EINDHOVEN UNIVERSITY OF TECHNOLOGY

Department of Mathematics and Computer Science
Data Science in Engineering

Applying higher-order correlation to optimize replacement of assets in stock portfolios

Master Thesis

Gregory Aerts

Supervisors:
dr. O. Papapetrou (TU/e)
prof. dr. D. Broeders (DNB)
dr. R. Heijmans (DNB)

version 1

Eindhoven, April 2021

Acknowledgements

The thesis has been written to fulfill my master degree Data Science in Engineering at the University of Technology in Eindhoven. First of all, I would like to thank all of my colleagues and all of the professors I have worked with during my masters degree. A special thanks to my direct supervisors dr. O. Papapetrou, prof. dr. D. Broeders and dr. R. Heijmans for their continuous support and guidance during this graduation project. I also want to thank my peers in the graduation group for their input.

Furthermore, a special thanks goes out to all of my friends and family for their support. They have always showed trust and optimism. In particular I would like to thank my brother Ir. T. Aerts, Msc. J. Van Sloun, Ing. N. Janssen, M. Ramaekers for their support and J. Aerts and T. Aerts for their enormous amount of patience with me. Last but not least F. Lodewijks deserves a special praise for her perpetual support.

Abstract

This thesis investigates optimal asset replacing strategies in portfolio management with minimal impact on a portfolio's risk characteristics. Changing investment opportunities, regulatory constraints or investor preferences might be reasons for replacing a small number of assets in an existing portfolio. Re-running a mean-variance optimization after deleting a limited number of assets could change the entire structure of a portfolio involving high transaction costs. Using the framework proposed in this thesis and an efficient algorithm implementation, it is possible to consider a large search space. The output of the algorithm only contains suitable sets of assets that can be included in the portfolio, according to user-defined requirements and the applied econometric methods. Using multiple correlation we aim to identify correlation patterns across the set of 'leaving' and 'incoming' assets simultaneously and match the characteristics of the leaving assets as well as possible. The algorithm outperforms, based on tracking errors, a random selection of up to ten replacing assets by a maximum of 62 percent. Larger replacing sets result in a lower outperformance, e.g. replacing 32 assets outperforms a random selection by a maximum of 50 percent. Our proposed framework allows portfolio managers to efficiently adapt their portfolio to changing investment opportunities using a large set of possible assets.

Contents

List of Figures

List of symbols

User-defined requirements:

- Correlation coefficient threshold τ
- Minimum jump δ
- Maximum degree of freedom ω
- Amount of solutions included in S

Important mathematical definitions of data:

- Current portfolio: $X = \{a_i\}_{i=1}^{N} \mid \forall a_i \in \mathbb{R}^m$
- Possibility set: $Y = \{y_j\}_{j=1}^{T} \mid \forall y_j \in \mathbb{R}^m$
- Single solution vector: sv
- total accumulated weight: AW
- Set of removed assets: $RA \subset X$
- Set of possible solutions: $CSV \subset \mathcal{P}(\mathcal{Y})$
- Each possible solution provided by algorithm: $sv \in CSV$
- Set of suitable candidate solutions: $S = \{S : \forall sv \in CSV : C(sv) \geq \tau\}$

Chapter 1

Introduction

Over time, portfolio managers dynamically replace assets in their portfolios. These replacements can be driven by innovations in the investment opportunity set, developments in regulatory constraints, or changes in their investment preferences. Changes in the expected risk and return characteristics of individual assets may motivate the asset manager to replace assets in his portfolio. They may also face regulatory constraints such as the exclusion of cluster ammunition manufacturers. Furthermore, the portfolio manager's investment preference might change, for instance, in favor of assets with specific characteristics such as sustainability. Investors, for example, move away from exposures to heavy polluting oil and gas companies and decide to reallocate their money to green companies. According to the research of [1] currently 75 percent of institutional investors and 77 percent of professional portfolio managers consider Environmental, Social and Governance (ESG) rated assets as a substantial part of their future portfolios. On top of this, [1] found that 71 percent of private investors prefer to have a positive social impact with their investments. Therefore, understanding the risk and return implications of replacing assets in a portfolio is of crucial importance.

Markowitz's modern portfolio theory [2] is a method to create mean-variance efficient portfolios. Consequently, replacing assets in such a mean-variance efficient portfolio may result in a sub-optimal solution. The portfolio manager might therefore decide to re-run the optimization algorithm again for a given set of assets. In theory, this could mean that replacing a single asset changes the entire composition of the portfolio. Rebalancing a complete portfolio is however not desirable for many reasons, the prime one being the high transaction costs involved. Researchers have accommodated transaction costs in their mathematical optimization models. However, this increases the complexity of the portfolio optimization algorithm substantially. The complexity of these algorithms might not be an issue for small sets of assets and small portfolios. Nevertheless, complex algorithms are not suitable and scalable to big data sets.

This thesis investigates a framework that is able to select replacing assets in a more extensive search space while approaching similar risk levels as the current portfolio, using advanced big data techniques. Multiple correlation is the measure used in the algorithm of this framework, which calculates similarity measures of aggregated vectors. Similarity measures can be any mathematical comparison mechanism, such as the Pearson correlation or the Minkowski

distance, between two numerical vectors. For applying Pearson to multiple time-series simultaneously, weighted averaging is used. Weighted averaging creates aggregated vectors of returns. This thesis investigates whether this framework could be an appropriate tool to replace assets in a portfolio without significantly changing the portfolio's risk and return balance. On top of this, this framework secures the remaining assets and their weights in the portfolio. With the help of higher-order correlation, portfolio managers are able to replace assets in a portfolio without re-calculating the efficient frontier. Using multiple correlation, we minimize the differences between the time series return characteristics of the 'leaving' assets and the 'incoming' assets. If the similarity is sufficiently high, then the risk characteristics of the overall portfolio will not change materially. In the end, the framework adapts an existing portfolio with a minimum loss in expected return while maintaining almost the same level of expected risks.

1.1 Problem definition

The goal of this thesis is the investigate the use of higher-order correlation in the context of portfolio management. This thesis creates a framework to replace a set of assets in a portfolio while keeping the remainder of the portfolio intact. With this proposed framework we investigate whether it is capable of identifying a suitable set of replacing assets by reviewing a large data set. This is done with the help of a 'discovery' algorithm and using quantitative and qualitative analyses.

The discovery algorithm used in this thesis is an adapted version of the work of [3] to detect higher-order relationships. With this algorithm we are able to review all possible combinations up to six assets, following an efficient implementation. This algorithm has been designed based on general idea behind the framework of Agrawal [4], which quantifies linear dependency between triangular relationships.

For the practical application we use an equity portfolio that closely represents the MSCI World Index. Due to disclosure agreements, all detailed information and reasoning is based on MSCI index. The portfolio we evaluate contains 1,000 different stocks weighted based on their market capitalisation and is considered to be mean-variance efficient for a given search space. Important to note is that in this research, only long positions in equities will be considered.

The input to our framework is a set of $N = 1,000$ assets, which together form the existing portfolio X. The weight of asset a_i in this portfolio is denoted by w_i, where $0 \leq w_i \leq 1$ with $i \in N$. Furthermore, each asset a_i has a m-dimensional vector, $\mathbb{R}^m$, where m represents the time-series over a given interval. Consequently, the portfolio is defined as

$$X = \{a_i\}_{i=1}^{N} \mid \forall a_i \in \mathbb{R}^m$$

A user-defined search space, the possibility set Y, contains assets y_j with time-series having the same lengths as the assets in the portfolio. The number of assets in the possibility set Y is indicated by T. Mathematically, we define this set as:

$$Y = \{y_j\}_{j=1}^{T} \mid \forall y_j \in \mathbb{R}^m$$

Then a user-defined set of assets RA, is removed from the current portfolio, where $RA \subset X$. With the help of multiple correlation measure we maximize the correlation coefficient between the aggregated 'leaving' vector and aggregated candidate solution vector SV, selected from the possibility sets $SV \subset Y$. Maximizing this coefficient provides us with a set of candidate solutions, CSV, which can be optimized further using other quantitative and qualitative techniques. All these steps are performed with the goal to minimize the distances between the return vectors of the set of leaving assets and the set of incoming assets. Small distances between these return vectors imply similar risk-return levels. In the end, with the solutions provided by the algorithm we aim to match the aggregated time-series pattern of the leaving set of assets as good as possible.

The metric represented in equation 1.1 is the multiple correlation measure used in the discovery algorithm of [3]. Using this measure, we calculate the coefficients for the multiple correlation between the RA and each possible solution $sv \in CSV$. The information gathered from the multiple correlation coefficients allows us to define a candidate solution set.

$$Correlation(Aggregation(\{RA\}), Aggregation(\{SV\})) \tag{1.1}$$

Each possible solution, $sv \subset Y$, is part of the set of all results in the output of the algorithm CSV which strictly contains subsets of Y. The power-set of Y, $\mathcal{P}(\mathcal{Y})$, contains all possible subsets of Y, which implies that the set of all possible solutions is a subset of the power-set of Y, hence $CSV \subset \mathcal{P}(\mathcal{Y})$. The multiple correlation formula 1.1 provides a higher-order correlation coefficient C for each possible candidate solution to replace the leaving set of assets. Furthermore, in the algorithm user-defined requirements UR are used which influencing the results. The solution set S only contains assets satisfying all these constraints, $UR(sv)$. For example, we define a minimum correlation threshold τ to solely consider interesting candidate solutions. UR can be seen as a collection of predicates, where $UR(sv)$ is true, if and only if all predicates are true. All in all, this creates a set S which only contains suitable candidate solutions:

$$S = \{S : \forall sv \in CSV \mid UR(sv)\}.$$

This thesis framework tries to define this S such that only suitable candidate solutions are provided. According to the working of the framework, every solution included in CSV automatically satisfies all user-defined requirements, which are elaborate on further in this research. All in all, it is important to note that we are not aiming to find the optimal solution, but try to deliver a set to the user with all possible candidate solutions satisfying the requirements.

1.2 Main results

Overall, we have conducted three different experiments: 1) random asset replacement, 2) replacing low sustainability rated assets with high sustainability rated assets and 3) replacement of all assets in the energy sector. From the results observed from these experiments, we have concluded that the framework has a decent performance. The framework outperforms

random asset selection in all cases. Improvements have been seen, based on the first moment tracking error, of 5-30 percent. Furthermore, we have observed improvements, based on the second moment tracking error, ranging from 10-65 percent depending on factors involved in the experiments, e.g., the size of the set of removing assets. Furthermore, we have discussed that the ranking strategies included in this thesis are necessary for defining a decent candidate solution set. If we are not using the full potential of the framework, meaning using the output of the algorithm directly, could in rare cases lead to unsuccessful replacements. This shows the importance of the post-processing phase in this framework.

1.3 Outline

The structure of this thesis is as follows. First, preliminaries and related work are presented in Chapter 2. In this chapter, relevant theories, such as modern portfolio theory and similarity measures, are brought to attention, and related papers are being discussed. Second, the framework and methodology are described in Chapter 3. Here the complete structure is discussed including the following three steps: 1) pre-processing, 2) main engine and 3) post-processing. Third, to evaluate the algorithmic performance and the framework, three different experiments are conducted and discussed in Chapter 4. Furthermore, the experimental setup is provided, and for each experiment, the expectations are postulated. Moreover, the results are being evaluated and compared with the expectations. Fourth, an overview of future work is provided in Chapter 5, such as applying different similarity measures in the framework or change other aspects of the framework. Similarity measures have a great impact on the performance, and the decision, which one to use, is based on knowledge gathered in the domain. However, using this particular similarity measure does not directly imply that other measures uncommon in the finance domain could not work. Last, a conclusion is provided in Chapter 6.

Chapter 2

Preliminaries & Related Work

In this section we present some preliminaries and introduce related work from different fields. Most importantly the fields of finance, risk management and data science. In order to structure the relevant information, this chapter is divided into five subsections: 1) portfolio management, 2) risk factors, 3) aggregation & similarity, 4) evaluation methods and 5) related work.
We attempt to replace assets from the current portfolio. As we assume an optimal allocation of assets in the current portfolio, relevant information regarding portfolio management is needed. First, we start by presenting returns and the most important topics of portfolio management, in particular modern portfolio theory and betas as a key risk measure. Modern portfolio theory is a mathematical optimization method to find an efficient portfolio according to the theory of Markowitz [5]. Explaining these theories and the mathematical interpretation behind this method, helps in understanding the problems related to replacing assets in a portfolio.

Second, we provide a short overview of risk measures important in portfolio management. An optimal portfolio is widely diversified according to mathematical optimization and exposure to risk factors. Only relevant risk factors for this particular research are explained: 1) market risk and 2) concentration risk.

Third, we elaborate on aggregation and similarity measures which could be used in this work. The aggregation measures are the key concept in the aggregated comparisons between the time-series of the returns, which allows us to use the multiple correlation algorithm. The aggregated vectors are used by the similarity measure, which quantifies a similarity between two vectors.

Fourth, to be able to asses whether proposed solutions are suitable replacement assets most commonly used evaluation method, tracking error, is explained. This evaluation method allow us to compare the current portfolio with the 'newly' designed portfolio.

Last, different works and papers are introduced to show some of the evolution related to the multiple correlation metric used in this research. The multiple correlation metric in this work is based on algorithms explained in relevant literature [6], [7] and [3].

2.1 Portfolio management

In portfolio management, asset managers aim to optimize the trade-off between portfolio risk and return by choosing the weights of the assets in the portfolio optimally. Investors benefit from diversifying their money over different asset classes and assets. The amount of invested money in an individual asset divided by the total amount of money invested determines the weight of this asset in the portfolio. Consequently, this weight is an indicator of the impact that the risk and return of this individual asset have on the portfolio. The rate of return of an asset can be tracked by analysing its logarithmic returns over time. Logarithmic returns allow us to compare returns of different assets simultaneously, due to several reasons e.g. time-additivity.

2.1.1 Logarithmic returns

Asset evaluation in portfolio management is done by time-series comparisons. In [8] several important aspects of logarithmic returns are described, which will be elaborated on by us. For portfolio managers, the exact price level is not important. By contrast, the change in an assets price level over time is important because this determines the revenue on their initial investment in this asset [1]. Therefore, portfolio managers consider returns instead of absolute asset prices. Considering a price P at time t, P_t, and a price P at time $t-1$, P_{t-1} the simple return of this asset can be described as [8]: $R_i = (P_t - P_{t-1})/P_{t-1}$.

Logarithmic returns is a form of normalization, where all variables, in this case time series of asset prices, are in a similar comparable metric. With this normalization, we enable the evaluation of relationships and comparisons between two or more assets. In finance it is common to use continuously compounded returns, or log returns [8], and not to evaluate simple returns. log returns are calculated as follows:

$$r_i = \frac{log(P_i)}{log(P_{i-1})}$$

log refers to the natural logarithm. There are various reasons for considering continuously compounded returns such as [8]:

- Log-normality
- Approximate raw-log equality
- Time-additivity
- Mathematical ease

First, log-normality of time-series of returns is convenient for conducting statistical analysis. If we assume that assets prices are log-normally distributed, then $log(1+r_i)$ is normally distributed.

[1] Investors may also receive cash flows in the form of dividends or coupons from investing in an asset. These cash flows are out of scope for this thesis.

The second reason might be the most important reason to portfolio managers. As portfolios contain many assets and each asset has different returns over time, portfolio managers should be able to look at total returns over time, meaning aggregate results. When considering an ordered sequence of m different changes to the portfolio, the compounding return is given by:

$$(1 + r_1)(1 + r_2) \dots (1 + r_m) = \prod_m (1 + r_i)$$

In probability theory it is known that the product of normally-distributed variables is not by definition normally distributed. With these results many statistical inferences can not be performed. However, it is known that the sum of normally-distributed variables is normally distributed and in combination with the logarithmic identity we find the following:

$$\sum_i log(1 + r_i) = log(1 + r_1) + log(1 + r_2) + \cdots + log(1 + r_m) = log(P_m) - log(P_0)$$

The compounding return of the initial investment can be calculated with the initial price at time 0 and the current price at time m. Furthermore, there is also algorithmic complexity improvement as this simplification reduces the O(m) multiplications to O(1) additions.

The third reason for applying logarithmic transformation is the concept of approximate raw-log equality. If returns between time i and i-1 are very small [2], the following is approximately true: $log(1 + r_i) \approx simplereturn_i$.

The fourth benefit of using continuously compounded returns instead of simple returns, is the mathematical ease of calculations. In the finance domain many mathematical calculations assume continuous time stochastic processes which rely on differentiation and integration. Multiplying small numbers, in our case daily returns, are subject to arithmetic underflow [3]. When using the logarithmic returns we eliminate this, as addition of logarithmic returns is not subject arithmetic underflow.

Logarithmic returns also have some limitations. For example, the normality assumption of price changes might not be accurate in real life cases, and therefore this 'advantage' might turn into a downside. However, the advantages presented above do outperform the downsides of using this transformation. So, the use of continuously compounded returns is appropriate and mostly standard in this domain, which is in line with the opinion of the domain experts in the fields. Therefore, discussion about other transformations are out of the scope for this project.

2.1.2 Modern Portfolio Theory

The primary goal of portfolio theory is to optimally allocate your invested money across different assets. For this purpose Markowitz introduced in 1952 mean-variance optimization, which is the most general portfolio allocation theory and the foundation for many other portfolio optimization theories. Modern Portfolio theory (MPT) is developed by Markowitz [5]

[2] between -0.5x2^{-127} and $-0.5x2^{-127}$ [9]

[3] Arithmetic underflow, also called floating point underflow, is a condition in computer science where the result of a calculation is too small for capturing in the memory of a CPU, see for example [10]

and can be seen as the evolved version of the efficient frontier method.

Across the efficient frontier, there is no better allocation of the investors money among the selected assets to achieve a better return for given level of risks. Or, the other way around, the risks could not be lowered while maintaining a similar level of expected return. Portfolios satisfying the optimal allocation are called efficient portfolios, located on the efficient frontier. This assumes that investors are risk-averse, in case an investor faces two different investments with the same expected return but two different risks, the investor prefers the lower risk.

For this thesis, we focus on single-period portfolio returns with N different stocks. To evaluate a portfolio, we need to calculate its expected rate of return and its expected risk. The expected return is the expected returns of the individual assets in the portfolio, weighted by the relative amount of money invested in each asset. Given portfolio, X, with N different assets and weights w_i, the expected return of a portfolio is denoted by:

$$E(R_p) = w_1 \cdot E(R_1) + w_2 \cdot E(R_2) + ... + w_N \cdot E(R_N)$$

Markowitz quantified risk with the help of statistical metrics: variance and co-variances. The expected risk is defined as the weighted sum of the variances of the returns of each individual assets plus the sum of weighted co-variances across these assets returns. The expected risk of a portfolio is given by:

$$\vartheta(R_p) = \sum_{i=1}^{N} w_i^2 \cdot \vartheta(R_i) + \sum_{i=1}^{N} \sum_{j \neq i}^{N} w_j \cdot w_k \cdot COV(R_j, R_k)$$

Using the fact that asset returns are not perfectly correlated, asset managers can diversify their portfolio. This is done by including assets from multiple asset classes or within an asset class using different assets. Markowitz's MPT is a primary strategy concerned with the degree of co-variance between asset returns in a portfolio. This diversification method intents to combine assets in a portfolio with returns that are less then perfectly positively correlated, this to lower total portfolio risk without sacrificing expected return. Diversification in this manner leads to an asset allocation that has the highest expected return for a given level of risk. These portfolios are called efficient portfolios, which are represented in figure 2.1. The efficient frontier is a linear programming optimization problem:

$$\min \sum_{i=1}^{N} \sum_{j=1}^{N} \sigma_{i,j} \cdot w_i \cdot w_j$$

Such that:

$$\sum_{i=1}^{N} u_i \cdot w_i = p$$

$$\sum_{i=1}^{N} w_i = 1$$

$$w_i \geq 0$$

One should take into account that in this evaluation method, investors take only two parameters into account, the expected return and the variance of returns. The set of optimal

portfolios, that provides the maximum expected return for a given level of risk, are all located on the efficient frontier. Portfolios not located on this hyperbole, shown in Figure 2.1 [4], are sub-optimal as they have a higher risks for similar returns. A combination of a risky asset portfolio with a risk-free asset is represented through the so called capital allocation line (CAL).

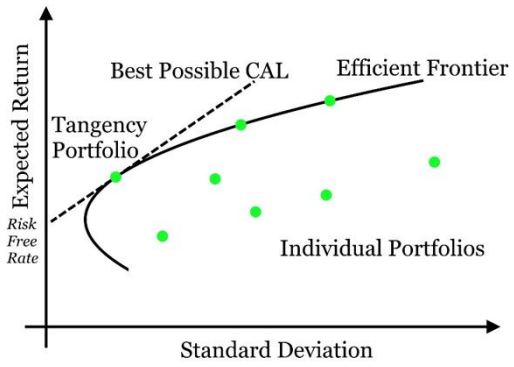

Figure 2.1: Markowitz frontier created by Scot Stockton

A general theory of financial markets, the Capital Asset Pricing Model (CAPM), assumes that all investors have homogeneous expectations on expected returns and risks. This leads to the Capital Market Line (CML) representing the optimal balance between risk and return. The CML is a special case of the CAL where the portfolio of risky assets is the market portfolio. The market portfolio is a portfolio consisting of a weighted sum of every asset in the market, with weights in the proportions that they exist in the market. The CAPM model was developed, amongst others, by William Sharpe in 1964. The slope of the CML is the Sharpe ratio of the market portfolio.

Limitations of MPT

One of the assumptions of MPT is that an investor has a one-period investment horizon. However, many private and institutional investors have a multi-period investment horizon and choose to dynamically update their portfolio over time. Reasons for changing a portfolio can be due to changes in the investment opportunity set, regulatory developments, or changes in their investment preferences. Several investors nowadays, e.g, try to invest money in assets which have a sustainable characteristics.

2.1.3 Beta

The MPT uses the variance and co-variance metrics for evaluating assets and then determining an optimal portfolio. A function of these two metrics is defined as beta and can help us better understand the behaviour of assets. Beta is a measure for systematic risk of an asset or a portfolio relative to the market portfolio in the CAPM [11]. A beta coefficient is able to capture the volatility of a single asset (for this thesis only stocks) in comparison to the market. Hence, we can use beta to compare the set of incoming assets with the set of leaving

[4] https://seekingalpha.com/article/4200744-difference-45-percent-return-and-28-efficient-frontier

assets and the complete portfolio. The calculation of beta is as follows [11]:

$$\beta = \frac{COV(R_a, R_m)}{\vartheta(R_m)} \tag{2.1}$$

$$\beta = \frac{\sqrt{\vartheta(R_a)}}{\sqrt{\vartheta(R_m)}} \times \rho(R_a, R_m)$$

with:

- R_a the return on a single asset
- R_m the return on the market portfolio
- COV the covariance as a measure of the joint variability of an asset's return and the market portfolio's return
- ϑ the variance of returns
- ρ the correlation or linear dependency between two variables

This formula shows that beta is a measures of return volatility, of an asset or portfolio, in comparison to the market portfolio. It provides us insights in the direction of a movement of a stock compared to the market. This volatility indicator provides information whether the current asset is more volatile and therefore more risky than the market. Ultimately, the investors is using beta to gauge the risk of adding this particular asset to a portfolio. Volatility and return mostly go side by side, as higher volatility also implies higher expected returns.

Beta can be interpreted best in combination with information regarding the correlation. Looking at Equation 2.1 we defined beta as a relation between the scaling of the variances and the correlation. Hence, for the purpose of this research it is important to address the following reasoning: given a correlation of 1 and a beta of 1, we can conclude that the volatility's of the compared assets are similar. Hence, a perfect replacement is found from a risk perspective. Therefore, with information regarding beta and correlation coefficient we are able to score equally good solutions better.

Beta provides us insights in the movement and the volatility of the investigated set of assets in comparison to a chosen benchmark. For this work only betas with positive coefficients are interesting. For example, a perfect beta equal to 1 would provide us with the information of a positive relationship. If it is known that the correlation coefficient is also 1, then we can conclude that the assets do move in similar direction and have a similar volatility. This is valuable information and quantifies characteristics of the time-series. On the contrary, if a beta equals 1 and the correlation might be 0.5, we conclude that this is not a good replacement solution. According to the coefficients of beta and correlation, we might be able to conclude that the volatility of the incoming assets might be twice as large as the leaving assets. If the linear dependency and the beta between the set of assets do match closely, we aim to identify the best combination of replacing assets.

2.2 Risk Factors

In the world of investments, many different risk are considered: market risk, liquidity risk, concentration risk, credit risk, reinvestment risk, inflation risk, horizon risk and foreign investment risks. In an optimal portfolio all these types of risks are well balanced. In this work only two different risk types are interesting to address: 1) market risk and 2) concentration risk. All other risk factors are obviously important in risk and portfolio management, but will not be touched upon in this research. In our work market risk is important in the selection of the data sets.

Market risk is the most general risk when investing money. The market risk is the exposure to the complete market and hence the exposure to the general economic developments and extreme events that affect the entire market. Therefore, when deleting and inserting new assets to an existing portfolio the exposure to market risk is changing accordingly. Therefore, when defining a collection of assets which might be suitable to replace, the exposure to market risk should be taken into account.

Furthermore, if we chose to adapt an existing portfolio by inserting and removing a set of assets does also impact the diversification. Inserting and deleting assets from an existing portfolio changes the risk characteristics of this portfolio. The risk related to these diversification aspects is captured best by concentration risk. This risk is associated with a lack of geographical, asset class or sector diversification in the portfolio. When investing in different asset classes, different industries and different locations the risk of a loss is spread.

With the help of these definitions we can justify decision in the data selection and use these definitions when reviewing solutions.

2.3 Similarity & aggregation

With all information regarding the modern portfolio theory and the risk factors, we are able to interpreted the mathematical formula of the optimal portfolio. Besides this, the risk factors give us insight in the diversification, which is needed when we change the asset allocation in the current portfolio. We select new incoming assets with the help of our framework, which uses similarity measures. The core of this thesis is based on the similarity between multiple time-series of returns. Using a similarity measure together with an aggregation method allows us to define this multiple correlation coefficient. Therefore it is important to elaborate on both the similarity measures and the aggregation methods useful for this research. In theory, any pair-wise similarity measurement in combination with an aggregation method could be applied for the use of multiple time-series comparisons simultaneously.

2.3.1 Similarity measures

For applying multiple correlation it is of vital importance to select a proper similarity measure. Every similarity measure which is capable of pair-wise comparisons is applicable to a specific implementation of multiple correlation. A similarity measure is a measure which determines

the abstract distance between time-series of returns of different assets. Time-series similarity has been extensively researched for years. For the purpose of this research we focus on two similarity measures from these distinct groups:

- Shape-based distance measures
- Stochastic information distances: more specifically information theory

The first category of time-series similarity measures is based on the shape of the time-series. This implies that these methods do directly compare the raw data of a pair of time-series. The most common similarity measure of the shape-based distance category is the Minkowski distance, also referred to as L_d norm. Given time-series T_1 and T_2 with length m this similarity measure is defined as:

$$d_{ld}(T_1, T_2) = \left(\sum_{i=0}^{m} (T_{1i} - T_{2i})^d \right)^{\frac{1}{d}}$$

Where d is a positive integer [12]. In case d=2 this formula is the well known Euclidean distance measure. The advantages of using these Minkowski distances is that the formula is intuitive, parameter free and has a linear complexity equal to the length of the time-series [13].

The Minkowski distance, and every variation of this, are also called lock-step measures as they are restricted to compare fixed pairs of data points. Elastic measures have been developed to solve this inflexibility, however these methods are much more computational complex. Dynamic time warping is an known elastic measure, but suffers from a complexity of $\mathcal{O}(\frac{t^2}{\log t})$ in time and space [14]. This is not considered to be suitable for this application and is therefore not used.

The second category, a method from information theory mutual information is introduced, as this method does not require any adaptation for implementation, making them interesting to investigate. This method does capture the information shared among two random variables [15]. The amount of shared information is quantified for one random variable by reviewing the other random variable. Also a generalization of the mutual information, total information, is shortly addressed. Both these theories are closely related to the entropy theories in information theory.

Pearson correlation

Pearson is a lock-step measures and widely used in finance. The Pearson correlation calculates a linear relation between two variables describing a linear dependency. This coefficient is defined between -1 to +1, [-1,1], where a zero indicates no linear dependence between the two variables. A positive value does indicate a positive association, the variables move mainly in the same positive direction. A negative coefficient indicates a movement in the opposite direction.

Formula 2.2 represents the calculation of the Pearson coefficient between 2 different random variables, in our case two different time series vectors T_1 and T_2. We define the daily

returns of the first time-series as r_i^1 and the second one as r_i^2. Then the Pearson correlation between these time-series of returns is defined as [5]:

$$r_{xy} = \frac{\sum_{i=1}^{m}(r_i^1 - \bar{r^1})(r_i^2 - \bar{r^2})}{\sqrt{\sum_{i=1}^{m}(r_i^1 - \bar{r^1})^2}\sqrt{\sum_{i=1}^{m}(r_i^2 - \bar{r^2})^2}} \tag{2.2}$$

With:

- $\bar{r^1}$ is the mean of the time-series T_1.
- $\bar{r^2}$ is the mean of the time-series T_2
- r_i^1 is the i^{th} value in T_1.
- r_i^2 is the i^{th} value in T_2.

To use the Pearson correlation in a proper way, the following assumptions are required:

- Homoscedasticity: equality of variances of multiple random variables
- Linearity
- Normality of variables
- Absence of outliers
- Related pairs: refers to the absence of missing values in pairs. In our case this would also holds for triple etc.
- Level of measurement: this refers to the fact that each variable should be continuous. If one of the variables are ordinal it should be better to use Spearman or Kendal's tau correlation.

Besides these assumptions, lock-step measures all have some restrictions regarding time shifts, handling outliers/noise and time warping. Furthermore, the time-series need to be of equal length.

Pearson correlation is very common to use in the field of finance and economics. By itself it is intuitive to interpret, but more important other statistical metrics work with same intuition and set-up. Variances, co-variance and betas can all be related with a correlation metric. To use this metric, the use of an aggregation method is required to compare multiple time-series simultaneously.

[5]Note that in the formulation m is the length of the time-series. In Pearson correlation, the length of the time-series needs to be equal.

Information theory

If we want to consider a non-linear dependency measure to quantify the relationship between multiple time-series, we should use methods originating from information theory. Information distances are based on empirical distributions and will use features from the fitted distributions to determine distances. One stochastic information distance, originating from information theory, is discussed briefly: mutual information. This distance measures is suited for implementing multiple correlation, as it is not designed for only pair-wise comparisons. It is unrestricted in the amount of time-series inputted simultaneously. This information criteria provides us with a certain 'similarity', which describes a relative distance. The higher the mutual information, the more information the time series share across them and are therefore more similar to each other.

Mutual information

Mutual information is a correlation measure based on probability theory and more specifically information theory [16]. Total correlation measures the amount of shared information across the set of variables. The adaptation for multivariate analysis implementation for relatively small data sets is already researched [16] and is very effective. However, the efficient implementation on big data sets has yet to be researched. Total correlation is a more specific implementation of mutual information and might be suitable for this project. Provided with a set containing a total of N random variables the total correlation can be defined as [16]:

$$C(X_1, X_2, ..., X_N) = \left[\sum_{i=1}^{N} H(X_i) \right] - H(X_1, X_2, ..., X_N) \tag{2.3}$$

where $H(X_i)$ is function capturing the information entropy of variable X_i [17]. The absolute difference between these information criteria defines the redundancy (in bits) present across all the variables. Hence, this is the quantitative measurement of shared structure across the variables.

The degree of dependence among group variables is the quantification of total correlation. As already implied by the formula, an information coefficient close to 0 indicates a statistically independent set of variables.

2.3.2 Aggregation Method

Next to a similarity measure, for the implementation of multiple correlation, a proper aggregation method is needed. This aggregation method allow us to capture information of multiple time-series into one 'aggregated' time series. This method is being applied for all possible combinations out there, and in the spectrum of big data, this should be implemented simply without any overhead.

An aggregation method is necessary to implement the multiple correlation. We need to capture information from multiple time-series and compress it to two single time-series, which can be compared. Different examples of aggregation methods are:

- Sum: the total sum of all raw data points of a variable

- Minimum: the minimum of all raw data points of a variable
- Maximum: the maximum of all raw data points of a variable
- Median: the median of all raw data points of a variable
- Average: the average of all raw data points of a variable

The minimum and maximum aggregation metrics allow us to accurately define outliers in time-series. However for this particular implementation the overall pattern of a time-series is important to define distances and similarities between time-series. Therefore, the average aggregation method is preferred for our particular implementation. This averaging method can be adapted to the weights of the replacing assets. For example, deleting stock A and B, with respective weights of 0.75% and 0.25%, the averaging method can average the according return time series in point to point comparisons with respectively: $075 * R_i^A + 0.25 * R_i^B$ for every point comparison. The information of asset A will be more dominant and therefore a better representation of the aggregated replacing vector of these assets is provided.

2.4 Evaluation methods

For performance evaluation of the framework proposed in this research, the most common evaluation method is defined. In the financial context the main evaluation method commonly used is tracking error [18].

2.4.1 Tracking error

A tracking error is defined between two different portfolios, where one of the portfolios will be defined as benchmark portfolio. This benchmark portfolio is often a benchmark, for example MSCI World Index. The iShares ETF closely matches this index. The tracking error is the divergence between the created portfolio and the benchmark portfolio. It is calculated as the standard deviation between two portfolio's. How far the portfolio deviates from the benchmark can be used to consider whether the additional risk is worth the additional pay-off. The tracking error between return on portfolio, R_p, and return on benchmark, R_b, is defined as follows:

$$TrackingError(R_p, R_b) = \sqrt{\vartheta(R_b, R_p)} \tag{2.4}$$

In our application we define the benchmark as our old portfolio. We are interested in differences in returns on daily basis between the old portfolio and the new portfolio, when assets have been replaced. With this we are able to evaluate every possible solution. First, we start by introducing the 'normal' tracking error, which is used for economical interpretation. Second, we define a relative tracking error, which accounts for the total removed weight, which allows us to compare tracking errors from different replacements.

Absolute tracking error

In the 'normal' tracking error, also called absolute tracking error, we do no account for the total weight of the removed assets RA. For economical interpretation, the absolute tracking error is used as it shows the deviation in comparison to the benchmark. A small deviation is

in economic evaluation always a good replacement, independent of the accumulated weight. For example, replacing set A with a weight of 10 percent with a tracking error of 1 percent is considered to be an as good as a replacement of set B with an accumulated weight of 0.1 percent and a tracking error of also 1 percent.

Relative tracking error

When evaluating multiple replacements in mathematical terms, the solutions should be scaled to similar dimensions. Therefore, considering the example above from the absolute tracking error, we should be scaling the tracking errors of the example by the total weight, resulting in:

- relative tracking error of $A = 1/0.1 = 10$
- relative tracking error of $B = 1/0.001 = 1000$

In mathematical terms, the first replacement is considered to be much more precise and therefore a better replacement.

2.5 Related Work

The framework created in this research is based on implementations which share similar reasoning about quantifying linear relationship between multiple time-series simultaneously. Many research areas benefit from understanding this complex relationships between time-series. For example, in the field of climate change, [19], [20] investigate the relationship between pressure dipoles time-series [6] in combination with other extreme climate events expressed in time-series, such as hurricanes or forest fires. Also in neuroscience relationships between time-series in highly complex systems have been detected in the work of [21]. These works all try to detect complex relationships between pairs of time-series.

S. Agrawal tried to capture a relationship between multiple time-series simultaneously, a higher-order relationship. in 2017 [6] defined a method to detect triangular relationships in time-series data. In 2019 [7] expanded this idea to a method capable of detecting higher-order relationships withing multiple time-series simultaneously, a multipole. A multipole is defined between a set of variables if: 1) the variables included in the set do show strong linear relationships and 2) each variable does contribute significantly to this linear relationship.

These works are some of the inspirations for the correlation detective created by [3]. The correlation detective algorithm is a more general implementation than the algorithms explained above. This algorithm allows us to detect higher-order relationships between any set of time-series simultaneously. Together with the efficient implementation and the generalized usability, this algorithm is suitable to apply in the financial context. The core algorithm used in this framework is an adaptive version of this work from [3].

[6] pairs of locations with a heavy negative dependency in sea level pressure, see for example [19]

Chapter 3

Data & Methodology

3.1 Data description

Our work is based upon two data sets: 1) the current portfolio including time-series of asset returns and 2) the possibility data set. The current portfolio data set consists of all assets present in the portfolio along with metrics such as the nominal investments and portfolio weights. This portfolio based on a sample of the MSCI World index. Investors can invest in this index via an exchange-traded fund (ETF) from iShares. All experiments are based on the assets and weights extracted from the MSCI index. This is done by extracting the weight distribution from the MSCI World ETF. We define the current portfolio as X containing N different assets. Each asset represented in the portfolio has a weight and a time-series of returns associated, with length m.

$$X = \{a_i\}_{i=0}^{N} \mid \forall a_i \in \mathbb{R}^m$$
$$w_i \in [0, 1] \mid \sum_{i=0}^{N} w_i = 1.$$

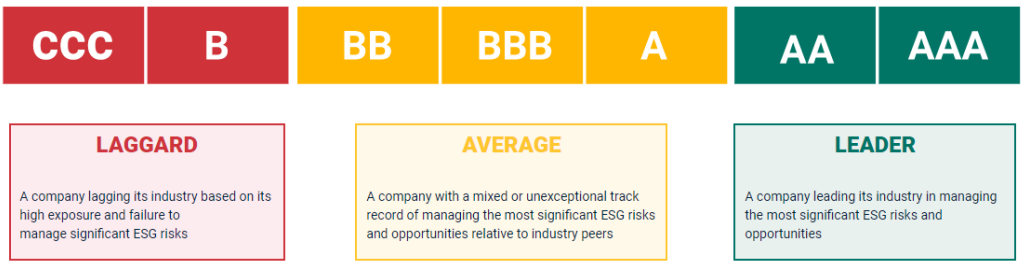

Figure 3.1: ESG Rating [1]

The possibility set Y contains time-series of returns of T different assets. Each return time-series, y_j, is a m-dimensional vector, where m represents all business days starting from 3-1-2017 up until 17-7-2020. The assets included in the possibility set are selected by the preferences of the end-user. The time-series associated with the assets are used by the algorithm to select suitable replacement assets. For the purpose of this research, only assets of sustainable companies are included. These assets are not necessarily new to the current

portfolio.

The possibility set can be changed according to different user-requirements, e.g. only European or tech sector companies, without changing the implementation of this framework. The possibility set contains 1,414 sustainable companies, all of them rated A, AA and AAA, according to the Environmental, Social and Governance (ESG) rating framework designed by MSCI illustrated in Figure 3.1. In Figure 3.1 the meaning of each rating is described and the order of the ratings is defined.

We define the possibility set as:

$$Y = \{y_j\}_{j=0}^{T} \mid \forall y_j \in \mathbb{R}^m. \tag{3.1}$$

Issuer Ticker	Name	Weight (%)	Price	Market Value	Sector	ISIN	Exchange	Location	Market Currency
AAPL	APPLE INC	3.69	127.35	219,355,153.65	Information Technology	US0378331005	NASDAQ	United States	USD
MSFT	MICROSOFT CORP	3.19	257.89	189,681,189.68	Information Technology	US5949181045	NASDAQ	United States	USD
AMZN	AMAZON COM INC	2.47	3,346.83	146,802,004.29	Consumer Discretionary	US0231351067	NASDAQ	United States	USD
FB	FACEBOOK CLASS A INC	1.37	331.26	81,380,975.46	Communication	US30303M1027	NASDAQ	United States	USD
GOOG	ALPHABET INC CLASS C	1.28	2,513.93	75,978,506.39	Communication	US02079K1079	NASDAQ	United States	USD
GOOGL	ALPHABET INC CLASS A	1.26	2,430.20	74,935,217.00	Communication	US02079K3059	NASDAQ	United States	USD
JPM	JPMORGAN CHASE & CO	0.84	160.29	49,861,891.17	Financials	US46625H1005	New York Stock Exchange Inc.	United States	USD
TSLA	TESLA INC	0.81	609.89	48,270,963.83	Consumer Discretionary	US88160R1014	NASDAQ	United States	USD
NVDA	NVIDIA CORP	0.76	713.01	45,144,941.16	Information Technology	US67066G1040	NASDAQ	United States	USD
JNJ	JOHNSON & JOHNSON	0.75	164.96	44,705,479.68	Health Care	US4781601046	New York Stock Exchange Inc.	United States	USD
V	VISA INC CLASS A	0.69	234.96	40,832,523.60	Information Technology	US92826C8394	New York Stock Exchange Inc.	United States	USD
BRKB	BERKSHIRE HATHAWAY INC CLASS B	0.68	286.82	40,346,395.76	Financials	US0846707026	New York Stock Exchange Inc.	United States	USD
UNH	UNITEDHEALTH GROUP INC	0.65	397.89	38,738,172.51	Health Care	US91324P1021	New York Stock Exchange Inc.	United States	USD
NESN	NESTLE SA	0.63	126.64	37,474,039.81	Consumer Staples	CH0038863350	SIX Swiss Exchange	Switzerland	CHF
HD	HOME DEPOT INC	0.58	310.77	34,451,962.20	Consumer Discretionary	US4370761029	New York Stock Exchange Inc.	United States	USD
PG	PROCTER & GAMBLE	0.58	134.86	34,389,974.30	Consumer Staples	US7427181091	New York Stock Exchange Inc.	United States	USD
MA	MASTERCARD INC CLASS A	0.56	365.5	33,389,521.50	Information Technology	US57636Q1040	New York Stock Exchange Inc.	United States	USD
BAC	BANK OF AMERICA CORP	0.56	41.86	33,385,484.86	Financials	US0605051046	New York Stock Exchange Inc.	United States	USD
DIS	WALT DISNEY	0.55	177.38	32,835,343.94	Communication	US2546871060	New York Stock Exchange Inc.	United States	USD
PYPL	PAYPAL HOLDINGS INC	0.52	271.45	31,071,524.25	Information Technology	US70450Y1038	NASDAQ	United States	USD
ASML	ASML HOLDING NV	0.5	694.66	29,871,695.36	Information Technology	NL0010273215	Euronext Amsterdam	Netherlands	EUR
ROG	ROCHE HOLDING PAR AG	0.46	380.53	27,410,229.76	Health Care	CH0012032048	SIX Swiss Exchange	Switzerland	CHF
XOM	EXXON MOBIL CORP	0.46	62.17	27,040,903.67	Energy	US30231G1022	New York Stock Exchange Inc.	United States	USD
ADBE	ADOBE INC	0.45	541.26	26,848,661.04	Information Technology	US00724F1012	NASDAQ	United States	USD
CMCSA	COMCAST CORP CLASS A	0.44	56.88	26,321,561.28	Communication	US20030N1019	NASDAQ	United States	USD
VZ	VERIZON COMMUNICATIONS INC	0.41	57.33	24,338,247.57	Communication	US92343V1044	New York Stock Exchange Inc.	United States	USD
INTC	INTEL CORPORATION CORP	0.41	57.85	24,206,001.95	Information Technology	US4581401001	NASDAQ	United States	USD
CSCO	CISCO SYSTEMS INC	0.4	54.77	23,708,016.05	Information Technology	US17275R1023	NASDAQ	United States	USD
KO	COCA-COLA	0.4	56.16	23,603,711.04	Consumer Staples	US1912161007	New York Stock Exchange Inc.	United States	USD
MC	LVMH	0.39	809.1	22,986,602.17	Consumer Discretionary	FR0000121014	Nyse Euronext - Euronext Paris	France	EUR
PFE	PFIZER INC	0.38	40.15	22,875,542.80	Health Care	US7170811035	New York Stock Exchange Inc.	United States	USD
CRM	SALESFORCE.COM INC	0.38	240.31	22,629,992.70	Information Technology	US79466L3024	New York Stock Exchange Inc.	United States	USD
WMT	WALMART INC	0.38	140.75	22,461,588.75	Consumer Staples	US9311421039	New York Stock Exchange Inc.	United States	USD
NFLX	NETFLIX INC	0.37	488.77	22,071,875.66	Communication	US64110L1061	NASDAQ	United States	USD
T	AT&T INC	0.36	29.32	21,407,323.64	Communication	US00206R1023	New York Stock Exchange Inc.	United States	USD

Figure 3.2: portfolio snapshot

Figure 3.2 presents a sample of the current portfolio, including all associated data fields. All included assets have a weight in the existing portfolio. Figure 3.2 also shows the weight distribution across countries and is diversified accordingly. Figure 3.3 provides a distribution of the number of assets over the sectors. In total the current portfolio has a size of approximately 1,000 assets and is widely diversified according to all risk factors described in Chapter 2.

From the current portfolio the 'leaving' assets will be selected. This set of leaving assets and the possibility set are used by this framework to apply the multiple correlation algorithm. Furthermore, all assets in this current portfolio data set are used for calculating the benchmark. This benchmark represents the current 'optimal' portfolio. Removing and adding assets will be evaluated according to a first and second moment tracking error. The second moment tracking error is defined in two ways. First, the tracking error between the returns of the leaving and incoming assets is determined. Second, the tracking error between the return of the old portfolio and the newly created portfolio is calculated.

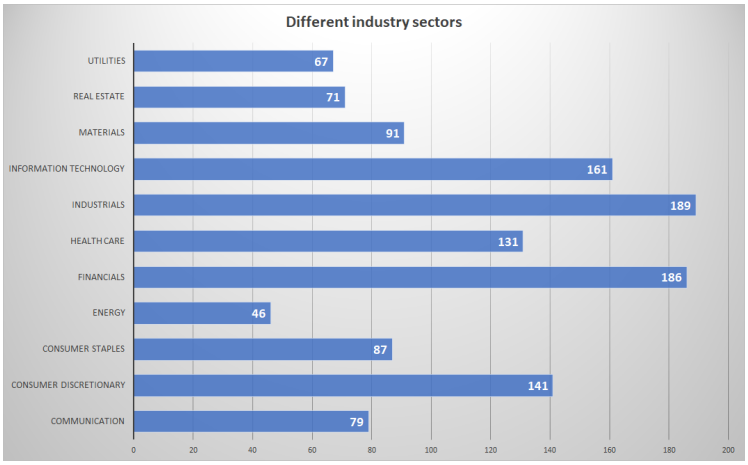

Figure 3.3: Amount of companies in different sectors

3.2 Framework and methodology

In order to detect the suitable replacing assets, we need to make some assumptions and restrictions: 1) the current portfolio is mean variance efficient according to Markowitz [2] (assumption), 2) stocks with a high sustainability rating will not be removed from the current portfolio (restriction), 3) we identify a user defined optimum of replacing assets instead of mathematical optimum (assumption). The first assumption states that our portfolio is mean variance efficient, where it is important to note that this optimal portfolio is created from a search space including the MSCI index and the complete set of sustainable assets. Hence, an improvement in return can only be achieved by having more exposure to risk. For a robust evaluation of the framework the second assumption is needed to eliminate replacements of assets by the same ones. In the third assumption we make sure to look for the optimum including the preferences of the investor, and not to look at the mathematical optimum. All in all, we aim for the newly created portfolio to be as close as possible to the tangency portfolio while accounting for these personal preferences.

Removing assets results in a sub-optimal solution, assuming a unique optimal allocation for a given level of risk. The set of removing assets RA is a subset of the current portfolio X, having a total removed weight, AW, which is the sum of the weights of the individual leaving assets. The number of assets removed, K, is the cardinality of RA. Therefore, the set of removing assets can be defined as:

$$RA = \{b_k\}_{k=0}^{K} \mid \forall b_k \in \mathbb{R}^m$$

$$AW = \sum_{k=0}^{K} w_K^b$$

After deleting subset RA from the portfolio, the remainder of the portfolio, Z is defined as $Z = X - RA$. The total removed weight needs to be distributed over the incoming assets. The total remaining weight of the portfolio equals Z: $1 - AW$.

In Figure 2.1 the optimal portfolio is the tangency portfolio. This portfolio is the best

trade-off between risk and return, taking the risk-free rate into account. Using this framework, new assets selected from the possibility set will be proposed to swap with the set of removing assets. This newly created portfolio will also result in a sub-optimal solution, as assets from the optimal allocation have been removed. To further minimize the loss of risk-return balance, the weights of the best assets proposed by the algorithm will be optimized. The best possible balance for the new portfolio represents a point in Figure 2.1 closest to the tangency portfolio. One of the most important assumptions is that the remainder of the portfolio must remain intact. Furthermore, we assume the asset manager does not want to re-run the overall mean-variance optimization of the portfolio to avoid trading costs. Therefore, minimizing the distance between the current portfolio (situated on the efficient frontier) and 'newly' created portfolio would be the best allocation we can theoretically find, taking the requirements and assumptions into account.

Issuer Ticker	Name	Weight (%)	Price	Market Value	Sector	ISIN	Exchange	Location	Market Currency
AAPL	APPLE INC	3.69	127.35	219,355,153.65	Information Technology	US0378331005	NASDAQ	United States	USD
MSFT	MICROSOFT CORP	3.19	257.89	189,681,189.68	Information Technology	US5949181045	NASDAQ	United States	USD
AMZN	AMAZON COM INC	2.47	3,346.83	146,802,004.29	Consumer Discretionary	US0231351067	NASDAQ	United States	USD
FB	FACEBOOK CLASS A INC	1.37	331.26	81,380,975.46	Communication	US30303M1027	NASDAQ	United States	USD
GOOG	ALPHABET INC CLASS C	1.28	2,513.93	75,978,506.39	Communication	US02079K1079	NASDAQ	United States	USD
GOOGL	ALPHABET INC CLASS A	1.26	2,430.20	74,935,217.00	Communication	US02079K3059	NASDAQ	United States	USD
JPM	JPMORGAN CHASE & CO	0.84	160.29	49,861,891.17	Financials	US46625H1005	New York Stock Exchange Inc.	United States	USD
TSLA	TESLA INC	0.81	609.89	48,270,963.83	Consumer Discretionary	US88160R1014	NASDAQ	United States	USD
NVDA	NVIDIA CORP	0.76	713.01	45,144,941.16	Information Technology	US67066G1040	NASDAQ	United States	USD
JNJ	JOHNSON & JOHNSON	0.75	164.96	44,705,479.68	Health Care	US4781601046	New York Stock Exchange Inc.	United States	USD
V	VISA INC CLASS A	0.69	234.96	40,832,523.60	Information Technology	US92826C8394	New York Stock Exchange Inc.	United States	USD
BRKB	BERKSHIRE HATHAWAY INC CLASS B	0.68	286.82	40,346,395.76	Financials	US0846707026	New York Stock Exchange Inc.	United States	USD
UNH	UNITEDHEALTH GROUP INC	0.65	397.89	38,738,172.51	Health Care	US91324P1021	New York Stock Exchange Inc.	United States	USD
NESN	NESTLE SA	0.63	126.64	37,474,039.81	Consumer Staples	CH0038863350	SIX Swiss Exchange	Switzerland	CHF
HD	HOME DEPOT INC	0.58	310.77	34,451,962.20	Consumer Discretionary	US4370761029	New York Stock Exchange Inc.	United States	USD
PG	PROCTER & GAMBLE	0.58	134.86	34,389,974.30	Consumer Staples	US7427181091	New York Stock Exchange Inc.	United States	USD
MA	MASTERCARD INC CLASS A	0.56	365.5	33,389,521.50	Information Technology	US57636Q1040	New York Stock Exchange Inc.	United States	USD
BAC	BANK OF AMERICA CORP	0.56	41.86	33,385,484.86	Financials	US0605051046	New York Stock Exchange Inc.	United States	USD
DIS	WALT DISNEY	0.55	177.38	32,835,343.94	Communication	US2546871060	New York Stock Exchange Inc.	United States	USD
PYPL	PAYPAL HOLDINGS INC	0.52	271.45	31,071,524.25	Information Technology	US70450Y1038	NASDAQ	United States	USD
ASML	ASML HOLDING NV	0.5	694.66	29,871,695.36	Information Technology	NL0010273215	Euronext Amsterdam	Netherlands	EUR
ROG	ROCHE HOLDING PAR AG	0.46	380.53	27,410,229.76	Health Care	CH0012032048	SIX Swiss Exchange	Switzerland	CHF
XOM	EXXON MOBIL CORP	0.46	62.17	27,040,903.67	Energy	US30231G1022	New York Stock Exchange Inc.	United States	USD
ADBE	ADOBE INC	0.45	541.26	26,848,661.04	Information Technology	US00724F1012	NASDAQ	United States	USD
CMCSA	COMCAST CORP CLASS A	0.44	56.88	26,321,561.28	Communication	US20030N1019	NASDAQ	United States	USD
VZ	VERIZON COMMUNICATIONS INC	0.41	57.33	24,338,247.57	Communication	US92343V1044	New York Stock Exchange Inc.	United States	USD
INTC	INTEL CORPORATION CORP	0.41	57.85	24,206,001.95	Information Technology	US4581401001	NASDAQ	United States	USD
CSCO	CISCO SYSTEMS INC	0.4	54.77	23,708,016.05	Information Technology	US17275R1023	NASDAQ	United States	USD
KO	COCA-COLA	0.4	56.16	23,603,711.04	Consumer Staples	US1912161007	New York Stock Exchange Inc.	United States	USD
MC	LVMH	0.39	809.1	22,986,602.17	Consumer Discretionary	FR0000121014	Nyse Euronext - Euronext Paris	France	EUR
PFE	PFIZER INC	0.38	40.15	22,875,542.80	Health Care	US7170811035	New York Stock Exchange Inc.	United States	USD
CRM	SALESFORCE.COM INC	0.38	240.31	22,629,992.70	Information Technology	US79466L3024	New York Stock Exchange Inc.	United States	USD
WMT	WALMART INC	0.38	140.75	22,461,588.75	Consumer Staples	US9311421039	New York Stock Exchange Inc.	United States	USD
NFLX	NETFLIX INC	0.37	488.77	22,071,875.66	Communication	US64110L1061	NASDAQ	United States	USD
T	AT&T INC	0.36	29.32	21,407,323.64	Communication	US00206R1023	New York Stock Exchange Inc.	United States	USD

Figure 3.4: Example of subset of removing assets

Figure 3.4 shows an example assets (RA) which could be removed from the portfolio. It is important to denote that several aspects are important when deleting a set of assets from the existing portfolio. For example, both the number of assets and the accumulated weight removed are of importance in this process. The number of assets removed impacts the diversification across the portfolio, such as diversification of countries, currencies and sectors. The total removed weight denotes the amount of impact of the changes applied to the portfolio. Removing, for example, a single asset with a large weight can have more impact on the overall risk, and therefore on tracking error, than removing multiple assets with a small accumulated weight.

For optimal replacement of the set of RA, the best candidate solution can be further improved by optimizing the weights of the incoming aggregated vector. This work proposes a framework, which aims to define a set of candidate solutions which provide similar risk levels,

with the help of multiple correlation. In general, the smaller the distances between the two return time-series, the higher the multiple correlation coefficient. This is the reason why we aim to maximize the correlation coefficient. Also, a smaller distance between the leaving and incoming assets implies a smaller deviation from the current characteristics of the portfolio, resulting in small tracking error. In the end, we know that the tangency portfolio is optimal and therefore aim to minimize the absolute distance from this tangency point. For finding the best replacements, a three step framework is created. These three main steps are illustrated in Figure 3.5.

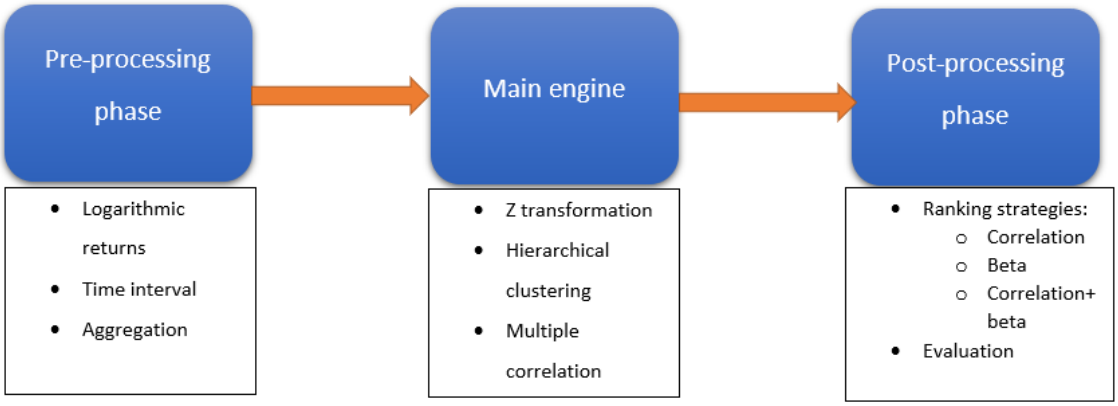

Figure 3.5: Thesis framework

In the pre-processing phase, time-series are tailored, for the main engine, with three different adaptations: Logarithmic returns, time interval and aggregation. For comparing assets in relative terms, the time-series are transformed according to log-transformations. Furthermore, all time-series must be aligned with each other before comparison can take place. On top of this, the time-series need to be aggregated to be able to execute the main engine. Therefore, an appropriate aggregation method is needed. All decision made and metrics chosen for the pre-processing phase are explained in more detail in Section 3.3.

Before executing the main engine, an appropriate similarity measure for applying multiple correlation has to be chosen. Pearson correlation is mostly used in the investment domain. With the help of Pearson correlation coefficient investors gain insight for diversification purposes. Investors will add assets with low or even negative correlation to an existing portfolio to reduce the exposure of the entire portfolio risk. We start with an optimal portfolio and

are not looking for further diversification benefits. Moreover, we aim to keep the portfolio's risk intact and hence we are not searching for low or negatively correlated assets.

In fact, when removing a single asset we would aim to find an asset with a strong similarity coefficient. Replacing an asset with a high correlation coefficient implies similar correlation dynamics in comparison to the leaving asset, which in turn is important for portfolio diversification. If nothing else but correlation is considered as similarity measurement, replacing a single asset with an asset having correlation coefficient of 1, would result in a 'perfect' replacement.

Therefore, if a set of leaving assets could be replaced by a set of incoming assets where the sets do have a correlation of 1, the replacement based on Pearson correlation would be perfect. The incoming set of assets would represent the similar diversification dynamics compared to the leaving set. The advantage of replacing a set of assets with an incoming set of assets in comparison to one-by-one replacement is that extreme events will be weighted according to the distribution and the aggregated pattern is easier to find in the solution set. The algorithm has the possibility to use complex combinations of the possibility set to match the outgoing assets. The algorithm provides solutions sets with different cardinalities. The core of this framework is described in detailed in the main engine section 3.4.

After the main engine, solutions will be ranked according to three ranking strategies: 1) multiple correlation coefficient, 2) beta coefficient and 3) a combination of multiple correlation and beta coefficients. Based on these ranking strategies the top ranked solutions will be selected, which should be closest to the removing set of assets. These top rated solutions will be post-optimized, according to optimal weight distribution. The top ranked solutions and the most optimal solutions will be evaluated according to tracking errors and beta and correlation coefficients. All of this will be described in detail in Section 3.5. But first all processes executed in the pre-processing is elaborated on.

3.3 Pre-processing

In the pre-processing phase the data will be prepared according the specified requirements for the multiple correlation algorithm. Three processes in the pre-processing phase are: Logarithmic returns, time interval and aggregation. All data needs to be transformed with a log transformation to review returns instead of prices, explained in Section 3.3.1. Furthermore, Pearson correlation requires equally length time-series which is elaborated on in Section 3.3.2. The multiple correlation coefficient will be determined based on aggregated vectors, explained in Section 3.3.3.

3.3.1 Log-returns

The first step is to determine continuously compounded returns. In this thesis, only assets with a market capitalisation bigger than 100,000,000 euros are considered. This reduces the exposure to liquidity risk.

The time-additivity of logarithmic returns is a powerful feature in portfolio management. This allows for easy comparison between different trades and adding and subtracting results from a sequences of different trades. If multiple changes have been made in different time-periods it allows for easy evaluation of the cumulative returns.

3.3.2 Time interval

For our applications, approximately 5 years of data based on daily granularity is considered to be an adequate length of the time-series. The interval defined is 03-01-2017 up until 17-08-2020. Quantitative analysis, such as volatility measures and correlation coefficients, are impacted by the length of the time-series. Therefore, the decision on the time interval is of importance. Overall, there is no consensus about the optimal length a time-series should have to be best evaluated and therefore it depends on human judgment. The length of a time-series has impact on correlation and volatility measures due to different reasons. Including more raw data points has direct impact on the calculation of the measures [4].

Analyzing a time-series which is too short might lead to an inaccurate analysis. Characteristics for long-term investment strategies, such as seasonality and trends, might not be captured accurately. Also extreme events, like COVID-19, have in shorter time-series more impact on e.g. volatility.

On the contrary, if the interval of a time-series is too large the impact of rare events might be too small. Too long time-series will adapt less effectively to changing situations over time. Additionally, optimising parameters for a long time-series becomes computational more complex without adding significant impact on the accuracy of the results.

The model is robust and reliable when merely reviewing common trading days and align the time-series accordingly. Therefore we decided to not apply interpolation to align missing values according to national holidays. This would only add additional noise to our data without providing value. Thus, assets listed on different exchanges over the world have dif-

ferent trading calendars. We determined different total common trading days per year. For 2017, we have 237 common trading days, 2018 there are 251 common trading days, 2019 respectively 252 common trading days. In 2020 we only considered 180 common trading days for an additional reasons. In 2020, we had more common trading days, however we allow the algorithm and the post-processing phase only to use data up until a certain point in 2020. The rest of the data is considered to be training data. We could use the remainder of 2020 and the beginning of 2021 to evaluate the chosen solutions not only based on 'historical' data but also on training data, also called out of sample data.

3.3.3 Aggregation

The time-series of the set of removed assets RA and all possible combinations of the time-series from the possibility set are aggregated by a weighted averaging method. This creates two single time-series capturing the returns of the associated sets. Every particular aggregation method has its advantages and disadvantages. The min and max aggregation methods capture outliers nicely, but are less capable of capturing the overall trends. We assume that investment strategy for the current portfolio is based on a long time horizon, which makes weighted aggregation most suitable. This allows to correct for the weighted distribution of these assets in the current portfolio.

More formally, given the time-series of the set of K leaving assets RA, with each assets a_k defined as $a_k \in \mathbb{R}^m$. Where we define w_K as the associated weight of the asset in the portfolio. The weight of each portfolio over the total removed weight determines the importance of each asset in the aggregation vector. Then the aggregated leaving vector, $R_l \in \mathbb{R}^m$, can be defined as follows:

$$AW = \sum_{k=0}^{K} w_k$$

$$R_l = \sum_{k=0}^{K} \frac{w_k}{w_l} \times a_k$$

The aggregated returns vector R_l is defined over the interval January 2017 up until August 2020. This R_l is used by the multiple correlation algorithm to quantify linear dependency between sets of assets. This due to the fact that this aggregation method is also being done for each single possible solution in the solution set. This results in having two single aggregated return vectors, which can be evaluated with the help of the main engine.

3.4 Main engine

After the pre-processing phase, the data is transformed such that it is possible to apply the multiple correlation algorithm properly. The multiple correlation algorithm is based on the work of Minartz and Papapetrou [3]. Their correlation detective algorithm has been implemented in this thesis. Given a set of time-series X and two subsets $z_1 \subset x$ & $z_2 \subset X$, the multiple correlation formula in the algorithm of [3] is defined as:

$$Pearson(Average(\{z_1\}), Average(\{z_1\})) \tag{3.2}$$

Where the average is calculated with the weighted aggregation method. The input vectors are transformed according to a z transformation, which is not a specific implementation for this thesis but a pre-requisite for the use of this algorithm [3].

In this Pearson correlation, it is convenient to denote a left hand side (LHS) and right hand side (RHS). Unlike the implementation of [3] where both the LHS and RHS are dynamic, the RHS is fixed for the purpose of this work. It is important to keep in mind that the application is in the financial domain and more specifically, financial assets and their associated time series. Therefore, the RHS in our application denotes the aggregated time-series of removing assets RA. The set of removing assets is not supposed to be dynamic, as replacing assets are being identified, and thats why the RHS should be fixed.

The LHS in the multiple correlation coefficient denotes a single possible solution. The algorithm is trying to find solutions which satisfy user-defined requirements. Each solution consisting of a subset of the possibility set Y, which is dynamic in cardinality.

This multiple correlation coefficient provides us with a single indicator for the linear dependency between the two sets of assets. With the help of aggregation and this multiple correlation metric, sets of time-series can be compressed to a single time-series and compared with another set of time-series. These single time-series capture the characteristics of all individual time-series of returns in the sets weighted accordingly. This allows us for comparisons between sets of different cardinalities.

3.4.1 Algorithmic design

The algorithm uses the input of the RA and the possibility set together with several user-defined parameters to decrease the search space for suitable solutions. Together with multi-threaded implementation, load-balancing optimizations and hierarchical clustering these user-defined parameters contribute to the efficient working of this algorithm. The user-defined parameters and input data sets can be defined as:

- Replacing assets, $RA \subset X = \{a_k\}_{k=1}^{K} \mid \forall a_k \in \mathbb{R}^m$.
- Possibility set, $Y = \{y_j\}_{j=1}^{T} \mid y_j \in \mathbb{R}^m$.
- Threshold, τ.

- Degrees of freedom for the RHS, ω.
- Minimum jump, δ.

A minimum threshold, τ, is imposed on the multiple correlation coefficient. This is the minimum coefficient for a single solution which needs to be satisfied before it can be included in the solution set. For the purpose of this research only high multiple correlations coefficients are interesting.

The degree of freedom ω is defined which denotes the maximum cardinality a single solutions can have. Meaning the maximum number of assets included in a single solution. The LHS is a fixed vector which will be aggregated only once. Therefore, in theory the LHS can have any cardinality. All possible solutions can be defined by the number of elements in the power set of a set. The power set of set Y, $\mathcal{P}(\mathcal{Y})$ is the set containing all subsets of Y including the empty set. Obviously, the empty set is not suitable as a possible solution. Since we have 1,414 stocks, the number of possible solutions is therefore:

$$\mid \mathcal{P}(\mathcal{Y}) \mid \approx 2^T = 2^{1,414} \approx \inf \tag{3.3}$$

Evaluating each single solution is therefore not possible. Hence, we defined δ which is the minimum jump to limit the number of solutions we have to evaluate. The minimum jump, δ, is the minimum increase in multiple correlation coefficient needed to justify additional complexity. The complexity in this context is defined as the amount of assets included in a single solution. Adding an extra assets to a single solution set increases the complexity. This involves more transaction and monitoring costs. Each solution is therefore bounded by a maximum cardinality, which we refer to as degree of freedom, ω.

With the help of these user-defined parameters and the provided data sets of the removing assets RA and the possibility set Y the algorithm can be executed. The algorithm tries to discover possible solutions with the help of these parameters. The efficient implementation is of importance for reviewing a large search space.

3.4.2 Efficient implementation

The efficient working of the algorithm is based on the following reasons:

1. The value of the minimum correlation threshold τ
2. The choice for the degrees of freedom ω
3. Hierarchical clustering
4. Multi-threading

First, τ creates a lower-bound for the multiple correlation coefficient. In the work of [22] the lower-bound for the time-series of all assets involved in the multiple correlation, is defined as a function of pair-wise correlations. The lower-bound in combination with τ eliminates the possible search space heavily and contributes to the ability of researching Big Data sets.

Second, ω limits the cardinality of each possibility set. Hence, equation 3.3 does not hold, for this implementation. The binomial coefficient formula defines the number of possible subsets of elements with a user-defined ω is limited to:

$$\#Subsets = \sum_{i=2}^{\omega} \frac{i!}{i!(N! - i!)} \tag{3.4}$$

Where the cardinality of a subset is minimal two.

Fourth, hierarchical clustering is implemented for efficient processing of the algorithm. This hierarchical clustering is build with cluster combinations, where the top represents a cluster containing all assets. The algorithm will prune through the hierarchy according the user-defined threshold and minimum jump. The algorithm defines and calculates upper and lower bounds on the multiple correlation between all different clusters, which it uses to decide whether to consider or not to consider the possibility cluster combinations.

Last, many calculations in this algorithm are processed in parallel. For example, the pair-wise correlations used for defining the lower-bound can be calculated in parallel. This speeds up the algorithm significantly.

3.5 Post processing

The algorithm returns all solutions found according to the user-specified requirements: 1) the minimum coefficient threshold, 2) the minimum jump and 3) the possibility set. This output will be post-processed in an additional module of the framework. Each solution also provides us with information of the maximum subset correlation. For each subset, included in a single solution set, the multiple correlation coefficient is calculated between the subset and the RA. The maximum of these subset coefficients is displayed in the maximum subset correlation. Displaying this provides the user the information on how much the minimum jump is exceeded to find this solution. With this information the cost of adding additional complexity is quantified.

The post processing method includes two main steps: ranking based on betas and an evaluation based on tracking errors.

3.5.1 Ranking based on correlations and betas

The solutions are ranked according to their multiple correlation coefficients, from high to low. By construction we know that all outputted solutions have a coefficient larger than the threshold τ. The number of found solutions depends therefore on τ. In general, the lower the threshold the higher the number of possible solutions. The hundred solutions with the highest correlation coefficient are selected, as candidate solutions for further analysis. First, for each solution included in the candidate solutions set, the beta is calculated. Then these top 20 candidate solutions are ranked, from high to low, according to the calculated betas.

To use this beta coefficient for scoring the solutions, we have calculated the beta between the set of incoming assets and the set of leaving assets RA. With the help of this additional ranking strategy, the candidate solution set proposed by this framework could be improved. We are not calculating the beta between the set of incoming assets and the complete portfolio on purpose. This can be seen as the benchmark in normal applications. Due to the fact that the leaving assets are considered optimal in the current portfolio, calculating the benchmark between the set of incoming assets and the portfolio would provide us different insights. For example, whether the set of incoming assets are in line with the characteristics of the complete portfolio. However, we want them to be in line with the characteristics of the RA.

3.5.2 Evaluation metric: tracking error

To evaluate our outputted results we need to compare the newly created portfolio with the portfolio before it changed. We conduct this analysis by using the tracking error. This tracking error defines the mathematical distance between two time series of returns. To review a candidate solution we need both the old portfolio and we need to create a new portfolio. The 'newly' created portfolio is designed by deleting all assets included in the RA and inserting all assets included in the candidate solution. Next, both portfolio will be represented as an aggregated vector of returns. Based on these two aggregated vector of returns we can calculate the tracking errors. The comparisons of these two portfolio is done by calculating both the first moment and the second moment tracking error, TE_1 and TE_2:

$$TE_1 = \frac{\sum_{i=0}^{m}(R_{ni} - R_{oi})}{m-1} \tag{3.5}$$

$$TE_2 = \sqrt{\frac{\sum_{i=0}^{m}(R_{ni} - R_{oi})^2}{m-1}} \tag{3.6}$$

R_n and R_o represent the aggregated historical return vectors for both the new and old portfolio. To correct for a bias in the calculations, a division by the length of time-series is applied minus 1, $N-1$. Using both moments of tracking error provide us different insights. The first moment tracking error shows us in which direction we have the deviation. The new portfolio has a positive or negative deviation on the overall returns. By definition the old portfolio was optimal, and hence we conclude with a positive deviation the exposure to risk is increased.

The second moment defines the squared deviation between the return vectors of the portfolios. This second moment tracking error measures the fluctuation of the newly created portfolio compared to the old portfolio, which is our benchmark. This gives us insights, in how good the new portfolio mimics the old one.

In the ideal situation, the tracking error would be 0. With this nothing has changed in the new portfolio compared to the old portfolio from a risk perspective.

Furthermore, the tracking errors can be measured in two different ways, with historical data and future 'unseen' data. The tracking error based on historical data directly determines how far the each 'newly' created portfolio deviates from the old portfolio. Every data point has already be seen by the algorithm and if the algorithm has good results the TE is by construction small. However, all the results are calculated based training data ranging

to September 2020. Therefore, it is possible to calculate the tracking error based on future 'unseen' data, the test data. The test data ranges from 17-09-2020 up until 31-03-2021. With this analysis method we can evaluate what would happen if we would have changed the leaving assets in September 2020 and replace them by the solution provided by this thesis framework.

Absolute tracking error vs relative tracking error

All tracking errors discussed above are in absolute terms. There is no correction for the impact the changes have on the new portfolio based on the accumulated removed weights. To compare tracking errors, they should be converted to a relative tracking error. Comparing absolute tracking errors is useful in terms of economic evaluation. However, to compare absolute tracking errors of different outputted experiments, we need to correct for the accumulated deleted weights, resulting in a relative tracking error.

For example, let the user define a RA with an accumulated weight of 1 percent, which results in a absolute tracking error of 10. Than, when another user defines a different RA with an accumulated weight of 10 percent, having a similar tracking error of 10, these solutions can not be compared directly. In economic terms, the pay-offs from replacing assets are equally good. However, to compare the replacements in mathematical terms, we need to correct for the accumulated weight by dividing the absolute tracking error over the total removed weight,

$$\frac{TE_{absolute}}{AW}$$

In this chapter we have discussed all three phases of this thesis framework. This allows us to conduct experiments and review them properly.

Chapter 4

Experiments & Results

To evaluate the quality of the solutions that are proposed by our algorithm, we will conduct three different experiments: 1) we randomly select assets which will be replaced, 2) we replace seven CCC rated assets to improve the overall sustainability rating of the portfolio, 3) we remove all the stocks from the energy sector and try to replace them to see the impact on the current portfolio. Furthermore, for each experiment we address expectations of the results before running the experiment. We aim to find results which are in line with our expectations. Otherwise, we elaborate on the differences between the results and the expectations.

Figure 4.1 presents the evolution of the benchmark over time. The benchmark is an ETF on the MSCI World Index. This graph can be intuitively seen as an aggregated weighted representation of all assets included in the MSCI World Index.

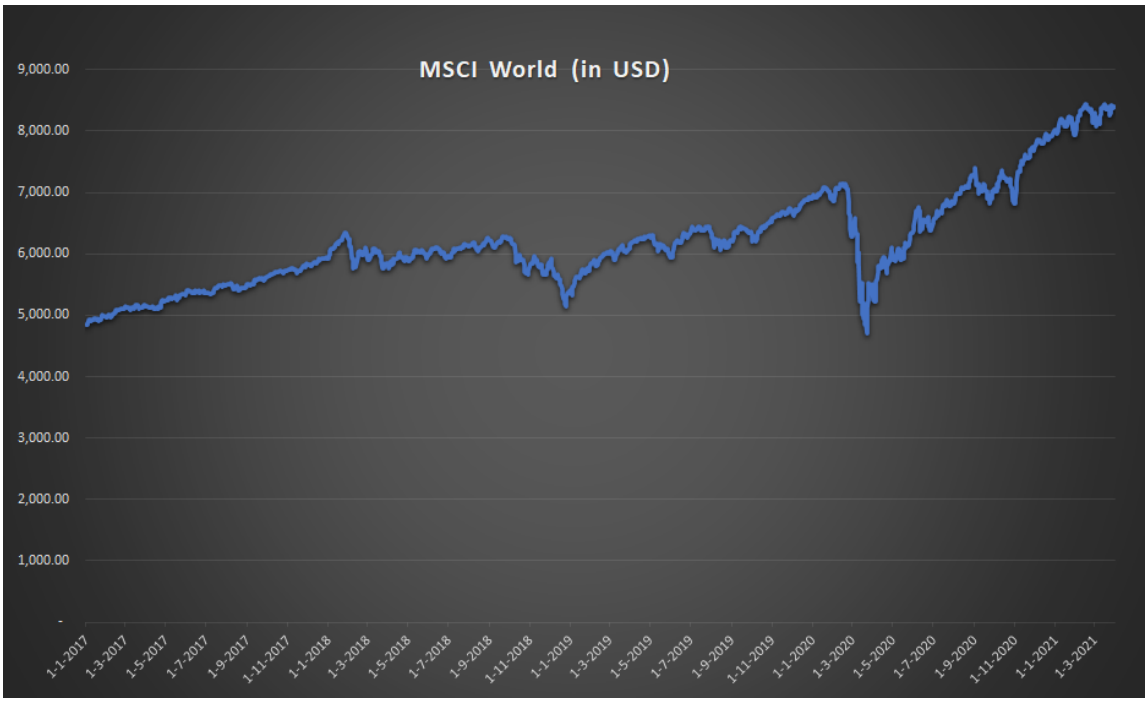

Figure 4.1: MSCI World ETF

4.1 Experimental setup

In each experiment we conduct three different methods to find the desired candidate solution set: 1) random selection, 2) algorithmic output and 3) algorithmic output and post-processing phase. Every method is evaluated, from which we can conclude the best performing method in terms of the lowest tracking errors.

The best performing method will define our solution candidate set S which needs to incorporate all desired user-defined requirements. The user-defined requirements are the set parameters for the algorithm: 1) minimum correlation threshold τ, 2) the minimum jump δ and 3) the maximum degree of freedom ω. Also included in the user-defined requirements is 4) the number of solutions included in the solution set S. The number of solutions included determines the amount of possible solutions the user has to evaluate according to all risk measures discussed in 2 and other econometric measures. For the purpose of this research we have set the amount of solutions included in the candidate solution set equal to 20. All of these four user-defined requirements do determine the complete candidate solution set S.

The first method, the random selection, is always evaluated at the beginning of each experiment. This method randomly selects a set of incoming assets to replace the set of leaving assets RA. The random selection method is considered to be the benchmark for each of the experiments proposed in this research. The solutions from the algorithm should at least be better then this random selected method. We conduct the random selection a 1,000 times. We take the mean of the tracking errors out of these 1,000 runs. This leaves us with a robust benchmark to compare our solutions with.

Second, the main engine is executed and the solutions provided by the algorithm are compared to the benchmark. In this method we want to evaluate whether all solutions provided by the algorithm are suitable for defining our candidate solution set S. If all solutions provided by the algorithm are suitable for defining this candidate solutions set S, this would mean that all solutions are suitable solutions and we could simply pick 20 assets from the solution set provided by the algorithm without using additional metrics.

In the third method, the provided solutions from the algorithm will be ranked according to three ranking strategies: 1) correlation coefficient, 2) beta coefficient and 3) combination of correlation and beta coefficient. Such additional metrics allows us to distinguish different solutions which are extremely close to each other. Ultimately, with each of these ranking strategies we try to define a suitable candidate solutions set S. We evaluate each of the ranking strategies in comparison to the benchmark. Moreover, we also want to decide which of the three ranking strategies provides us with the best candidate solution set. For comparing the ranking strategies the tracking errors need to be compared amongst each other.

Furthermore, all experiments have been executed on a local computer with the following specifications:

- CPU Ryzen 5 3600 3,6GHz (4,2GHz overclock), 8 cores and 16 threads

With the local runs we are able to evaluate whether this framework is able to provide solutions which can replace leaving assets without re-running the efficient frontier optimiz-

ation. Running on a computing cluster is only required when exceeding $\omega > 5$. Moreover, the computational complexity increases so substantially, resulting in an extreme increase in running time. This does not fit the purpose of this framework. Adding additional complexity is not desired by the user after all and this is the reason why a minimum jump is added as an additional learning parameter.

4.1.1 Overview of different experiments

In total we conduct three different experiments, where the experimental setup from above will apply to. We start by evaluating the random asset experiment, experiment 1 (see section 4.2.2). In this experiment, we randomly select assets to be removed from the portfolio and replace them with the help of the three methods explained above (only non-sustainable assets can be selected). This experiment is purely to evaluate the overall performance of the algorithm. In the second experiment, experiment 2 (see section 4.3.4), we replace 7 very unsustainable assets, all rated with ESG rating CCC. We assume here that the portfolio manager tries to improve sustainability in the current portfolio, so removing the worst assets and replacing them by better ESG rated assets would be viable to investigate. In the third experiment, the complete energy sector is deleted from the portfolio, (see section 4.5). This is a useful experiment to evaluate for portfolio managers, as it makes them more flexible in their strategies. Moreover, it is good to evaluate whether it is possible to remove a set of approximately 30 assets.

4.2 Experiment 1: Random asset replacement

The random asset replacement investigates the overall working of the framework for replacing assets in portfolios. This is done by evaluating each method proposed in the experiment setup 4.1. We aim to find significant differences between the tracking errors of the benchmark and the algorithmic solutions. With this we can argue that this framework is a tool which can be used for replacing assets in an existing portfolio. Similar reasoning holds for the post-optimization results.

We randomly select a set of 10 assets to be replaced and try to replace it with a set of 4 assets, hence $\omega = 4$. ω is similar to the one used in the other experiments. The RA [1] has a total weight of 1.1.

This experiment is solely to investigate the overall working of replacing assets with the help of this framework. Given that each solution included in the post-processing phase is included in the solutions provided by the algorithm, we only compare the random replacements and the solution provided by the algorithm.

4.2.1 Benchmark: random asset selection

We have repeated the random asset selections with $\omega = 4$ 1,000 times.

[1] RA = Standard Chartered PLC, Telefonica SA, Coca-Cola HBC AG-DI, Berkshire Hath-B, CAE Inc, Globe Life Inc, Hasbro Inc, ADV Micro Device, Banco Bilbao Vizcaya Argenta, Basf SE.

We have defined the mean of the tracking errors which is set as the benchmark. The overall mean of the first moment relative tracking error is equal to: $-5.02 * 10^{-04}$ and the overall mean of the second moment relative tracking error: $2.96 * 10^{-13}$. From this we conclude that overall the replacements results in a slightly negative deviation from the current portfolio. Therefore, a random replacement is decreasing the returns slightly.

4.2.2 Algorithmic solution

For running the main engine of the algorithm the following parameters have been set [2]:

- Threshold $\tau = 0.85$
- Maximum degree of freedom $\omega = 4$
- Minimum jump $\delta = 0.045$
- Total number of solutions sv in $S = 20$

Expectations: We expect to outperform the randomly selected assets (baseline solution) which has an relative tracking error of approximately $2.96 * 10^{-13}$. We expect that both our candidate solution sets created with the help of solely the algorithm and the algorithmic solutions plus the post-processing phase, do outperform this.

Results of main engine

For the replacement of this RA, in total 1,853 unique solutions, sv, have been found by the algorithm. We have selected candidate 20 solutions out of all possible solutions CSV. These 20 solutions together form the candidate solution set S. The mean of the first moment relative tracking error of S is equal to: $-4.14 * 10^{-04}$ and the mean of the second moment relative tracking error of S is equal to: $1.68 * 10^{-13}$. The relative tracking errors have been defined for comparison further in this chapter. The differences between the tracking errors of the randomized selection and the algorithmic selection quantifies the performance of the algorithm.

First, we observe, similar to the random replacement, a minor negative deviation for the first moment tracking error, which indicates a small loss in total returns based on the mean of S. However, the deviation is smaller than the deviation of the randomly selected assets. We observe approximately 18 percent improvement in first moment deviation.

Second, we observe a mean of the second moment relative tracking error of S which is smaller than the one observed by randomly selecting replacing assets. We find a 43 percent lower tracking error with the use of our algorithm. Consequently, the solutions according to all the phases of the framework will have at least an improvement of 43 percent in comparison to the benchmark.

[2] see list of symbols

The results outputted by the algorithm and ranked according to their multiple correlation coefficient provides us with a lower bound. Adding other ranking methods and post-optimization methods is only improving the proposed results. As we observe that the lower-bound of the solution set is outperforming the random asset allocation, it is not needed to conduct all metrics for this particular experiment. In the next experiments these additional ranking metrics are used and the post-optimization procedure is also applied.

4.3 Experiment 2: Replacement of unsustainable assets

This second experiment is conducted to improve the total sustainability rating of the portfolio. The leaving assets will be selected based on the worst ESG ratings currently available in the portfolio (MSCI index). The current portfolio has seven assets rated with the lowest possible sustainability rating (CCC). This experiment attempts to replace them with the help of this framework. The possibility set contains 1,414 sustainable assets with a minimum rating of A.

The *RA* includes seven CCC rated assets from MSCI index [3] with a total weight of 0.2 percent. Figure 4.2 displays all relevant information of the assets leaving the portfolio (MSCI index), such as weight distribution, sector distribution and geographical locations. These diversification factors are subject to the goal of the user when selecting the optimal solution out of our proposed candidate solution set S.

For visual interpretation, the z-transformed return aggregated vector of the set of leaving assets is displayed in Figure 4.3. This basically provides us visual insights in the behaviours of the leaving CCC assets over time. The goal of the replacing assets is to match these behaviours as good as possible. We try to do this by mimicking the monthly aggregated log returns illustrated in Figure 4.4. The relation between the z-transformed pattern and the monthly aggregated vector is clearly visible, for example the COVID drop in 2020.

Issuer Ticker	Name	Asset Class	Weight (%)	Price	Nominal	Market Value	Notional Value	Sector	Exchange	Location	Market Curr
MNST	MONSTER BEVERAGE CORP	Equity	0.06	93.5	40,112.00	3,750,472.00	3,750,472.00	Consumer Staples	NASDAQ	United States	USD
CCL	CARNIVAL CORP	Equity	0.04	29.93	79,751.00	2,386,947.43	2,386,947.43	Consumer Discretionary	New York Stock Exchange Inc.	United States	USD
FFH	FAIRFAX FINANCIAL HOLDINGS SUB VOT	Equity	0.02	460.33	2,819.00	1,297,661.45	1,297,661.45	Financials	Toronto Stock Exchange	Canada	CAD
UHS	UNIVERSAL HEALTH SERVICES INC CLAS	Equity	0.02	160.38	7,756.00	1,243,907.28	1,243,907.28	Health Care	New York Stock Exchange Inc.	United States	USD
DISH	DISH NETWORK CORP CLASS A	Equity	0.02	40.07	28,905.00	1,158,223.35	1,158,223.35	Communication	NASDAQ	United States	USD
NWSA	NEWS CORP CLASS A	Equity	0.02	26.84	38,707.00	1,038,895.88	1,038,895.88	Communication	NASDAQ	United States	USD
VER	VEREIT INC	Equity	0.01	48.85	17,212.00	840,806.20	840,806.20	Real Estate	New York Stock Exchange Inc.	United States	USD

Figure 4.2: CCC assets

4.3.1 Benchmark: random asset selection

We start by conducting the random selection method to define the set of incoming assets. We randomly select assets from the possibility set with $\omega = 4$. Hence, each solution sv, the set of 'incoming' assets, can take up to a maximum of 4 different assets. Furthermore, for the

[3] 1) Carnival Corporation, 2) Monster Beverage, 3) Universal HLTH-B, 4) Fairfax FINL Holding, 5) Vereit inc, 6) Dish network-A and 7) News corp-cl A.

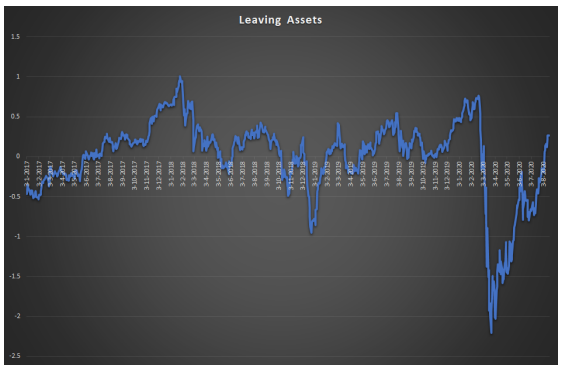

Figure 4.3: Aggregated set of leaving assets (z-score)

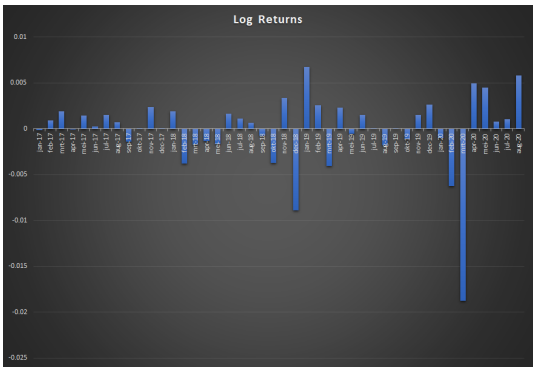

Figure 4.4: Aggregated set of leaving assets (monthly log returns)

random allocation, no threshold and minimum jump is defined as the main algorithm is not being executed.

In total, we collected 1,000 random sets of incoming assets. In Figure 4.5 we see a sample of the solutions of degree 3. This to illustrate that with a bigger ω the algorithm can still detect solutions, sv, with a lower cardinality satisfying all user-defined requirements. Similar to other cases, the first and second moment tracking error have been calculated between the old portfolio and the newly created portfolio, including the randomly selected assets.

Random Asset 1	Random Asset 2	Random Asset 3	Tracking error Random degree2
LLOYDS BANKING G	CHEMED CORP	GENERAL MOTORS C	2.5638091709642833e-17,
QUANTA SERVICES	CSL LTD-SPON ADR	DARDEN RESTAURAN	2.5681149382917614e-17,
NUANCE COMMUNICA	THOR INDUSTRIES	TWILIO INC - A	2.5699221136590126e-17,
CAN NATL RAILWAY	GRUPO AVAL ACCIO	JEFFERIES FINANC	2.5805002424748297e-17,
TRINET GROUP INC	NICE LTD -SP ADR	TIMKEN CO	2.6022534165644366e-17,
POST HOLDINGS IN	ALLEGHANY CORP	CATERPILLAR INC	2.6026267428414248e-17,
CYRUSONE INC	RELX PLC - ADR	PENN NATL GAMING	2.6233029194015784e-17,
ALASKA AIR GROUP	AUTOHOME INC-ADR	BRASKEM SA-ADR	2.6281277204694897e-17,
DISCOVERY INC -	OVINTIV INC	VIACOMCBS INC-B	8.941454210155835e-16,
REPLIGEN CORP	VIACOMCBS INC-B	SASOL LTD-SP ADR	9.34231206675154e-16,
TAKEDA PHARM-ADR	BANK MANDIRI-ADR	VIACOMCBS INC-B	9.708674329750897e-16,

Figure 4.5: Example of random solutions

For each random solution, the set of assets have been added to the current portfolio, which defines the 'newly' created portfolio. To determine the weights of each assets, the total removed weight AW has been uniformly distributed accordingly. Then, the second moment tracking error between the old portfolio and the new portfolio is calculated. This is used as baseline indication to compare with the candidate solutions from the algorithm and the solutions provided by the post-process phase.

For this particular RA, we found a mean of the second moment tracking error of $2.97 *$

10^{-17}, resulting in a mean of the relative second moment tracking error of: $1.49 * 10^{-13}$. After setting such benchmark tracking errors, we continue with the execution of the main engine. In this main engine we evaluate a candidate solution set of the results of the algorithm. After the main engine, we continue using the post-processing phase to rank the results according to the three ranking strategies: 1) the correlation coefficient, 2) the beta coefficient and 3) a combination of the correlation and beta coefficient, as described in 4.1. This is done to help defining the a better candidate solution set S in the post-processing phase of this framework.

4.3.2 Algorithmic solution

For this second experiment the following parameters have been set:

- Threshold $\tau = 0.87$
- Degree of freedom $\omega = 4$
- Minimum jump $\delta = 0.045$
- Total number of solutions sv in $S = 20$

These parameters limit the amount of results and should therefore be set correctly. The evaluation will be more precise and the interpretation in comparison to bigger CSV will not be different. Other aspects of this procedure are similar to the algorithmic solutions procedure in the previous experiment.

Expectations: we expect to outperform the randomly selected assets (baseline solution) of approximately $3.0 * 10^{-17}$ with the candidate solution set provided by the algorithm and therefore also with the addition of the post-processing phase, where we apply three ranking methods. We expect ranking strategy based on betas and correlation together to provide us with the best candidate solution set, in terms of second moment tracking errors. The combination of these metrics does provide us additional insights over the volatility and the directions of the movements.

Results of main engine

The output of the algorithm provides us with a solution set CSV, containing single solutions, sv, with set of 'incoming' assets. We observe solutions sv in the solution set CSV having different cardinalities, respectively three and four. Furthermore, with the current parameter settings, the algorithm has found 4,943 unique solutions sv, which denotes the total size of CSV. Every result provided as output by the algorithm do exceed the threshold τ of 0.87.

We have plotted the solutions in figures 4.6 and 4.7. In these figures the x-axis represents the multiple correlation coefficient and the y-axis the beta for each solution. These plots are identical except for the additional point plotted with an x and y coordinate of 1. This point is representing the optimal replacement according to the reviewed metrics in this thesis. It is important to note that the perfect solution is represented at the location where the multiple correlation coefficient equals one and the beta equals one. This point can be seen as a replacement of assets with themselves.

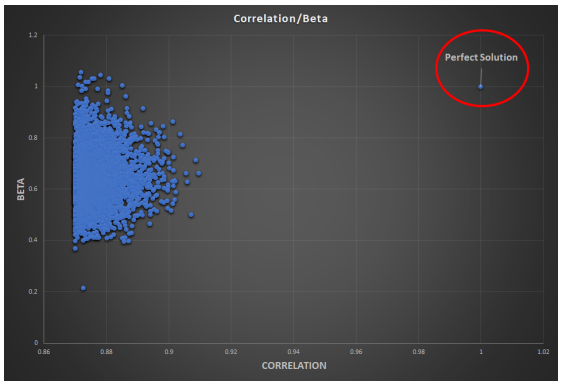

Figure 4.6: Ranked solutions degree 3 and 4 correlation vs Beta. (optimal solution included)

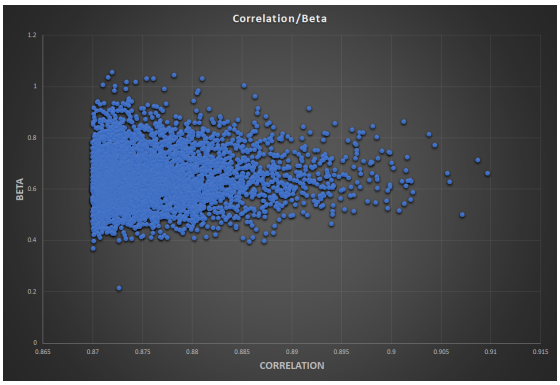

Figure 4.7: Ranked solutions degree 3 and 4 correlation vs Beta. (not scaled to optimal solution)

Furthermore, it should be noted that due to the set threshold, no solutions with correlations coefficients lower than 0.87 are considered. A beta coefficient cannot be accurately interpreted without information regarding the correlation coefficient or the scaling of the variances. Due to the relations defined in the equations in 2.1, where we see that beta is a function of the scaled variances and the correlation.

We have presented the candidate solution set CSV for this experiment in the appendix A. Overall, we have observed that most of the solutions provided by the algorithm do outperform the baseline indication. However, it is interesting to note that not all solutions outperform the benchmark, this means that randomly selecting the assets from the solutions does not always lead to better replacements. If we would do this it is very important to observe that the S would end up with assets included which do satisfy the user-defined requirements but do have worse second moment tracking errors than the baseline. Some of these 'bad' solutions are visible in figure 4.8.

RHS1	RHS2	RHS3	RHS4	Correlation	betas	variances	beta_correlation	tracking_error_degree1_portfolio	tracking_error_degree2_portfolio
PPL CORP	AUST & NZ BK-ADR	HILL-ROM HOLDING	VIACOMCBS INC-B	0.8980723	0.662569691	1.75E-05	1.560641991	5.95E-07	9.09E-16
MOTOROLA SOLUTIO	AUST & NZ BK-ADR	HILL-ROM HOLDING	VIACOMCBS INC-B	0.905637993	0.662007952	1.87E-05	1.567645945	-1.88E-07	9.16E-16
AUST & NZ BK-ADR	WM MORRISON-ADR	HILL-ROM HOLDING	VIACOMCBS INC-B	0.895731554	0.6436084	2.06E-05	1.539339954	-1.12E-07	9.47E-16
MOTOROLA SOLUTIO	HILL-ROM HOLDING	SCIENCE APPLICAT	VIACOMCBS INC-B	0.89534763	0.51984664	2.18E-05	1.41519427	1.28E-06	9.52E-16
PEOPLE'S UNITED	VIACOMCBS INC-B	AUST & NZ BK-ADR	HILL-ROM HOLDING	0.894893427	0.563465871	2.58E-05	1.458359298	-8.45E-09	9.73E-16
MOTOROLA SOLUTIO	HILL-ROM HOLDING	UNIVERSAL DISPLA	VIACOMCBS INC-B	0.898444304	0.546893947	2.36E-05	1.445338251	9.54E-07	1.01E-15
AUST & NZ BK-ADR	HILL-ROM HOLDING	UNIVERSAL DISPLA	VIACOMCBS INC-B	0.900802534	0.515780603	2.72E-05	1.416583137	6.32E-07	1.02E-15

Figure 4.8: Bad replacements

The information provided in figure 4.8 shows us that the betas of these incoming sets of assets are relatively low. This directly leads to a low combination of the correlation coefficient and beta. Due to the fact that we have information regarding the correlation coefficient we can reason why these solutions are poor. As these solutions all have a correlation coefficient higher than the set threshold τ, we can conclude that all of these proposed incoming sets of assets have much higher volatility, which could clarify the worse tracking error.

This immediately identifies a limitation of solely using provided solutions of the algorithm, without analysis based on other metrics like beta. Hence, reviewing solely based on the mul-

tiple correlation coefficient is not accurate, as there is a small chance of having a proposed solution sv included in S having worse tracking error than the random selection method.

4.3.3 Algorithmic solution and post-processing phase

The complete thesis framework also consist of a post-processing phase, where additional metrics are used to evaluate results provided by the algorithm. The algorithm here functions as a pre-selection engine, providing us with possible solutions while reviewing this large search space.

Betas are therefore used to provide additional insights in the candidate results. We limit this research to those two factors. To help identifying the relation between the ranking metrics and the proposed solutions, the solution sets have been ranked according to three different ranking strategies: 1) the correlation coefficient, 2) the beta coefficient and 3) by a combination of the correlation and beta coefficient. In this combination we gave equal importance to the beta and correlation coefficient, meaning that we can simply add them.

In the figures below the solutions provided by the algorithm have been ranked according to the three different ranking strategies. Selecting the top 20 solution here provides us with candidate solution set S. The top 20 ranked correlation solutions are illustrated in figure 4.9. In figure 4.10 the top 20 solutions ranked according to the beta coefficient are shown. Last, in figure 4.11, the combination of both coefficients have been used to rank them.

RHS1	RHS2	RHS3	RHS4	Correlation	betas	variances	te_1_assets	te_2_assets	beta_correlation	tracking_error_degree1_portfolio	tracking_error_degree2_portfolio
AUST & NZ BK-ADR	CAMPBELL SOUP CO	UNIVERSAL DISPLA	US FOODS HOLDING	0.909697894	0.661474	2.06E-05	0.000220682	1.35E-10	1.57117231	2.41E-06	2.03E-17
LITHIA MOTORS-A	AUST & NZ BK-ADR	HILL-ROM HOLDING	AXA -ADR	0.908713766	0.713444	1.80E-05	0.00011149	6.66E-11	1.622157701	1.81E-06	1.54E-17
AUST & NZ BK-ADR	HILL-ROM HOLDING	UNIVERSAL DISPLA	ANALOG DEVICES	0.907160815	0.500401	3.01E-05	-0.000131543	1.17E-10	1.40756134	5.89E-07	2.96E-17
MOTOROLA SOLUTIO	AUST & NZ BK-ADR	HILL-ROM HOLDING	ANALOG DEVICES	0.905905122	0.627067	2.13E-05	0.000396176	1.22E-10	1.532971923	-2.30E-07	1.88E-17
SUN LIFE FINANCI	AUST & NZ BK-ADR	HILL-ROM HOLDING	SMC CORP	0.904373992	0.771327	1.36E-05	-0.000200963	1.11E-10	1.675700548	2.44E-06	2.21E-17
SUN LIFE FINANCI	AUST & NZ BK-ADR	HILL-ROM HOLDING	ENN ENERGY H-ADR	0.903808323	0.812946	1.39E-05	0.000844984	1.53E-10	1.716754294	5.91E-07	2.01E-17
EXPEDITORS INTL	UNIVERSAL DISPLA	AUST & NZ BK-ADR	HILL-ROM HOLDING	0.902180077	0.587358	2.30E-05	0.000504641	1.38E-10	1.489537868	1.04E-06	2.95E-17
PHILLIPS 66	UNIVERSAL DISPLA	LITHIA MOTORS-A	AUST & NZ BK-ADR	0.902035571	0.629635	1.92E-05	-1.63E-05	1.67E-10	1.531670141	2.11E-06	1.76E-17
FOOT LOCKER INC	AUST & NZ BK-ADR	HILL-ROM HOLDING	ANALOG DEVICES	0.901947643	0.558906	2.53E-05	0.000151596	1.67E-10	1.460853914	-6.75E-07	2.25E-17
AUST & NZ BK-ADR	HILL-ROM HOLDING	ANALOG DEVICES	KERING-UNSP ADR	0.901907941	0.634331	2.02E-05	0.000952897	2.47E-10	1.536239271	-3.95E-07	1.77E-17
SUN LIFE FINANCI	DOLBY LABORATO-A	LITHIA MOTORS-A	AUST & NZ BK-ADR	0.901639809	0.631456	2.00E-05	-0.000189707	1.22E-10	1.533095995	2.27E-06	2.03E-17
SUN LIFE FINANCI	LITHIA MOTORS-A	AUST & NZ BK-ADR	HILL-ROM HOLDING	0.901601574	0.72343	1.60E-05	0.001094045	2.25E-10	1.625031111	1.46E-06	1.90E-17
LITHIA MOTORS-A	UNIVERSAL DISPLA	AUST & NZ BK-ADR	CAMPBELL SOUP CO	0.901493677	0.611258	2.26E-05	0.000703716	1.53E-10	1.512752168	2.76E-06	1.78E-17
EXPEDITORS INTL	LITHIA MOTORS-A	AUST & NZ BK-ADR	HILL-ROM HOLDING	0.901486067	0.673471	1.84E-05	-0.00024995	1.03E-10	1.574957557	8.48E-07	1.86E-17
PEOPLE'S UNITED	AUST & NZ BK-ADR	HILL-ROM HOLDING	ANALOG DEVICES	0.901306367	0.542415	2.86E-05	0.000160084	7.42E-11	1.443721501	-5.11E-08	2.64E-17
PPL CORP	SUN LIFE FINANCI	AUST & NZ BK-ADR	HILL-ROM HOLDING	0.90128494	0.861647	1.19E-05	-0.00051326	2.37E-10	1.762932117	1.61E-06	2.06E-17
PPL CORP	ANALOG DEVICES	AUST & NZ BK-ADR	HILL-ROM HOLDING	0.901116448	0.630311	2.00E-05	0.00039568	1.02E-10	1.531427898	5.52E-07	1.98E-17
LITHIA MOTORS-A	HILL-ROM HOLDING	STEEL DYNAMICS	AUST & NZ BK-ADR	0.900251557	0.681851	1.92E-05	0.000436822	6.75E-11	1.582102855	1.18E-06	2.06E-17
OGE ENERGY CORP	HILL-ROM HOLDING	LITHIA MOTORS-A	AUST & NZ BK-ADR	0.900045914	0.700811	1.61E-05	0.000497773	1.88E-10	1.600856843	9.86E-07	1.64E-17

Figure 4.9: TOP 20 rated assets by algorithm ranked on correlation

Each of these top 20 solutions ranked according to the different ranking strategies can be seen as an individual candidate solution set S.

4.3.4 Discussion

From figures 4.9, 4.10 and 4.11 we observe that each solution, which is included in one of these ranking strategies, outperforms the benchmark.

By looking at the top ranked solutions based on correlations (figure 4.9) an average improvement in comparison to the benchmark is detected of 29.81 percent. Interesting to note

RHS1	RHS2	RHS3	RHS4	Correlation	betas	variances	beta_correlation	tracking_error_degree1_portfolio	tracking_error_degree2_portfolio
PHILLIPS 66	AGILENT TECH INC	TRACTOR SUPPLY	MITSUB ELEC-ADR	0.871956892	1.055383	7.65E-06	1.927339471	9.75E-07	1.34E-17
WHIRLPOOL CORP	TRACTOR SUPPLY	AUST & NZ BK-ADR	AGILENT TECH INC	0.878214596	1.044492	7.42E-06	1.922707019	6.38E-07	1.20E-17
TRACTOR SUPPLY	AUST & NZ BK-ADR	SCHNEIDER -ADR	AGILENT TECH INC	0.875444807	1.030082	7.41E-06	1.905526797	1.06E-06	1.09E-17
WHIRLPOOL CORP	LHC GROUP INC	AUST & NZ BK-ADR	US FOODS HOLDING	0.88100012	1.029481	9.75E-06	1.910481243	1.20E-06	1.96E-17
TRACTOR SUPPLY	AGILENT TECH INC	ADOBE INC	AUST & NZ BK-ADR	0.873391909	1.016773	6.54E-06	1.890164945	8.44E-07	1.19E-17
ATOS ORIGIN-ADR	GALAPAGOS NV-ADR	US FOODS HOLDING	WM MORRISON-ADR	0.871050809	1.006786	9.06E-06	1.877836943	7.50E-07	1.53E-17
SUN LIFE FINANCI	FAIR ISAAC CORP	AUST & NZ BK-ADR	VEREIT INC	0.873342122	0.989738	8.47E-06	1.863080527	-1.42E-08	1.60E-17
TRACTOR SUPPLY	SUN LIFE FINANCI	AUST & NZ BK-ADR	FAIR ISAAC CORP	0.880625666	0.984507	8.68E-06	1.865132675	7.60E-07	1.55E-17
NATURGY ENER-ADR	MOTOROLA SOLUTIO	HILL-ROM HOLDING	KERING-UNSP ADR	0.872177976	0.982713	9.08E-06	1.8548914	1.29E-06	1.32E-17
AGILENT TECH INC	TRACTOR SUPPLY	LIBERTY BR-C	GOLD FIELDS-ADR	0.873619596	0.951358	9.47E-06	1.824977893	6.65E-07	1.76E-17
SUN LIFE FINANCI	AUST & NZ BK-ADR	VEREIT INC	WM MORRISON-ADR	0.880162239	0.943661	1.10E-05	1.823823398	4.37E-07	1.59E-17
INTERPUBLIC GRP	GALAPAGOS NV-ADR	US FOODS HOLDING	WM MORRISON-ADR	0.870536902	0.941564	9.66E-06	1.812101144	1.28E-06	1.65E-17
EXPEDITORS INTL	TECHTRONIC I-ADR	HILL-ROM HOLDING	KERING-UNSP ADR	0.873897716	0.941473	8.84E-06	1.815370304	1.21E-06	1.90E-17
TRACTOR SUPPLY	MITSUB ELEC-ADR	HILL-ROM HOLDING	KERING-UNSP ADR	0.873574377	0.940941	9.75E-06	1.814515618	1.16E-06	1.50E-17
TRACTOR SUPPLY	SUN LIFE FINANCI	AUST & NZ BK-ADR	AGILENT TECH INC	0.873044563	0.934729	8.79E-06	1.807773077	9.42E-07	1.51E-17
PEARSON PLC-ADR	AUST & NZ BK-ADR	HILL-ROM HOLDING	KERING-UNSP ADR	0.87212086	0.933444	1.02E-05	1.805565222	1.52E-07	1.57E-17
PHILLIPS 66	AGILENT TECH INC	TRACTOR SUPPLY	AUST & NZ BK-ADR	0.871145357	0.933311	9.23E-06	1.804456563	6.29E-07	1.20E-17
SUN LIFE FINANCI	FAIR ISAAC CORP	AUST & NZ BK-ADR	BCE INC	0.877358296	0.932933	9.29E-06	1.810291634	9.64E-07	1.51E-17
AUST & NZ BK-ADR	US FOODS HOLDING	WESTPAC BANK-ADR	NORSK HYDRO-ADR	0.873271286	0.930952	1.02E-05	1.804222806	1.79E-06	1.24E-17
OGE ENERGY CORP	HILL-ROM HOLDING	TEXAS PACIFIC LA	FRANCO-NEVADA CO	0.870757882	0.930475	1.15E-05	1.801232958	2.48E-06	1.80E-17

Figure 4.10: TOP 20 rated assets by algorithm ranked on betas

RHS1	RHS2	RHS3	RHS4	Correlation	betas	variances	beta_correlation	tracking_error_degree1_portfolio	tracking_error_degree2_portfolio
TRACTOR SUPPLY	AGILENT TECH INC	AUST & NZ BK-ADR	SCHNEIDER -ADR	0.875444807	1.030082	7.41E-06	1.905526797	1.06E-06	1.09E-17
TRACTOR SUPPLY	ADOBE INC	AUST & NZ BK-ADR	AGILENT TECH INC	0.873391909	1.016773	6.54E-06	1.890164945	8.44E-07	1.19E-17
PHILLIPS 66	AGILENT TECH INC	TRACTOR SUPPLY	AUST & NZ BK-ADR	0.871145357	0.933311	9.23E-06	1.804456563	6.29E-07	1.20E-17
WHIRLPOOL CORP	TRACTOR SUPPLY	AUST & NZ BK-ADR	AGILENT TECH INC	0.878214596	1.044492	7.42E-06	1.922707019	6.38E-07	1.20E-17
TRACTOR SUPPLY	AUST & NZ BK-ADR	US FOODS HOLDING	AGILENT TECH INC	0.870126141	0.917401	8.38E-06	1.787527605	1.21E-06	1.21E-17
AUST & NZ BK-ADR	MMC NORILSK ADR	WESTPAC BANK-ADR	US FOODS HOLDING	0.87616843	0.92256	9.25E-06	1.798728424	1.99E-06	1.22E-17
TRACTOR SUPPLY	AUST & NZ BK-ADR	INTERACTIVE BROK	AGILENT TECH INC	0.87043354	0.904631	1.06E-05	1.775064201	3.27E-07	1.23E-17
WHIRLPOOL CORP	SANOFI-ADR	LITHIA MOTORS-A	AUST & NZ BK-ADR	0.878773583	0.879215	1.07E-05	1.75798899	1.43E-06	1.24E-17
LITHIA MOTORS-A	AUST & NZ BK-ADR	MITSUBISHI U-ADR	INTERCONTIN-ADR	0.873905719	0.906157	1.08E-05	1.780062607	1.57E-06	1.24E-17
AUST & NZ BK-ADR	US FOODS HOLDING	WESTPAC BANK-ADR	NORSK HYDRO-ADR	0.873271286	0.930952	1.02E-05	1.804222806	1.79E-06	1.24E-17
TRACTOR SUPPLY	AUST & NZ BK-ADR	KKR & CO INC	NORSK HYDRO-ADR	0.87204181	0.895441	1.08E-05	1.767482987	1.78E-06	1.26E-17
HUMANA INC	US FOODS HOLDING	LITHIA MOTORS-A	AUST & NZ BK-ADR	0.882811724	0.882854	1.11E-05	1.76566525	1.56E-06	1.26E-17
TRACTOR SUPPLY	AGILENT TECH INC	HUNTINGTON BANC	AUST & NZ BK-ADR	0.875804806	0.894785	9.04E-06	1.770589963	3.19E-07	1.27E-17
CAMECO CORP	TRACTOR SUPPLY	AUST & NZ BK-ADR	FAIR ISAAC CORP	0.871667762	0.907224	9.73E-06	1.77889215	3.51E-07	1.31E-17
AUST & NZ BK-ADR	KERING-UNSP ADR	SCHNEIDER -ADR	HILL-ROM HOLDING	0.886573982	0.896586	1.10E-05	1.783160359	7.84E-07	1.31E-17
CAMECO CORP	AUST & NZ BK-ADR	NORSK HYDRO-ADR	WM MORRISON-ADR	0.872242249	0.885424	1.24E-05	1.757666597	1.17E-06	1.32E-17
ATOS ORIGIN-ADR	TRACTOR SUPPLY	AUST & NZ BK-ADR	AGILENT TECH INC	0.873625366	0.925035	9.60E-06	1.798659905	2.68E-07	1.32E-17
NATURGY ENER-ADR	MOTOROLA SOLUTIO	HILL-ROM HOLDING	KERING-UNSP ADR	0.872177976	0.982713	9.08E-06	1.8548914	1.29E-06	1.32E-17
TRACTOR SUPPLY	AGILENT TECH INC	AUST & NZ BK-ADR	DOVER CORP	0.873538147	0.8984	8.36E-06	1.771938554	1.52E-06	1.33E-17
CAMECO CORP	AUST & NZ BK-ADR	SCHNEIDER -ADR	WM MORRISON-ADR	0.876550398	0.885603	1.23E-05	1.762153527	7.74E-07	1.33E-17

Figure 4.11: TOP 20 rated assets by algorithm ranked on betas+correlation

is that the mean of the first moment tracking error is slightly positive with: $9.50 * 10^{-07}$. Furthermore, the top ranked solution [4] ranked based on correlation has a 2^{nd} moment tracking error of: $2.03 * 10^{-17}$, which is an improvement of 32 percent. The relative tracking error of this solution is $1.02 * 10^{-13}$.

Furthermore, one interesting solution needs to be highlighted in 4.9. This is the solution with almost identical tracking error compared to the mean of random asset allocation, represented in the third row. This solution has a higher variance than any other solution in this table, sometimes twice the variance of other solutions. Besides a much higher variance, this solution also the lowest beta of the complete table. This is in line with the information found in the bad replacements, and highlights the fact that additional metrics are needed for evaluating the set of possible solutions CSV.

Reviewing the top ranked solutions based on betas (figure 4.11) shows an average improvement in comparison to the baseline of 51.11 percent, which is considerably higher than the mean of the top correlations. The best outputted solution [5] where beta is closest to 1, is the solution located at the 6^{th} line , with a correlation coefficient of: 0.87, beta: 1.01, variance: $9.06 * 10^{-06}$, beta+correlation: 1.88, 1^{st} moment tracking error: $7.50 * 10^{-07}$ and 2^{nd} moment

[4] The top ranked solution is known to be : Aust NZ BK-ADR, Campbell SOUP CO, UniversalL DISPLA, US Foods Holding

[5] Atos origin, Galapagos NV, US Foods Holding, WM Morrison

tracking error: $1.53 * 10^{-17}$. an improvement of 48.57 in comparison to the baseline. This is lower than the average improvement in this section. Hence, this indicates that both the correlation and beta metric do play an important role [6].

By looking at the top ranked solutions based on betas and correlation (figure 4.11) an average improvement in comparison to the baseline of 57.83 percent is detected, which is considerably higher than the mean of the top correlations. Also a slight improvement in comparison to the average of betas is found, which is in line with the expectations. Figure 4.11 shows that we only detect small deviations to the previous result only ranked on betas. However, the minor differences are visible in the average improvements results according to the ranking strategy of betas and correlations are slightly outperforming most solutions. This is visible in the second moment tracking errors. Another visible result are the lower variances, which can justify the positive impact on the second moment tracking error in comparison to the solely beta ranking strategy.

Summary of discussion

To sum up, not all solutions included in the possible solution set CSV outperform the benchmark. However, most of the solutions included outperform the benchmark. Consequently, we can not define a proper candidate set S without the help of additional metrics. Moreover, we observed that all top ranked candidate solutions provided by the algorithm outperformed the benchmark. Therefore, it can be concluded that with the help of the algorithm and the post-processing phase we would be able to define a proper candidate set S. On top of these, out of the three different ranking methods, it is clearly demonstrated that, based on the derivations, the ranking strategy based on correlation and beta combined performs the best. Therefore, the best possible candidate solution set S should be build upon this strategy. With this candidate solution set, we can successfully replace the seven unsustainable assets with assets having minimal ESG rating of A.

4.4 Experiment 3: Replacement of energy sector

In the third experiment, a complete sector in the portfolio (MSCI index) is deleted, the energy industry. The portfolio managers could consider to remove a complete sector in the portfolio, and are interested in seeing the impact by replacing it by other assets. Removing all energy related assets will greening the portfolio for sure. Currently, the set of 'leaving' assets RA does not include assets with A, AA or AAA ESG ratings. Due to the fact that the 'incoming' set of assets is selected from the sustainable possibility set, the 'incoming' assets have by definition these sustainable ratings. We will remove in total 32 assets with a total weight of 2.22 percent.[7]

[6] Please note that all provided solutions by the algorithm do exceed the correlation threshold defined by the user. It is therefore not possible to find solutions with high beta and low correlation coefficients. This makes the evaluation biased. Furthermore, betas higher than 1 indicate more volatility is added to the portfolio, which can be undesirable for an investor.

[7] MSCI energy assets: Exxon mobil corp, Chevron corp, Enbridge inc, TC Energy group, Oneok inc, Hess corp, Comeco corp, BP plc, Repsol sa, OMV AG, Galp energia SGPS SA, Parkland corp, Keyera corp, Tenaris SA, EOG Resources, Kinder Morgan In, Marathon Petrole, Phillips 66, ENI SPA, Pioneer natural, Williams cos

Different than in the second experiment, we have decided to also review the out of sample data set, the future 'unseen' data, for this experiment.

Expectations: Removing a much bigger set, with a much higher accumulated weight, will most likely result in a heavier deviation from the old portfolio. This will lead to larger absolute tracking errors. Nonetheless, we expect that the algorithm will still be capable of selecting suitable sets of 'incoming' assets for all ranking methods. According to the findings in the first experiment, the best solution will be provided by the combination of ranking on betas and correlation. Due to the fact that a set of 32 assets is replaced with a set of 4-5 assets, it might be hard to capture all relevant characteristics of the 'leaving' set of assets. This could lead to higher relative tracking errors. Besides this, it is important to note that we are deleting a complete sector, i.e. the energy sector. Assets within a certain sector are correlated, and therefore it might be interesting whether assets in the possibility set do match closely. Furthermore, the energy sector is not considered to be the most sustainable sector, so replacing it with sustainable assets might lead to adding assets from different sectors which makes it harder to find good candidate solutions. On top of this, deleting the complete energy sector has impact on the market risk and concentration risks, as we are removing a diversification aspect of the portfolio.

4.4.1 Benchmark: random asset selection

In Figure 4.12 a snapshot of random solutions is illustrated. The procedure for this baseline experiment is similar to the previous experiments. Here, we find a mean of the second moment tracking error equal to $4.91 * 10^{-13}$. In absolute terms we detect a much higher tracking error for the baseline, compared to benchmark second moment tracking errors for previous experiments, which was expected. In relative terms we have a second moment tracking error of: $2.23 * 10^{-12}$. We clearly see that this relative tracking error is much higher in comparison to the relative tracking of the benchmark the other experiments. This could indicate that indicate that replacing a large set of data (≥ 15) provides us with significantly worse results.

4.4.2 Algorithmic solution

For this experiment the parameters are defined as follows:

- $\tau = 0.9$
- $\omega = 4$
- $\delta = 0.045$
- Total number of solutions sv in $S = 20$

An interesting point to discuss here is the choice for similar ω compared to previous experiments. It seems reasonable to increase ω as the set of removing assets RA has also increased substantially. However, we argue that it is reasonable to define a ω not bigger than 4. This,

inc, Valero energy, Neste OYJ, Occidental Pete, Equinor ASA, Chenniere energy, Pembina pipeline, Halliburton co, Baker Hughes co, Lundin energy AB, Inter Pipeline L, Cabot Oil and Gas, which have an accumulated weight of 2.22 percent.

Random Asset 1	Random Asset 2	Random Asset 3	Random Asset 4	Tracking error Random degree1
ICON PLC	EXP WORLD HOLDIN	ARCHER-DANIELS	VIACOMCBS INC-B	5.52E-12
TERADYNE INC	COMERICA INC	MAXIMUS INC	CHEGG INC	8.65E-13
TERADYNE INC	JOYY INC	CONSTELLATION-A	ASCENDIS PHA-ADR	7.91E-13
EASTGROUP PROP	GRACO INC	NIKE INC -CL B	INPHI CORP	6.57E-13
NEXSTAR MEDIA-A	UNUM GROUP	TRIP.COM GRO-ADR	SCOTTS MIRACLE	6.34E-13
L BRANDS INC	CSL LTD-SPON ADR	EXACT SCIENCES	ENN ENERGY H-ADR	6.23E-13
BLUEPRINT MEDICI	MAXIMUS INC	SYNCHRONY FINANC	BASF SE-ADR	6.04E-13
LOCKHEED MARTIN	SHARP CORP-ADR	UBIQUITI INC	II-VI INC	5.96E-13
EXP WORLD HOLDIN	BUILDERS FIRSTSO	ADV MICRO DEVICE	GOLD FIELDS-ADR	5.88E-13
GENUINE PARTS CO	ROYAL CARIBBEAN	UBISOFT-UNS ADR	DISCOVER FINANCI	5.71E-13
COCA-COLA AM-ADR	SGS SA-UNSP ADR	JONES LANG LASAL	DIAMONDBACK ENER	5.65E-13
MARATHON OIL	NICE LTD -SP ADR	INTESA SAN- ADR	ZILLOW GRO-C	5.61E-13
MARATHON OIL	PERFORMANCE FOOD	COMERICA INC	WESTERN ALLIANCE	5.61E-13
MARRIOTT INTL-A	RAKUTEN INC-ADR	NEW YORK TIMES-A	ROYAL CARIBBEAN	5.55E-13
RPM INTL INC	CLEARWAY ENERG-A	L BRANDS INC	AAC TECHNOLO-ADR	5.53E-13
VULCAN MATERIALS	SL GREEN REALTY	OLLIE'S BARGAIN	ECOLAB INC	5.50E-13
UNUM GROUP	CLOROX CO	HUAZHU GROUP LTD	DONALDSON CO INC	5.44E-13
TRIP.COM GRO-ADR	BORGWARNER INC	MAGELLAN MIDSTRE	ORANGE-SPON ADR	5.38E-13
CORELOGIC INC	NATL AUSTR-ADR	MATTEL INC	SID NACIONAL-ADR	5.37E-13
UNUM GROUP	MOBILE TELES-ADR	ENCOMPASS HEALTH	PRA HEALTH SCIEN	5.25E-13
ENN ENERGY H-ADR	SPLUNK INC	BALL CORP	OLLIE'S BARGAIN	5.25E-13
LATTICE SEMICOND	MITSUB ELEC-ADR	TERADYNE INC	FIRST FIN BANKSH	5.20E-13
ELECTRONIC ARTS	MATTEL INC	G4S PLC-UNS ADR	PEARSON PLC-ADR	5.12E-13
RELIANCE STEEL	COMPAGNIE DE-ADR	BOOZ ALLEN HAMIL	BOSTON BEER-A	5.11E-13
CULLEN/FROST	MAXIMUS INC	ALLEGION PLC	JACK HENRY	5.06E-13
TRAVEL + LEISURE	ASGN INC	BOYD GAMING CORP	IDEX CORP	5.02E-13
DR. REDDY'S LABO	PROOFPOINT INC	STRYKER CORP	FIRST AMERICAN F	5.01E-13
OLYMPUS CORP-ADR	NETAPP INC	TAPESTRY INC	CARNIVAL PLC-ADR	5.00E-13
AES CORP	MAXIMUS INC	PETROBR-SP P ADR	BRISTOL-MYER SQB	4.99E-13

Figure 4.12: Random asset allocation for energy sector

due to running time complications and the need of using of cloud computing clusters which is very costly [8]. More importantly, this allows us to compare results between the different experiments and clearly show the impact of increasing the size of the RA by using this framework. Increasing ω to 5 could help us in this particular case to define slightly better 'incoming' sets of assets, but the solutions are not scalable. Moreover, an ω greater than 6 would not be feasible, even by using cloud computing clusters, as the complexity grows exponentially.

For this particular experiment it is true that a candidate solution set S can be defined without having bad replacements included. However, we are aiming for a robust evaluation of the framework. According to the solutions found in the previous experiment we would not suggest creating this candidate set without applying the post-processing phase.

4.4.3 Algorithmic solutions and post-processing phase

Solutions provided by the algorithm have been ranked according to the three different ranking strategies. Selecting the top 20 solution here provides us with candidate solution set S. The top 20 ranked correlation solutions are illustrated in Figure 4.9. In figure 4.10 the top 20 solutions ranked according to the beta coefficient are shown. Last, in figure 4.11, the com-

[8]This is not feasible for the use of this framework. The small gain we would create, by using this framework in comparison to reviewing smaller possibility sets, is offset by the cost involved by solving these complications

bination of both coefficients have been used to rank them.

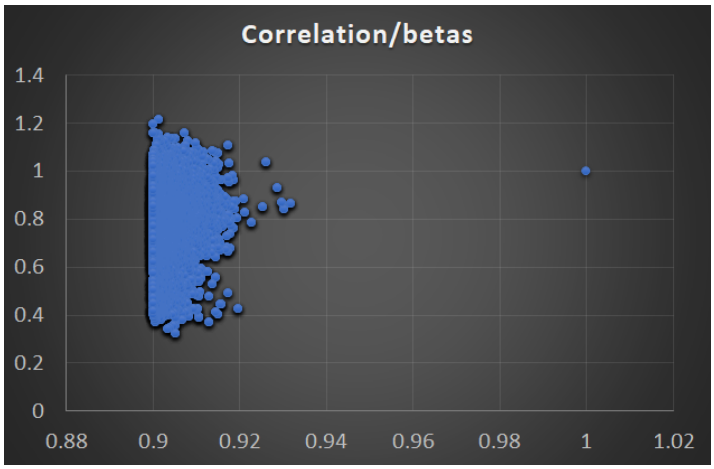

Figure 4.13: Ranked solutions degree 3 and 4 correlation vs Beta

RHS1	RHS2	RHS3	RHS4	CORR	betas	tracking_error_degree1	tracking_error_degree2	betas_correlation	te_1_portfolio	te_2_portfolio
MARATHON PETROLE	EXELIXIS INC	KKR & CO INC	NORSK HYDRO-ADR	0.93	0.93	0.000846277	8.73E-11	1.86	3.03E-05	3.28E-13
ROYAL DUTCH-ADR.1	MARATHON PETROLE	NORSK HYDRO-ADR	WM MORRISON-ADR	0.93	0.84	0.000425694	6.00E-11	1.77	1.47E-05	3.40E-13
BANK RAKYAT-ADR	CARREFOUR SA-ADR	ALLIANT ENERGY	FRANCO-NEVADA CO	0.93	0.85	0.000510949	5.34E-11	1.78	1.66E-05	2.76E-13
AMERICAN ELECTRI	MARATHON PETROLE	NORSK HYDRO-ADR	WM MORRISON-ADR	0.93	0.87	0.000471009	5.80E-11	1.8	1.60E-05	3.19E-13
AIA GROUP LT-ADR	MARATHON PETROLE	NORSK HYDRO-ADR	WM MORRISON-ADR	0.93	0.87	0.000457811	5.92E-11	1.8	1.60E-05	3.31E-13
DEUTSCHE BOE-ADR	WM MORRISON-ADR	MARATHON PETROLE	NORSK HYDRO-ADR	0.93	1.04	0.000585074	4.20E-11	1.97	2.00E-05	2.75E-13
WHIRLPOOL CORP	FRANCO-NEVADA CO	MARATHON PETROLE	EXELIXIS INC	0.92	1.03	0.000559719	5.67E-11	1.95	1.92E-05	2.89E-13
MARATHON PETROLE	EXELIXIS INC	CAMPBELL SOUP CO	FRANCO-NEVADA CO	0.92	0.89	0.000744763	5.65E-11	1.81	2.63E-05	2.80E-13
AU OPTR-SPON ADR	MMC NORILSK ADR	VALLEY NATL BANC	TRACTOR SUPPLY	0.92	0.77	0.000409147	8.88E-11	1.69	1.33E-05	3.66E-13
WHIRLPOOL CORP	VALE SA-SP ADR	MARATHON PETROLE	RALPH LAUREN COR	0.92	0.96	0.000339923	6.13E-11	1.88	1.18E-05	3.97E-13
MARATHON PETROLE	WM MORRISON-ADR	GOLD FIELDS-ADR	NORSK HYDRO-ADR	0.92	0.96	0.000560541	4.43E-11	1.88	1.78E-05	3.72E-13
WHIRLPOOL CORP	CHINA SHENH-ADR	BLUEPRINT MEDICI	FRANCO-NEVADA CO	0.92	0.4	0.000736465	3.61E-10	1.32	2.43E-05	5.66E-13
AIA GROUP LT-ADR	NORSK HYDRO-ADR	MARATHON PETROLE	RALPH LAUREN COR	0.92	0.82	0.000626936	7.65E-11	1.74	2.31E-05	3.89E-13
GARTNER INC	ARROWHEAD PHARMA	BURBERRY GRO-ADR	FRANCO-NEVADA CO	0.92	1.02	0.000710382	6.01E-11	1.94	2.46E-05	2.69E-13
WHIRLPOOL CORP	WM MORRISON-ADR	MARATHON PETROLE	MOTOROLA SOLUTIO	0.92	1.03	0.00045732	4.73E-11	1.95	1.51E-05	2.83E-13
WHIRLPOOL CORP	CHEVRON CORP	ARROWHEAD PHARMA	WM MORRISON-ADR	0.92	0.96	0.000498536	4.39E-11	1.88	1.61E-05	2.59E-13
AMERICAN ELECTRI	NORSK HYDRO-ADR	MARATHON PETROLE	RALPH LAUREN COR	0.92	0.83	0.000640134	7.47E-11	1.75	2.31E-05	3.83E-13
AU OPTR-SPON ADR	MARATHON PETROLE	MOTOROLA SOLUTIO	HILL-ROM HOLDING	0.92	0.85	0.000346116	6.70E-11	1.77	1.12E-05	3.74E-13
AMERICAN ELECTRI	ARROWHEAD PHARMA	KKR & CO INC	NORSK HYDRO-ADR	0.92	0.82	0.000784537	8.83E-11	1.74	2.78E-05	2.74E-13
MARATHON PETROLE	UNIVERSAL DISPLA	MOTOROLA SOLUTIO	UGI CORP	0.92	0.72	0.000994266	1.01E-10	1.64	3.49E-05	3.85E-13

Figure 4.14: TOP 20 rated assets by algorithm ranked on correlation

RHS1	RHS2	RHS3	RHS4	CORR	betas	tracking_error_degree1	tracking_error_degree2	betas_correlation	te_1_portfolio	te_2_portfolio
WHIRLPOOL CORP	WM MORRISON-ADR	TRACTOR SUPPLY	AUST & NZ BK-ADR	0.9	1.22	0.00059342	5.23E-11	2.12	1.88E-05	2.20E-13
WHIRLPOOL CORP	WM MORRISON-ADR	TRACTOR SUPPLY	TELADOC HEALTH I	0.9	1.2	0.000586606	5.92E-11	2.1	1.91E-05	2.47E-13
ALLSTATE CORP	TRACTOR SUPPLY	HEALTHPEAK PROPE	MMC NORILSK ADR	0.9	1.16	0.001102575	7.10E-11	2.06	3.93E-05	2.16E-13
BANK RAKYAT-ADR	MOTOROLA SOLUTIO	LAM RESEARCH	WESTPAC BANK-ADR	0.9	1.16	0.000635743	6.68E-11	2.06	2.18E-05	3.01E-13
WHIRLPOOL CORP	TRACTOR SUPPLY	ANGLO AM-SP ADR	KKR & CO INC	0.9	1.16	0.00079888	8.29E-11	2.06	2.77E-05	2.34E-13
WHIRLPOOL CORP	UNITED MICRO-ADR	MARATHON PETROLE	MOTOROLA SOLUTIO	0.91	1.16	0.000562589	4.92E-11	2.07	1.96E-05	2.72E-13
MARATHON PETROLE	UNITED MICRO-ADR	MOTOROLA SOLUTIO	KIMBERLY-CLARK	0.9	1.14	0.000930797	6.57E-11	2.04	3.26E-05	2.56E-13
WHIRLPOOL CORP	CATERPILLAR INC	RPM INTL INC	WM MORRISON-ADR	0.9	1.14	0.000625748	7.20E-11	2.04	2.06E-05	2.54E-13
ARROWHEAD PHARMA	KERING-UNSP ADR	LAM RESEARCH	DAVITA INC	0.9	1.14	0.000818556	7.40E-11	2.04	2.85E-05	2.57E-13
MEDTRONIC PLC	TRACTOR SUPPLY	MARATHON PETROLE	MMC NORILSK ADR	0.9	1.14	0.000986811	7.00E-11	2.04	3.55E-05	2.52E-13
MARATHON PETROLE	UNITED MICRO-ADR	MOTOROLA SOLUTIO	NORSK HYDRO-ADR	0.9	1.14	0.000809045	6.30E-11	2.04	2.86E-05	2.50E-13
WHIRLPOOL CORP	FRANCO-NEVADA CO	CHINA SHENH-ADR	UNICHARM CORP	0.91	1.14	0.000633549	6.00E-11	2.05	2.06E-05	2.47E-13
WHIRLPOOL CORP	MMC NORILSK ADR	TRACTOR SUPPLY	MARATHON PETROLE	0.91	1.13	0.000679091	5.07E-11	2.04	2.42E-05	2.69E-13
WHIRLPOOL CORP	MEDTRONIC PLC	MARATHON PETROLE	MMC NORILSK ADR	0.9	1.13	0.000842041	6.23E-11	2.03	3.05E-05	2.68E-13
WHIRLPOOL CORP	TRACTOR SUPPLY	BRIDGESTONE-ADR	MMC NORILSK ADR	0.9	1.13	0.001236153	8.84E-11	2.03	4.41E-05	2.68E-13
UNITED PARCEL-B	TRACTOR SUPPLY	ARROWHEAD PHARMA	MMC NORILSK ADR	0.9	1.13	0.001182038	7.91E-11	2.03	4.24E-05	2.18E-13
WHIRLPOOL CORP	FRANCO-NEVADA CO	TENET HEALTHCARE	ARROWHEAD PHARMA	0.9	1.12	0.000676658	4.68E-11	2.02	2.25E-05	2.46E-13
ELECTRONIC ARTS	MMC NORILSK ADR	TRACTOR SUPPLY	ARROWHEAD PHARMA	0.9	1.12	0.000927018	5.81E-11	2.02	3.26E-05	2.00E-13
WHIRLPOOL CORP	WM MORRISON-ADR	MARATHON PETROLE	ILLINOIS TOOL WO	0.91	1.11	0.00054618	5.11E-11	2.02	1.86E-05	2.84E-13
WHIRLPOOL CORP	ARROWHEAD PHARMA	OMRON CORP-ADR	AUST & NZ BK-ADR	0.9	1.11	0.00072949	4.45E-11	2.01	2.48E-05	2.14E-13

Figure 4.15: TOP 20 rated assets by algorithm ranked on betas

RHS1	RHS2	RHS3	RHS4	CORR	betas	tracking_error_degree1	tracking_error_degree2	betas_correlation	te_1_portfolio	te_2_portfolio
WHIRLPOOL CORP	MOTOROLA SOLUTIO	MARATHON PETROLE	RPM INTL INC	0.92	1.11	0.000554712	6.24E-11	2.03	1.95E-05	2.75E-13
MARATHON PETROLE	WM MORRISON-ADR	E.ON SE-ADR	NORSK HYDRO-ADR	0.92	1.08	0.00055367	5.64E-11	2	1.82E-05	2.30E-13
DEUTSCHE BOE-ADR	WM MORRISON-ADR	MARATHON PETROLE	NORSK HYDRO-ADR	0.93	1.04	0.000585074	4.20E-11	1.97	2.00E-05	2.75E-13
WHIRLPOOL CORP	FRANCO-NEVADA CO	MARATHON PETROLE	EXELIXIS INC	0.92	1.03	0.000559719	5.67E-11	1.95	1.92E-05	2.89E-13
WHIRLPOOL CORP	WM MORRISON-ADR	MARATHON PETROLE	MOTOROLA SOLUTIO	0.92	1.03	0.00045732	4.73E-11	1.95	1.51E-05	2.83E-13
GARTNER INC	ARROWHEAD PHARMA	BURBERRY GRO-ADR	FRANCO-NEVADA CO	0.92	1.02	0.000710382	6.01E-11	1.94	2.46E-05	2.69E-13
MACQUARIE GR-ADR	MARATHON PETROLE	NORSK HYDRO-ADR	WM MORRISON-ADR	0.92	0.98	0.000744623	5.45E-11	1.9	2.59E-05	2.92E-13
WHIRLPOOL CORP	MARATHON PETROLE	RALPH LAUREN COR	FRANCO-NEVADA CO	0.92	0.97	0.000627234	6.87E-11	1.89	2.17E-05	3.16E-13
GARTNER INC	ARROWHEAD PHARMA	BURBERRY GRO-ADR	BRITISH LAND-ADR	0.92	0.97	0.000591245	7.31E-11	1.89	2.10E-05	2.79E-13
WHIRLPOOL CORP	VALE SA-SP ADR	MARATHON PETROLE	RALPH LAUREN COR	0.92	0.96	0.000339923	6.13E-11	1.88	1.18E-05	3.97E-13
MARATHON PETROLE	WM MORRISON-ADR	GOLD FIELDS-ADR	NORSK HYDRO-ADR	0.92	0.96	0.000560541	4.43E-11	1.88	1.78E-05	3.72E-13
WHIRLPOOL CORP	CHEVRON CORP	ARROWHEAD PHARMA	WM MORRISON-ADR	0.92	0.96	0.000498536	4.39E-11	1.88	1.61E-05	2.59E-13
INTERPUBLIC GRP	MARATHON PETROLE	NORSK HYDRO-ADR	WM MORRISON-ADR	0.92	0.95	0.000679475	5.66E-11	1.87	2.28E-05	3.22E-13
MARATHON PETROLE	EXELIXIS INC	KKR & CO INC	NORSK HYDRO-ADR	0.93	0.93	0.000846277	8.73E-11	1.86	3.03E-05	3.28E-13
GARTNER INC	MS&AD INSUR-ADR	MARATHON PETROLE	MOTOROLA SOLUTIO	0.92	0.91	0.000634294	6.14E-11	1.83	2.14E-05	3.56E-13
MARATHON PETROLE	EXELIXIS INC	CAMPBELL SOUP CO	FRANCO-NEVADA CO	0.92	0.89	0.000744763	5.65E-11	1.81	2.63E-05	2.80E-13
AMERICAN ELECTRI	MARATHON PETROLE	NORSK HYDRO-ADR	WM MORRISON-ADR	0.93	0.87	0.000471009	5.80E-11	1.8	1.60E-05	3.19E-13
AIA GROUP LT-ADR	MARATHON PETROLE	NORSK HYDRO-ADR	WM MORRISON-ADR	0.93	0.87	0.000457811	5.92E-11	1.8	1.60E-05	3.31E-13
WM MORRISON-ADR	MARATHON PETROLE	CHEVRON CORP	NORSK HYDRO-ADR	0.92	0.88	0.000417304	6.03E-11	1.8	1.39E-05	3.10E-13
WHIRLPOOL CORP	NATIONAL RETAIL	MARATHON PETROLE	RALPH LAUREN COR	0.92	0.88	0.000756407	7.37E-11	1.8	2.65E-05	3.46E-13

Figure 4.16: TOP 20 rated assets by algorithm ranked on betas+correlation

Future evaluation

To see the impact of a change in a current portfolio, we are using a out of sample data. With the help of this data set we can quantify the loss or gain resulting from a change made at a certain point in time.

Our possibility set and portfolio data is both defined on the interval up until 17^{th} August 2020. We have data ranging from 17^{th} up until 1^{st} of June 2021 to evaluate a change made at this crossing point, 17^{th} of August 2020. To evaluate 'future' data, we replace the set of 'leaving' assets RA by two different set of 'incoming' assets and compare them. Obviously, one of these set of 'incoming' assets is defined by the random asset selection. The other set of 'incoming' assets is defined by the top solution sv selected from the best candidate solution set S, which is defined by using the ranking strategy of correlation and beta combined.

This RA has been replaced at 17^{th} of August 2020 by the randomly defined set of 'incoming' assets: Icon PLC, Exp World holdin, Archer Daniels, Viacoms. The first moment tracking error is defined as: $4.3 * 10^{-5}$ and the second moment tracking error is defined as: $3.4 * 10^{-13}$. Furthermore, the top ranked solution according to candidate solution set S equals: Zoetis Inc, Tractor Supply Company, Arrowhead Pharma, MMC Norilsk ADR. This solution having a first moment tracking error of $-7.3 * 10^{-06}$ and a second moment tracking error of $1.6 * 10^{-13}$.

From this we can observe that the top ranked solution has an improvement of 53 percent in comparison to the random selection of the 'incoming' set of assets. Furthermore, we observe that the random selection of assets has a small increase in the payoff. In contrast to the top ranked solution which has a slightly negative deviation. The top ranked solution has an improvement of 83 percent in comparison of the random selection. It should however be noted that the deviation from the random asset selection is positive. When reviewing this in terms of money, we would end up with an increase of risk exposure. We could also have an slight increase of the risk exposure with the proposed algorithm, which should be evaluated according to more econometric methods. Without the help of additional information, no clear conclusion can be drawn here.

4.4.4 Discussion

Similar to the previous experiment we can observe that all top ranked solutions are outperforming the baseline indication according to the Figures 4.14, 4.15 and 4.16. Also here we have create a plot with the results provided by the algorithm where the x-axis represents the correlation coefficient and the y-axis represents the beta coefficient.

The top 20 results ranked according to correlation only are displayed in figure 4.14. The top rated solution [9] has the following properties: correlation:0.93, beta:0.93, beta+correlation:1.86, 1^{st} moment tracking error: $3.03 * 10^{-05}$, 2^{nd} moment tracking error: $3.28 * 10^{-13}$, which implies an improvement of 33.26 in comparison to the baseline model.

Second, the top ranked solution according to ranking solely based on betas is [10]. All top 20 results are begin displayed in figure 4.15. The top ranked solution has the following properties: correlation: 0.90, beta: 1.22, beta and correlation: 2.12, 1^{st} moment tracking error: $1.88 * 10^{-05}$, 2^{nd} moment tracking error: $2.20 * 10^{-13}$, which is an improvement to the baseline of 55.26 percent.

Third, ranking the solutions based on a combination of correlation and beta coefficient is illustrated in 4.16. The top solutions [11] has the following coefficients: correlation:0.91, beta: 1.09, beta and correlation: 2, 1^{st} moment tracking error: $2.44 * 10^{-05}$, 2^{nd} moment tracking error: $2.54 * 10^{-13}$. This is an improvement of 48 percent compare to the baseline indication.

These results are mostly in line with the expectations. We expected that our framework outperforms the random asset allocation with all ranking methods. All ranking methods have proved to outperform the random asset allocation methods by at least 30 percent.

We also expected that the relative tracking errors compared to the previous experiments is expected to be much higher. deleting a complete sector is impacting the characteristics of an portfolio heavily. The total number of removing assets seems to be to high and should be capped at approximately 15 assets at once.

Summary of discussion

All in all, the replacement of the complete energy sector has successfully been processed by this framework. We are able to provide decent candidate sets based on all three ranking metrics. Moreover, the results we found are in line with the expectations we had for replacing such big set of assets. Therefore, we should cap the cardinality of the RA, which could be for example 15. This should however be investigated further to define a accurate number for the maximum cardinality of RA, most likely in a relationship with ω. Furthermore, most of the results are in line with the previous experiment, which indicates the best candidate solutions set based on both ranking metrics.

[9] Marathon Petrole, Exelixis Inc, KKR & CO inc, Norsk Hydro-ADR,
[10] Whirlpool corp, WM Morrisson-ADR, Tractor supply and Aust NZ bank−ADR
[11] Zoetis Inc, WM Morrison-ADR, Tractor Supply, Arrowhead Pharma

For both these reasons, it might be better to replace 32 assets in packages of 8, resulting in lower relative tracking error. This could be reasoned from the solutions provided in the second experiment, where we show a better relative performance in comparison to this method.

4.5 Summary of all results

All experiments have been conducted in line with the experimental setup presented in 4.1. For all experiments improvements have been observed in comparison to the benchmark. The found improvements are, based on the first and second moment tracking errors, of respectively 10-20 percent for the first moment and 30-50 percent for the second moment tracking errors.

In the second experiment we have replaced seven CCC rated assets and showed that the use of the algorithm, more specifically the algorithm plus the post-processing phase ranking strategies, outperformed the random selection of replacement assets. The take-away of this experiment was that the post-processing phase of this framework is of vital importance to not end up with bad replacement solutions.

In the third experiment, a complete sector has been replaced. Although, we have found a succesfull canidate solution set for replacing this set, the key thing to note here is that the cardinality of RA does have significant impact on the performance. Therefore, there should be additional research to investigate the optimal relationship between RA and ω.

Chapter 5

Future work

Although we have observed a decent performance of the framework, additional research could improve it further. The core of the framework is based on a chosen similarity measure and an aggregation method. The implementation of other similarity measures and aggregation methods could be interesting to evaluate. For example, using the Minkowski distance with p=3 would have major impact on the evaluation of time-series, which have outliers. These outliers will be penalized more, by using the L_3 norm, and this could help to clean the output of the algorithm. This could result in less bad replacement selections, which have been observed in experiment 1. Moreover, the volatility differences between the assets could be limited with the help of this method. Furthermore, it could be interesting to combine different aggregation metrics, by combining aggregation methods which do perform well in detecting the overall pattern and the outliers.

Besides this, we observed that our proposed candidate solutions sets are best ranked with the help of a combination of both correlation and beta coefficients. The help of additional metrics, such as co-variance and variance analysis, could provide us with even better ranking methods. This could in term lead to better candidate solution sets. On top of this, it could be interesting to swap order applied by the algorithm and let the algorithm first select on the basis of beta, and second rank them based on correlation. This swap might be interesting, as beta is already a function of co-variance, variance and correlations. Moreover, including both in the core algorithm phase could lead to even better candidate solution sets.

Last it could be worth researching the optimal maximal cardinality function for RA in combination with ω. Finding this optimal balance between these two factors can provide us with more accurate results and more importantly result in a more robust framework.

Chapter 6

Conclusions

This thesis investigated the possibility of defining a proper candidate solution set for replacing a set of assets in an optimal portfolio. We have defined a proper candidate solution set as a set containing all solutions satisfying all user-defined requirements: 1) minimum correlation threshold τ, 2) minimum jump δ, 3) maximum degree of freedom ω and 4) the number of solutions included in the candidate set. Furthermore, with the help of these user-defined requirements we aimed to replace the set of 'leaving' assets while approaching similar risk-return levels. This allows portfolio managers to anticipate to new investment opportunities, regulatory constraints and investor preferences.

Overall, we can conclude that our proposed framework has decent performance based on the results found in the experiments. We have seen significant improvements of over 50 percent depending on the 'leaving' set of assets. Furthermore, all candidate solutions sets S, provided by the post-processing phase, do out performed the random asset selection for each experiment. However, it is important to note that the ranking strategies are needed. We have observed a small set of bad replacement assets when not applying ranking strategies and simply randomly select a set of assets from the results of the algorithm. having said this, it was clearly visible that solutions ranked according to a combination of the multiple correlation and beta coefficients had the strongest performance. This has result in the biggest improvements based on percentages calculated on the second moment tracking errors. This tells us that the help of beta does improve the quality of the candidate solution set proposed by this framework.

From the conclusions drawn above, we should also be critical towards the framework. Adding this extra econometric metric, beta, has improved the quality of the candidate solution set S significantly, hinting on the fact that with the help of other econometric metrics we could create an even better solution set S. Furthermore, this framework does not account for most of the risk measures involved in the field of finance and risk management. For now, this is left for the user. The user will select the most optimal solution based on the proposed candidate solution set S, by evaluating according to all risk factors and diversification measures. This framework could improve, by incorporating some of these risk factors in the model. For example, in stead of defining a single possibility set, it would be interesting to evaluate multiple possibility set each with including assets with totally different diversification aspects like: geographical location or currency.

Bibliography

[1] MSCI, "2021 esg investor insight report: Natixis investment managers." vii, 1, 17

[2] H. Markowitz, "Portfolio selection," *The Journal of Finance*, vol. 7, pp. 77–91, Mar. 1952. 1, 19

[3] K. Minartz and O. Papapetrou, "Correlation detective: Efficient multivariate correlation mining," 2020. 2, 3, 5, 16, 25

[4] R. J. Hyndman, "George athanasopoulos," *Forecasting: Principles and Practice. Monash University, Australia*, 2018. 2, 23

[5] H. Markowitz, "Modern portfolio theory," *Journal of Finance*, vol. 7, no. 11, pp. 77–91, 1952. 5, 7

[6] S. Agrawal, G. Atluri, A. Karpatne, W. Haltom, S. Liess, S. Chatterjee, and V. Kumar, "Tripoles: A new class of relationships in time series data," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 697–706, 2017. 5, 16

[7] S. Agrawal, M. Steinbach, D. Boley, S. Chatterjee, G. Atluri, A. T. Dang, S. Liess, and V. Kumar, "Mining novel multivariate relationships in time series data using correlation networks," *IEEE Transactions on Knowledge and Data Engineering*, vol. 32, no. 9, pp. 1798–1811, 2019. 5, 16

[8] R. S. Hudson and A. Gregoriou, "Calculating and comparing security returns is harder than you think: A comparison between logarithmic and simple returns," *International Review of Financial Analysis*, vol. 38, pp. 151–162, 2015. 6

[9] M. Schmalz, "Computer arithmetic." 7

[10] Coonen, "Special feature an implementation guide to a proposed standard for floating-point arithmetic," *Computer*, vol. 13, no. 1, pp. 68–79, 1980. 7

[11] F. J. Fabozzi and J. C. Francis, "Beta as a random coefficient," *Journal of Financial and Quantitative Analysis*, vol. 13, no. 1, pp. 101–116, 1978. 9, 10

[12] B.-K. Yi and C. Faloutsos, "Fast time sequence indexing for arbitrary lp norms," 2000. 12

[13] L. Ferreira and L. Zhao, "Time series clustering via community detection in networks," *Information Sciences*, vol. 326, 08 2015. 12

[14] O. Gold and M. Sharir, "Dynamic time warping and geometric edit distance: Breaking the quadratic barrier," *ACM Transactions on Algorithms (TALG)*, vol. 14, no. 4, pp. 1–17, 2018. 12

[15] T. Cover and J. Thomas, "Elements of information theory,(pp 33-36) john wiley and sons," *Inc, NY*, 1991. 12

[16] T. E. Duncan, "On the calculation of mutual information," *SIAM Journal on Applied Mathematics*, vol. 19, no. 1, pp. 215–220, 1970. 14

[17] S. Watanabe, "Information theoretical analysis of multivariate correlation," *IBM Journal of research and development*, vol. 4, no. 1, pp. 66–82, 1960. 14

[18] R. Roll, "A mean/variance analysis of tracking error," *The Journal of Portfolio Management*, vol. 18, no. 4, pp. 13–22, 1992. 15

[19] A. S. Taschetto, A. S. Gupta, N. C. Jourdain, A. Santoso, C. C. Ummenhofer, and M. H. England, "Cold tongue and warm pool enso events in cmip5: Mean state and future projections," *Journal of Climate*, vol. 27, no. 8, pp. 2861–2885, 2014. 16

[20] J. M. Wallace and D. S. Gutzler, "Teleconnections in the geopotential height field during the northern hemisphere winter," *Monthly weather review*, vol. 109, no. 4, pp. 784–812, 1981. 16

[21] G. Atluri, A. MacDonald III, K. O. Lim, and V. Kumar, "The brain-network paradigm: using functional imaging data to study how the brain works," *Computer*, vol. 49, no. 10, pp. 65–71, 2016. 16

[22] K. Minartz, "Correlation detective: Scalable higher-order correlation discovery in big data." 26

Appendix A

Appendix

LHS	RHS1	RHS2	RHS3	RHS4	Correlation	betas	beta_correlation	tracking_error_degree1_portfolio	tracking_error_degree2_portfolio	outperform random
replacing_assets	AUST & NZ BK-ADR	CAMPBELL SOUP CO	UNIVERSAL DISPLA	US FOODS HOLDING	0.909697894	0.661474416	1.57117231	2.41E-06	2.03E-17	good
replacing_assets	LITHIA MOTORS-A	AUST & NZ BK-ADR	HILL-ROM HOLDING	AXA -ADR	0.908713766	0.713443935	1.622157701	1.81E-06	1.54E-17	good
replacing_assets	AUST & NZ BK-ADR	HILL-ROM HOLDING	UNIVERSAL DISPLA	ANALOG DEVICES	0.907160815	0.500400525	1.40756134	5.89E-07	2.96E-17	good
replacing_assets	MOTOROLA SOLUTIO	AUST & NZ BK-ADR	HILL-ROM HOLDING	ANALOG DEVICES	0.905905122	0.627066801	1.532971923	-2.30E-07	1.88E-17	good
replacing_assets	MOTOROLA SOLUTIO	AUST & NZ BK-ADR	HILL-ROM HOLDING	VIACOMCBS INC-B	0.905637993	0.662007952	1.567645945	-1.88E-07	9.16E-16	bad
replacing_assets	SUN LIFE FINANCI	AUST & NZ BK-ADR	HILL-ROM HOLDING	SMC CORP	0.904373992	0.771326556	1.675700548	2.44E-06	2.21E-17	good
replacing_assets	SUN LIFE FINANCI	AUST & NZ BK-ADR	HILL-ROM HOLDING	ENN ENERGY H-ADR	0.903808323	0.812945971	1.716754294	5.91E-07	2.01E-17	good
replacing_assets	EXPEDITORS INTL	UNIVERSAL DISPLA	AUST & NZ BK-ADR	HILL-ROM HOLDING	0.902180077	0.587357791	1.489537868	1.04E-06	2.95E-17	good
replacing_assets	PHILLIPS 66	UNIVERSAL DISPLA	LITHIA MOTORS-A	AUST & NZ BK-ADR	0.902035571	0.62963457	1.531670141	2.11E-06	1.76E-17	good
replacing_assets	FOOT LOCKER INC	AUST & NZ BK-ADR	HILL-ROM HOLDING	ANALOG DEVICES	0.901947643	0.558906271	1.460853914	-6.75E-07	2.25E-17	good
replacing_assets	AUST & NZ BK-ADR	HILL-ROM HOLDING	ANALOG DEVICES	KERING-UNSP ADR	0.901907941	0.63433133	1.536239271	-3.95E-07	1.77E-17	good
replacing_assets	SUN LIFE FINANCI	DOLBY LABORATO-A	LITHIA MOTORS-A	AUST & NZ BK-ADR	0.901639809	0.631456186	1.533095995	2.27E-06	2.03E-17	good
replacing_assets	SUN LIFE FINANCI	LITHIA MOTORS-A	AUST & NZ BK-ADR	HILL-ROM HOLDING	0.901601574	0.723429537	1.625031111	1.46E-06	1.90E-17	good
replacing_assets	LITHIA MOTORS-A	UNIVERSAL DISPLA	AUST & NZ BK-ADR	CAMPBELL SOUP CO	0.901493677	0.611258491	1.512752168	2.76E-06	1.78E-17	good
replacing_assets	EXPEDITORS INTL	LITHIA MOTORS-A	AUST & NZ BK-ADR	HILL-ROM HOLDING	0.901486067	0.67347149	1.574957557	8.48E-07	1.86E-17	good
replacing_assets	PEOPLE'S UNITED	AUST & NZ BK-ADR	HILL-ROM HOLDING	ANALOG DEVICES	0.901306367	0.542415134	1.443721501	-5.11E-08	2.64E-17	good
replacing_assets	PPL CORP	SUN LIFE FINANCI	AUST & NZ BK-ADR	HILL-ROM HOLDING	0.90128494	0.861647177	1.762932117	1.61E-06	2.06E-17	good
replacing_assets	PPL CORP	ANALOG DEVICES	AUST & NZ BK-ADR	HILL-ROM HOLDING	0.901116448	0.63031145	1.531427898	5.52E-07	1.98E-17	good
replacing_assets	AUST & NZ BK-ADR	HILL-ROM HOLDING	UNIVERSAL DISPLA	VIACOMCBS INC-B	0.900802534	0.515780603	1.416583137	6.32E-07	1.02E-15	bad
replacing_assets	LITHIA MOTORS-A	HILL-ROM HOLDING	STEEL DYNAMICS	AUST & NZ BK-ADR	0.900251557	0.681851298	1.582102855	1.18E-06	2.06E-17	good
replacing_assets	OGE ENERGY CORP	HILL-ROM HOLDING	LITHIA MOTORS-A	AUST & NZ BK-ADR	0.900045914	0.700810929	1.600856843	9.86E-07	1.64E-17	good
replacing_assets	AUST & NZ BK-ADR	HILL-ROM HOLDING	ANALOG DEVICES	WM MORRISON-ADR	0.899906672	0.616356769	1.516263441	-1.55E-07	1.98E-17	good
replacing_assets	AUST & NZ BK-ADR	HILL-ROM HOLDING	ENERGY TRANSFER	ALLIANT ENERGY	0.899896861	0.607950707	1.507847568	1.51E-06	2.55E-17	good
replacing_assets	FOOT LOCKER INC	AUST & NZ BK-ADR	HILL-ROM HOLDING	NINTENDO CO-ADR	0.899874121	0.741855854	1.641729975	7.12E-07	1.78E-17	good
replacing_assets	SANTEN PHARM-ADR	GAMESTOP CORP-A	UNIVERSAL DISPLA	HONDA MOTOR-ADR	0.899770376	0.744636172	1.644406548	2.33E-06	2.55E-17	good
replacing_assets	SUN LIFE FINANCI	AUST & NZ BK-ADR	HILL-ROM HOLDING	ANALOG DEVICES	0.89964407	0.593885173	1.493529243	-2.13E-07	2.67E-17	good
replacing_assets	MOTOROLA SOLUTIO	HILL-ROM HOLDING	UNIVERSAL DISPLA	ANALOG DEVICES	0.899628844	0.525891041	1.425519885	9.11E-07	2.86E-17	good
replacing_assets	LITHIA MOTORS-A	AUST & NZ BK-ADR	HILL-ROM HOLDING	ANALOG DEVICES	0.899530659	0.553802957	1.453333616	4.01E-07	1.85E-17	good
replacing_assets	AUST & NZ BK-ADR	US FOODS HOLDING	HILL-ROM HOLDING	AXA -ADR	0.89909522	0.749125293	1.648220513	1.46E-06	1.82E-17	good
replacing_assets	PHILLIPS 66	TELEFLEX INC	HILL-ROM HOLDING	UNIVERSAL DISPLA	0.898663341	0.718647658	1.617310999	1.93E-06	2.58E-17	good
replacing_assets	AUST & NZ BK-ADR	CAMPBELL SOUP CO	HILL-ROM HOLDING	UNIVERSAL DISPLA	0.898612873	0.603813796	1.502426669	1.99E-06	2.37E-17	good
replacing_assets	TIM SA-ADR	AUST & NZ BK-ADR	US FOODS HOLDING	FREEPORT-MCMORAN	0.8985931	0.636023853	1.534616953	-8.23E-07	2.52E-17	good
replacing_assets	SUN LIFE FINANCI	AGCO CORP	AUST & NZ BK-ADR	HILL-ROM HOLDING	0.898554967	0.802798731	1.701353698	6.81E-07	1.97E-17	good
replacing_assets	AUST & NZ BK-ADR	HILL-ROM HOLDING	UNIVERSAL DISPLA	KERING-UNSP ADR	0.898541466	0.646992314	1.54553378	1.47E-06	2.08E-17	good

Figure A.1: *CSV* of algorithm solutions of experiment 2.V1

LHS	RHS1	RHS2	RHS3	RHS4	Correlation	betas	beta_correlation	tracking_error_degree1_portfolio	tracking_error_degree2_portfolio	outperform random
replacing_assets	AUST & NZ BK-ADR	HILL-ROM HOLDING	UNIVERSAL DISPLA	KERING-UNSP ADR	0.898541466	0.646992314	1.54553378	1.47E-06	2.08E-17	good
replacing_assets	MOTOROLA SOLUTIO	HILL-ROM HOLDING	UNIVERSAL DISPLA	VIACOMCBS INC-B	0.898444304	0.546893947	1.445338251	9.54E-07	1.01E-15	bad
replacing_assets	PHILLIPS 66	PPL CORP	AUST & NZ BK-ADR	HILL-ROM HOLDING	0.898165157	0.845846372	1.744011529	1.30E-06	1.65E-17	good
replacing_assets	TIM SA-ADR	MOTOROLA SOLUTIO	AUST & NZ BK-ADR	HILL-ROM HOLDING	0.898088409	0.704011614	1.602100023	-1.61E-07	2.18E-17	good
replacing_assets	PPL CORP	AUST & NZ BK-ADR	HILL-ROM HOLDING	VIACOMCBS INC-B	0.8980723	0.662569691	1.560641991	5.95E-07	9.09E-16	bad
replacing_assets	ALLEGION PLC	AUST & NZ BK-ADR	US FOODS HOLDING	FREEPORT-MCMORAN	0.897951828	0.707172835	1.605124663	-6.39E-07	1.58E-17	good
replacing_assets	AUST & NZ BK-ADR	US FOODS HOLDING	HILL-ROM HOLDING	ALLIANT ENERGY	0.897910133	0.699017685	1.596927818	8.37E-07	2.10E-17	good
replacing_assets	MOTOROLA SOLUTIO	AUST & NZ BK-ADR	HILL-ROM HOLDING	UNIVERSAL DISPLA	0.897794522	0.635521089	1.533315611	1.63E-06	2.16E-17	good
replacing_assets	LITHIA MOTORS-A	ANALOG DEVICES	FOOT LOCKER INC	AUST & NZ BK-ADR	0.897625108	0.551518963	1.449144071	9.84E-08	1.66E-17	good
replacing_assets	PHILLIPS 66	FOOT LOCKER INC	HILL-ROM HOLDING	RALPH LAUREN COR	0.897245774	0.708418202	1.605663976	1.11E-06	2.31E-17	good
replacing_assets	SANOFI-ADR	SUN LIFE FINANCI	AUST & NZ BK-ADR	HILL-ROM HOLDING	0.897189832	0.820212657	1.717402489	9.59E-07	2.12E-17	good
replacing_assets	ROYAL DUTCH-ADR.1	LITHIA MOTORS-A	AUST & NZ BK-ADR	HILL-ROM HOLDING	0.896944051	0.659981688	1.556925739	5.87E-07	1.75E-17	good
replacing_assets	AIA GROUP LT-ADR	HILL-ROM HOLDING	LITHIA MOTORS-A	AUST & NZ BK-ADR	0.896906488	0.677617914	1.574524402	7.03E-07	1.71E-17	good
replacing_assets	SANOFI-ADR	AUST & NZ BK-ADR	HILL-ROM HOLDING	ANALOG DEVICES	0.896854275	0.608566123	1.505420398	-1.02E-07	2.09E-17	good
replacing_assets	SUN LIFE FINANCI	AUST & NZ BK-ADR	HILL-ROM HOLDING	TEXAS PACIFIC LA	0.896840284	0.819544285	1.716384569	2.69E-06	2.29E-17	good
replacing_assets	SUN LIFE FINANCI	AUST & NZ BK-ADR	HILL-ROM HOLDING	WM MORRISON-ADR	0.896769156	0.8029131	1.699682256	9.07E-07	1.99E-17	good
replacing_assets	PHILLIPS 66	UNIVERSAL DISPLA	AUST & NZ BK-ADR	HILL-ROM HOLDING	0.896597438	0.62523678	1.521834218	1.34E-06	2.43E-17	good
replacing_assets	EXPEDITORS INTL	HILL-ROM HOLDING	SANOFI-ADR	AUST & NZ BK-ADR	0.896538203	0.75142606	1.647964263	3.45E-07	2.07E-17	good
replacing_assets	SUN LIFE FINANCI	AUST & NZ BK-ADR	HILL-ROM HOLDING	ACCELERON PHARMA	0.896517364	0.737359222	1.633876586	6.26E-07	2.25E-17	good
replacing_assets	OGE ENERGY CORP	LITHIA MOTORS-A	MOTOROLA SOLUTIO	HILL-ROM HOLDING	0.896473195	0.773511339	1.669984534	1.31E-06	1.60E-17	good
replacing_assets	PHILLIPS 66	HILL-ROM HOLDING	LITHIA MOTORS-A	AUST & NZ BK-ADR	0.896426619	0.718779479	1.615206098	1.15E-06	1.52E-17	good
replacing_assets	EXPEDITORS INTL	HILL-ROM HOLDING	PPL CORP	AUST & NZ BK-ADR	0.89638537	0.778624069	1.675009439	9.99E-07	2.03E-17	good
replacing_assets	EPAM SYSTEMS INC	LITHIA MOTORS-A	AUST & NZ BK-ADR	HILL-ROM HOLDING	0.896264937	0.514224694	1.410489631	7.68E-07	3.07E-17	bad
replacing_assets	CAMPBELL SOUP CO	VIRTU FINANCIA-A	UNIVERSAL DISPLA	HONDA MOTOR-ADR	0.896262101	0.685652554	1.581914655	4.91E-06	2.76E-17	good
replacing_assets	TIM SA-ADR	MOTOROLA SOLUTIO	HILL-ROM HOLDING	SCIENCE APPLICAT	0.896221538	0.576620169	1.472841707	1.30E-06	2.83E-17	good
replacing_assets	OGE ENERGY CORP	MOTOROLA SOLUTIO	HILL-ROM HOLDING	US FOODS HOLDING	0.896080037	0.832461928	1.728541965	9.58E-07	1.83E-17	good
replacing_assets	ALGONQUIN POWER	LITHIA MOTORS-A	AUST & NZ BK-ADR	HILL-ROM HOLDING	0.895973635	0.607834202	1.503807837	7.03E-07	1.89E-17	good
replacing_assets	AUST & NZ BK-ADR	HILL-ROM HOLDING	ANALOG DEVICES	AXA -ADR	0.89587972	0.57382962	1.46970934	1.31E-07	2.12E-17	good
replacing_assets	AUST & NZ BK-ADR	WM MORRISON-ADR	HILL-ROM HOLDING	VIACOMCBS INC-B	0.895731554	0.6436084	1.539339954	-1.12E-07	9.47E-16	bad
replacing_assets	SUN LIFE FINANCI	AUST & NZ BK-ADR	HILL-ROM HOLDING	SCOTTS MIRACLE	0.895654568	0.78917116	1.684825728	3.00E-06	2.33E-17	good
replacing_assets	AUST & NZ BK-ADR	ANALOG DEVICES	US FOODS HOLDING	WM MORRISON-ADR	0.895439024	0.65381279	1.549251814	2.69E-07	1.72E-17	good
replacing_assets	STEEL DYNAMICS	US FOODS HOLDING	AUST & NZ BK-ADR	HILL-ROM HOLDING	0.895422828	0.723035945	1.618458773	8.33E-07	2.49E-17	good
replacing_assets	MOTOROLA SOLUTIO	HILL-ROM HOLDING	SCIENCE APPLICAT	VIACOMCBS INC-B	0.89534763	0.51984664	1.41519427	1.28E-06	9.52E-16	bad
replacing_assets	PHILLIPS 66	ANALOG DEVICES	LITHIA MOTORS-A	AUST & NZ BK-ADR	0.895209408	0.598828403	1.494037811	2.46E-07	1.57E-17	good
replacing_assets	TIM SA-ADR	LITHIA MOTORS-A	AUST & NZ BK-ADR	SANTEN PHARM-ADR	0.895174976	0.710906659	1.606081635	4.38E-07	1.94E-17	good
replacing_assets	EXPEDITORS INTL	FOOT LOCKER INC	HILL-ROM HOLDING	RALPH LAUREN COR	0.895124524	0.656932446	1.55205697	8.11E-07	2.78E-17	good
replacing_assets	TIM SA-ADR	LITHIA MOTORS-A	AUST & NZ BK-ADR	FREEPORT-MCMORAN	0.894989086	0.594573439	1.489562525	-4.73E-07	2.12E-17	good
replacing_assets	PEOPLE'S UNITED	VIACOMCBS INC-B	AUST & NZ BK-ADR	HILL-ROM HOLDING	0.894893427	0.563465871	1.458359298	-8.45E-09	9.73E-16	bad
replacing_assets	PEOPLE'S UNITED	HILL-ROM HOLDING	EXPEDITORS INTL	AUST & NZ BK-ADR	0.894773054	0.647447115	1.542220169	3.96E-07	2.60E-17	good
replacing_assets	SUN LIFE FINANCI	AUST & NZ BK-ADR	HILL-ROM HOLDING	KERING-UNSP ADR	0.894369246	0.855232371	1.749601617	6.67E-07	1.84E-17	good
replacing_assets	PHILLIPS 66	MOTOROLA SOLUTIO	HILL-ROM HOLDING	UNIVERSAL DISPLA	0.894333404	0.675484231	1.569817635	1.66E-06	2.35E-17	good
replacing_assets	SANOFI-ADR	AUST & NZ BK-ADR	HILL-ROM HOLDING	BIOGEN INC	0.894279265	0.655211754	1.549491019	2.30E-06	1.93E-17	good

Figure A.2: *CSV* of algorithm solutions of experiment 2.V2