	0.0182	0.0273	0.0251	0.0701	0.0355	0.1926

Table 13: Annualized volatility of Risk Parity Portfolios

consider the annualized volatility of a benchmark, which is set at 0.20. It is observed that, with the of the portfolio composed of sets of 8 equal-weighted cointegrated stocks, which exhibits elevated volatility, all other portfolios closely match the benchmark's volatility.

In addition to annualized volatility, calculating the Max Drawdown (MD) is crucial. The MD represents the maximum percentage loss that an investor might suffer if they purchase an asset at its peak value and sell it at its lowest point. This measure is particularly useful as it provides a more tangible perspective for understanding compatibility with investment strategies. The MD can be mathematically represented as:

$$\text{MD} = \frac{\text{Bottom t-value}}{\text{Peak i-value}} \tag{27}$$

The following tables demonstrate the max drawdowns of the portfolios. Comparing these with the benchmark's max drawdown of -0.57, it is noted that, again, the portfolio consisting of sets of 8 cointegrated stocks with equal weight stands out.

Set of 2	Set of 3	Set of 4	Set of 5	Set of 6	Set of 7	Set of 8	Set of 9	Set of 10
-0.4033	-0.5519	-0.1839	-0.1137	-0.2506	-0.3374	-0.9796	-0.1377	-0.2734

Table 14: Max Drawdown of Equal Weighted Portfolios

Set of 2	Set of 3	Set of 4	Set of 5	Set of 6	Set of 7	Set of 8	Set of 9	Set of 10
-0.0826	-0.0371	-0.0408	-0.0364	-0.0489	-0.0561	-0.2858	-0.0303	-0.2734

Table 15: Max Drawdown of Risk Parity Portfolios

For equal-weighted portfolios, the portfolio with the lowest max drawdown is the one composed of sets of 5 stocks cointegrated, with an MD of -0.1137. Conversely, for risk parity portfolios, the one with the best performance is the one composed of sets of 9, with an MD of -0.1376.

In each case, it is observed that equal-weighted portfolios have a higher max drawdown than risk parity portfolios, reflecting the greater prudence of the latter type compared to equal-weighted portfolios.

4.3 Kurtosis and Skewness

In addition to measuring the volatility of returns, studying the distribution of returns by calculating kurtosis and skewness can provide deeper insights into the behavior of financial instruments.

Kurtosis is a statistical measure used to describe the heaviness of the tails of a probability distribution. More specifically, it measures the extent to which the tails of the distribution differ from those of a normal distribution, which has a kurtosis value of 3. Formally, kurtosis is the fourth standardized moment of the data.

Given a random variable X with mean μ and standard deviation σ, and x_i as the observations of X, the kurtosis is calculated as:

$$\text{Kurtosis} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \mu}{\sigma}\right)^4 \tag{28}$$

Where:

- n is the total number of observations.

- x_i is the i-th observation.
- μ is the mean of the observations.
- σ is the standard deviation of the observations.

The formula for kurtosis typically yields a value of 3 for a normal distribution. To assess the extent to which a distribution deviates from normality, the concept of excess kurtosis is used. This is obtained by subtracting 3 from the kurtosis value:

$$\text{Excess Kurtosis} = \text{Kurtosis} - 3 \tag{29}$$

A value of 0 in excess kurtosis indicates a normal distribution. A value greater than 0 suggests fatter tails compared to a normal distribution, while a value less than 0 indicates lighter tails.

Skewness is a crucial statistical measure that captures the degree of asymmetry of a probability distribution relative to its mean. It is a key concept in understanding the distribution characteristics of financial returns and other data sets.

Skewness quantifies both the direction and the extent of asymmetry. A distribution that is symmetric will have a skewness of 0. In contrast, distributions with asymmetric tails will exhibit non-zero skewness values.

Formally, the skewness of a distribution is akin to the standardized third moment of the data. Given a random variable X with mean μ and standard deviation σ, where x_i represents the observations of X, the skewness is calculated as:

$$\text{Skewness} = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i - \mu}{\sigma} \right)^3 \tag{30}$$

Where:

- n is the total number of observations.
- x_i is the i-th observation.
- μ is the mean of the observations.
- σ is the standard deviation of the observations.

The sign of the skewness value indicates the direction of the asymmetry:

- A **positive skewness** indicates a distribution with a longer right tail.
- A **negative skewness** indicates a distribution with a longer left tail.

This information is particularly important in financial analysis as it can indicate a propensity for extreme values in one direction over the other.

Set of 2	Set of 3	Set of 4	Set of 5	Set of 6	Set of 7	Set of 8	Set of 9	Set of 10
178.7118	142.9806	34.1045	18.8456	32.9462	83.9041	3853.6398	16.5741	34.7323

Table 16: Kurtosis of Equal Weighted Portfolios

Set of 2	Set of 3	Set of 4	Set of 5	Set of 6	Set of 7	Set of 8	Set of 9	Set of 10
79.8635	26.4472	35.6849	25.5518	65.5922	16.3217	83.1992	7.9411	35.0017

Table 17: Kurtosis of Risk Parity Portfolios

Set of 2	Set of 3	Set of 4	Set of 5	Set of 6	Set of 7	Set of 8	Set of 9	Set of 10
-3.0100	-4.6048	1.1804	0.2948	1.0194	2.1063	62.0346	0.1985	-0.0726

Table 18: Skewness of Equal Weighted Portfolios

Set of 2	Set of 3	Set of 4	Set of 5	Set of 6	Set of 7	Set of 8	Set of 9	Set of 10
-2.6892	0.5268	0.7948	0.5282	2.4499	-0.0789	-2.1084	0.0978	-0.0737

Table 19: Skewness of Risk Parity Portfolios

Among the equal-weighted portfolios, the portfolio composed of sets of 8 cointegrated stocks exhibits the highest kurtosis. This is immediately followed by the portfolios composed of 2 stocks, and then by those of 3 stocks. In each of these cases, the tails are much larger than in a normal distribution. For risk parity portfolios, kurtosis is also high, indicating that these portfolios have fatter tails in their distribution. As shown in Table 18, equal-weighted portfolios with sets of 2, 3, and 10 stocks have negative skewness, i.e., the left tail is larger than the right tail. Conversely, the skewness is positive for the other portfolios. In risk parity portfolios, those with sets of 2, 7, 8, and 10 stocks exhibit negative skewness. With the exception of portfolios consisting of 10 stocks, both skewness and kurtosis values are lower for risk parity portfolios compared to equal-weighted portfolios, indicating a lower risk exposure. The skewness and kurtosis of the benchmark are 11.49 and -0.24, respectively. This suggests that the distribution has larger tails than a normal distribution, with the left tail being larger than the right tail. Among the equal-weighted portfolios, those composed of sets of 4, 5, 6, 7, 8, 9, and 10 stocks have higher kurtosis and skewness than the benchmark. This indicates a greater presence of positive returns relative to the benchmark. For risk parity portfolios, all but those composed of 2 and 8 stocks show different skewness and kurtosis characteristics compared to the benchmark.

4.4 Sharpe Ratio

To make different portfolios comparable under a broader plan, it is important to consider not only returns but also volatility. This leads to the consideration of the trade-off between these two characteristics. Consequently, the Sharpe Ratio is calculated, which is the most widely used measure of the risk-return ratio. The Sharpe Ratio is formally defined as the average excess return over the portfolio's annualized volatility. The formula for the Sharpe Ratio is given by:

$$\text{Sharpe Ratio} = \frac{R_p - R_f}{\sigma_p}$$

where:

- R_p is the expected return of the investment portfolio.
- R_f is the return on a risk-free asset, such as the 3-month U.S. Treasury Bill.
- σ_p is the standard deviation of the portfolio's excess returns, representing the risk of the portfolio.

The Sharpe ratio of the benchmark is 0.22. As seen in Table 20, the equal-weighted portfolios with 2, 3, 4, 5, 6, 7, and 9 stocks are preferable to the benchmark. Among them, those with 4, 5, and 9 stocks are preferable to the cointegrated pairs portfolio. The

best among these is the portfolio with 4 stocks, which has a Sharpe Ratio of 0.77. This is 3.5 times better than the benchmark and about 1.5 times that of the pairwise portfolio.

As per Table 21, the Sharpe Ratios of the portfolios with 2, 3, 4, 5, and 9 stocks are higher than that of the S&P 500.

Set of 2	Set of 3	Set of 4	Set of 5	Set of 6	Set of 7	Set of 8	Set of 9	Set of 10
0.49	0.46	0.77	0.64	0.40	0.28	0.01	0.59	0.01

Table 20: Sharpe Ratio of Equal Weighted Portfolios

Set of 2	Set of 3	Set of 4	Set of 5	Set of 6	Set of 7	Set of 8	Set of 9	Set of 10
0.56	1.05	0.69	0.51	0.13	-0.18	-0.45	0.47	0.02

Table 21: Sharpe Ratio of Risk Parity Portfolios

5 Conclusion

This study commenced by laying the theoretical foundations, initiating with the long/short strategy employed by hedge funds and its economic-financial attributes. Subsequently, pair trading was introduced, followed by a comprehensive exploration of cointegration, encompassing stationarity and cointegration tests - notably the ADF and Johansen tests - essential in identifying cointegrated stocks.

The primary objective of the thesis was to develop and simulate the application of a statistical arbitrage strategy predicated on cointegration among various stock groups. This aimed at evaluating its practicality and observing the outcomes. For this purpose, a program was devised that initially selects the most profitable sets based on the quantity of cointegrated stocks and validates them. Based on cointegration, criteria for opening and closing positions were then established. Subsequently, the program simulated the portfolios' performance over test periods, gathering data for both equal weighted and risk parity portfolios composed of cointegrated sets ranging from 2 to 10 stocks.

The portfolios underwent testing over a period extending from September 2007 to February 2023, divided into 31 datasets. The equal weighted portfolios generally outperformed both the risk parity portfolios and the benchmark in terms of returns. These portfolios also exhibited comparatively low volatility and maximum drawdowns relative to the S&P 500. Observations of kurtosis and skewness indicated that the equal weighted portfolios generally maintained a robust profile against the benchmark index. Conversely, the risk parity portfolios, while demonstrating significantly limited volatility and maximum drawdown, were unable to sustain a satisfactory return profile. This was evident from the observed skewness and kurtosis, as the risk parity portfolios appeared to constrain the positive asymmetry of returns. Although the Sharpe ratios were predominantly positive and frequently surpassed the benchmark's ratio, portfolios comprising larger cointegrated sets, such as 8, 9, and 10, faced computational challenges in identifying an adequate number of sets, resulting in somewhat unstable portfolios across various metrics, particularly noticeable in those with 8 and 10 cointegrated stocks.

Excluding these, the remaining portfolios performed commendably across almost all parameters, affirming the efficacy of a multidimensional cointegration-based approach to pair trading. The most outstanding portfolio was the equal weighted one comprising 4 cointegrated stocks, which boasted the highest Sharpe ratio. Several risk parity portfolios also surpassed the benchmark's Sharpe ratio but were not as competitive in terms of returns. The study posits that in this strategy, the use of risk parity weights to diminish volatility and equitably distribute risk excessively impacts the returns, leading to inefficient capital allocation. Nevertheless, all strategies demonstrated market neutrality, as evidenced by their beta values being closely aligned with zero.

Despite the profitability of applying a multidimensional pair trading strategy, various factors warrant consideration. The computational constraint in set selection emerged as a significant challenge, complicating the identification of viable sets. As the size of the cointegrated sets increased, so did the complexity of portfolio management. This was particularly evident in larger sets where trading commissions had a more pronounced impact. Although trading commissions were estimated satisfactorily, major costs associated with short operations were overlooked. Moreover, risks inherent to short operations, such as margin risk and recall risk, were present. Despite the apparent insulation from market risk, the specific risk of holding individual stocks and potential limited market liquidity, mitigated by initial stock selection, remained pertinent concerns. These factors could potentially be managed through diverse money management measures or precautions like implementing a stop loss. Ultimately, this thesis did not extensively delve into the operational nuances of the strategy but rather demonstrated the feasibility of expanding the dimensionality of the cointegration approach within the statistical arbitrage framework.

Bibliography

- Bossaerts, Peter. "Common Nonstationary Components of Asset Prices". *Journal of Economic Dynamics and Control*, vol. 12, 1988, pp. 347–364.
- Chen, Joseph, Hong, Harrison, and Stein, Jeremy. "Breadth of Ownership and Stock Returns". *Journal of Financial Economics*, vol. 66, 2002, pp. 171–205.
- Conrad, Jennifer, and Kaul, Gautam. "Mean Reversion in Short-horizon Expected Returns". *Review of Financial Studies*, vol. 2, 1989, pp. 225–240.
- Engle, Robert, and Granger, Clive. "Co-integration and Error Correction: Representation, Estimation, and Testing". *Econometrica*, vol. 55, 1987, pp. 251–276.
- Goetzmann, William N., et al. "Sharpening Sharpe Ratios". Working paper, Yale School of Management, 2002.
- Hansell, Saul. "Inside Morgan Stanley's Black Box". *Institutional Investor*, May 1989, p. 204.
- Lhabitant, François-Serge. *Handbook of Hedge Funds*. John Wiley & Sons, 2007.
- Vidyamurthy, Ganapathy. *Pairs Trading: Quantitative Methods and Analysis*. John Wiley & Sons, 2004.
- Alexander, Carol. *Market Models: A Guide to Financial Data Analysis*. John Wiley & Sons, 2001.
- Tsay, Ruey S. *Analysis of Financial Time Series*. Wiley-Interscience, 2005.
- Dunis, Christian, Laws, Jason, and Naïm, Patrick. *Applied Quantitative Methods for Trading and Investment*. John Wiley & Sons, 2001.