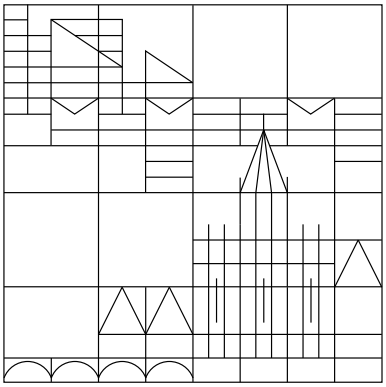
Universität Konstanz



Master's Thesis

Widened Learning of Portfolio Selection for Index Tracking

A thesis submitted in partial fulfillment of the requirements for the degree of *Master of Social and Economic Data Analysis* at the Department of Economics of the University of Konstanz

by

Iuliia Gavriushina

August 27, 2018

1st assessor: PD Dr.-Ing. habil. Christian Borgelt
2nd assessor: Prof. Dr. Winfried Pohlmeier

Universität Konstanz
D–78457 Konstanz
Germany

Abstract

This master's thesis considers index tracking from the perspective of solution space exploration. Several search space heuristics are used in combination with different portfolio optimization models in order to select a tracking portfolio with returns that mimic a benchmark index. Even with the fastest hardware and the most massively parallel systems available today, it is infeasible to conduct an exhaustive search for the large solution space in a reasonable time. Instead of increasing the number of parallel resources with the aim to traverse as much solution space as possible, we try to obtain the best use of every parallel resource. With this aim we introduce several portfolio diversity measures. Experimental results conducted on real-world datasets show that adding diversity to the set of parallel search paths can provide a better solution (tracking portfolio) due to exploration of disparate solution space regions. However, the choice of the diversity measure plays an important role. Poor path diversification can hinder the progress towards a better solution.

Acknowledgments

I would like to thank the entire Chair for Bioinformatics and Information Mining and Chair for Economics for their support during the completion of this thesis. In particular, special thanks to my advisers PD Dr.-Ing. habil. Christian Borgelt and Prof. Dr. Winfried Pohlmeier for their guidance, ideas, and inspiration for this work; to Oliver Sampson for his time, patience, understanding for many questions, interesting discussions, and every day guidance in the world of Computer Science; to Peter Burger for setting up and administering servers to run the experiments on; to Sebastian Bayer for his assistance in using Thomson Reuters Eikon Datastream. Furthermore, I would like to say special thanks to my parents for always believing in me and supporting me in everything that I am doing.

Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

Introduction

Money. Simply keeping it in the pocket, we will not increase our financial wealth. Saving can be a good strategy for small and short-term goals. In the long-term, inflation can significantly degrade the value of cash savings, whereas financial investing has more potential for high returns. Nowadays there are many investment opportunities, such as mutual funds, options, futures, bonds, stocks, equities, and real estate. Wise investing is the key to building wealth. The method of investing money and the investment horizon are essential when choosing an investment strategy.

1.1 Tracking an Index

The stock market is an important instrument of finance and economic development. It is a mechanism for allocating capital resources from investors to producers. The stock indices of a country are one of the indicators of the general trend in its economy. They are tools used by investors for describing the market and for characterizing direction, speed, and efficiency of the market segments.

Stock market analysis is a popular research area in data analytics, whereby the main aim is to automate the processing and analysis of data without intermediate human judgments. Stock market indices are one of the most popular investment goals. Several studies show that in the long-term many investors fail to outperform the market. A significant part of the underperformance is often related to the high fees associated with active trading [Ferri and Benke, 2013; Maringer and Oyewumi, 2007]. Any new information on the market is rapidly incorporated into stock prices. Therefore, to beat the market, an investor should obtain superior and/or faster information. *Stock index tracking* is a very popular strategy among those investors, who instead of trying to beat the market, wish to match the market's returns. The main advantage of *index investing* is lower (in comparison to *active investment* strategies) expenses due to less frequent trading. This master's thesis aims to explore the ways of selecting a *tracking portfolio*—a portfolio of assets with returns that mimic a certain investment target. To find the tracking portfolio, we have to solve two problems simultaneously: choose

assets for the portfolio and determine investing weights. There are many approaches to portfolio selection, which use different types of optimizations. The choice of the optimization model often depends on available resources.

Tracking portfolio selection can be considered to be a common solution space exploration problem in Machine Learning, where the aim is to find a model (in our case a portfolio) which accurately represents a set of observed data (an index) in order to predict its future performance. The number of possible solutions can be prohibitively large to be searched exhaustively. To limit the search space, some heuristic-based optimizations can be applied. Hill-Climbing[1] is the most straightforward greedy search heuristic, which iteratively chooses a specification for further exploration that provides the most benefit at the current step. Because a greedy heuristic does not consider the problem as a whole but makes a decision based on local optimization, it imposes a bias on the search.

Exploration of several different solution paths can improve the accuracy of the best-first search. Sequential exploration of the solution paths can significantly slow down the search process. *Parallel computing* in Machine Learning is traditionally associated with accelerating the analysis process. However, it has the other very import implications. Parallel resources can be used to improve the accuracy of data mining algorithms, i.e., to find better (more representative) patterns of a studied dataset in the same time [Akbar et al., 2012]. In other words, parallel computing can be used to provide better solution space exploration and, therefore, possibly better performance with respect to some measure. Moreover, the use of parallel resources allows significantly reducing computational time in comparison to serial processing. Beam Search is a common heuristic which iteratively picks several most promising models for further exploration. The bigger the search width k, the more likely it is to find a better solution. Similar to Hill-Climbing, Beam Search risks getting trapped in a local optimum. A locally optimal choice may in fact not lead towards the globally optimal solution. Choice of the k best solutions at each iteration does not ensure exploration of *different* regions of the search space. In contrast, it is likely that we are exploring only closely related variations of the locally best model. Therefore, instead of increasing the search width with attempts to traverse as much solution space as possible, we have to obtain the best use of every parallel resource at each stage. Inducing a *diversity measure* δ between solution paths can help to explore disparate regions of the solution space (See Figure 1.1) [Sampson, 2013].

Diversity is an important subject in bio- and chemoinformatics regarding protein and molecule dissimilarity [Meinl, 2010]. In Data Mining the effect of diversity on the parallel exploration of the solution space was studied in [Akbar et al., 2012; Fillbrunn and Berthold, 2015; Fillbrunn et al., 2017; Ivanova and Berthold, 2013; Sampson, 2013; Sampson and Berthold, 2014, 2016; Sampson et al., 2018]. Finding a diversity measure is not trivial. What is diversity in Portfolio Theory? What makes portfolios dissimilar

[1] In this thesis, we use Steepest-Ascent Hill-Climbing (or Best-First search). In contrast to Simple Hill-Climbing, which selects the first improvement of the current solution for further exploration, Steepest-Ascent Hill-Climbing examines all successors and chooses the best improvement.

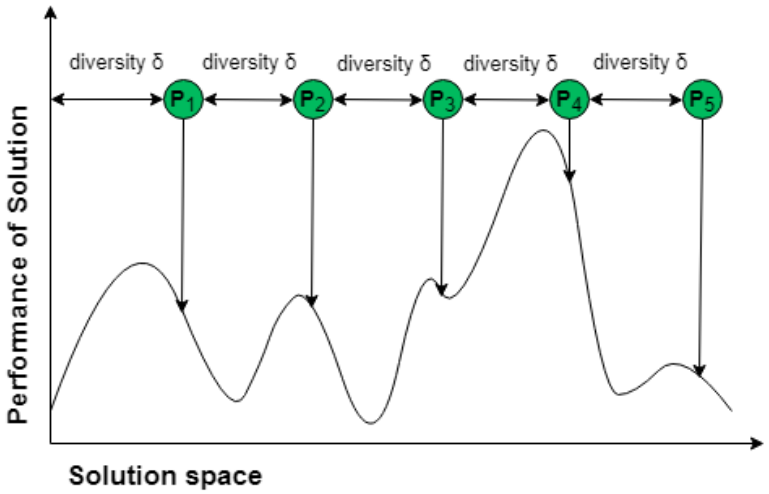


Figure 1.1: Diversity-driven solution space exploration. Inducing a diversity measure δ between solution paths (between portfolios) can help to explore disparate regions of the solution space.

in the solution space? There are many statistics which characterize a portfolio. Which metric and which distance measure can steer the solution paths away from each other, guaranteeing exploration of distinct regions and at the same time provide "the best" possible solution, e.g., the minimum divergence between an index and its tracking portfolio?

Using several datasets of different sizes, we will answer the following questions and check the hypotheses:

- Can we find the best set of assets which will replicate an index?
- Given an index with hundreds of constituents, is it possible to mimic its behavior with a few assets?
- Can we find a useful notion of diversity for selecting a tracking portfolio?
- H1: Using diverse parallel search in the solution space allows finding a potentially better solution than that obtained by a standard greedy search.
- H2: Increasing the number of parallel resources allows for broader exploration of the solution space and, therefore, is more likely to find a better solution than that obtained by a standard greedy search.

1.2 Overview

This master's thesis considers tracking portfolio selection as a solution space exploration problem. Chapter 2 provides an overview of different types of investments, portfolio

management strategies, and motivation for index replication. Additionally, it covers a comparison of different index-tracking approaches, such as physical and synthetic replication. A comparative analysis of different strategies suggests using sample physical replication for tracking portfolio construction. Chapter 3 describes traditional portfolio optimization models, which in this work will be used in combination with heuristics suggested in Chapter 4. Here we describe three solution space exploration algorithms, which help to choose a set of assets for the tracking portfolio: Hill-Climbing, Beam Search, and Widened Index Tracking. The last two utilize parallel resources to provide broader solution space exploration. Adding diversity to the set of parallel search paths helps to explore disparate regions of the solution space. In Chapter 4 we also discuss diversity measuring and distance metrics which can be used for tracking portfolio selection. Chapter 5 provides experimental results obtained for different datasets. This document focuses primarily on tracking portfolio construction (known as an investing or creation phase), i.e., on single-stage index tracking, and aims to show an advantage of using Widening—a framework for utilizing parallel recourses, which allows simultaneous exploration of different solutions [Akbar et al., 2012]. The other part of the index replication problem is portfolio rebalancing that incorporates transaction costs—the costs associated with trading (embedded after the creation phase) and requires development of a rebalancing strategy. We discuss how to modify suggested algorithms in order to apply them in the multiperiod perspective. In Chapter 6 we summarize the main results of this thesis and discuss how this work can be extended.

Chapter 2

Index-Based Investing

Investing creates wealth. Following the rule "a penny saved is a penny earned"[1] some people prefer to save money instead of taking the risks of investing. Saving, when considered as postponed consumption, can be a good strategy over a short-term period. However, if we put money in a shoebox under the bed for a long time, the interest which our "savings account" earns does not keep up with inflation. Investing makes money work for us and potentially increases our wealth. Suppose every month we put aside 400€. After 10 years we will have $400€ \times 12 \times 10 = 48,000€$. If instead this 400€ is deposited each month in an account which pays 3% annual interest, after 10 years our account balance will be about 55,896€:

$$FV = PV * \frac{\left(\left(1+\frac{r}{m}\right)^{mT}-1\right)}{\frac{r}{m}} = 400€ \times \frac{\left(\left(1+\frac{0.03}{12}\right)^{12\times 10}-1\right)}{\frac{0.03}{12}} = 55,896.57€ \tag{2.1}$$

where FV is the future value of return, r is the annual interest rate, T is the number of years, and m is the number of periods based on compounding frequency [Ross et al., 2008].

Financial investing can increase our wealth in different ways.

Definition 2.1. ***Capital gain*** *is a profit which an investor can receive from a sale of a security for a price higher than its purchase price* [Sullivan and Sheffrin, 2003].

In other words, the capital gain is the positive difference in the price of a security due to price appreciation. For example, if an investor buys a share of stock for 100€, and a year later it is worth 115€, then the stock has appreciated 15€.

Definition 2.2. ***Dividends*** *are payments that are distributed among the shareholders based on the company's earnings* [Sullivan and Sheffrin, 2003].

[1] This quote is often attributed to Benjamin Franklin, one of the most famous American polymath.

Dividends are usually paid in the form of cash payments. However, they can be issued as shares of stock or other property. Assuming a dividend of 4€ per share, an investor who owns 100 shares would receive $100 \times 4€ = 400€$ in dividend income.

Definition 2.3. ***Interest*** *is a fee an institution pays to an investor as compensation for loaning the investor's principal (through the purchase of a bond or certificate of deposit)* [Sullivan and Sheffrin, 2003].

For example, having 10,000€ in a government savings bond which pays 2% interest annually will bring 200€ a year.

Because interest rates offered by banks are usually less than the inflation rate, many people prefer to invest in mutual funds, bonds, stocks, or real estate. One of the most popular investment choices are *equity indices* (or *stock market indices*).

Definition 2.4. ***Equity indices*** *are measurements of an equity market (or a stock market), which characterizes direction, speed, and efficiency of the market as a whole or its segments* [Bosworth et al., 1975].

There are many indices which are designed to track a stock market in general. They consist of a broad basket of assets. Therefore, *stock index investing* may be an effective method of diversification for those investors who are willing to limit the risks. For example, one of the most famous equity indices is the *Standard and Poor's 500 Stock Index* (S&P 500), which contains 500 of the largest companies in the US. It is also one of the leading benchmarks for the market. The year 2013 would have been be very successful for those whose portfolio tracked the performance of the S&P 500—the index had its largest annual jump since the 1990s (a 29.60% rate of return).[2] It means investors could significantly increase their wealth within only one year. However, investing always has certain risks. When one invests in a portfolio which closely follows a stock market index, and the market (as represented by the index) falls, the portfolio will inevitably bring a loss. In 2007 the subprime mortgage crisis spread to the US financial sector, resulting in unusual market volatility and bringing with it one of the greatest year-to-date loss for S&P 500 in 2008 (a -38.49% rate of return).[1]

It is important to have an understanding of our investments. The money one invests in securities, mutual funds, and other similar investments is not typically federally insured. As a rule, investments that pay a high rate of return are subject to the higher risk and volatility. The way of investing money and the investment horizon are fundamental. The more time an investor has, the more risk the investor could take. The more risk one is able to take, the more potential for making more money. Hence, it is essential to choose the right investment strategy.

Definition 2.5. *The art of selecting an investment policy is called* ***portfolio management*** [Reilly and Brown, 2011].

[2] `https://ycharts.com/indicators/sandp_500_total_return_annual`

2.1 Portfolio Management Strategies

There are two main approaches to financial investing: *active* and *passive*.

Definition 2.6. ***Active investing*** *is a portfolio management strategy according to which an investor attempts to outperform an investment benchmark index or, in other words, to "beat the market"* [Jeurissen and van den Berg, 2005].

Proponents of this strategy believe that pricing inefficiencies in the market create investing opportunities. There are various techniques to construct an actively managed portfolio. Because active investing doesn't imply following a specific index, investors may pick for the portfolio only those assets which they believe have good growth prospects. Timing the ups and downs of the market can help to make the right decision about buying or selling of assets. Risk management is one of the benefits of active investing. When risks become too high, an investor can sell undesirable stocks. Moreover, an investor may apply different techniques to hedge against losses [Beasley et al., 2003].

Even though active investing is a very flexible investment strategy, it comes with high transaction costs caused by active trading.

Definition 2.7. ***Transaction costs*** *are expenses incurred during the process of buying and selling securities (including commissions, fees, and taxes)* [Beasley et al., 2003].

While active investing can show positive gains over a short-term period, several studies show that in a long-term comparison, many actively managed investment funds underperform market indices, and a significant part of the underperformance is often related to the high fees [Maringer and Oyewumi, 2007]. For example, [Ferri and Benke, 2013] showed that it is difficult to outperform index funds in both the short term and the long term. According to the study conducted in 2016 by *S&P Dow Jones Indices*, about 90% of active stock managers were not able to beat their index targets over the previous one-year, 5-year and 10-year periods [Soe and Poirier, 2016].

The second approach to investing is a *passive portfolio strategy* (or *index investing*).

Definition 2.8. ***Index investing*** *is a portfolio management strategy which implies holding a portfolio of securities without attempting to outperform an investment benchmark index* [Jeurissen and van den Berg, 2005].

In this case, an investor attempts to approximate the market's returns. This approach is based on the *efficient-market hypothesis*, which states that the current stock prices fully reflect all information available to the market [Malkiel and Fama, 1970]. It is difficult to tell in advance which stocks will outperform the market. As new information becomes available, market prices adjust in response reflecting a security's true value. Proponents of passive portfolio management believe that it is difficult if not impossible to beat the market in the long-term and to gain an advantage over any other investor [Maringer and Oyewumi, 2007].

The main advantage of passive investing is lower (in comparison to active investing) expenses due to less frequent trading. Active investors spend a lot of time and money to conduct exhaustive research and analysis of market trends, the economy, and companies in order to obtain timely information, gather unique insights and make a valuable investment decision. As a rule, even experienced investors seek assistance from fund managers, who do this job for an extra fee. However, it turns out, beating the market is not easy. Millions of people invest in the same relatively small set of opportunities, and to outperform the market, one should have the edge over most of them and obtain superior and/or faster information. Therefore, index investing, which merely seeks to achieve market returns, can be more effective than most active management strategies.

2.2 Index Construction

An index is a basket of securities which represents a whole market or its submarket. There are different types of indices: fixed-income indices, which are composed of government or corporate bonds; commodity indices, which represent an investment in commodities such as gold, oil, wheat; real estate indices and so on. The most widely tracked indices are equity indices (See Definition 2.4).

Indices are constructed according to the predefined rules, designed by *index providers*. The most famous and largest index providers are S&P Dow Jones, MSCI, FTSE Russell, and STOXX.[3] They define which securities are included in the basket. Depending on the goal, the decision can be based on a market capitalization size, industry (e.g., energy, finance), geographic region or value and growth investment style. In the case of stock indices, the rules are particularly often based on the two properties: the stock exchange on which a company is listed and the size of the company. For example, the DAX is a stock market index which represents 30 of the largest German companies listed on the *Frankfurt Stock Exchange*.

In addition to securities selection, an index provider determines the proportions (weights) of the individual companies in the index. There are different ways of *index weighting*:

- price weighting,
- market capitalization weighting,
- equal weighting,
- fundamental weighting.

[3] *S&P Dow Jones* is an American index provider established in 2012 as a result of the merge of Standard & Poors and Dow Jones Indices; *MSCI* is an American index provider founded in 1968 by the investment bank Morgan Stanley and Capital Group International; *FTSE Russell* is an index provider owned by the London Stock Exchange unifying the two biggest index providers the Financial Times Stock Exchange and Russell; *STOXX* is an index provider which belongs to Deutsche Börse Group [Miziołek, 2018].

The oldest and most straightforward weighting approach is *price-weighting*.

Definition 2.9. ***Price-weighted indices*** *are indices whose constituents are weighted in proportion to their price per share* [Karlow, 2013].

For example, the *Dow Jones Industrial Average* (DJIA) is a price-weighted stock market index that represents 30 large publicly owned companies based in the US. The main disadvantage of the price-weighting schema is that a company's stock price does not represent its market value. It means a large company with a low stock price influences the index less than a small company with a high stock price.

Most common indices are *market capitalization-weighted*.

Definition 2.10. ***(Market-)capitalization-weighted indices*** *are indices whose components are weighted according to their total market value size* [Karlow, 2013].

According to this approach, the largest companies receive greater weight in the index. For example, the *Swiss Market Index* SMI, which represents 20 of the largest and most liquid large- and mid-cap stocks in Switzerland, is weighted by market capitalization. The main drawback of this weighting schema is that it causes a shift towards overpriced stocks.

Definition 2.11. ***Equal-weighted indices*** *are indices whose shares have the same weight* [Karlow, 2013].

Equal weighting is the most straightforward approach, which implies that small companies influence the index the same as large companies.

Definition 2.12. ***Fundamentally-weighted indices*** *are indices whose components are chosen based on a fundamental criterion* [Karlow, 2013].

A *fundamentally-weighted* approach has appeared recently. This method assumes that there are measures which better estimate a company's intrinsic value than market capitalization. The constituents are weighted based on the specific descriptors such as dividends, sales, and cash flow. In this way, overpriced stocks have less influence on the index value.

The weights of the securities in indices change over time. For example, in the case of market capitalization weighting, if the share price of a company increases, its weight in the index also rises. Companies can leave the index and be replaced by new ones. Because indices are not static, index providers make a periodic *index rebalancing*.

Definition 2.13. ***Index rebalancing*** *is a regular check of the index components, which takes place at fixed time intervals determined by the index provider in the index rules* [Karlow, 2013].

For example, the EURO STOXX 50, a stock index representing the largest and most liquid stocks of the *Eurozone*, is reviewed annually. Furthermore, the index provider can place additional restrictions on the weights of index constituents, which can limit the influence of individual companies on the index. For example, to ensure tradability of the DAX index, the weight of an individual share is capped at 10%. The knowledge and understanding of the index construction is crucial when applying an index replication strategy.

2.3 Index-Tracking Approaches

Because it is not possible to buy an index itself directly, to "*invest in the index*," one needs to approximate its performance. One way is to construct and maintain a *tracking portfolio*.

Definition 2.14. ***Tracking portfolio*** is a portfolio of assets with returns that mimic a benchmark index [Ruiz-Torrubiano and Suárez, 2009].

One can invest in an index fund, which is created to replicate the index performance. In this case, the fund manager has to solve the tracking problem for an extra fee [Karlow, 2013].

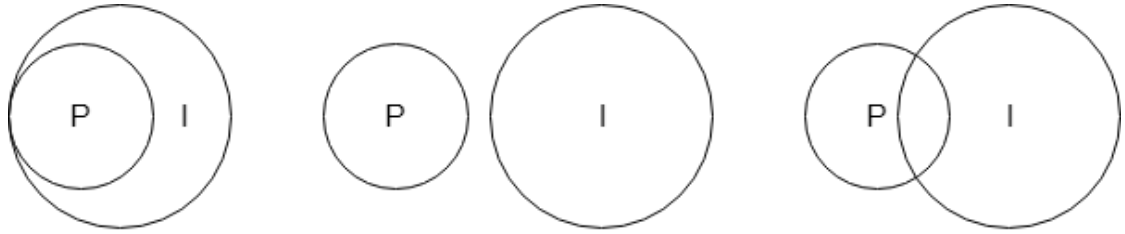
There are two approaches to track an index: *physical* and *synthetic* replication. First implies direct investment in the assets and aims to mimic the performance of the target index by holding all (*full replication*) or a representative sample (*sample replication*) of the underlying securities that make up the index. Instead of physically holding the equities in the constituents of the benchmark, *synthetic replication* relies on derivatives which are linked to those equities.

2.3.1 Physical Index Replication

Physical replication aims to mimic the index performance by trading the underlying shares. There are two ways of direct index replication. The first method is known as *full replication*.

Definition 2.15. ***Full replication*** *is a type of physical index replication which implies purchasing all of the index constituents* [Jeurissen and van den Berg, 2005].

Consider the NASDAQ 100, a stock market index which represents 100 of the largest domestic and international non-financial companies listed on the *Nasdaq Stock Market* (an American stock exchange). According to the full replication approach, an investor should buy all 100 stocks in proportion to their weights in the index. Full replication is the most transparent and easy to understand way for constructing a tracking portfolio which ensures close index replication. However, it is one of the most expensive methods. If any changes are made to the index, for example, when it is rebalanced or reconstituted,



1) 2) 3)

Figure 2.1: Structure of the sample universe. $\boldsymbol{P}$ is a set of assets in the tracking portfolio, $\boldsymbol{I}$ is a set of index constituents. 1) The tracking portfolio $\boldsymbol{P}$ is a proper subset of the index $\boldsymbol{I}$. 2) The tracking portfolio $\boldsymbol{P}$ and the index $\boldsymbol{I}$ are disjoint subsets. 3) The tracking portfolio $\boldsymbol{P}$ contains some elements of the index $\boldsymbol{I}$.

a tracking portfolio has to adjust its holdings accordingly. *Portfolio rebalancing* is not free of charge and could involve trading fees, taxes on capital gains and so on.

Definition 2.16. ***Portfolio rebalancing*** *is the process of adjusting the portfolio holdings with the aim of reducing the deviation between the portfolio performance and its target* [Jeurissen and van den Berg, 2005].

Accordingly, when an index includes a large number of constituents, full physical replication involves high transaction costs (See Definition 2.7). Furthermore, in practice, it is often infeasible to buy all of the index constituents because not all markets are always liquid.

When a stock index has hundreds of assets or contains illiquid constituents, an investor can use *sampling* (or *partial replication*).

Definition 2.17. ***Sampling*** *is a type of physical index replication which involves investing in only a subset of the assets* [Jeurissen and van den Berg, 2005].

When applying sample replication, it is important to choose the structure of the sampling universe (See Figure 2.1)—when a tracking portfolio is:

1) a proper subset of the benchmark,

2) not a subset of the benchmark,

3) not a proper subset of the benchmark but contains some elements of the benchmark.

The choice of the sample universe depends on the structure of the index to be tracked. A typical choice for indices with transparent and investable structures is investing in only a subset of the benchmark's assets. Adding assets which are not constituents of the index can extend the flexibility of the replication at the expense of transparency. For some

exotic indices, which do not have investable structures, choosing arbitrary assets can be a good solution to the problem [Karlow, 2013].

The advantage of the sample replication strategy compared to full replication is that the trading costs can be significantly reduced, especially for indices with many securities. However, one should take into account a risk that in this case portfolio performance can deviate from the index over time, i.e., have higher *tracking error*.

2.3.2 Synthetic Index Replication

Instead of direct investment in the underlying members of the index, one can use *synthetic replication*.

Definition 2.18. ***Synthetic replication*** *is an indirect index replication strategy which suggests using a derivative based on the target index, such as swaps, futures, and options* [Karlow, 2013].

This method involves entering into contractual agreements with a counterparty (usually a financial institution) which is obliged to pay the return on the benchmark in exchange for a fee. In a simple example case, if the EURO STOXX 50 gains 10% over a year, then the counterparty (e.g., a bank) concerned to pay off that amount according to the derivative contract.

The main advantage of the synthetic replication method is that there are no issues with buying and selling assets whenever the index is rebalanced. There is only one trade that has to be made—the derivative contract. On the other hand, this approach is not entirely transparent because it hides all the details of the tracking process and can mislead an investor. Furthermore, it involves counterparty risk.

Definition 2.19. ***Counterparty risk*** *is the risk that a counterparty will not meet its obligations, for example, due to its bankruptcy* [Karlow, 2013].

Therefore, it is important that the financial institution selling a derivative contract has solid profitability and a robust balance sheet. Because holding the underlying securities of the benchmark is no longer a prerequisite, synthetic replication makes possible for investors to track those markets which are difficult to access due to trading restrictions.

2.4 Summary

Index investing has become a popular investment approach that can outperform active portfolio strategies. Infrequent trading allows for avoiding high (in comparison to active investing) transaction costs, making it attractive as a long-term perspective for many investors.

There are different approaches to index tracking. Physical replication implies investing directly in assets, while synthetic replication uses derivative contracts. The choice

of a strategy depends on many factors, such as index structure, tracking error, transaction costs, and involved risks. The full physical replication approach is applied if the index has an investable and transparent structure, i.e., securities of the index and their quantities are known. However, in this case a portfolio should precisely follow its benchmark. On the one hand, it implies relatively low tracking error, but, on the other hand, it means that an investor should keep all of the index constituents irrespective of the market situation. Full physical replication is the most expensive method of index tracking. Index providers make a periodic rebalancing of their indices. Hence, the tracking portfolio must also be rebalanced. This process involves transaction costs whose size depends on the frequency of rebalancing, a number of constituents, the liquidity of assets and taxation system. Sampling and synthetic replication allow for reducing transaction costs. Additionally, in comparison to full replication, they are very flexible methods that can be applied to any index. The sample and synthetic replication could be used when the index does not have an investable structure, the assets in the index are not liquid, or the number of assets is substantial. However, sample replication causes the deviation between portfolio and index returns, as in this case the tracking portfolio does not hold all of the assets from the index, whereas synthetic replication, which involves additional counterparty risk, is not as transparent.

Applying an index replication approach, it is crucial that the reward is high enough to mitigate the risks undertaken. Each method of replication has its strengths and weaknesses. The full replication approach is the most transparent but an expensive and inflexible method. Synthetic replication is a very flexible but opaque approach. Sampling is right in between and characterized by being both sufficiently transparent and sufficiently flexible. Additionally, it is a good way to reduce expenses. This in total makes sample replication to be a good alternative when constructing a tracking portfolio.

Chapter 3

Index Replication with Sampling

Sample replication is a very attractive investment strategy, because it may significantly reduce expenses associated with management of the tracking portfolio. However, the way sample replication is implemented is crucial: the more aggressively portfolio optimization is performed, the more undesired variation between the portfolio and index could appear over time. When applying sampling to construct the tracking portfolio, one should solve two problems: define which assets will compose the portfolio (i.e., *asset selection*) and their proportions within it (i.e., *asset weighting*).

There are several approaches to *asset selection*. The first uses a certain *selection criterion*. One can analyze each asset's contribution to the index. For example, to construct a portfolio which tracks the price-weighted index (See Definition 2.9), the assets with low prices could be removed from the sample universe, as they have the smallest influence on the index. Another possible criterion is based on a certain statistic. For example, one may choose those assets for the tracking portfolio which are highly correlated with the index. The second asset selection approach aims to *mimic the structure* of the index. One may want to get rid of duplicated information. For example, if there are strongly correlated assets, they will influence the tracking portfolio in the same direction, i.e., if the value of one stock decreases, the value of the correlated asset will decrease as well. Thus, one of the two assets can be omitted. This approach allows for the creation of a diversified portfolio, which can help to achieve more consistent returns over time and reduce overall investment risk. One could use stratified sampling or factor analysis. The third approach to asset selection is *optimization-based*, where the tracking portfolio is a solution to an optimization problem. To select assets for the portfolio, one can use different tracking quality measures. For example, we can calculate the tracking error for all possible combinations of assets and select a portfolio with the smallest tracking error. However, depending on the size of the sample universe, the number of possible combinations can be enormous [Karlow, 2013].

The weighting approaches can be divided into *heuristic weighting* and *optimized weighting*. *Heuristic weighting* can be applied if the structure of the index is known. For example, if the index is price-weighted, assets in the portfolio can be weighted ac-

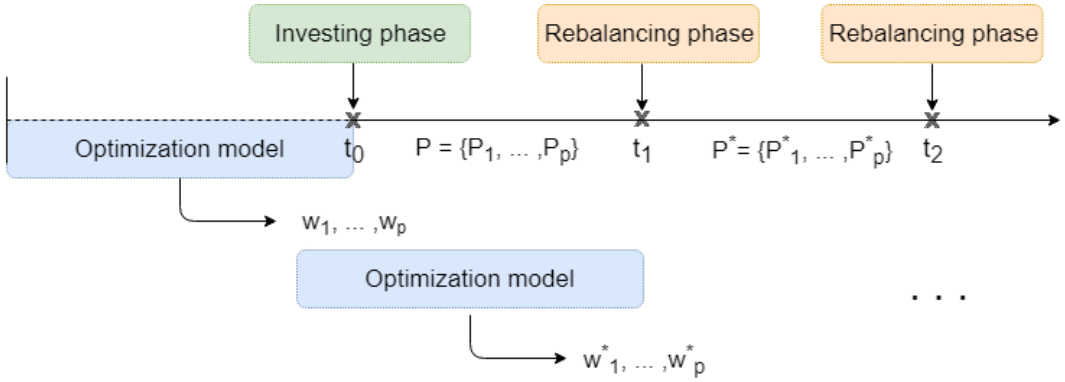


Figure 3.1: Index replication process. Applying an optimization model we find investing weights and create a tracking portfolio for the period $[t_0, t_1]$. When the tracking portfolio should be rebalanced (at t_1), we apply the optimization model to find new investing weights and form a new tracking portfolio for the next period.

cording to their price. The advantage of this method is its simplicity. However, to compensate the use of a smaller number of assets, the weights of the assets in the portfolio should be increased proportionally to the weights in the index. *Optimized weighting* is the type of approach when the weights are determined by solving an optimization problem [Karlow, 2013]. Mathematical modeling of the portfolio optimization problem attempts to address the issue considering different quality measure parameters.

There are various approaches to portfolio selection, which use different kinds of mathematical optimization models. There is no uniquely defined model. Choice of the model often depends on a purpose and capability to solve it using available resources. The models that are used for the index-tracking problem are often one-period (in [Di Tollo and Maringer, 2009] referred to as *cash endowed*). In order to apply a multiperiod perspective, the optimization problem is often solved sequentially for each period. The advantage of a multiperiod model that it can consider rebalancing strategies and include transaction costs. The main disadvantage is computational complexity [Karlow, 2013]. This master's thesis aims to find an appropriate strategy for the *investing phase* (in [Beasley et al., 2003] referred to as *creation phase*), where finding a minimum number of assets for the tracking portfolio is the primary goal (See Figure 3.1). Finding the right rebalancing strategy is another type of the problem and is partly explored.

Consider a market index $\boldsymbol{I}$ with $|\boldsymbol{I}| = n$ components and each component stock being I_i. Denote the return of an asset I_i with $i \in \{1, \ldots, n\}$ during a trading period $t \in \{1, \ldots, T\}$ by $r_t(I_i)$. Then the return of an asset I_i during a studied time interval is $\boldsymbol{r}(I_i) = [r_1(I_i), \ldots, r_T(I_i)]'$, where 'prime' denotes the transposition of the vector. According to the sample replication strategy, a tracking portfolio $\boldsymbol{P} \subseteq \boldsymbol{I}$, where $|\boldsymbol{P}| = p$ assets. Let P_j with $j \in \{1, \ldots, p\}$ be a component stock of $\boldsymbol{P}$ with $\boldsymbol{r}(P_j) = [r_1(P_j), \ldots, r_T(P_j)]'$. A portfolio return is $\boldsymbol{r}(\boldsymbol{P}) = \sum_{j=1}^{p} \omega_j \boldsymbol{r}(P_j)$, where $\boldsymbol{\omega}$=$[\omega_1, \ldots, \omega_p]'$ is the $p \times 1$ vector of the portfolio weights reflecting the investment decision, and $\sum_{j=1}^{p} \omega_j = 1$.

3.1 Markowitz Modern Portfolio Theory

The fundamental breakthrough towards managing financial investments was provided by Harry Markowitz who suggested solving a portfolio optimization problem using the mean-variance analysis [Dangi, 2013]. When constructing the optimal portfolio, [Markowitz, 1952] suggested taking into account both expected return and underlying risk. His model based on the Modern Portfolio Theory (MPT), an investment theory, which supports the concept of diversification and trade-offs between risk and return. It assumes that an investor is rational and risk-averse, meaning that between two portfolios with the same return an investor will prefer the less risky one. Risk is an inherent part of the higher reward, and, therefore, an investor will take on increased risk only if compensated by potentially higher expected returns. Apart from that, MPT assumes that financial markets are efficient, and there are no transaction costs. The Mean-Variance portfolio optimization model (MV) considers a single period of investment and aims to obtain an *efficient portfolio*—a portfolio which provides the highest level of expected return for a given level of risk or the lowest portfolio risk for given target expected return.

According to Markowitz, $E(\boldsymbol{r}(\boldsymbol{P})) = \sum_{j=1}^{p} \omega_j E(\boldsymbol{r}(P_j)) = \sum_{j=1}^{p} \omega_j \mu_j = \boldsymbol{\mu}^{'}\boldsymbol{\omega}$ is the expected return on the portfolio, $Var(\boldsymbol{r}(\boldsymbol{P})) = \sum_{j=1}^{p}\sum_{i=1}^{p} \omega_j\omega_i\sigma_{ji} = \boldsymbol{\omega}^{'}\boldsymbol{\Sigma}\boldsymbol{\omega}$ is the variance of return on the portfolio, where $\boldsymbol{\Sigma}$ is the variance-covariance matrix of assets returns and $\sigma_{ji} = cov(\boldsymbol{r}(P_j), \boldsymbol{r}(P_i))$ for $j, i \in \{1, \ldots, p\}$.

An expected return maximization problem has the following representation:

$$\begin{aligned} \max(E(\boldsymbol{r}(\boldsymbol{P}))) &= \max(\boldsymbol{\mu}^{'}\boldsymbol{\omega}) \\ \boldsymbol{\omega}'\boldsymbol{\Sigma}\boldsymbol{\omega} &= \sigma_0 \\ \sum_{j=1}^{p} \omega_j &= 1 \end{aligned} \tag{3.1}$$

where $\boldsymbol{\mu}^{'}$ is the transposed vector of $\boldsymbol{\mu} = [\mu_1, \ldots, \mu_p]^{'} = [E(\boldsymbol{r}(P_1)), \ldots, E(\boldsymbol{r}(P_p))]^{'}$, $\boldsymbol{\omega}^{'}$ is the transposed vector of $\boldsymbol{\omega}$, and σ_0 is a specified by the investor level of risk.

Often investors prefer to specify target expected returns rather than target risk levels. An equivalent representation which minimizes portfolio variance has the following representation:

$$\begin{aligned} \min(Var(\boldsymbol{r}(\boldsymbol{P}))) &= \min(\boldsymbol{\omega}^{'}\boldsymbol{\Sigma}\boldsymbol{\omega}) \\ \boldsymbol{\mu}^{'}\boldsymbol{\omega} &= r_0 \\ \sum_{j=1}^{p} \omega_j &= 1 \end{aligned} \tag{3.2}$$

where r_0 is a specified by the investor desirable expected return.

The third representation of the MV model combines the expected return and risk in

the objective function:

$$\max(\tau E(\boldsymbol{r}(\boldsymbol{P})) - Var(\boldsymbol{r}(\boldsymbol{P}))) = \max(\tau \boldsymbol{\mu}^{'} \boldsymbol{\omega} - \boldsymbol{\omega}^{'} \mathbf{V} \boldsymbol{\omega})$$
$$\sum_{j=1}^{p} \omega_j = 1 \tag{3.3}$$

where $\tau \geq 0$ is the risk tolerance coefficient—a measurement of an individual's willingness to accept the risk [Markowitz, 1987].

In order to avoid slow or infeasible convergence of the above optimization formulations, the optimizations with soft return/risk constraints are sometimes used instead.

The classical mean-variance portfolio optimization model remains one of the most important benchmarks in the literature on asset allocation and in the asset management industry. Modern portfolio theory is widely regarded as one of the major theories in financial economics. Optimization models based on some form of the mean-variance analysis are still used in many studies. In [Yu et al., 2006] a modified Markowitz model was used for a single-stage index tracking problem. [Edirisinghe, 2013] suggested an extension of the Markowitz mean-variance portfolio in order to apply it to index replication.

3.2 Risk-based Asset Allocation Strategies

A growing amount of the literature on portfolio construction approaches focuses on risks and asset diversification. Risk-based asset allocation strategies received significant attention in the marketplace especially after the global financial crisis of 2008 [Lee, 2011]. Different portfolios which are obtained using these strategies do not focus on the expected return but aim to benefit from diversification using different meanings of diversification [Braga, 2016].

Equally-Weighted Portfolio

The Equally-Weighted Portfolio (EWP) is the portfolio which aims to achieve diversification by allocating equal weights to the assets ignoring their characteristics:

$$\omega_j = \frac{1}{p} : \quad \forall j \in \{1, \ldots, p\} \tag{3.4}$$

With the larger amount of assets in the portfolio, the weights become smaller. It is one of the most straightforward portfolio construction approaches and does not require investigation of the asset returns distribution and does not use any objective function in the optimization model. It is the best choice when all the assets in the sample universe have the same return, volatility, and correlation, which is in the real financial world rare. Management of turnover and transaction costs can become critical especially in the case of illiquid securities [Lee, 2011]. Despite this, it is an interesting fact that, in reality, an equal-weighted portfolio can outperform value- and price-weighted portfolios [Plyakha et al., 2012].

Global Minimum Variance Portfolio

The Global Minimum Variance Portfolio (GMVP) is the portfolio that is expected to have the lowest possible volatility:

$$\begin{aligned} \min(Var(\boldsymbol{r}(\boldsymbol{P}))) &= \min(\boldsymbol{\omega}^{'}\boldsymbol{\Sigma}\boldsymbol{\omega}) \\ \sum_{j=1}^{p} \omega_j &= 1 \end{aligned} \tag{3.5}$$

This optimization model aims to achieve diversification by minimizing portfolio variance (the lowest risk) without considering the expected return. It tends to allocate higher weights to the low volatility assets and is very sensitive to the covariance matrix.

Most-Diversified Portfolio

The Most-Diversified Portfolio (MDP) is the portfolio that is expected to have the highest *Diversification Ratio* (DR) [Choueifaty and Coignard, 2008].

$$DR = \frac{\sum_{j=1}^{p} \omega_j \sigma_j}{\sigma_P} \tag{3.6}$$

where σ_j is the volatility of the asset P_j and σ_P is the portfolio volatility.

The optimization model aims to maximize the entire diversification of wealth across all assets in the portfolio [Dangi, 2013].

$$\begin{aligned} \max(DR) &= \max\left(\frac{\boldsymbol{\omega}^{'}\boldsymbol{\sigma}}{\sqrt{\boldsymbol{\omega}^{'}\boldsymbol{\Sigma}\boldsymbol{\omega}}}\right) \\ \sum_{j=1}^{p} \omega_j &= 1 \end{aligned} \tag{3.7}$$

where $\boldsymbol{\sigma} = [\sigma_1, \ldots, \sigma_p]^{'}$ is a $p \times 1$ vector of the portfolio assets volatilities.

Sharpe ratio is the measure of risk-adjusted return of a portfolio. The higher the Sharpe ratio is, the more return an investor gets per unit of risk. When all assets have the same Sharpe ratio then maximizing the diversification ratio is equivalent to maximizing the Sharpe ratio [Dangi, 2013]:

$$\begin{aligned} \max(SR) &= \max\left(\frac{\boldsymbol{\omega}^{'}\boldsymbol{\mu} - r_f}{\sqrt{\boldsymbol{\omega}^{'}\boldsymbol{\Sigma}\boldsymbol{\omega}}}\right) \\ \sum_{j=1}^{p} \omega_j &= 1 \end{aligned} \tag{3.8}$$

where r_f is a risk-free rate.

In [Jagannathan and Ma, 2003] it is shown that the tangency portfolio—the portfolio that maximizes the Sharpe ratio, whether constrained or not—does not perform as well as the Global Minimum Variance Portfolio in terms of the out-of-sample Sharpe ratio.

Risk Contribution Portfolio

The Risk Contribution Portfolio (RCP) portfolio generally refers to the portfolio which is constructed to achieve a predetermined profile of risk contributions by an asset [Lee, 2011; Maillard et al., 2010]. When the risk contributions of all assets in the portfolio are equalized, the strategy is known as the Risk Parity approach. According to [Lee, 2011], a Risk Parity portfolio should satisfy:

$$\omega_j \beta_j = \omega_j \beta_i = \frac{1}{n} \quad \text{for } \forall j, i \in \{1, \ldots, p\} \tag{3.9}$$

where $\beta_j = \frac{Cov(\boldsymbol{r}(P_j), \boldsymbol{r}(\boldsymbol{I}))}{\sigma_I^2}$ is the beta of an asset P_j, which measures the volatility of the asset P_j relative to the benchmark index $\boldsymbol{I}$ [Edirisinghe, 2013].

In other words, the higher the volatility or the correlation of an asset with other assets, the lower its weight in the Risk Parity portfolio. The Risk Parity approach does not have an analytical solution as portfolio weights are endogenous in determining the risk contribution of an asset in the portfolio. Finding the solution numerically becomes tricky when the number of assets in the sample universe increases and may require additional analysis [Lee, 2011].

3.3 Tracking Error Portfolio Models

The main focus of this master's thesis is index replication. The optimization methods for tracking error portfolio models involves minimizing a measure of tracking error.

Portfolio $\boldsymbol{P}$ reproduces the performance of the benchmark index $\boldsymbol{I}$ if the portfolio return $\boldsymbol{r}(\boldsymbol{P})$ follows closely the return of the index $\boldsymbol{r}(\boldsymbol{I})$ at every unit time period. The majority of measures of tracking quality are based on the *tracking error*.

Definition 3.1. ***Tracking error*** *is the measure of the deviation between the index and tracking portfolio returns* [Jeurissen and van den Berg, 2005]:

$$\boldsymbol{TE} = \boldsymbol{r}(\boldsymbol{P}) - \boldsymbol{r}(\boldsymbol{I}) = \boldsymbol{r}(\boldsymbol{P}) - \sum_{j=1}^{p} \omega_j \boldsymbol{r}(P_j) \tag{3.10}$$

There are various tracking-error-based measures, which are often used not only for the evaluation of the tracking portfolio performance but also in predictive modeling. One of them is the mean squared error (MSE):

$$MSE = \frac{1}{T} \sum_{t=1}^{T} (r_t(\boldsymbol{P}) - r_t(\boldsymbol{I}))^2 = \frac{1}{T} \sum_{t=1}^{T} (TE_t)^2 \tag{3.11}$$

Another measure is the root mean square error (RMSE), which has the same units as the estimated variable:

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{T} \sum_{t=1}^{T} (r_t(\boldsymbol{P}) - r_t(\boldsymbol{I}))^2} = \sqrt{\frac{1}{T} \sum_{t=1}^{T} (TE_t)^2} \tag{3.12}$$

A frequently used performance measure, which is more robust to outliers than MSE and RMSE, is the mean absolute error (MAE):

$$MAE = \frac{1}{T}\sum_{t=1}^{T}|r_t(\boldsymbol{P}) - r_t(\boldsymbol{I})| = \frac{1}{T}\sum_{t=1}^{T}|TE_t| \tag{3.13}$$

One of the most popular tracking quality measures is the tracking error variance (TEV) [Derigs and Nickel, 2004; Edirisinghe, 2013]:

$$\begin{aligned} TEV &= Var(\boldsymbol{r}(\boldsymbol{P}) - \boldsymbol{r}(\boldsymbol{I})) \\ &= \sigma_P^2 + \sigma_I^2 - 2Cov(\boldsymbol{r}(\boldsymbol{P}), \boldsymbol{r}(\boldsymbol{I})) \\ &= \sigma_P^2 + \sigma_I^2 - 2\sigma_I^2\beta_P, \end{aligned} \tag{3.14}$$

where $\beta_P = \frac{Cov(\boldsymbol{r}(\boldsymbol{P}), \boldsymbol{r}(\boldsymbol{I}))}{\sigma_I^2}$ is the portfolio beta, which measures portfolio volatility relative to the benchmark index or, in other words, sensitivity relative to the fluctuations of the index.

$$\begin{aligned} \beta_P &= \frac{Cov(\boldsymbol{r}(\boldsymbol{P}), \boldsymbol{r}(\boldsymbol{I}))}{\sigma_I^2} = \frac{Cov(\sum_{j=1}^{p}\omega_j\boldsymbol{r}(P_j), \boldsymbol{r}(\boldsymbol{I}))}{\sigma_I^2} \\ &= \sum_{j=1}^{p}\frac{\omega_j Cov(\boldsymbol{r}(P_j), \boldsymbol{r}(\boldsymbol{I}))}{\sigma_I^2} = \sum_{j=1}^{p}\beta_j\omega_j = \boldsymbol{\beta}'\boldsymbol{\omega}, \end{aligned} \tag{3.15}$$

where $\boldsymbol{\beta}'$ is the transposed vector of $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_p]'$ [Edirisinghe, 2013].

The general optimization model for the portfolio selection problem is often formulated as the minimization of a tracking quality measure, which is based on the $\boldsymbol{TE}$ [Karlow, 2013]:

$$\begin{aligned} &\min(\text{measure}(\boldsymbol{TE})) \\ &\sum_{j=1}^{p}\omega_j = 1 \end{aligned} \tag{3.16}$$

In the case of TEV (See Equation 3.14), the goal is to find the set $\boldsymbol{P}$ that minimizes the variance of the difference between $\boldsymbol{r}(\boldsymbol{P})$ and $\boldsymbol{r}(\boldsymbol{I})$. Since the index variance σ_I^2 is independent of portfolio positions:

$$\begin{aligned} \min(\text{measure}(\boldsymbol{TE})) &= \min(Var(\boldsymbol{r}(\boldsymbol{P}) - \boldsymbol{r}(\boldsymbol{I}))) \\ &= \min(\sigma_P^2 - 2\sigma_I^2\beta_P) \\ &= \min(\boldsymbol{\omega}'\boldsymbol{\Sigma}\boldsymbol{\omega} - 2\sigma_I^2\boldsymbol{\beta}'\boldsymbol{\omega}), \end{aligned} \tag{3.17}$$

where $\boldsymbol{\Sigma}$ is the variance-covariance matrix of assets returns, $\boldsymbol{\omega}'$ is the transposed vector of $\boldsymbol{\omega}$.

According to Equation 3.16 and Equation 3.17 the TEV optimization model has the following representation:

$$\begin{aligned} \min(TEV) &= \min(\boldsymbol{\omega}^{'}\boldsymbol{\Sigma}\boldsymbol{\omega} - 2\sigma_I^2\boldsymbol{\beta}^{'}\boldsymbol{\omega}) \\ \sum_{j=1}^{p}\omega_j &= 1 \end{aligned} \tag{3.18}$$

Based on the MV optimization (See Equation 3.2) [Edirisinghe, 2013] suggested using the above optimization problem by specifying the target expected return of the portfolio. Mean Constrained TEV optimization model (TEMV) has the following representation:

$$\begin{aligned} \min(TEV) &= \min(\boldsymbol{\omega}^{'}\boldsymbol{\Sigma}\boldsymbol{\omega} \ - \ 2\sigma_I^2\boldsymbol{\beta}^{'}\boldsymbol{\omega}) \\ \boldsymbol{\mu}^{'}\boldsymbol{\omega} &= r_0 \\ \sum_{j=1}^{p}\omega_j &= 1 \end{aligned} \tag{3.19}$$

where $\boldsymbol{\mu}^{'}$ is the transposed vector of $\boldsymbol{\mu} = [\mu_1, \ldots, \mu_p]^{'} = [E(\boldsymbol{r}(P_1)), \ldots, E(\boldsymbol{r}(P_p))]^{'}$ and r_0 is a desirable expected return specified by the investor (in [Edirisinghe, 2013] it is equal to the index mean return).

While the TEV is used very often, it has the disadvantage of including a shift: if the difference in returns is constant, then the $\boldsymbol{TE}$ is a constant and its variance is zero, although there could be a deviation from the index.

In [Ruiz-Torrubiano and Suárez, 2009] the index tracking optimization problem is formulated by minimizing MSE. Without using extra constraints:

$$\begin{aligned} \min(MSE) &= \min(\boldsymbol{\omega}^{'}\mathbf{H}\boldsymbol{\omega} - 2\boldsymbol{q}^{'}\boldsymbol{\omega}) \\ \sum_{j=1}^{p}\omega_j &= 1 \end{aligned} \tag{3.20}$$

where $\mathbf{H}$ is the matrix with $H_{ji} = \frac{1}{T}\sum_{t=1}^{T} r_t(P_j)r_t(P_i)$ and $\boldsymbol{q} = [q_1, \ldots, q_p]^{'}$ with $q_j = \frac{1}{T}\sum_{t=1}^{T} r_t(P_j)r_t(\boldsymbol{I})$ for $j, i \in \{1, \ldots, p\}$.

In the tracking optimization perspective, we wish to avoid significant divergence between the index and portfolio returns. That is why MSE can be a good tracking metric for the objective function in the optimization problem. It is used more often than MAE in the literature on index replication. MSE is characterized by faster convergence. It is continuously differentiable and, therefore, allows for gradient-based methods. To use linear programming, MAE optimization requires more complicated tools or additional modifications [Liu, 2011; Mansini et al., 2015; Rudolf et al., 1999].

There are other less popular metrics, which are found in the literature. Trying to minimize undesirable negative deviation rather than deviation in general, [Rudolf et al., 1999] used mean absolute downside deviation (MADD) and maximal absolute downside

deviation (MAXDD). The conditional value-at-risk measure (CVAR) is often used to control the risk of the tracking portfolio [Goel et al., 2018; Wang et al., 2012; Wilt et al., 2010]. The most popular tracking measures are quadratic in weights: tracking error variance (See Equation 3.18 and Equation 3.19) and mean squared error (See Equation 3.20) and, therefore, will be utilized in this document.

3.4 Portfolio Optimization Constraints

Wishing to obtain a precise solution with specific characteristics, fund managers impose a variety of constraints on the traditional formulations.

Definition 3.2. ***Budget constraint*** *is a typical constraint which implies that all the available budget must be invested* [Ruiz-Torrubiano and Suárez, 2009]:

$$\sum_{j=1}^{p} \omega_j = 1 \tag{3.21}$$

Definition 3.3. ***Long-only constraint*** *is one of the most popular constraints which implies that short selling is not allowed* [Di Tollo and Maringer, 2009]:

$$w_j \geqslant 0: \quad \forall j \in \{1, \ldots, p\} \tag{3.22}$$

In finance, a short sale is the sale of an asset that the seller does not own. Shorting stocks has pluses and minuses. On the one hand, short-selling is a way to earn money when a market's value is going to decline. It may be difficult to make money from a declining market without short selling. On the other hand, keeping a short position for a long time is often associated with extra risk and additional loss when the stock price goes up. Moreover, borrowing the shares, an investor is obliged to pay interest to a lender (broker) for this loan. The interest rate depends on the availability of the borrowed shares on the market. There are many discussions about the use of this constraint. Indeed, it is a common practice to sell assets that are not yet owned by the investor [Di Tollo and Maringer, 2009]. In addition, some studies show that such portfolios perform better. "When short sales are allowed, minimum variance portfolios and minimum tracking error portfolios constructed using the daily return sample covariance matrix performs the best" [Jagannathan and Ma, 2003].

Definition 3.4. ***Cardinality constraint*** *is a portfolio optimization constraint which restricts the number of assets in the portfolio* [Jin et al., 2016]:

$$\sum_{i=1}^{n} z_i = p: \quad \forall z_i \in \{0, 1\} \tag{3.23}$$

Often in the literature, one can see the cardinality constraint in the inequality form [Coleman and Li, 2006]:

$$k_{min} \leq \sum_{i=1}^{n} z_i \leq \ k_{max} : \ \ \forall z_i \in \{0, 1\} \tag{3.24}$$

This constraint is imposed to facilitate the portfolio management and to reduce the management costs [Di Tollo and Maringer, 2009]. Moreover, it reduces the number of the possible solutions and makes the optimization problem easier to solve. From the other side, trying to obtain a minimum number of assets for the tracking portfolio, this constraint can be too restrictive.

Definition 3.5. ***Quantity constraint*** *is a portfolio optimization constraint which restricts the allocated proportions to each asset in the portfolio* [Jin et al., 2016]:

$$L_j \leq \omega_j \leq U_j : \ \ \forall j \in \{1, \ldots, p\} \tag{3.25}$$

where L_j and U_j are lower and upper proportion bounds.

Upper bound constraints are introduced to prevent excessive exposure to a specific asset. Lower bounds are used to avoid the cost of administrating very small portions of assets [Di Tollo and Maringer, 2009]. Thus, quantity constraint (or floor and ceiling constraint) aims to diversify a portfolio and to minimize negligible holding of assets.

When trying to gain a diversified portfolio, some investors wish to restrict the proportion invested in specific sets of assets.

Definition 3.6. ***Class constraint*** *is a constraint which limits the total proportion of assets with common characteristics* [Jin et al., 2016]:

$$L_m \leq \sum_{j \in C_m} \omega_{j_m} \leq U_m : \ \ \forall m \in \{1, \ldots, M\} \tag{3.26}$$

where L_m and U_m are lower and upper proportion bounds for a class (or sector) C_m and M is the total number of classes.

Specific constraints can be imposed by law or regulations in several countries [Di Tollo and Maringer, 2009].

Definition 3.7. ***Round lot constraint*** *is a constraint which defines that the investment of any asset in the portfolio should be exact multiple units of a minimum lot* [Jin et al., 2016].

$$\omega_j = y_j \times l_j : \ \ \forall j = \{1, \ldots, p\} \tag{3.27}$$

where y_j is an integer variable and l_j is the minimum lot for the asset P_j.

Because the capital can not be always represented as exact multiple of trading lot for all the assets, this constraint might cause the budget constraint (See Equation 3.21) not to be strictly satisfied [Jin et al., 2016].

Definition 3.8. ***Turnover constraint*** *is a constraint that allows for reducing the turnover rate* [Di Tollo and Maringer, 2009].

High turnover is associated with higher transaction costs. To reduce them, some investors set a constraint which does not allow an asset weight to deviate by more than ξ:

$$|\omega_{j_1} - \omega_{j_0}| \leqslant \xi : \quad \forall j \in \{1, \ldots, p\} \tag{3.28}$$

where ω_{j_0} is a current portfolio weight for the asset P_j, ω_{j_1} is a new portfolio weight for the asset P_j, and ξ is a specified turnover deviation level.

Definition 3.9. ***Transaction costs constraint*** *is a constraint which allows for reducing the costs associated with trading* [Di Tollo and Maringer, 2009].

Transaction costs can appear due to brokerage costs, taxes, or bid-ask spreads. Because total transaction costs are associated with buying and selling assets (See Definition 2.7), they could be modeled as:

$$TC = f(\sum_{j=1}^{p} TO_j) = f(\sum_{j=1}^{p} |\omega_{j_1} - \omega_{j_0}|) \tag{3.29}$$

where ω_{j_0} is a current portfolio weight for the asset P_j, ω_{j_1} is a new portfolio weight for the asset P_j

The most straightforward approach is to consider transaction costs proportional to the total turnover:

$$TC_{proportional} = c \sum_{j=1}^{p} TO_j = c \sum_{j=1}^{p} |\omega_{j_1} - \omega_{j_0}| \tag{3.30}$$

where c—a positive coefficient for proportional transaction cost—often in the literature specified as 50 basis points [DeMiguel et al., 2009; Mei and Nogales, 2018; Olivares-Nadal and DeMiguel, 2018].

The transaction cost constraint can have the following representation:

$$TC \leq TC_{max} \tag{3.31}$$

TC_{max} can be chosen arbitrarily or as a certain fraction of the portfolio value.

Imposing additional constraints increases the complexity of the associated optimization problem and demands a lot of computational time and memory. In this document only *budget constraint* will be used. A more detailed problem with transaction costs will be explored in Chapter 5 in the context of portfolio rebalancing.

3.5 Portfolio Performance Assessment

In financial literature, there are many different portfolio performance measures. Tracking quality measures which are used in optimization problems for constructing a tracking portfolio can be used as well to report on its performance. While *TEV* is often used as a tracking quality measure, it has the disadvantage of including a possible shift. That is why this measure cannot be considered separately. *MSE* and *MAE* (See Equation 3.11 and Equation 3.13) will be used to measure how well a portfolio tracks an index.

Portfolio beta β is often used as a measure of a portfolio's risk relative to its benchmark (See Equation 3.15). $\beta = 1$ implies that a portfolio's value moves with the market. $\beta > 1$ shows that a portfolio's value outperforms the market, and $\beta < 1$—that a portfolio's value underperforms the market. For example, $\beta = 0.85$ implies that the portfolio is 15% less volatile than the index.

Excess return and *certainty equivalent* are the other important measurements which used for the portfolio performance evaluation.

Definition 3.10. ***Excess return*** *is a portfolio performance measure which shows if a portfolio underperforms/overperforms the index return over time* [Di Tollo and Maringer, 2009; Goel et al., 2018]:

$$ER = \frac{1}{T}\sum_{t=1}^{T}(r_t(\boldsymbol{P}) - r_t(\boldsymbol{I})) \tag{3.32}$$

Definition 3.11. ***Certainty equivalent*** *is a portfolio performance measure defined as a return which makes the investor indifferent between investing into the risky portfolio or receiving the certainty equivalent return* [DeMiguel et al., 2009].

$$CE(\boldsymbol{P}) = \mu_P - \frac{1}{2}\gamma\sigma_P \tag{3.33}$$

where γ is a risk aversion coefficient of the investor (the inverse of risk tolerance coefficient τ in Equation 3.3), μ_P is the portfolio mean, and σ_P is the portfolio variance.

A larger γ characterize bigger aversion to risk. Following [DeMiguel et al., 2009] we set $\gamma = 1$. This specification is often used in the financial literature. It allows us to compare different strategies for an investor who is willing to take some risks.

Certainty equivalent is a popular portfolio evaluation criterion. Given two different portfolios, an investor will prefer the one with the bigger certainty equivalent.

Definition 3.12. ***The CE difference*** *is a portfolio performance measure defined as*:

$$\Delta_{CE}(\boldsymbol{P}_1, \boldsymbol{P}_2) = CE(\boldsymbol{P}_1) - CE(\boldsymbol{P}_2) \tag{3.34}$$

Portfolio $\boldsymbol{P}_1$ is said to outperform portfolio $\boldsymbol{P}_2$ if the difference in certainty equivalents is positive [Kazak and Pohlmeier, 2017]. Similarly, we can define *the difference in certainty equivalents between the portfolio and index*:

$$\Delta_{CE}(\boldsymbol{P}, \boldsymbol{I}) = CE(\boldsymbol{P}) - CE(\boldsymbol{I}) \tag{3.35}$$

When the index $\boldsymbol{I}$ and the tracking portfolio $\boldsymbol{P}$ perform equally, this performance measure will converge to 0.

Another important portfolio measure is *turnover rate*. High turnover leads to extra trading costs, which significantly affect the portfolio return.

Definition 3.13. ***Total turnover*** *is a portfolio performance metric which allows quantifying the volume of rebalancing* [Shen and Wang, 2017]:

$$TO = \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{p}|\omega_{j_t} - \omega_{j_{t-1}}| : \quad \forall j \in \{1, \ldots, p\} \tag{3.36}$$

There are several others measures, which can assist in decision-making. For example, [Goel et al., 2018] used *Information ratio*, which shows how well a portfolio outperforms the index. *Maximum Loss*, used in the same work, focuses on negative deviations of a portfolio from the index. Similarly, *Roy's Safety-first ratio* is based on the minimization of the chances that the portfolio return will fall below the threshold (index return). It is clear that positive deviations from the index are more desirable than negative. However, this thesis focuses not on the methods which attempt to outperform the index but rather on the techniques which help to *replicate* its performance.

3.6 Potential Problems and Possible Solutions

A review of the literature on index replication shows that there is no consensus on data frequency and length of the estimation period that should be used. Even though there is a certain criticism of using daily data (e.g., daily prices contain noise and are autocorrelated), it is mostly used in studies. Monthly and weekly data contains less noise and is less autocorrelated. However, there is a limited number of points (12 points of monthly data and 52 for weekly per year). Daily data contains more information. To carry out the estimation, one needs a sufficiently large dataset. To provide a stable estimation of the weights of the tracking portfolio using monthly or weekly data, we have to obtain data from several years, which cause possible bias if we go too far back in the past [Karlow, 2013]. Therefore, for this thesis, we chose daily data.

In theory, it is possible to obtain the best portfolio for a chosen optimization model using an exhaustive search. The number of all possible portfolios, constructed from the composed index assets can be found using the following equation:

$$\sum_{p=1}^{n}\binom{n}{p} = \sum_{p=1}^{n}\frac{n!}{p!(n-p)!} \tag{3.37}$$

For example, an index with $n = 10$ assets, we need to try out 1022 combinations:

$$\begin{aligned}\sum_{p=1}^{n}\binom{n}{p} &= \sum_{p=1}^{n}\frac{n!}{p!(n-p)!} = \sum_{p=1}^{10}\binom{n}{p} = \sum_{p=1}^{10}\frac{10!}{p!(10-p)!} \\ &= \frac{10!}{1!(10-1)!} + \frac{10!}{2!(10-2)!} + \cdots + \frac{10!}{10!(10-10)!} \\ &= 10 + 45 + \cdots + 1 = 1022\end{aligned} \tag{3.38}$$

But for an index with $n = 50$ constituents we need to check $\sum_{p=1}^{50}\binom{n}{p} = \sum_{p=1}^{50}\frac{50!}{p!(50-p)!} = 1.1259 \times 10^{15}$ combinations. In this case the space of potential solutions to a problem is too large for an exhaustive search. Assuming that each (quadratic) optimization for $p \in \{1, \ldots, n\}$ takes 1 ms, then the algorithm would need more than $\frac{1.1259\times10^{15}\text{ combinations}}{31556\times10^{6}\text{ ms/year}} = 35,680$ years to find the optimal solution. Identifying the appropriate assets for the tracking portfolio is an NP-hard problem and can be approximated with the use of heuristic algorithms [Ruiz-Torrubiano and Suárez, 2009].

There are a lot of different approaches to index replication. The most popular are evolutionary-based algorithms. An *Evolutionary algorithm* (EA) is a population-based metaheuristic optimization algorithm, which uses mechanisms inspired by Darwinian evolution [Canakgoz and Beasley, 2009]. The *solution space*, i.e., the set of all possible solutions to the optimization problem, represents a population, and a candidate solution from the solution space plays the role of an individual in the population. In each generation, the features of candidates are evaluated based on a fitness function and the best individuals are selected for further reproduction. Least-fit individuals are replaced with the individuals generated through crossover and mutations [Edelkamp and Schroedl, 2011].

A *Genetic algorithm* is the most popular type of EA, where the solution to the optimization problem is encoded in the form of strings of numbers (usually binary). Genetic algorithms found their application to index tracking-problem in many studies [Oh et al., 2005; Rafaely and Bennell, 2006; Shapcott, 1992]. In [Jeurissen and van den Berg, 2008; Ruiz-Torrubiano and Suárez, 2009; Sant'Anna et al., 2017] the authors use a hybrid strategy which combines two algorithms—a genetic algorithm to define a subset of assets that will compose the portfolio and a quadratic optimization model (minimization of TE) to define weights for chosen assets. Regarding the speed of convergence, genetic algorithms are not always efficient.

Differential evolution (DE) is another type of EA, which often outperforms genetic algorithms in multiobjective optimization [Hegerty et al., 2009; Tušar and Filipič, 2007]. It is often used to solve the constrained index-tracking problem [Faizliev et al., 2016; Krink et al., 2009; Maringer and Oyewumi, 2007]. Similar to genetic algorithms this population-based strategy optimizes a problem by iteratively trying to improve a candidate solution with regards to a certain quality measure. Differential evolution uses a different type of encoding, which is based on vector differences and leads to a different mutation and crossover process.

The most straightforward local search algorithm, which was used in several studies for the index-tracking problem, is Hill-Climbing (a type of a greedy search) [Churchill and Buro, 2013; Derigs and Nickel, 2004; Faizliev et al., 2016]. Iteratively adding an asset from the sample universe to a portfolio, the Hill-Climbing algorithm attempts to find a tracking portfolio which closely follows the index. [Faizliev et al., 2016] applied this heuristic together with the cardinality constraint, which prescribes how many assets can be included in the portfolio.

There are less popular approaches applied to the tracking index problem. For example, [Focardi and Fabozzi, 2004] suggested using Euclidean distances between stock price series as a basis for *hierarchical clustering*. To form a portfolio, the authors suggest selecting a stock from each cluster. [Yu et al., 2006] used a Markowitz model for a single-stage index-tracking problem. The authors suggest using a model with downside risk constraint to reduce the probability that the portfolio return falls below the index return. [Scozzari et al., 2013] suggested the mixed-integer quadratic programming formulation for the constrained index tracking problem. [Stoyan and Kwon, 2010] defined the index tracking model in a two-stage stochastic mixed-integer programming framework.

There is no one specific algorithm which should be used to obtain the tracking portfolio. Some authors combine different approaches; others impose different types of constraints to limit the search space and be able to solve a problem with available computational resources.

3.7 Summary

There are many approaches to portfolio selection, which use different types of optimization problems. The choice of the optimization model often depends on available resources. Mathematical modeling of the portfolio optimization problem attempts to address the issue considering different quality measure parameters. The most popular measures are TEV and MSE. The objective function in these cases is quadratic in weights of the assets, and one needs to solve the constrained convex quadratic programming problem. The optimization models based on these measures will be explored in this thesis (See Equation 3.18, Equation 3.19, Equation 3.20). Additionally, two other important optimization models will be considered: MV and GMVP (See Equation 3.2, Equation 3.5). The classical mean-variance portfolio optimization model remains one of the most important benchmarks in the literature on asset allocation and in the asset management industry. GMVP, which belongs to the Risk-based Asset Allocation strategies, also received significant attention in the financial literature.

The above mentioned optimization models allows finding the optimal weights for a portfolio with specified assets. However, how do we choose assets for the portfolio? Applying sampling, one should solve the asset selection problem, choosing the assets which lead to the best possible solution. An exhaustive search would provide the best solution. The bigger the solution space, which in case of tracking portfolio depends on the sample universe, the more difficult it is to find the best portfolio in terms of

a certain measure. Even with the fastest hardware and the most massively parallel systems available today, it is infeasible to conduct an exhaustive search for the large solution space in a reasonable time. Therefore, this thesis suggests using some of the traditional optimization models in combination with different search space heuristics, which are described in Chapter 4.

Chapter 4

Tracking Portfolio Heuristics

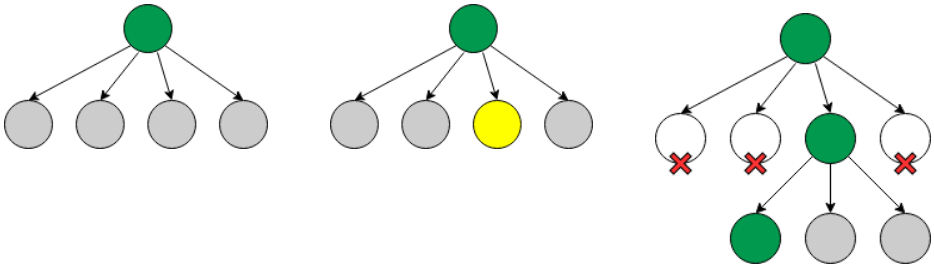
As shown in Section 3.6 the number of possible solutions can be prohibitively large and their generation expensive in terms of computational time and memory. When an exhaustive search is not possible or computationally expensive, different *heuristics* are used to limit the *search space*. Exploring the search space often corresponds to a search for the shortest path in an underlying problem graph. For this purpose, some heuristic-based algorithms evaluate intermediate models and discard those which do not meet a certain performance criterion. Most space searching algorithms are build either on *Breadth-First Search* (BFS) or *Depth-First search* (DFS), which differs in search order. The first one explores the neighbor nodes before moving to the next level. The second, in contrast, explores as far as possible along each branch [Edelkamp and Schroedl, 2011].

Heuristic algorithms are designed to find a solution in a faster and more efficient way than traditional methods sacrificing optimality for operationality. They are often applied to NP-hard problems, where there is no known efficient way to find a solution quickly and accurately.

4.1 Hill-Climbing Index Tracking

The number of possible solutions increases rapidly with the number of assets in the sample universe, which is in this work equal to the set of all index constituents. A *greedy heuristic* can solve the combinatorial problem of portfolio selection. The name greedy comes from the fact that the decision about a local specification is based on a single criterion: the specification that seems to be most promising is chosen for further exploration [Hromkovic, 2004].

Hill-Climbing is a greedy evaluation heuristic algorithm which adds the best feature in each round until a certain criterion is met. Applying the Hill-Climbing algorithm for tracking portfolio selection, it should stop when adding a new feature (asset) to the portfolio, generated in the previous round, does not improve the current solution or when all the assets from the sample universe are included in the tracking portfolio. [Faizliev et al., 2016] used Hill-Climbing for the index-tracking problem, restricting



(a) Generating refinements for the current model

(b) Selection of the best refinement

(c) Generating refinements for the selected at the previous level model

Figure 4.1: The refine-and-select process for the HILL-CLIMBING algorithm

the number of assets in the portfolio. In this thesis we omit the cardinality constraint, aiming to find a minimum number of assets for the tracking portfolio which leads the most accurate possible index replication.

Choosing a subset from $n = 10$ assets, the HILL-CLIMBING algorithm needs to try out at most $\sum_{p=1}^{n} p = \frac{n(n+1)}{2} = 55$ combination instead of the above-mentioned 1022 (See Equation 3.38). Often a greedy search does not produce an optimal solution, but nevertheless, a greedy heuristic may yield locally optimal solutions that approximate a global optimal solution in a reasonable time [Edelkamp and Schroedl, 2011].

Let $\mathcal{M}$ be the solution space and $m \in \mathcal{M}$ is a model or, in other words, one of the possible solutions from the solution space. In this thesis, a model plays the role of a possible tracking portfolio ($m_i = \boldsymbol{P}_i$). HILL-CLIMBING is a type of breadth-first search, which can be presented as an iterative application of two operators: *refinement* $r(\cdot)$ and *selection* $s(\cdot)$.

Definition 4.1. *A* ***refinement operator*** $r(\cdot)$ *generates a set of potentially better models (known as refinements) than those generated in the previous step* [Ivanova and Berthold, 2013].

$$r(m) = \{m_1, \ldots, m_l\} = \boldsymbol{M} \tag{4.1}$$

where $\boldsymbol{M} = \{\boldsymbol{P}_1, \ldots, \boldsymbol{P}_l\}$ is a set of l different portfolios from the solution space ($\boldsymbol{M} \subseteq \mathcal{M}$).

Definition 4.2. *A* ***selection operator*** $s(\cdot)$ *chooses the locally best model from all possible refinements* [Ivanova and Berthold, 2013]:

$$m' = s_{best}(r(m)) = s_{best}(\{m_1, \ldots, m_l\}) = s_{best}(\boldsymbol{M}) \tag{4.2}$$

In the context of tracking portfolio selection, refinement process at step i consists of adding to the obtained at $i-1$ step portfolio a new asset from the sample universe. This way we generate several new portfolios (refinements), each with the size i. Using

a tracking quality measure ϕ, we evaluate all of the new refinements, and, applying selection operator $s(\cdot)$, we choose the best portfolio at step i. The number of refinements decreases at each step of the algorithm because the number of assets (from the sample universe) which are not yet added in the tracking portfolio decreases. Figure 4.1 shows *the refine-and-select process* for the HILL-CLIMBING algorithm.

Trying to obtain the least possible divergence between the index and tracking portfolio, we set ϕ to be MSE (See Equation 3.11). Therefore, the refine-and-select process for index tracking requires one of the optimization models, defined in Chapter 3. We choose MV (See Equation 3.2), GMVP (See Equation 3.5), and tracking-error-based optimization models (See Equation 3.18, Equation 3.19, and Equation 3.20). A portfolio which tracks index performance in the best way, i.e., has the smallest MSE, is chosen for further refinement.

Let $\boldsymbol{R}(\boldsymbol{I}) = [\boldsymbol{r}(I_1), \ldots, \boldsymbol{r}(I_n)]'$ be the set of returns of n assets of the index $\boldsymbol{I}$, $\boldsymbol{R}(\boldsymbol{P}) \subseteq \boldsymbol{R}(\boldsymbol{I})$, where $\boldsymbol{R}(\boldsymbol{P}) = [\boldsymbol{r}(P_1), \ldots, \boldsymbol{r}(P_p)]'$ is the set of returns of p assets of $\boldsymbol{R}(\boldsymbol{P})$ with $p \in \{1, \ldots, n\}$. HILL-CLIMBING INDEX TRACKING can be found in Algorithm 4.1:

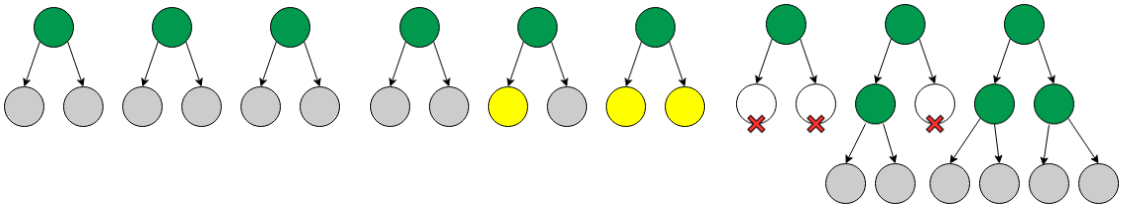
Algorithm 4.1 HILL-CLIMBING INDEX TRACKING

Input : $\boldsymbol{R}(\boldsymbol{I})$ is the set of returns of n index constituents
Output: $\boldsymbol{R}(\boldsymbol{P})$ is the set of returns of p assets in the tracking portfolio and
$\boldsymbol{\omega}_{\boldsymbol{R}(\boldsymbol{P})}$ are investing weights

$\boldsymbol{R}(\boldsymbol{P}) = \varnothing$
while $\phi_j < \phi_{j-1}$ *and* $j \leqslant n$ **do**
  execute **refine-and-select procedure**:
    $\forall \boldsymbol{r}(I_i) \in \boldsymbol{R}(\boldsymbol{I}) \backslash \boldsymbol{R}(\boldsymbol{P})$ calculate $\boldsymbol{\omega}_{\boldsymbol{R}(\boldsymbol{I}) \bigcup \boldsymbol{r}(I_i)}$
    using a portfolio optimization model
    (See Equation 3.2, Equation 3.5, Equation 3.18, Equation 3.19, Equation 3.20)
    select $\boldsymbol{r}(P_j) = \underset{\boldsymbol{r}(I_i) \in \boldsymbol{R}(\boldsymbol{I}) \backslash \boldsymbol{R}(\boldsymbol{P})}{\arg\min} \phi(\boldsymbol{R}(\boldsymbol{P}) \bigcup \boldsymbol{r}(I_i))$
    set $\boldsymbol{R}(\boldsymbol{P}) = \boldsymbol{R}(\boldsymbol{P}) \bigcup \boldsymbol{r}(P_j)$
    set $\boldsymbol{\omega}_{\boldsymbol{R}(\boldsymbol{P})} = \boldsymbol{\omega}_{\boldsymbol{R}(\boldsymbol{P}) \bigcup \boldsymbol{r}(P_j)}$
  j = j + 1
end

4.2 Beam Search Index Tracking

The drawback of the HILL-CLIMBING algorithm is its stopping when reaching a local optimum which may not be either the global optimum. The use of parallel resources helps to find a better solution, investigating more alternatives in parallel. A common heuristic for prohibitively large search spaces is BEAM SEARCH. It is is a type of breadth-first search algorithm, where several of the most promising models at each step of the search are retained for further branching. This heuristic compares all of the solutions



(a) Generating refinements for the current models

(b) Selection of the best $k = 3$ refinements

(c) Generating refinements for the selected at the previous level models

Figure 4.2: The refine-and-select process for Beam Search with width $k = 3$

found at a particular depth and stores a defined number (*beam width*) of "the best" ones (regarding a quality measure ϕ). The rest of the solutions is discarded and is not explored anymore. Beam Search with width k iteratively explores solution paths in parallel. The bigger the beam width k, the more likely it is to find a better solution. Beam Search with width $k = 1$ is equivalent to the Hill-Climbing algorithm. With $k = \infty$ it is identical to the implicit search. By restricting the width of the search space to k, the complexity of the search becomes linear, $O(kd)$ where k is the width and d is the depth of the search [Sampson, 2013].

Similar to the Hill-Climbing algorithm, Beam Search can be presented using the refine-and-select process (See Equation 4.2). The difference is that after a model is refined, a selection operator *s(·)* is applied to choose *the best k models* at each step until a stopping criterion met:

$$\{m_1^s, \ldots, m_k^s\} = s_{best_k}(\{r(m_1), \ldots, r(m_l)\}) \tag{4.3}$$

Figure 4.2 shows the refine-and-select process for Beam Search with the search width $k = 3$. Applying the Beam Search algorithm, one should take care of duplicate deletion in order to produce k different solutions. Beam Search stops when it is not possible to improve any of the currently k best solutions or when all the assets from the sample universe are included in one of the k best portfolios.

To define Beam Search Index Tracking, we introduce a *List of Selected Portfolios*, where k models are stored to be used as the next refinements. Successors are not examined until the rest of the previous k-best models are expanded. In terms of tracking portfolio selection each model m_j at step i is a portfolio $\boldsymbol{P}_j$ with i assets $\|\boldsymbol{P}_j\| = i$. Let $L_{selected}$ be a current sorted list of k selected models (portfolios):

$$L_{selected} = \{m_1, \ldots, m_k\} = \{\boldsymbol{P}_1, \ldots, \boldsymbol{P}_k\}, \tag{4.4}$$

where $\phi(m_1) \leq \ldots \leq \phi(m_k)$ and $L_{refined}$ be a list of refinements for a model m_j:

$$L_{refined} = r(m_j) = \{m_1^{'}, \ldots, m_l^{'}\} \tag{4.5}$$

Beam Search Index Tracking can be found in Algorithm 4.2.

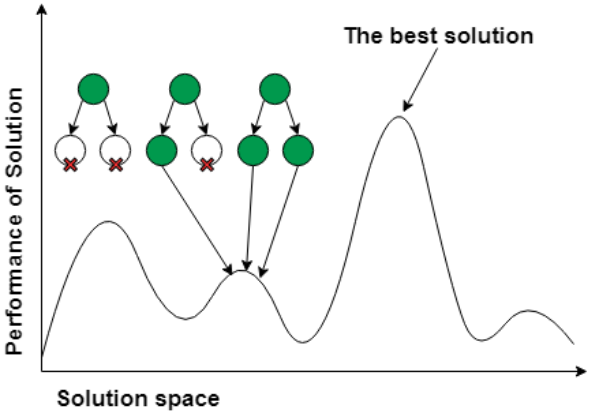


Figure 4.3: Exploitation effect of Beam Search. Obtained by Beam Search solutions can be closely related variations of one model.

Algorithm 4.2 Beam Search Index Tracking

Input : $\boldsymbol{R}(\boldsymbol{I})$ is the set of returns of n index constituents
Output: $\boldsymbol{R}(\boldsymbol{P})$ is the set of returns of p assets in the tracking portfolio and $\boldsymbol{\omega}_{\boldsymbol{R}(\boldsymbol{P})}$ are investing weights

$L_{selected} = s_{best_k}(\boldsymbol{I})$
$update = true$ // improvement of at least one of the k best models is possible
while $update = true$ **do**
  $update = false$ // reset
  foreach $m_i \in L_{selected}$ **do**
    $L_{refined} = \emptyset$
    $L_{refined}.append(r(m_i))$
    foreach $m_j^{'} \in L_{refined}$ **do**
      if $(m_j^{'} \notin L_{selected}$ *and* $\phi(m_j^{'}) < \phi(m_k)$, **then**
        Priority Queue:
        $L_{selected}.poll(m_k)$
        $L_{selected}.add(m_j^{'})$
        $update = true$ // improvement is possible
      end
    end
  end
end
$\boldsymbol{P_{top}} = L_{selected}.get(m_1)$, where $m_1 = top(L_{selected})$
$\boldsymbol{R}(\boldsymbol{P}) = \boldsymbol{R}(\boldsymbol{P}_{top})$
using a portfolio optimization model
(See Equation 3.2, Equation 3.5, Equation 3.18, Equation 3.19, Equation 3.20)
$\boldsymbol{\omega}_{\boldsymbol{R}(\boldsymbol{P})} = \boldsymbol{\omega}_{\boldsymbol{R}(\boldsymbol{P}_{top})}$

Beam Search is not optimal: there is no guarantee that it will find the best solution. Choice of the k best solutions at each iteration, based on the quality measure ϕ, does not ensure exploration of *different* regions of the search space. In contrast, it is likely that we are exploring only closely related variation of the locally best model [Ivanova and Berthold, 2013]. This effect is known as *exploitation*—improving the quality of solutions inside the neighborhood (See Figure 4.3) [Arslan et al., 2010]. Inappropriate balance between exploration and exploitation leads to an inefficient search [Al-Naqi et al., 2013].

4.3 Widened Index Tracking

The number of parallel resources needed for the implicit search increase with the size of the solution space. Even with the fastest hardware and the most massively parallel systems available today, it is infeasible to conduct an exhaustive search for the large solution space in a reasonable time. Instead of trying to traverse all of the solution space, we need to make the best use of every parallel resource. The goal of cost-effective use of parallel resources is closely related to the concept of *diverse exploration* of the search space [Ivanova and Berthold, 2013].

The iterative refine-and-select process is conceptually similar to the Beam Search (See Equation 4.3). The difference is in applying a diversity metric δ when refining models, which ensures that each of the solution paths is significantly different from the other solutions being explored [Sampson, 2013]. A diversity metric δ describes a difference between the resulting refined models $\{m_1, \ldots, m_l\}$.

Let $L_{diverse}$ be a list of diverse refinements:

$$L_{diverse} = r_\delta(\{m_1, \ldots, m_l\}) = r_\delta(\mathbf{M}) \tag{4.6}$$

A selection operator $s_{best_k}(\cdot)$ is applied to choose the k best models at each step until a stopping criterion met. Let $L_{selected}$ be a current sorted list of k selected models (portfolios):

$$L_{selected} = \{m_1, \ldots, m_k\} = s_{best_k}(\{r_\delta(m_1), \ldots, r(m_l)\}) = s_{best_k}(r_\delta(\mathbf{M})), \tag{4.7}$$

where $\phi(m_1) \leq \ldots \leq \phi(m_k)$. Widened Index Tracking can be found in Algorithm 4.3.

Diversity-Driven Widening allows the model refinement paths, separated by some measure of diversity, to be followed through the solution space until some stopping criterion is met [Sampson and Berthold, 2016]. In other words, inducing a *diversity measure* δ between solution paths (portfolios) can help to explore disparate regions of the solution space (See Figure 4.4). However, finding a metric, which is able to describe the difference *between* the portfolios, is not trivial.

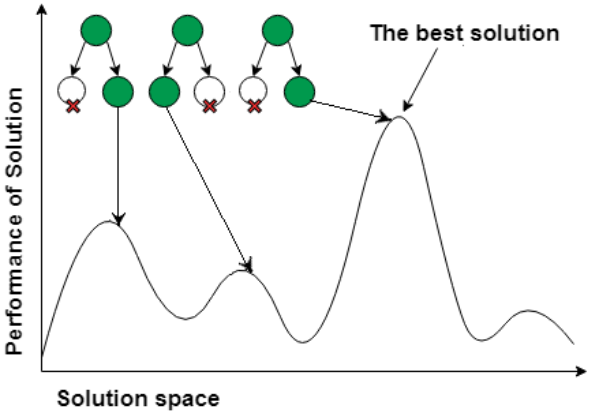


Figure 4.4: DIVERSITY-DRIVEN WIDENING. Inducing a diversity measure between solution paths can help to explore disparate regions of the solution space.

Algorithm 4.3 WIDENED INDEX TRACKING

Input : $\boldsymbol{R}(\boldsymbol{I})$ is the set of returns of n index constituents
Output: $\boldsymbol{R}(\boldsymbol{P})$ is the set of returns of p assets in the tracking portfolio and
$\boldsymbol{\omega}_{\boldsymbol{R}(\boldsymbol{P})}$ are investing weights

$L_{selected} = s_{best_k}(\boldsymbol{I})$
$update = true$ `// improvement of at least one of the` k `best models is possible`
while $update = true$ **do**
  $L_{diverse} = r_{\delta}(L_{selected})$
  $update = false$ `// reset`
  foreach $m_j^{'} \in L_{diverse}$ **do**
    if $(m_j^{'} \notin L_{selected}$ *and* $\phi(m_j^{'}) < \phi(m_k)$, **then**
      Priority Queue:
      $L_{selected}.poll(m_k)$
      $L_{selected}.add(m_j^{'})$
      $update = true$ `// improvement is possible`
    end
  end
end
$\boldsymbol{P_{top}} = L_{selected}.get(m_1)$, where $m_1 = top(L_{selected})$
$\boldsymbol{R(P)} = \boldsymbol{R(P_{top})}$
using a portfolio optimization model
(See Equation 3.2, Equation 3.5, Equation 3.18, Equation 3.19, Equation 3.20)
$\boldsymbol{\omega_{R(P)}} = \boldsymbol{\omega_{R(P_{top})}}$

4.3.1 Measuring Diversity

Measuring diversity and selecting a diverse subset are the main issues of DIVERSITY-DRIVEN WIDENING. The problem of selecting a subset of points from a larger set,

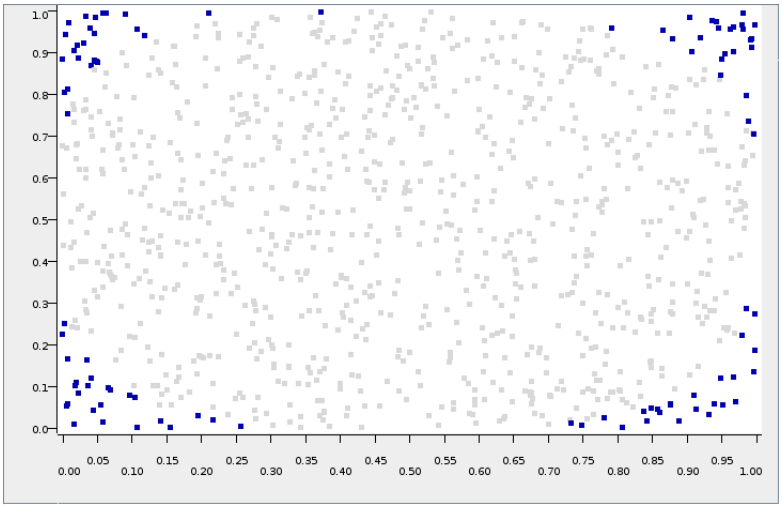


Figure 4.5: A subset of 100 points taken from a set of 1,000 randomly distributed points, representing a near-optimal solution for the p-DISPERSION-SUM problem

.

where some measure of diversity is maximized known as the *p-diversity problem* [Erkut, 1990; Erkut et al., 1994]. Diversity is an important issue in bio- and chemoinformatics and has been studied regarding protein and molecular similarity in [Meinl, 2010]. The author describes the maximum-score diversity selection (MSDS) problem on molecules, where two objectives should be optimized at the same time: select molecules with the highest activity and select a diverse subset. In Data Mining the effect of diversity on the parallel exploration of the solution space was studied in [Akbar et al., 2012; Fillbrunn and Berthold, 2015; Fillbrunn et al., 2017; Ivanova and Berthold, 2013; Sampson, 2013; Sampson and Berthold, 2014, 2016; Sampson et al., 2018].

Diversity does not have a single definition. It is often defined based on the distances between two objects. The further two points are apart, the more dissimilar they are. [Meinl, 2010] describes six diversity measures, which yield different types of diverse sets. Two diversity measures are described here—*p-dispersion-sum* and *p-dispersion-min-sum*.

Definition 4.3. *Given a set* $\boldsymbol{M} = \{\boldsymbol{P}_1, \ldots, \boldsymbol{P}_l\}$ *with* $|\boldsymbol{M}| = l$ *items, the* ***p-dispersion-sum*** *problem is defined as selection of the set* $\boldsymbol{P} \subseteq \boldsymbol{M}$ *of* $p \leq l$ *items* ($|\boldsymbol{P}| = p$) *by maximizing the sum of all pairwise distances* $d(\boldsymbol{P}_i, \boldsymbol{P}_j)$ *with* $\boldsymbol{P}_i, \boldsymbol{P}_j \in \boldsymbol{M}$ [Meinl, 2010]:

$$\boldsymbol{P} = \max \delta_s, \text{ where } \delta_s = \sum_{i=1}^{p} \sum_{j=1}^{i-1} d(\boldsymbol{P}_i, \boldsymbol{P}_j) \tag{4.8}$$

The optimization of *p-dispersion-sum* measure provides a subset whose elements located maximally far away from each other. In this case, the selected points are forced away from each other and are concentrated on the corners (See Figure 4.6). Hence, there is a high probability that such set will not represent a desirable diverse distribution.

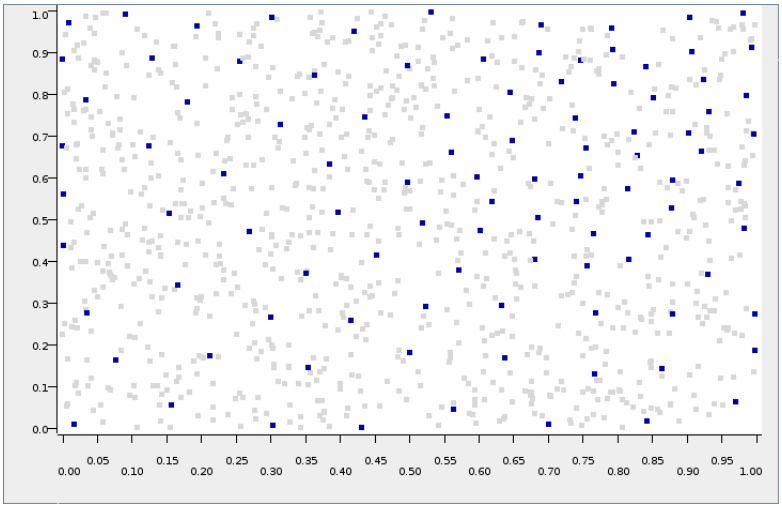


Figure 4.6: A subset of 100 points taken from a set of 1,000 randomly distributed points, representing a near-optimal solution for the p-DISPERSION-MIN-SUM problem

.

Single small distances, which occur using *p-dispersion-sum*, can ruin diversity. As an alternative, another measure which is known as *p-dispersion-min-sum* can be applied.

Definition 4.4. *Given a set* $\boldsymbol{M} = \{\boldsymbol{P}_1, \ldots, \boldsymbol{P}_l\}$ *with* $|\boldsymbol{M}| = l$ *items, the* ***p-dispersion-min-sum*** *problem is defined as selection of the set* $\boldsymbol{P} \subseteq \boldsymbol{M}$ *of* $p \leq l$ *items* ($|\boldsymbol{P}| = p$) *by maximizing the sum of minimal distances* $d(\boldsymbol{P}_i, \boldsymbol{P}_j)$ *with* $\boldsymbol{P}_i, \boldsymbol{P}_j \in \boldsymbol{M}$ [Meinl, 2010]:

$$\boldsymbol{P} = \max \delta_{ms}, \text{ where } \delta_{ms} = \sum_{j=1}^{p} \min_{1 \leq j \leq p, i \neq j} d(\boldsymbol{P}_i, \boldsymbol{P}_j) \tag{4.9}$$

It means that the distances from each point to its nearest neighbor are summed up. This way the influence of small distances on diversity is reduced. This method provides better coverage of the space (See Figure 4.5).

4.3.2 Distance Measurements

Distance measures are numerical measurements which describe how far apart objects are. The distance between items in a set can be measured in different ways. The choice often depends on the structure of the dataset [Berthold et al., 2010].

One of the most widely used measures for *non-numerical attributes* is the Jaccard distance (or the Jaccard dissimilarity measure). It is often used for market-basket data and data which consists of many sparse binary attributes.

$$d_{Jaccard}(x, y) = 1 - \frac{|X \bigcap Y|}{|X \bigcup Y|} \tag{4.10}$$

The elementary and most intuitive distance measurements for *numerical attributes* is the Euclidean distance. For an n-dimensional space the Euclidean distance has the form:

$$d_{Euclidean}(x,y) = \sqrt{(x_1 - y_1)^2 + \ldots + (x_n - y_n)^2} = \sqrt{\sum_{j=1}^{p}(x_j - y_j)^2} \tag{4.11}$$

To track an index, [Focardi and Fabozzi, 2004] suggested using Euclidean distances between stock price series as a basis for hierarchical clustering.

The Squared Euclidean Distance, which is also known as the sum of squared difference (SSD), has the following representation:

$$d_{SSD}(x,y) = \sum_{j=1}^{p}(x_j - y_j)^2 \tag{4.12}$$

This is the fundamental metric in least squares problems and linear algebra. The squares cause it to be sensitive to outliers. The Manhattan Distance, also known as sum of absolute difference is more robust than SSD:

$$d_{Manhatten}(x,y) = |x_1 - y_1| + \ldots + |x_n - y_n| = \sum_{j=1}^{p}|x_j - y_j| \tag{4.13}$$

The Chebyshev distance, also known as Chessboard distance, has the following representation:

$$d_{Chebyshev}(x,y) = \max(|x_j - y_j|) \tag{4.14}$$

Mentioned above distances is generalized by the Minkowski distance:

$$d_{Minkowski}(x,y) = \sqrt[m]{\sum_{j=1}^{p}|x_j - y_j|^m} \tag{4.15}$$

With $m = 1$ this distance is equal to the Manhattan distance, $m = 2$ to the Euclidean distance and $m \to \infty$ to the Chebyshev distance.

The Canberra distance is a weighted version of the Manhattan distance and is often used for data scattered around an origin:

$$d_{Canberra}(x,y) = \sum_{j=1}^{p}\frac{|x_j - y_j|}{|x_j| - |y_j|} \tag{4.16}$$

The use of this metric is limited because it is sensitive for values close to zero.

The Pearson distance is a correlation distance which is based on the Pearson's product-momentum correlation coefficient of the two sample vectors. Because the correlation coefficient can take a range of values between [-1, 1], the Pearson distance is bounded between [0, 2] and measures the linear relationship between the two vectors:

$$d_{Pearson}(x, y) = 1 - Corr(x, y) \tag{4.17}$$

The Pearson distance was described in [Focardi and Fabozzi, 2004] for the assets diversification in a portfolio. [Papenbrock, 2011; Ren et al., 2017] describe a similar correlation based metric, which was applied for portfolio construction using hierarchical clustering:

$$d_{Pearson_{sqr}}(p, q) = \sqrt{2(1 - Corr(x, y))} \tag{4.18}$$

4.3.3 Portfolio Diversity

Implementation of the Widened Index Tracking algorithm requires a diversity measure. However, what in terms of portfolio selection could be a good measure that would lead better solution space exploration than a greedy search. What makes portfolios different from each other in the search space? There are many statistics which characterize a portfolio, such as mean, variance, certainty equivalent, Sharpe ratio. An investor often wishes to minimize the portfolio variance, maximize the expected return, Sharpe ratio or certainty equivalent. That is why these measures may be not the best for the diverse solution space exploration.

Trying to get a diversified portfolio, several studies suggested using the Pearson distance (See Equation 4.17) *between stock prices* as a basis for hierarchical clustering [Focardi and Fabozzi, 2004; Papenbrock, 2011; Ren et al., 2017]. Asset clustering is used to identify assets with different return structures. We will apply this distance metric *between portfolios returns* of the same size, hoping to obtain diverse solutions for broader exploration of the search space.

Each portfolio is characterized with a certain investment weight. In [Goetzmann and Kumar, 2008] the sum of squared portfolio weights was used to measure the diversification of retail investors:

$$SSPW = \sum_{j=1}^{p} (\omega_j - \omega_m)^2 = \sum_{j=1}^{p} (\omega_j - \frac{1}{n_m})^2 = \sum_{j=1}^{p} \omega_j^2 \tag{4.19}$$

where n_m is the number of stocks in the market portfolio, ω_j is the portfolio weight assigned to stock P_j in the investor's portfolio, and ω_m is the weight assigned to a stock in the market portfolio, which is this paper $\omega_m = \frac{1}{n_m}$ (the weight of each security in the market portfolio is very small) [Goetzmann and Kumar, 2008]. In this thesis, we apply SSPW with the Euclidean distance for the portfolio solution space exploration.

4.4 Summary

Index replication can be seen as a solution space exploration problem. When the number of assets in the sample universe gets large, it is not possible to apply an exhaustive search. To reduce the search space, we can use a heuristic-based optimization. One of the simplest algorithms is Hill-Climbing, which iteratively adds a new asset in the portfolio until it is possible to improve the current solution. However, this greedy heuristic has the disadvantage of stopping when reaching a local optimum which may not be either the global optimum. To expand our search and, therefore, possibly increase the accuracy of the algorithm, we use Beam Search, which allows exploration of several "best" portfolios in parallel. However, there is no guarantee that we do not explore closely related variation of the locally best model. In other words, we have to make sure that the algorithm does not step around the local model, trying to improve the quality of the solution. Increasing the number of parallel resources can be computationally expensive. To overcome this issue and investigate disparate regions of the solution space, we apply Widened Index Tracking. Here the main problem is to define the right distance metric and diversity measure with regards to tracking portfolio selection. There are several studies which use clustering to gain a diversified portfolio, i.e., to obtain diverse assets it the portfolio. We try to apply some of the used in the literature metrics for the "*within*" portfolio diversification to the "*between*" portfolios diversification.

Chapter 5

Experimental Results

The algorithms presented in Chapter 4 are implemented in KNIME[1] [Berthold et al., 2007] and tested on 6 datasets of different sizes. The one-period daily returns (or simple returns) for the index $\boldsymbol{I}$ and each component of the index I_i with $i \in \{1, \ldots, n\}$ are computed:

$$r_t^*(\boldsymbol{I}) = \frac{r_{t+1}(\boldsymbol{I}) - r_t(\boldsymbol{I})}{r_t(\boldsymbol{I})},\ r_t^*(I_i) = \frac{r_{t+1}(I_i) - r_t(I_i)}{r_t(I_i)} \tag{5.1}$$

The experiments are carried out with the real-world market indices (See Table 5.1). Daily returns are obtained for 01.01.15–31.01.18 from *Thomson Reuters Datastream*.[2] The stocks with missing historical data from the beginning of the used testing periods were excluded.

Index	**Area**	$\boldsymbol{n}$	$\boldsymbol{n^*}$
SMI	Switzerland	20	20
DAX	Germany	30	28
DJIA	USA	30	29
STOXX50	Eurozone	50	50
NASDAQ	USA	103	101
S&P500	USA	505	495

Table 5.1: Summary of the replicated indices. The number of index constituents n is taken based on 31.01.2018. n^* denotes the number of index constituents after data preprocessing, i.e., n^* is the size of the sample universe.

The composition of an index is often reviewed annually or quarterly. That is why the *validation period*, as well as the *testing period* is chosen to be not bigger than 3 months $H = 60$ data points.[3] For the in-sample period, we chose $T = 500$ as the sample size, where 440 data points are used for training and 60 data points are chosen for validation.

[1] KNIME Analytics Platform version 3.5.0

[2] `https://eikon.thomsonreuters.com`

[3] 3 months is approximately 60 trading days.

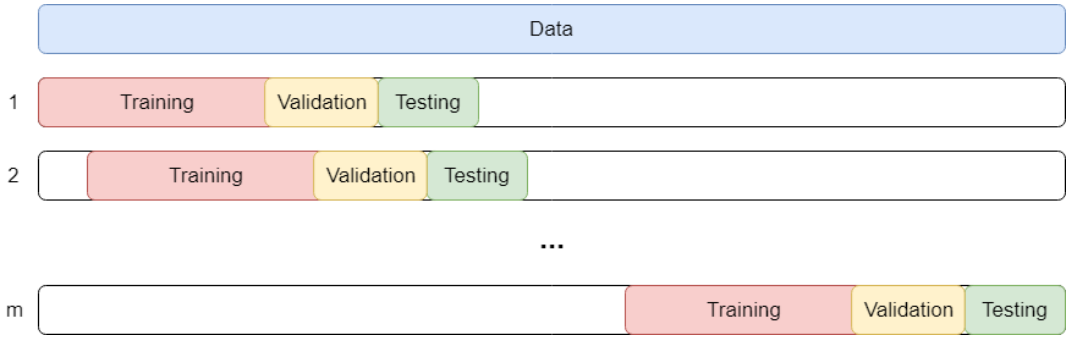


Figure 5.1: Design of the experiment

For out-of sample performance evaluation we use rolling (moving) window with the step size $h = 20$ (See Figure 5.1).

As discussed in Chapter 3 we choose several portfolio optimization models:

- Markowitz Mean Variance Portfolio optimization (See Equation 3.2)
- Global Minimum Variance Portfolio optimization (See Equation 3.5)
- Tracking Error Portfolio optimization which minimizes TEV (See Equation 3.18)
- Tracking Error Portfolio optimization which minimizes MSE (See Equation 3.20)
- Mean Constrained Tracking Error Portfolio optimization which minimizes TEV (See Equation 3.19)

These models are used in combination with the search space heuristics described in Chapter 4 :

- Hill-Climbing Index Tracking (See Algorithm 4.1)
- Beam Search Index Tracking (See Algorithm 4.2)
- Widened Index Tracking (See Algorithm 4.3)

5.1 Investing Phase

The experiments are carried out on two types of datasets: small datasets with a number of index constituents $n < 50$ and relatively big datasets with $n \geqslant 50$. The small datasets are used to test several diversity measures. To find the best tracking portfolio with the minimum possible number of assets, we use the search space heuristics in combination with several portfolio optimization models. In this case, Widening becomes computationally expensive for big datasets. Therefore, for indices with the number of constituents $n \geqslant 50$, we choose a diversity measure which shows the best performance on small datasets.

5.1.1 Replication of Small Indices

From Table 5.2 we can see that independent of the used heuristic, the tracking-error-based optimizations provide better results than MV and GMVP (detailed out-of-sample statistics is presented in Table A.1 and Table A.2 of Appendix A). As expected TEV, MSE, TEMV are more appropriate for the index replication problem because the objective function in the optimization model in these cases aims to reduce the tracking error. In the case of MV and GMVP models, minimization of the tracking error is applied using a fitness function at each step of the algorithm, but the investing weights for the portfolio are obtained by minimizing the portfolio variance (See Equation 3.2 and Equation 3.5). This can be the reason of larger divergence between an index and the portfolio, especially when the index is volatile. From the other side, when the index is volatile, an investor may wish to apply these models to reduce the risks and transaction costs. The use of TEV and MSE optimization models in combination with the chosen search space heuristics shows very similar results. When the difference between two models is statistically insignificant, by the law of parsimony the model with fewer assets in the portfolio is preferable.[4]

The experimental results show that the use of parallel resources allows finding a better solution than that obtained by Hill-Climbing. Beam Search with width $k = 5$ provides better results (in terms of MSE) for each optimization model. In most of the cases, the number of assets in the tracking portfolio increases as well. However, on the example of SMI replication, we can notice that Beam Search (in combination with TEMV) is able to find a better result (than that obtained by Hill-Climbing) with the *lower* number of assets in the portfolio.

Widened Index Tracking shows better results than those obtained by Beam Search in combination with TEV and MSE models if we use the Pearson distance between the portfolio returns (See Equation 4.17) and for TEV, MSE, TEMV if we use the Euclidean distance with the squared sum of portfolio weights (See Equation 4.11 and Equation 4.19). The use of these measures for MV and GMVP mostly is not able to provide better results than Beam Search. For these portfolio optimization models, chosen metrics provide poor path diversification and hinder the progress towards a better solution. In Table 5.2 the results of Widened Index Tracking presented for the Euclidean distance (See detailed out-of-sample statistics for all diversity measures in Table A.2 of Appendix A). Relatively small MSE between the index and portfolio returns for the Widened Index Tracking, which is used in combination with the tracking-error-based optimization models, shows that the search width $k = 5$ is big enough for such small datasets to provide a better solution than a simple greedy heuristic. The use of *p-dispersion-min-sum* (See Definition 4.4) and *p-dispersion-sum* (See Definition 4.3) measures often shows very similar results. However, in the case of SMI replication, Widened Index Tracking with the use of *p-dispersion-sum* measure shows in general worse performance than Beam Search.

[4]The law of parsimony, or Occam's razor principle, refers to a problem-solving principle which implies that "plurality should not be posited without necessity."

Index	Method	Hill-Climbing		Beam Search		Widened Index Tracking			
						p-sum		p-min-sum	
	Model	p	MSE $\times 10^4$	p	MSE $\times 10^4$	p	MSE $\times 10^4$	p	MSE $\times 10^4$
SMI ($n = 20$)	GMVP	6.82	0.0967	6.91	0.0895	9.27	0.1212	9.64	0.1220
	MV	6.09	0.1050	6.55	0.1019	8.73	0.1248	10.09	0.0950
	TEV	15.09	0.0439	17.09	0.0031	19.18	0.0102	19.27	**0.0023**
	MSE	15.09	0.0440	17.09	0.0031	19.00	0.0102	19.09	0.0024
	TEMV	17.36	**0.0032**	16.82	**0.0030**	18.45	**0.0028**	18.82	0.0028
DAX ($n = 28$)	GMVP	6.91	0.1063	7.55	0.0862	8.09	0.2250	9.45	0.2338
	MV	6.82	0.1120	8.27	0.0637	8.64	0.0851	10.27	0.1427
	TEV	21.27	0.0158	25.00	**0.0081**	27.18	**0.0052**	26.73	**0.0053**
	MSE	21.27	**0.0157**	25.09	0.0081	26.73	0.0053	27.18	0.0053
	TEMV	19.73	0.0537	24.09	0.0104	27.18	0.0057	27.00	0.0056
DJIA ($n = 29$)	GMVP	8.82	0.1103	9.36	0.0911	11.09	0.1012	13.82	0.1204
	MV	8.36	0.0975	9.27	0.0911	11.82	0.1156	13.91	0.0995
	TEV	20.64	0.0157	24.45	0.0058	28.09	0.0022	28.09	0.0023
	MSE	20.09	0.0165	24.45	0.0058	27.91	**0.0022**	27.82	**0.0023**
	TEMV	21.82	**0.0124**	24.36	**0.0047**	27.45	0.0036	27.18	0.0037

Table 5.2: Out-of-sample statistics for small datasets. The search width for Beam Search and Widened Index Tracking is $k = 5$. The smallest average MSE between the index and portfolio returns for each search space heuristic appears in bold. Underlined text shows the minimum average number of assets which is included in the best tracking portfolio.

It is important to notice that Widened Index Tracking with width $k = 5$ increases the number of the assets in the tracking portfolio for the small datasets, so that it is very close to the number of index constituents. It may happen because we did not put any restrictions on tracking error convergence, i.e., we tried to get the smallest possible tracking error. In this way, the tracking error continued to decrease insignificantly, whereas the number of assets continued to grow. Additionally, another reason for this could be not "diverse enough" paths in the solution space, meaning that models with the smaller number of assets and may be even smaller tracking error were discarded and not explored due to "diversification criteria," or, in other words, because of the chosen diversity measure. We used small datasets to test different diversity metrics (for example, using portfolio variation, Sharpe ratio, highest/lowest portfolio returns) but did not obtained better results than the described ones and, therefore, we did not include them in this thesis.

5.1.2 Replication of Big Indices

In Table 5.3 we can see the experimental results provided for the larger datasets (See detailed out-of sample statistics in Table A.3 of Appendix A). For Widened Index Tracking we used the sum of the squared weights of portfolio as a diversity measure, because it showed better performance (among tested diversity measures) for the small datasets. Similar to the results obtained for the small datasets, we can see that

Index	Method	Hill-Climbing		Beam Search		Widened Index Tracking			
						p-sum		p-min-sum	
	Model	p	MSE $\times 10^4$	p	MSE $\times 10^4$	p	MSE $\times 10^4$	p	MSE $\times 10^4$
STOXX50 ($n = 50$)	GMVP	7.18	0.1467	7.82	0.1548	10.09	0.1281	9.27	0.1531
	MV	7.73	0.1447	8.55	0.1356	9.00	0.1383	8.00	0.1694
	TEV	25.27	0.0168	28.36	0.0193	44.73	0.0045	44.91	0.0055
	MSE	25.36	**0.0161**	29.55	**0.0135**	44.64	**0.0045**	47.45	**0.0032**
	TEMV	22.55	0.0662	27.09	0.0198	42.00	0.0092	35.36	0.0570
NASDAQ ($n = 101$)	GMVP	14.27	0.1424	16.09	0.0844	16.27	0.2178	16.36	0.3119
	MV	13.73	0.1092	16.55	0.0791	13.73	0.2736	17.45	0.2587
	TEV	23.91	**0.0483**	30.36	**0.0282**	65.18	**0.0070**	55.64	**0.0112**
	MSE	26.18	0.0487	29.36	0.0292	51.27	0.0108	50.55	0.0116
	TEMV	28.91	0.0387	30.73	0.0390	49.27	0.0230	52.64	0.0206
S&P500 ($n = 495$)	GMVP	27.82	0.1158	29.27	0.0962	21.00	0.1090	10.80	0.1646
	MV	23.64	0.1021	31.31	0.0838	18.00	0.1278	16.40	0.1534
	TEV	33.27	0.0813	35.64	0.0673	61.00	0.0132	57.20	**0.0154**
	MSE	31.82	**0.0763**	37.17	**0.0616**	59.60	**0.0124**	53.80	0.0174
	TEMV	28.91	0.0958	37.55	0.0689	56.00	0.0175	51.60	0.0207

Table 5.3: Out-of-sample statistics for the datasets with $n \geqslant 50$. The search width for Beam Search and Widened Index Tracking is $k = 5$. The smallest average MSE between the index and portfolio returns for each search space heuristic appears in bold. Underlined text shows the minimum average number of assets which is included in the best tracking portfolio.

the use of parallel resources provides better index replication. Beam Search shows better results than the Hill-Climbing algorithm. Moreover, Widened Index Tracking in combination with tracking-error-based optimization models allows wider exploration of the solution space and is able to provide a better solution than a greedy search. Therefore, Hypothesis 1 is not rejected by these experiments. For indices with hundreds of assets, such as NASDAQ and S&P500, Widened Index Tracking (using TEV and MSE models) found a set of assets for the tracking portfolio with a significantly smaller number of constituents than indices and at the same time with the smaller MSE than that obtained using a greedy heuristic. For example, Widened Index Tracking in combination with MSE portfolio optimization model with the search width $k = 5$ (using *p-dispersion-sum* measure) suggest replicating the S&P500 index with about 60 assets.

The use of Widened Index Tracking with the GMVP and MV mostly did not provide results better than that obtained by a standard greedy search. The same results we obtained replicating small indices (See Table 5.2). As discussed in Section 5.1.1, for these models another diversity measure should be used.

5.1.3 The Effect of Width

When replicating the STOXX50 index, Beam Search with width $k = 5$ was not able to find better tracking portfolios for all optimization models, i.e, for GMVP and TEV the results became slightly worse (See Table 5.3). Therefore, on the example of STOXX50,

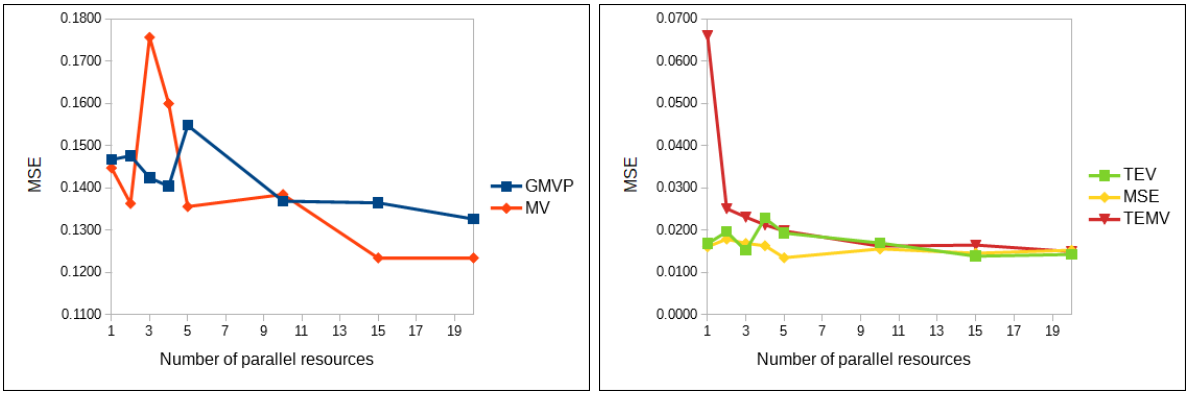


Figure 5.2: Replication of the STOXX50 index with Beam Search Index Tracking. Tracking portfolio performance while increasing the number of parallel resources. MSE $\times 10^4$ versus the search width k.

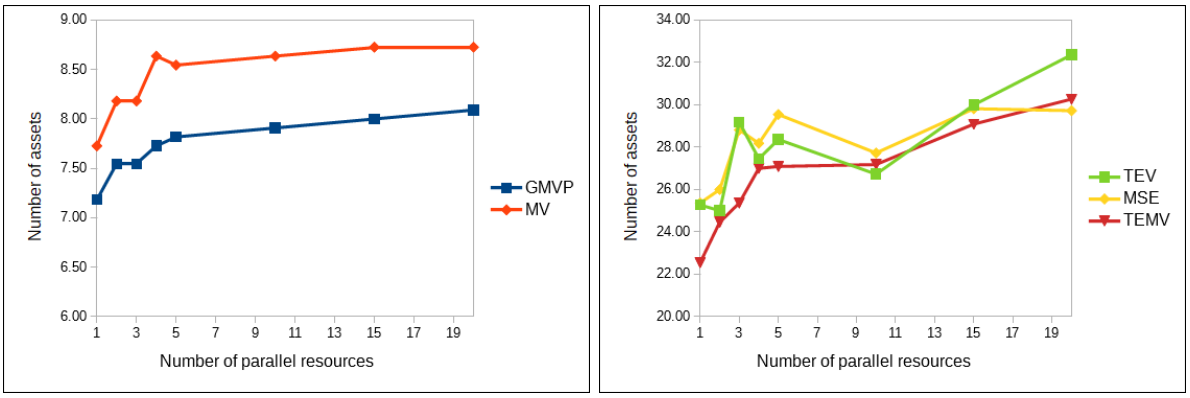


Figure 5.3: Replication of the STOXX50 index with Beam Search Index Tracking. The number of assets in the tracking portfolio p while increasing the number of parallel resources (the search width k).

we test whether increasing number of parallel resources provides better index replication. Figure 5.2 shows that a bigger search width k allows finding a portfolio with smaller MSE. However, the number of assets in the tracking portfolio has tendency to increase (See Figure 5.3).

We conducted the same analysis on the STOXX50 index replication for Widened Index Tracking (on the example of *p-dispersion-min-sum* problem). Figure 5.4 shows the performance of each optimization model while increasing the number of used parallel resources. Similar to Beam Search (See Figure 5.2 and Figure 5.3), for the tracking-error-based optimization models, divergence between an index returns and portfolio returns (in this work MSE) has general tendency to decrease with the bigger search width k. Therefore, Hypothesis 2 is not rejected by these experiments. The number of assets in the tracking portfolio has general tendency to increase with the wider search (See Figure 5.5). With the search width $k = 15$ this number for the TEV and MSE optimization

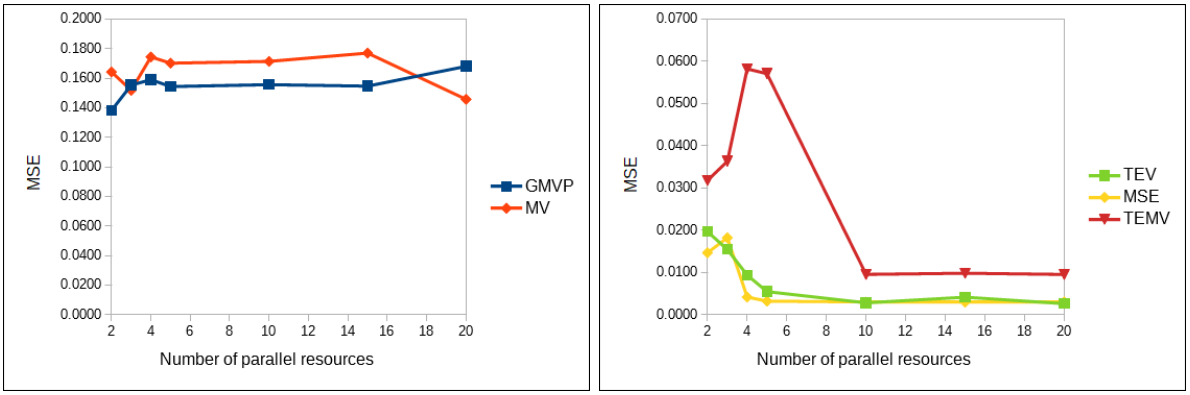


Figure 5.4: Replication of the STOXX50 index with Widened Index Tracking. Tracking portfolio performance while increasing the number of parallel resources. MSE $\times 10^4$ versus the search width k.

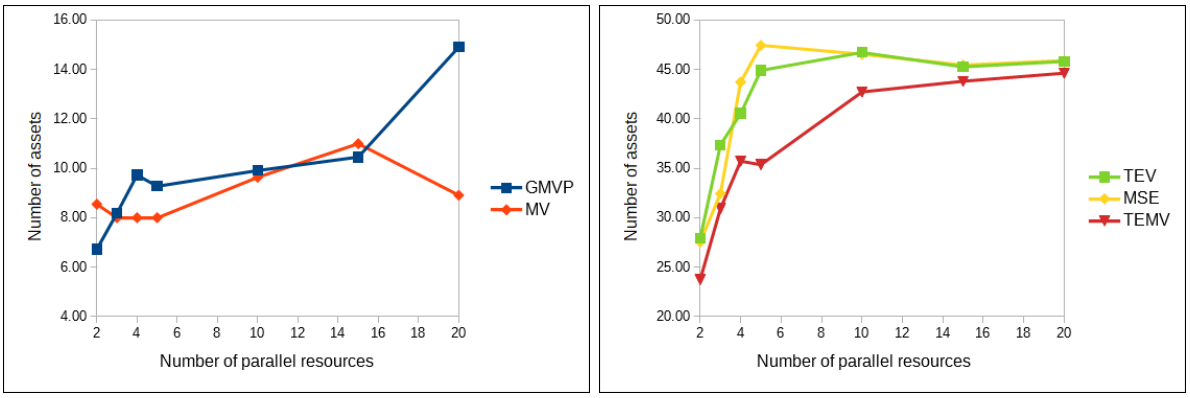


Figure 5.5: Replication of the STOXX50 index with Widened Index Tracking. The number of assets in the tracking portfolio p while increasing the number of parallel resources (the search width k).

models slightly decreases, whereas the difference between an index and portfolio returns increases insignificantly. In this case, increasing the number of parallel resources allows Widened Index Tracking to find a smaller number of assets for the tracking portfolio.

For the MV and GMVP models (used in Widened Index Tracking), MSE between the index and portfolio returns does not decrease while increasing the number of parallel resources, whereas the number of assets in the tracking portfolio has tendency to increase. This shows once again that used diversity measure is not applicable for these two models. In this case, another diversity measure should be found. As noticed before, used in this work heuristics in combination with the MV and GMVP optimization models showed the worst results (in comparison with the tracking-error-based optimization models) and even increasing the number of parallel resources can not provide better index tracking.

5.2 Rebalancing Phase

A lot of studies avoid using of transaction costs (See Definition 2.7), assuming that there is none. However, it is a very important factor in the index-tracking problem. The influence of transaction costs on portfolio performance in practice can be very significant. If the transaction costs are not considered during rebalancing, the performance of the portfolio could be poor [Fabozzi and Pachamanova, 2016].

Several studies suggest to incorporate transaction costs including them in the objective function in the following way:

$$\min\left(\rho \text{ measure}(\boldsymbol{TE}) + (1-\rho)TC\right) \tag{5.2}$$

where $\rho \in [0,1]$ is the weighting parameter [Adcock and Meade, 1994; Derigs and Nickel, 2004; Di Tollo and Maringer, 2009]. If $\rho = 1$, the fitness function is equivalent to minimizing a tracking error measure (See Equation 3.16).

Including transaction cost in the objective function or in the constraint (See Equation 3.31) make the optimization problem more complex. Transaction cost models can involve complicated non-linear functions. Although software for general nonlinear optimization problems exist, the computational time required for solving such problems is often relatively high. To account for transaction costs in this work, we include them in the decision making process in the fitness function. Of course, it may not be the most optimal solution, but ignoring the influence of transaction costs on the index replication (on portfolio returns) makes the analysis incomplete. For a single period problem, we used minimization of MSE as a fitness function. We stay with the same measure and use Equation 5.2 as a fitness function at each step of the algorithm with $\rho = \frac{1}{2}$.

According to the definition (See Definition 2.7), transaction costs can be represented as a function of turnover rate in the following way (See Equation 3.29):

$$TC = \|\boldsymbol{\Lambda}(\boldsymbol{\omega}_1 - \boldsymbol{\omega}_0)\|_\ell^\ell \tag{5.3}$$

where $\boldsymbol{\omega}_0$ is a current portfolio weight vector, $\boldsymbol{\omega}_1$ is a new portfolio weight vector, $\boldsymbol{\Lambda}$ is the transaction cost matrix and ℓ is norm.

The simplest and most common way is to set $\ell = 1$ and consider transaction costs as a linear function:

$$TC = \|\boldsymbol{\Lambda}(\boldsymbol{\omega}_1 - \boldsymbol{\omega}_0)\|_1^1 = |\boldsymbol{\Lambda}(\boldsymbol{\omega}_1 - \boldsymbol{\omega}_0)| \tag{5.4}$$

In this case, transaction costs have a L1-norm representation. This type of transaction costs are often used in research with a constant transaction costs factor $\boldsymbol{\Lambda} = c\mathbf{I}$, where c is a nominal proportional transaction cost term (often c set to be 50 basis points per transaction, i.e., 0.5% of trade volume) [DeMiguel et al., 2009; Mei and Nogales, 2018; Olivares-Nadal and DeMiguel, 2018].

The use of linear transaction costs in Equation 5.2 may not be the best choice. In this case, two parts of the function have a different magnitude. The use of transaction cost in the L1-norm representation will mean simply minimizing the transaction costs function

Method	Hill-Climbing				Beam Search				Widened Index Tracking			
Model	p	TO	TC $\times 10^4$	MSE $\times 10^4$	p	TO	TC $\times 10^4$	MSE $\times 10^4$	p	TO	TC $\times 10^4$	MSE $\times 10^4$
GMVP	8.4	0.473	0.9742	0.1604	13.4	0.348	0.1385	0.1183	12.4	1.620	0.8525	0.3262
MV	10.6	0.513	0.9273	0.1306	12.8	0.303	0.0591	0.0786	12.4	1.663	0.7991	0.2792
TEV	12.4	0.830	0.2161	0.0689	27.4	**0.140**	<u>0.0091</u>	0.0380	32.2	0.816	0.2686	0.0512
MSE	17.0	0.499	0.2979	0.0819	27.2	0.165	0.0120	0.0313	41.2	**0.571**	<u>0.1526</u>	0.0153
TEMV	21.4	**0.253**	<u>0.0598</u>	0.0733	28.2	0.216	0.0171	0.0454	43.4	0.573	0.1536	0.0298

Table 5.4: Replication of the NASDAQ index. Out-of-sample statistics after rebalancing phase. Text in bold shows the lowest turnover for each search space heuristic. Underlined text shows the lowest transaction costs calculated using Equation 5.6.

without taking into account the error term. Several studies suggest using quadratic transaction costs [Gârleanu and Pedersen, 2013; Mei et al., 2016; Moallemi and Salam, 2017; Taylor, 2015]. In this case transaction costs has a L2-norm representation:

$$TC = \|\mathbf{\Lambda}(\boldsymbol{\omega}_1 - \boldsymbol{\omega}_0)\|_2^2 \tag{5.5}$$

Often in the literature $\mathbf{\Lambda}$ is proportional to a measure of the risk, e.g., covariance matrix. Then $\mathbf{\Lambda} = \gamma\mathbf{\Sigma}$, where γ is a risk aversion parameter. Given $\gamma = 1$, [Takeda et al., 2013] suggested to use the following representation of quadratic transaction costs:

$$TC = \|\mathbf{\Lambda}(\boldsymbol{\omega}_1 - \boldsymbol{\omega}_0)\|_2^2 = \|\mathbf{\Sigma}^{1/2}(\boldsymbol{\omega}_1 - \boldsymbol{\omega}_0)\|_2^2 = (\boldsymbol{\omega}_1 - \boldsymbol{\omega}_0)\mathbf{\Sigma}(\boldsymbol{\omega}_1 - \boldsymbol{\omega}_0) \tag{5.6}$$

Such representation of transaction costs can be interpreted as a compensation for the risk [Taylor, 2015]. For our work, we chose such type of transaction costs, defining a new fitness function as:

$$\min\left(\text{MSE} + (\boldsymbol{\omega}_1 - \boldsymbol{\omega}_0)\mathbf{\Sigma}(\boldsymbol{\omega}_1 - \boldsymbol{\omega}_0)\right) \tag{5.7}$$

Applying the described heuristics with the modified fitness function, we demonstrate the rebalancing process on the example of the NASDAQ index replication, which contains 101 assets. Table 5.4 shows that Beam Search allows reducing turnover rate and transaction costs for each optimization model used in this work (See detailed out-of-sample statistics in Table A.4 of Appendix A). Similar to the investing phase (See Section 5.1), the use of GMVP and MV in combination with Widened Index Tracking for rebalancing provides worse results in terms of MSE than a greedy search. Turnover rate and transaction costs size, which are modeled as a quadratic function (See Equation 5.6), are very high. As discussed in Section 5.1 for these optimization models, another diversity measure should be applied. For the tracking-error-based optimization models, Widened Index Tracking shows the lowest MSE between the index and portfolio returns. The number of assets increases accordingly. However, the turnover rate and the size of transaction costs is mostly higher than that obtained by a greedy search. One of the reasons for poor rebalancing performance in this case can be an inappropriate rebalancing method, which in this work combines MSE and transaction costs, which are modeled as a quadratic function, in the fitness function. Therefore, for the Widened Index Tracking another rebalancing strategy should be developed.

5.3 Summary

Index tracking can be considered as a combination of two phases: investing and rebalancing. This thesis focuses primarily on the investing phase, aiming to find a minimum number of assets which is able to replicate an index. For analysis we chose the datasets of different sizes. Independently of the size of the sample universe, we were able to find a set of assets which is able to replicate an index behavior. To track an index we used different portfolio optimization models in combination with several search space heuristics.

The use of parallel resources supports broader exploration of the search space. We did not find a *unique* portfolio diversity measure which is able to provide a diverse space exploration for *all* of the portfolio optimization models. Among the chosen optimization models, Widened Index Tracking in combination with the tracking-error-based optimization models shows the best results, finding a better solution than that obtained by a standard greedy search. Moreover, we showed that increasing the number of parallel resources, we can obtain a better portfolio. However, a broader search increases the number of assets in the tracking portfolio as well. One of the reasons for this is that we tried to get the smallest possible tracking error. In this way, the tracking error continued to decrease insignificantly, whereas the number of assets continued to grow.

The rebalancing phase allows incorporating transaction costs in the index replication process. We discussed how to modify suggested algorithms and tested them on a real-word dataset. Beam Search provides the smallest turnover rate and allows for reducing transaction costs. Widened Index Tracking shows the best results in terms of MSE. However, rebalancing using this heuristic mostly shows higher turnover rate and higher transaction costs than a greedy search. For this heuristic, another rebalancing method should be developed or better measure of diversity should be found.

Chapter 6

Conclusion

In this master's thesis we explore the ways of finding a tracking portfolio. Including all of the index constituents in the portfolio can be quite expensive if not impossible for the indices which have hundreds of constituents or contains illiquid assets. To replicate an index with the smaller number of assets, we applied several portfolio optimization models in combination with different search space heuristics.

The experimental results show that it is possible to find a smaller set of assets which is able to mimic the index performance. Among the chosen portfolio optimization models, tracking error models revealed the smallest divergence between the index and tracking portfolio returns. The experimental results show that using Widening for the index replication allows finding better tracking portfolios than the most straightforward Hill-Climbing algorithm. Adding diversity to the set of parallel search paths can help to explore disparate regions of the solution space and improve the results (in terms of a certain metric) obtained by a greedy search heuristic. However, the choice of the diversity measure plays an important role. Poor path diversification can hinder the progress towards a better solution.

6.1 Future Work

As seen in the experiments the use of diversity can improve the solution found by a greedy search. Future work can be dedicated to finding a better (or may be even unique for all portfolio optimization models) diversity measure which is able to provide broader exploration of disparate regions of the portfolio solution space. For example, does industry/sector or resource allocation of a company as a diversity measure provide better results? One could use different portfolio optimization models and different portfolio measurements which may yield better results.

Increasing the width of a search can explore more solution space and provide a better solution. However, we can expect that at some point in time, e.g., when a "good enough" approximation of the global solution is found, further traversing of the solution space will not improve the results significantly. As shown by the experiments, increasing

the width of a search often increase the number of assets in the portfolio. Future work could aim to find an optimal search width for the index tracking. At what point is no longer beneficial to increase the search width? When does an obtained solution does not became significantly better but the number of assets is growing? In this case, one can develop a multi-objective optimization problem where the objectives of minimizing the portfolio size and the tracking error are taken into consideration.

In this work, we set the sample universe to be equal to the set of index constituents. However, different combination are possible. One can consider the sample universe as a part of the index universe or choose completely arbitrary assets. Future work could explore whether in these cases it is still possible to replicate an index. If so, future work can explore the relationship to the size of the portfolio solution. Furthermore, different indices can be considered. For example, future work can explore whether it is possible to track commodity indices using Widening.

Rebalancing is an important part of index tracking which allows incorporating transaction costs. In this thesis, we consider rebalancing at a fixed point in time. However, one can analyze the difference between the index and portfolio returns over time and make portfolio rebalancing when it is necessary. Future work could enrich this thesis by developing a rebalancing strategy and embedding it in Beam Search Index Tracking and Widened Index Tracking for the multiperiod index tracking.

Appendix A

Out-of-Sample Statistics

Index	Method	Hill-Climbing					Beam Search (beam width $k = 5$)				
	Measure	GMVP	MV	TEV	MSE	TEMV	GMVP	MV	TEV	MSE	TEMV
SMI ($n = 20$)	p	6.82	6.09	15.09	15.09	17.36	6.91	6.55	17.09	17.09	16.82
	MAE $\times 10^2$	0.2229	0.2155	0.1093	0.1094	0.0406	0.2148	0.2115	0.0389	0.0390	0.0399
	RMSE $\times 10^2$	0.3038	0.3085	0.1562	0.1564	0.0543	0.2931	0.3042	0.0532	0.0533	0.0537
	MSE $\times 10^4$	0.0967	0.1050	0.0439	0.0440	0.0032	0.0895	0.1019	0.0031	0.0031	0.0030
	TE VAR $\times 10^4$	0.0977	0.1065	0.0437	0.0438	0.0032	0.0902	0.1032	0.0030	0.0030	0.0030
	ER $\times 10^4$	0.0064	0.4802	1.4819	1.4716	-0.0421	-0.0991	0.0430	0.0557	0.0544	0.0226
	$CE(\boldsymbol{P})(\gamma = 1) \times 10^4$	4.4704	4.9259	5.9587	5.9484	4.4480	4.3695	4.4935	4.5458	4.5446	4.5128
	$\Delta_{CE}(\boldsymbol{P}, \boldsymbol{I})(\gamma = 1) \times 10^4$	-0.0235	0.4320	1.4648	1.4545	-0.0459	-0.1244	-0.0004	0.0519	0.0507	0.0190
	β_P	0.9358	0.9728	0.9843	0.9840	1.0085	0.9387	0.9656	1.0091	1.0091	1.0081
DAX ($n = 28$)	p	6.91	6.82	21.27	21.27	19.73	7.55	8.27	25.00	25.09	24.09
	MAE $\times 10^2$	0.2391	0.2376	0.0881	0.0881	0.1132	0.2134	0.1920	0.0584	0.0584	0.0668
	RMSE $\times 10^2$	0.3205	0.3105	0.1169	0.1169	0.1572	0.2834	0.2487	0.0823	0.0821	0.0918
	MSE $\times 10^4$	0.1063	0.1120	0.0158	0.0157	0.0537	0.0862	0.0637	0.0081	0.0081	0.0104
	TE VAR $\times 10^4$	0.1061	0.1119	0.0155	0.0155	0.0543	0.0855	0.0623	0.0078	0.0078	0.0100
	ER $\times 10^4$	-1.3571	-1.6794	-1.2767	-1.2778	-0.9073	-1.2545	-2.4245	-1.2696	-1.2616	-1.8664
	$CE(\boldsymbol{P})(\gamma = 1) \times 10^4$	2.9712	2.6603	3.1131	3.1120	3.4588	3.0945	1.9431	3.1235	3.1318	2.5260
	$\Delta_{CE}(\boldsymbol{P}, \boldsymbol{I})(\gamma = 1) \times 10^4$	-1.4212	-1.7321	-1.2793	-1.2804	-0.9336	-1.2979	-2.4493	-1.2689	-1.2606	-1.8664
	β_P	1.0269	0.9891	0.9917	0.9916	0.9969	1.0076	0.9802	0.9881	0.9876	0.9878
DJIA ($n = 29$)	p	8.82	8.36	20.64	20.09	21.82	9.36	9.27	24.45	24.45	24.36
	MAE $\times 10^2$	0.2414	0.2254	0.0865	0.0906	0.0749	0.2158	0.2237	0.0513	0.0515	0.0531
	RMSE $\times 10^2$	0.3234	0.3078	0.1160	0.1213	0.0992	0.2956	0.2962	0.0667	0.0670	0.0675
	MSE $\times 10^4$	0.1103	0.0975	0.0157	0.0165	0.0124	0.0911	0.0911	0.0058	0.0058	0.0047
	TE VAR $\times 10^4$	0.1089	0.0980	0.0155	0.0163	0.0123	0.0894	0.0900	0.0056	0.0056	0.0047
	ER $\times 10^4$	-4.0130	-1.3336	-1.1702	-1.2113	-0.5582	-2.6767	-3.4742	-0.4144	-0.4270	-0.1165
	$CE(\boldsymbol{P})(\gamma = 1) \times 10^4$	5.2743	7.9488	8.1424	8.1018	8.7522	6.6217	5.8107	8.9010	8.8884	9.2010
	$\Delta_{CE}(\boldsymbol{P}, \boldsymbol{I})(\gamma = 1) \times 10^4$	-4.0379	-1.3634	-1.1698	-1.2104	-0.5600	-2.6905	-3.5015	-0.4112	-0.4238	-0.1112
	β_P	0.8312	0.8897	0.9518	0.9473	0.9786	0.8243	0.8932	0.9659	0.9657	0.9596

Table A.1: Investing phase: detailed out-of-sample statistics for small datasets.

Index	Diversity measure	$d_{Pearson}(\boldsymbol{r}(\boldsymbol{P}_i), \boldsymbol{r}(\boldsymbol{P}_j))$										$d_{Euclidean}(SSPW_i, SSPW_j)$									
		p-dispersion-sum					p-dispersion-min-sum					p-dispersion-sum					p-dispersion-min-sum				
	Measure	GMVP	MV	TEV	MSE	TEMV	GMVP	MV	TEV	MSE	TEMV	GMVP	MV	TEV	MSE	TEMV	GMVP	MV	TEV	MSE	TEMV
SMI ($n = 20$)	p	11.45	2.27	19.00	19.09	5.45	12.00	4.64	19.09	19.09	9.00	9.27	8.73	19.18	19.00	18.45	9.64	10.09	19.27	19.09	18.82
	MAE $\times 10^2$	0.3111	0.4040	0.0348	0.0351	0.3497	0.3061	0.4130	0.0353	0.0354	0.3112	0.2623	0.2482	0.0506	0.0500	0.0382	0.2616	0.2219	0.0345	0.0343	0.0387
	RMSE $\times 10^2$	0.4199	0.5863	0.0472	0.0477	0.5128	0.3977	0.5895	0.0479	0.0481	0.4563	0.3377	0.3344	0.0707	0.0698	0.0514	0.3471	0.2995	0.0467	0.0471	0.0519
	MSE $\times 10^4$	0.1855	0.3863	0.0024	0.0024	0.3546	0.1642	0.4173	0.0025	0.0025	0.3596	0.1212	0.1248	0.0102	0.0102	0.0028	0.1220	0.0950	0.0023	0.0024	0.0028
	TE VAR $\times 10^4$	0.1873	0.3901	0.0024	0.0024	0.3582	0.1648	0.4209	0.0024	0.0025	0.3625	0.1227	0.1262	0.0102	0.0101	0.0028	0.1237	0.0954	0.0023	0.0024	0.0028
	ER $\times 10^4$	1.0041	0.0671	-0.0056	-0.0116	-0.1317	1.7659	-0.5387	-0.0259	-0.0074	-1.2407	0.8979	-0.4396	0.4065	0.4131	-0.0640	0.8661	0.2444	-0.0140	0.0934	-0.0358
	$CE(\boldsymbol{P})$ ($\gamma = 1$) $\times 10^4$	5.4441	4.3439	4.4849	4.4788	4.1493	6.2193	3.7755	4.4637	4.4823	3.0807	5.3804	4.0128	4.8942	4.9005	4.4263	5.3335	4.7067	4.4763	4.5828	4.4547
	$\Delta_{CE}(\boldsymbol{P}, \boldsymbol{I})$ ($\gamma = 1$) $\times 10^4$	0.9502	-0.1500	-0.0090	-0.0151	-0.3446	1.7255	-0.7184	-0.0302	-0.0116	-1.4132	0.8865	-0.4811	0.4003	0.4066	-0.0676	0.8396	0.2128	-0.0176	0.0889	-0.0391
	β_P	0.8634	1.0689	1.0083	1.0085	1.1153	0.8576	0.8873	1.0113	1.0108	0.9630	0.8543	0.9334	1.0062	1.0067	1.0080	0.8818	0.9441	1.0088	1.0117	1.0070
DAX ($n = 28$)	p	6.82	6.18	26.91	26.91	24.45	11.27	4.64	27.27	27.27	15.45	8.09	8.64	27.18	26.73	27.18	9.45	10.27	26.73	27.18	27.00
	MAE $\times 10^2$	0.3844	0.4060	0.0440	0.0441	0.1237	0.4216	0.4706	0.0472	0.0468	0.3008	0.3465	0.2180	0.0442	0.0447	0.0463	0.3430	0.2882	0.0446	0.0448	0.0464
	RMSE $\times 10^2$	0.5317	0.5789	0.0662	0.0667	0.2039	0.5599	0.6317	0.0683	0.0677	0.4265	0.4650	0.2885	0.0664	0.0665	0.0697	0.4693	0.3712	0.0667	0.0664	0.0697
	MSE $\times 10^4$	0.2954	0.4396	0.0051	0.0052	0.2259	0.3201	0.4465	0.0054	0.0054	0.3314	0.2250	0.0851	0.0052	0.0053	0.0057	0.2338	0.1427	0.0053	0.0053	0.0056
	TE VAR $\times 10^4$	0.2964	0.4411	0.0048	0.0049	0.2266	0.3215	0.4447	0.0051	0.0051	0.3309	0.2248	0.0851	0.0049	0.0050	0.0054	0.2359	0.1438	0.0049	0.0050	0.0052
	ER $\times 10^4$	0.1734	2.2568	-1.2655	-1.2646	0.6293	2.6737	0.5524	-1.2458	-1.2352	-1.2019	-1.4037	-0.3982	-1.1566	-1.2117	-1.2590	0.1623	-1.4144	-1.3233	-1.1579	-1.3705
	$CE(\boldsymbol{P})$ ($\gamma = 1$) $\times 10^4$	4.4631	6.4764	3.1319	3.1322	4.9239	6.9540	4.7725	3.1506	3.1610	3.0484	2.9136	3.9676	3.2396	3.1853	3.1388	4.4794	2.9389	3.0732	3.2385	3.0258
	$\Delta_{CE}(\boldsymbol{P}, \boldsymbol{I})$ ($\gamma = 1$) $\times 10^4$	0.0707	2.0840	-1.2606	-1.2603	0.5315	2.5615	0.3801	-1.2418	-1.2314	-1.3440	-1.4788	-0.4248	-1.1529	-1.2071	-1.2537	0.0870	-1.4535	-1.3193	-1.1539	-1.3666
	β_P	0.8972	0.9017	0.9827	0.9839	0.9654	0.9040	0.8809	0.9842	0.9842	0.9351	0.9080	0.9587	0.9848	0.9827	0.9817	0.8996	0.9233	0.9846	0.9845	0.9849
DJIA ($n = 29$)	p	10.82	9.55	27.73	27.55	25.00	11.73	2.45	28.27	28.45	5.73	11.09	11.82	28.09	27.91	27.45	13.82	13.91	28.09	27.82	27.18
	MAE $\times 10^2$	0.2997	0.3084	0.0372	0.0375	0.0913	0.2951	0.4935	0.0367	0.0374	0.4535	0.2432	0.2477	0.0367	0.0365	0.0468	0.2594	0.2367	0.0369	0.0375	0.0478
	RMSE $\times 10^2$	0.3944	0.4105	0.0477	0.0488	0.1285	0.3845	0.6892	0.0466	0.0473	0.6369	0.3156	0.3275	0.0469	0.0467	0.0595	0.3397	0.3108	0.0475	0.0477	0.0601
	MSE $\times 10^4$	0.1576	0.1869	0.0023	0.0024	0.0645	0.1490	0.5157	0.0022	0.0023	0.4944	0.1012	0.1156	0.0022	0.0022	0.0036	0.1204	0.0995	0.0023	0.0023	0.0037
	TE VAR $\times 10^4$	0.1568	0.1840	0.0023	0.0024	0.0608	0.1485	0.5090	0.0022	0.0023	0.4878	0.1017	0.1159	0.0022	0.0022	0.0035	0.1186	0.0997	0.0023	0.0023	0.0037
	ER $\times 10^4$	-1.6742	1.8893	-0.4425	-0.3345	1.2844	-1.7981	-1.8614	-0.3493	-0.3888	-1.4645	-0.8586	-2.2522	-0.1742	-0.2815	-0.8755	-2.8413	-0.6668	-0.4542	-0.3377	-0.4940
	$CE(\boldsymbol{P})$ ($\gamma = 1$) $\times 10^4$	7.6092	11.1599	8.8762	8.9839	10.5762	7.5005	7.2740	8.9695	8.9302	7.6708	8.4307	7.0390	9.1446	9.0372	8.4442	6.4555	8.6383	8.8650	8.9809	8.8249
	$\Delta_{CE}(\boldsymbol{P}, \boldsymbol{I})$ ($\gamma = 1$) $\times 10^4$	-1.7030	1.8477	-0.4360	-0.3283	1.2640	-1.8117	-2.0382	-0.3427	-0.3820	-1.6414	-0.8815	-2.2731	-0.1676	-0.2749	-0.8679	-2.8567	-0.6739	-0.4472	-0.3313	-0.4872
	β_P	0.7212	0.7184	0.9585	0.9603	0.9484	0.6600	0.5924	0.9585	0.9577	0.6591	0.8343	0.7909	0.9585	0.9587	0.9505	0.7520	0.7614	0.9559	0.9591	0.9546

Table A.2: Investing phase. DIVERSITY-DRIVEN WIDENING: detailed out-of-sample statistics for small datasets. The search width for WIDENED INDEX TRACKING is $k = 5$.

Index	Method	Hill-Climbing					Beam Search					Widened Index Tracking									
												p-sum					p-min-sum				
	Measure	GMVP	MV	TEV	MSE	TEMV	GMVP	MV	TEV	MSE	TEMV	GMVP	MV	TEV	MSE	TEMV	GMVP	MV	TEV	MSE	TEMV
STOXX50 (n = 50)	p	7.1818	7.73	25.27	25.36	22.55	7.82	8.55	28.36	29.55	27.09	10.09	9.00	44.73	44.64	42.00	9.27	8.00	44.91	47.45	35.36
	MAE $\times 10^2$	0.2898	0.2698	0.0970	0.0971	0.1520	0.2885	0.2541	0.0884	0.0855	0.1044	0.2714	0.2815	0.0487	0.0497	0.0718	0.2878	0.2987	0.0453	0.0426	0.1203
	RMSE $\times 10^2$	0.3813	0.3634	0.1249	0.1228	0.2050	0.3887	0.3553	0.1195	0.1100	0.1374	0.3541	0.3672	0.0623	0.0634	0.0953	0.3818	0.3965	0.0635	0.0553	0.1597
	MSE $\times 10^4$	0.1467	0.1447	0.0168	0.0161	0.0662	0.1548	0.1356	0.0193	0.0135	0.0198	0.1281	0.1383	0.0045	0.0045	0.0092	0.1531	0.1694	0.0055	0.0032	0.0570
	TE VAR $\times 10^4$	0.1478	0.1453	0.0168	0.0162	0.0666	0.1566	0.1362	0.0193	0.0137	0.0199	0.1283	0.1377	0.0046	0.0045	0.0093	0.1544	0.1702	0.0055	0.0032	0.0571
	ER $\times 10^4$	-0.2812	-2.2003	0.5749	0.0555	-0.2746	0.3097	-0.9103	0.6509	-0.0285	-0.2747	0.8173	-2.8451	-0.1595	-0.0084	0.0323	1.0927	-1.2961	0.2300	-0.0085	0.6354
	$CE(\boldsymbol{P})(\gamma = 1) \times 10^4$	2.6425	0.7353	3.5473	3.0319	2.6768	3.2292	2.0328	3.6255	2.9461	2.7048	3.7810	0.1424	2.8193	2.9729	3.0123	4.0315	1.6498	3.2093	2.9708	3.5858
	$\Delta_{CE}(\boldsymbol{P}, \boldsymbol{I})(\gamma = 1) \times 10^4$	-0.3378	-2.2449	0.5670	0.0516	-0.3034	0.2489	-0.9475	0.6452	-0.0342	-0.2754	0.8007	-2.8378	-0.1610	-0.0074	0.0321	1.0512	-1.3304	0.2291	-0.0094	0.6055
	β_P	0.9661	0.9203	0.9954	0.9848	0.9797	0.9575	0.9172	0.9861	0.9922	0.9708	0.8962	0.8032	0.9956	0.9890	0.9825	0.9218	0.8609	0.9932	0.9946	0.9970
NASDAQ (n = 101)	p	14.27	13.73	23.91	26.18	28.91	16.09	16.55	30.36	29.36	30.73	16.27	13.73	65.18	51.27	49.27	16.36	17.45	55.64	50.55	52.64
	MAE $\times 10^2$	0.2658	0.2405	0.1625	0.1580	0.1494	0.2096	0.2094	0.1263	0.1277	0.1478	0.3252	0.3289	0.0624	0.0792	0.1093	0.3882	0.3650	0.0759	0.0802	0.1042
	RMSE $\times 10^2$	0.3739	0.3275	0.2151	0.2103	0.1927	0.2871	0.2787	0.1644	0.1673	0.1938	0.4537	0.4780	0.0828	0.1017	0.1444	0.5527	0.4948	0.1011	0.1064	0.1408
	MSE $\times 10^4$	0.1424	0.1092	0.0483	0.0487	0.0387	0.0844	0.0791	0.0282	0.0292	0.0390	0.2178	0.2736	0.0070	0.0108	0.0230	0.3119	0.2587	0.0112	0.0116	0.0206
	TE VAR $\times 10^4$	0.1427	0.1087	0.0485	0.0489	0.0385	0.0828	0.0785	0.0282	0.0292	0.0392	0.2149	0.2743	0.0070	0.0108	0.0230	0.3158	0.2607	0.0112	0.0117	0.0206
	ER $\times 10^4$	0.2272	2.8551	0.6027	0.6817	0.0509	2.8369	1.1908	0.3017	0.3609	-0.4628	-4.5366	-1.9185	-0.1189	-0.5210	-0.5305	-2.1696	-1.5773	-0.3466	-0.5399	-1.1492
	$CE(\boldsymbol{P})(\gamma 1) \times 10^4$	10.2675	12.9013	10.6805	10.7590	10.1396	12.8758	11.2534	10.3889	10.4467	9.6330	5.5599	8.1168	9.9845	9.5800	9.5784	7.9283	8.5157	9.7512	9.5579	8.9640
	$\Delta_{CE}(\boldsymbol{P}, \boldsymbol{I})(\gamma = 1) \times 10^4$	0.1858	2.8196	0.5987	0.6773	0.0578	2.7941	1.1717	0.3072	0.3649	-0.4487	-4.5218	-1.9649	-0.0972	-0.5018	-0.5033	-2.1534	-1.5660	-0.3305	-0.5238	-1.1177
	β_P	0.9433	0.9625	0.9713	0.9698	0.9470	1.0035	0.9527	0.9636	0.9656	0.9248	0.7326	0.8116	0.9461	0.9443	0.9116	0.6129	0.6949	0.9505	0.9568	0.9077
S&P (n = 495)	p	27.82	23.64	33.27	31.82	28.91	29.27	31.31	35.64	37.17	37.55	21.00	18.00	61.00	59.60	56.00	10.80	16.40	57.20	53.80	51.60
	MAE $\times 10^2$	0.2533	0.2392	0.2108	0.2045	0.2357	0.2274	0.2145	0.1822	0.1760	0.1978	0.2561	0.2730	0.0902	0.0858	0.1004	0.2849	0.2864	0.0923	0.0975	0.1073
	RMSE $\times 10^2$	0.3384	0.3123	0.2764	0.2691	0.3042	0.3056	0.2855	0.2508	0.2391	0.2534	0.3226	0.3488	0.1138	0.1092	0.1299	0.4011	0.3840	0.1226	0.1290	0.1424
	MSE $\times 10^4$	0.1158	0.1021	0.0813	0.0763	0.0958	0.0962	0.0838	0.0673	0.0616	0.0689	0.1090	0.1278	0.0132	0.0124	0.0175	0.1646	0.1534	0.0154	0.0174	0.0207
	TE VAR $\times 10^4$	0.1166	0.1019	0.0809	0.0766	0.0962	0.0963	0.0840	0.0672	0.0619	0.0696	0.1089	0.1287	0.0133	0.0124	0.0177	0.1654	0.1559	0.0156	0.0176	0.0206
	ER $\times 10^4$	-0.8790	-2.6329	-1.0069	-1.1775	-1.0728	-0.4209	0.5709	0.6182	-0.8645	-0.1538	-1.8156	-2.7908	-0.0702	-0.5198	-0.2246	0.4121	-0.1521	0.3280	0.5802	-1.5890
	$CE(\boldsymbol{P})(\gamma = 1) \times 10^4$	6.1368	4.3918	6.0129	5.8425	5.9488	6.5913	7.0057	7.6289	6.0771	6.8753	3.3233	2.3285	5.0889	4.6394	4.9384	5.5506	4.9561	5.4837	5.7356	3.5727
	$\Delta_{CE}(\boldsymbol{P}, \boldsymbol{I})(\gamma = 1) \times 10^4$	-0.9074	-2.6524	-1.0313	-1.2016	-1.0953	-0.4529	0.5512	0.5848	-0.8939	-0.1689	-1.8338	-2.8286	-0.0682	-0.5178	-0.2187	0.3934	-0.2010	0.3265	0.5784	-1.5844
	β_P	0.8250	0.8256	0.9108	0.9224	0.8728	0.9098	0.8852	0.9941	0.9852	0.8919	0.8116	0.8750	0.9577	0.9599	0.9295	0.6788	0.8641	0.9721	0.9650	0.9269

Table A.3: Investing phase: detailed out-of-sample statistics for the datasets with $n \geqslant 50$. The search width for Beam Search and Widened Index Tracking is $k = 5$. Diversity is measured using Euclidean distance between the sums of squared portfolio weights.

Method	Hill-Climbing					Beam Search					Widened Index Tracking				
Model	GMVP	MV	TEV	MSE	TEMV	GMVP	MV	TEV	MSE	TEMV	GMVP	MV	TEV	MSE	TEMV
p	8.40	10.60	12.40	17.00	21.40	13.40	12.80	27.40	27.20	28.20	12.40	12.40	32.20	41.20	43.40
TO	0.473	0.514	0.830	0.499	0.253	0.348	0.303	0.140	0.165	0.216	1.620	1.663	0.816	0.571	0.573
MAE $\times 10^2$	0.2888	0.2611	0.1950	0.2059	0.1791	0.2362	0.2064	0.1430	0.1322	0.1549	0.3864	0.3647	0.1506	0.0923	0.1173
RMSE $\times 10^2$	0.3965	0.3468	0.2567	0.2729	0.2499	0.3424	0.2755	0.1908	0.1749	0.2034	0.5614	0.5173	0.1992	0.1222	0.1706
MSE $\times 10^4$	0.1604	0.1306	0.0689	0.0819	0.0733	0.1183	0.0786	0.0380	0.0313	0.0454	0.3262	0.2792	0.0512	0.0153	0.0298
TE VAR $\times 10^4$	0.1613	0.1307	0.0700	0.0825	0.0732	0.1195	0.0786	0.0383	0.0315	0.0444	0.3280	0.2820	0.0519	0.0153	0.0300
ER $\times 10^4$	0.0117	-2.9816	-0.2706	-1.6856	-2.0833	-1.4969	-1.3700	-0.6670	-0.2506	-3.4596	-2.8388	-0.8933	1.0101	-0.4886	-0.9934
$CE(\boldsymbol{P})(\gamma = 1)$ $\times 10^2$	11.1445	8.2482	10.9699	9.5514	9.1824	9.7010	9.8319	10.5740	10.9920	7.8182	8.4535	10.4067	12.2342	10.7840	10.3088
$\Delta_{CE}(\boldsymbol{P}, \boldsymbol{I})$ $(\gamma = 1)$ $\times 10^4$	-0.1058	-3.0021	-0.2803	-1.6988	-2.0678	-1.5492	-1.4183	-0.6762	-0.2582	-3.4321	-2.7967	-0.8435	0.9839	-0.4663	-0.9414
β_P	1.0795	0.9133	0.9466	0.9380	0.8959	0.9852	1.0199	0.9771	0.9814	0.8957	0.5540	0.5815	0.9992	0.9355	0.8537

Table A.4: Rebalancing phase: detailed out-of-sample statistics for the NASDAQ index tracking. The search width for Beam Search and Widened Index Tracking is $k = 5$. Diversity is measured using Euclidean distance between the sums of squared portfolio weights.

Bibliography

Adcock, C. and Meade, N. (1994). A simple algorithm to incorporate transactions costs in quadratic optimisation. *European Journal of Operational Research*, 79(1):85 – 94.

Akbar, Z., Ivanova, V. N., and Berthold, M. R. (2012). Parallel data mining revisited. better, not faster. In Hollmén, J., Klawonn, F., and Tucker, A., editors, *Advances in Intelligent Data Analysis XI - 11th International Symposium, IDA 2012, Helsinki, Finland, October 25-27, 2012. Proceedings*, volume 7619 of *Lecture Notes in Computer Science*, pages 23–34. Springer.

Al-Naqi, A., Erdogan, A. T., and Arslan, T. (2013). Adaptive three-dimensional cellular genetic algorithm for balancing exploration and exploitation processes. *Soft Comput.*, 17(7):1145–1157.

Arslan, T., Keymeulen, D., Merodio, D., Benkrid, K., Erdogan, A. T., and Patel, U. D., editors (2010). *2010 NASA/ESA Conference on Adaptive Hardware and Systems, AHS 2010, Anaheim, California, USA, June 15-18, 2010*. IEEE.

Beasley, J. E., Meade, N., and Chang, T.-J. (2003). An evolutionary heuristic for the index tracking problem. *European Journal of Operational Research*, 148(3):621–643.

Berthold, M. R., Borgelt, C., Höppner, F., and Klawonn, F. (2010). *Guide to intelligent data analysis: how to intelligently make sense of real data*. Springer Science & Business Media.

Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Kötter, T., Meinl, T., Ohl, P., Sieb, C., Thiel, K., and Wiswedel, B. (2007). KNIME: The Konstanz Information Miner. In *Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007)*. Springer.

Bosworth, B., Hymans, S., and Modigliani, F. (1975). The stock market and the economy. *Brookings Papers on Economic Activity*, 1975(2):257–300.

Braga, M. D. (2016). The different risk-based approaches to asset allocation. In *Risk-Based Approaches to Asset Allocation*, pages 43–55. Springer.

Canakgoz, N. A. and Beasley, J. E. (2009). Mixed-integer programming approaches for index tracking and enhanced indexation. *European Journal of Operational Research*, 196(1):384–399.

Choueifaty, Y. and Coignard, Y. (2008). Toward maximum diversification. *Journal of Portfolio Management*, 35(1):40.

Churchill, D. and Buro, M. (2013). Portfolio greedy search and simulation for large-scale combat in starcraft. In *Computational Intelligence in Games (CIG), 2013 IEEE Conference on*, pages 1–8. IEEE.

Coleman, T. F. and Li, Y. (2006). Minimizing tracking error while restricting the number of assets. *The journal of risk*, 8(4):33–55.

Dangi, A. (2013). Financial portfolio optimization: Computationally guided agents to investigate, analyse and invest!? Master's thesis, The Center for Modeling and Simulation, University of Pune.

DeMiguel, V., Garlappi, L., and Uppal, R. (2009). Optimal versus naive diversification: How inefficient is the 1-n portfolio strategy? *Review of Financial Studies*, 22(5):1915–1953.

Derigs, U. and Nickel, N.-H. (2004). On a local-search heuristic for a class of tracking error minimization problems in portfolio management. *Annals of Operations Research*, 131(1-4):45–77.

Di Tollo, G. and Maringer, D. (2009). Metaheuristics for the index tracking problem. In *Metaheuristics in the service industry*, pages 127–154. Springer.

Edelkamp, S. and Schroedl, S. (2011). *Heuristic search: theory and applications*. Elsevier.

Edirisinghe, N. C. P. (2013). Index-tracking optimal portfolio selection. *Quantitative Finance Letters*, 1(1):16–20.

Erkut, E. (1990). The discrete p-dispersion problem. *European Journal of Operational Research*, 46(1):48–60.

Erkut, E., Ülküsal, Y., and Yenicerioğlu, O. (1994). A comparison of p-dispersion heuristics. *Computers & operations research*, 21(10):1103–1113.

Fabozzi, F. J. and Pachamanova, D. A. (2016). *Portfolio construction and analytics*. John Wiley & Sons.

Faizliev, A. R., Sidorov, S. P., Mironov, S. V., and Khomchenko, A. A. (2016). Empirical analysis of index tracking error minimization algorithms based on stochastic dominance principle. In *CEUR Workshop Proceedings*, volume 1726, pages 23–34.

Ferri, R. and Benke, A. (2013). A case for index fund portfolios. Investors holding only index funds have a better chance for success. White paper. Portfolio Solutions & Betterment.

Fillbrunn, A. and Berthold, M. R. (2015). Diversity-driven widening of hierarchical agglomerative clustering. In Fromont, É., Bie, T. D., and van Leeuwen, M., editors, *Advances in Intelligent Data Analysis XIV - 14th International Symposium, IDA 2015, Saint Etienne, France, October 22-24, 2015, Proceedings*, volume 9385 of *Lecture Notes in Computer Science*, pages 84–94. Springer.

Fillbrunn, A., Wörteler, L., Grossniklaus, M., and Berthold, M. R. (2017). Bucket selection: A model-independent diverse selection strategy for widening. In *International Symposium on Intelligent Data Analysis*, pages 87–98. Springer.

Focardi, S. M. and Fabozzi, F. J. (2004). A methodology for index tracking based on time-series clustering. *Quantitative Finance*, 4(4):417–425.

Gârleanu, N. and Pedersen, L. H. (2013). Dynamic trading with predictable returns and transaction costs. *The Journal of Finance*, 68(6):2309–2340.

Goel, A., Sharma, A., and Mehra, A. (2018). Index tracking and enhanced indexing using mixed conditional value-at-risk. *Journal of Computational and Applied Mathematics*, 335:361–380.

Goetzmann, W. N. and Kumar, A. (2008). Equity portfolio diversification. *Review of Finance*, 12(3):433–463.

Hegerty, B., Hung, C.-C., and Kasprak, K. (2009). A comparative study on differential evolution and genetic algorithms for some combinatorial problems. In *Proceedings of 8th Mexican International Conference on Artificial Intelligence*, pages 9–13.

Hromkovic, J. (2004). *Algorithmics for Hard Problems - Introduction to Combinatorial Optimization, Randomization, Approximation, and Heuristics, Second Edition*. Texts in Theoretical Computer Science. An EATCS Series. Springer.

Ivanova, V. N. and Berthold, M. R. (2013). Diversity-driven widening. In Tucker, A., Höppner, F., Siebes, A., and Swift, S., editors, *Advances in Intelligent Data Analysis XII - 12th International Symposium, IDA 2013, London, UK, October 17-19, 2013. Proceedings*, volume 8207 of *Lecture Notes in Computer Science*, pages 223–236. Springer.

Jagannathan, R. and Ma, T. (2003). Risk reduction in large portfolios: Why imposing the wrong constraints helps. *The Journal of Finance*, 58(4):1651–1684.

Jeurissen, R. and van den Berg, J. (2005). Index tracking using a hybrid genetic algorithm. In *Computational Intelligence Methods and Applications, 2005 ICSC Congress on*, pages 6–pp. IEEE.

Jeurissen, R. and van den Berg, J. (2008). Optimized index tracking using a hybrid genetic algorithm. In *Proceedings of the IEEE Congress on Evolutionary Computation, CEC 2008, June 1-6, 2008, Hong Kong, China*, pages 2327–2334.

Jin, Y., Qu, R., and Atkin, J. A. D. (2016). Constrained portfolio optimisation: The state-of-the-art Markowitz models. In Vitoriano, B., Parlier, G. H., and de Werra, D., editors, *Proceedings of 5th the International Conference on Operations Research and Enterprise Systems (ICORES 2016), Rome, Italy, February 23-25, 2016.*, pages 388–395. SciTePress.

Karlow, D. (2013). *Comparison and Development of Methods for Index Tracking*. PhD thesis, Frankfurt School of Finance & Management.

Kazak, E. and Pohlmeier, W. (2017). Testing out-of-sample portfolio performance. Technical report, GSDS Working Paper Series.

Krink, T., Mittnik, S., and Paterlini, S. (2009). Differential evolution and combinatorial search for constrained index-tracking. *Annals of Operations Research*, 172(1):153.

Lee, W. (2011). Risk-based asset allocation: A new answer to an old question? *Journal of Portfolio Management*, 37(4):11.

Liu, S.-T. (2011). The mean-absolute deviation portfolio selection problem with interval-valued returns. *Journal of Computational and Applied Mathematics*, 235(14):4149–4157.

Maillard, S., Roncalli, T., and Teïletche, J. (2010). The properties of equally weighted risk contribution portfolios. *The Journal of Portfolio Management*, 36(4):60–70.

Malkiel, B. G. and Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, 25(2):383–417.

Mansini, R., Ogryczak, W., and Speranza, M. G. (2015). Linear models for portfolio optimization. In *Linear and Mixed Integer Programming for Portfolio Optimization*, pages 19–45. Springer.

Maringer, D. and Oyewumi, O. (2007). Index tracking with constrained portfolios. *Int. Syst. in Accounting, Finance and Management*, 15(1-2):57–71.

Markowitz, H. (1952). Portfolio selection. *The Journal of Finance*, 7(1):77–91.

Markowitz, H. (1987). *Mean-variance analysis in portfolio choice and capital markets*. Blackwell.

Mei, X., DeMiguel, V., and Nogales, F. J. (2016). Multiperiod portfolio optimization with multiple risky assets and general transaction costs. *Journal of Banking & Finance*, 69:108–120.

Mei, X. and Nogales, F. (2018). Portfolio selection with proportional transaction costs and predictability. *Journal of Banking & Finance*, 94:131–151.

Meinl, T. (2010). *Maximum-score diversity selection*. PhD thesis, University of Konstanz.

Miziołek, T. (2018). Index providers in the global financial market. *Acta Universitatis Lodziensis. Folia oeconomica*, 3(335):139–152.

Moallemi, C. C. and Salam, M. (2017). Dynamic portfolio choice with linear rebalancing rules. *Journal of Financial and Quantitative Analysis*, 52(3):12471278.

Oh, K. J., Kim, T. Y., and Min, S. (2005). Using genetic algorithm to support portfolio optimization for index fund management. *Expert Systems with Applications*, 28(2):371–379.

Olivares-Nadal, A. V. and DeMiguel, V. (2018). A robust perspective on transaction costs in portfolio optimization. *Operations Research*, 66(3):733–739.

Papenbrock, J. (2011). *Asset Clusters and Asset Networks in Financial Risk Management and Portfolio Optimization*. PhD thesis, Karlsruhe Institute of Technology.

Plyakha, Y., Uppal, R., and Vilkov, G. (2012). Why does an equal-weighted portfolio outperform value-and price-weighted portfolios? Working paper, EDHEC-Risk Institute.

Rafaely, B. and Bennell, J. A. (2006). Optimisation of ftse 100 tracker funds: A comparison of genetic algorithms and quadratic programming. *Managerial Finance*, 32(6):477–492.

Reilly, F. K. and Brown, K. C. (2011). *Investment analysis and portfolio management*. Cengage Learning.

Ren, F., Lu, Y.-N., Li, S.-P., Jiang, X.-F., Zhong, L.-X., and Qiu, T. (2017). Dynamic portfolio strategy using clustering approach. *PloS one*, 12(1):e0169299.

Ross, S. A., Westerfield, R., and Jordan, B. D. (2008). *Fundamentals of corporate finance*. Tata McGraw-Hill Education.

Rudolf, M., Wolter, H.-J., and Zimmermann, H. (1999). A linear model for tracking error minimization. *Journal of Banking & Finance*, 23(1):85 – 103.

Ruiz-Torrubiano, R. and Suárez, A. (2009). A hybrid optimization approach to index tracking. *Annals of Operations Research*, 166(1):57–71.

Sampson, O. (2013). Diversity driven parallel data mining. Master's thesis, University of Konstanz.

Sampson, O. and Berthold, M. R. (2014). Widened KRIMP: better performance through diverse parallelism. In Blockeel, H., van Leeuwen, M., and Vinciotti, V., editors, *Advances in Intelligent Data Analysis XIII - 13th International Symposium, IDA 2014, Leuven, Belgium, October 30 - November 1, 2014. Proceedings*, volume 8819 of *Lecture Notes in Computer Science*, pages 276–285. Springer.

Sampson, O. R. and Berthold, M. R. (2016). Widened learning of Bayesian network classifiers. In Boström, H., Knobbe, A. J., Soares, C., and Papapetrou, P., editors, *Advances in Intelligent Data Analysis XV - 15th International Symposium, IDA 2016, Stockholm, Sweden, October 13-15, 2016, Proceedings*, volume 9897 of *Lecture Notes in Computer Science*, pages 215–225.

Sampson, O. R., Borgelt, C., and Berthold, M. R. (2018). Communication-free widened learning of Bayesian network classifiers using hashed Fiedler vectors. In *Advances in Intelligent Data Analysis XVII*. Springer.

Sant'Anna, L. R., Filomena, T. P., Guedes, P. C., and Borenstein, D. (2017). Index tracking with controlled number of assets using a hybrid heuristic combining genetic algorithm and non-linear programming. *Annals of Operations Research*, 258(2):849–867.

Scozzari, A., Tardella, F., Paterlini, S., and Krink, T. (2013). Exact and heuristic approaches for the index tracking problem with ucits constraints. *Annals of Operations Research*, 205(1):235–250.

Shapcott, J. (1992). Index tracking: genetic algorithms for investment portfolio selection. *Edinburgh Parallel Computing Centre*.

Shen, W. and Wang, J. (2017). Portfolio selection via subset resampling. In Singh, S. P. and Markovitch, S., editors, *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA.*, pages 1517–1523. AAAI Press.

Soe, A. M. and Poirier, R. (2016). SPIVA US Scorecard. *S&P Dow Jones Indices, McGraw Hill Financial*.

Stoyan, S. J. and Kwon, R. H. (2010). A two-stage stochastic mixed-integer programming approach to the index tracking problem. *Optimization and Engineering*, 11(2):247–275.

Sullivan, A. and Sheffrin, S. M. (2003). Economics: Principles in action. *Upper Saddle River, New Jersey*, 7458:29.

Takeda, A., Niranjan, M., Gotoh, J.-y., and Kawahara, Y. (2013). Simultaneous pursuit of out-of-sample performance and sparsity in index tracking portfolios. *Computational Management Science*, 10(1):21–49.

Taylor, N. (2015). Managed portfolio performance and transaction costs. *Applied Economics Letters*, 22(4):272–280.

Tušar, T. and Filipič, B. (2007). Differential evolution versus genetic algorithms in multiobjective optimization. In *International Conference on Evolutionary Multi-Criterion Optimization*, pages 257–271. Springer.

Wang, M., Xu, C., Xu, F., and Xue, H. (2012). A mixed 0–1 lp for index tracking problem with cvar risk constraints. *Annals of Operations Research*, 196(1):591–609.

Wilt, C. M., Thayer, J. T., and Ruml, W. (2010). A comparison of greedy search algorithms. In Felner, A. and Sturtevant, N. R., editors, *Proceedings of the Third Annual Symposium on Combinatorial Search, SOCS 2010, Stone Mountain, Atlanta, Georgia, USA, July 8-10, 2010*. AAAI Press.

Yu, L., Zhang, S., and Zhou, X. Y. (2006). A downside risk analysis based on financial index tracking models. In *Stochastic finance*, pages 213–236. Springer.