Faculty of Mathematics and Informatics

VILNIUS UNIVERSITY
FACULTY OF MATHEMATICS AND INFORMATICS
MODELLING AND DATA ANALYSIS
MASTER'S STUDY PROGRAMME

A MODERN APPROACH TO LARGE-SCALE PORTFOLIO OPTIMIZATION

Master's thesis

Author: Mindaugas Kazimieras Stagys
VU email address: mindaugas.stagys@mif.stud.vu.lt
Supervisor: prof. habil. dr. Remigijus Leipus

Vilnius

2023

Contents

Abstract

Following the global financial crisis, researchers and practitioners have paid close attention to risk-based asset allocation strategies, which do not depend on the calculation of expected returns and are therefore viewed as more stable than the standard mean-variance framework. However, as the number of investable assets grows, so does the complexity of the optimum asset allocation problem. Traditional risk-based allocation techniques involve the inversion of a potentially ill-conditioned covariance matrix, which in turn results in an amplification of estimation errors. The optimization problem also becomes computationally challenging when the portfolio consists of more than a few hundred assets. In order to solve the aforementioned problems, modern portfolio optimization methods have introduced clustering-based allocation approaches. In this thesis, we examine various risk-based optimization strategies, clustering algorithms, and covariance matrix estimation methods in terms of their contribution to portfolio risk and risk-adjusted returns. The empirical study is performed on 100 randomly sampled 350-asset portfolios featuring realistic diversification across 11 sectors. Based on overall risk characteristics and risk-adjusted performance, this thesis suggests a combination of nested global minimum variance optimization, partitioning around medoids with dynamic time warping distance, and a Markov switching multifractal model with dynamic conditional correlation type structure and nonlinear shrinkage. Nonetheless, this choice heavily depends on the investor's risk profile as well as desired portfolio turnover and weight concentration.

Keywords: portfolio optimization, covariance matrix, clustering

1 Introduction

1.1 Related Works

Optimal portfolio construction refers to the process of efficiently distributing capital among a given set of financial assets [62]. The field of portfolio construction has been extensively studied by both academics and practitioners since the 1950s, when Markowitz first introduced his pioneering mean-variance approach to portfolio construction [65, 66], and continues to be the subject of extensive research. While the mean-variance portfolio is theoretically optimal, nowadays it is often considered flawed and unreliable in practice [48]. In essence, there are two fundamental weaknesses with Markowitz's approach to portfolio construction [2]:

1. The tendency to amplify the effects of errors in the input assumptions causing significant portfolio misallocations [72];

2. The possibility of substantial portfolio weight changes between investment periods, resulting in hefty transaction costs. Many researchers proposed penalizing the mean-variance optimization function with transaction costs as the remedy to the problem [8, 56, 74, 59, 68]. However, such methods tend to be opaque in practice [2].

While the covariances of a few assets can be adequately assessed, predicting expected returns with decent accuracy appears to be considerably more challenging [69, 21, 44]. This issue has led to a rise in the usage of risk-based optimization techniques. In particular, Markowitz's mean-variance portfolio was heavily criticized following the global financial crisis, as most institutional portfolios were mean-variance optimal yet nonetheless demonstrated poor performance, resulting in severe losses for investors all over the world [90]. Since then, a sizable proportion of investors have preferred risk-based asset allocation strategies over methods that rely on estimating expected returns. As a result, a substantial amount of research has been conducted on the performance of these alternative portfolio construction strategies that do not take returns into account. Some of the notable and more popular risk-based methods include the minimum variance [24], maximum diversification [22], equal risk contribution [62], and risk budgeting [13].

Despite the fact that risk-based allocation techniques do not rely on the assessment of expected returns, they nevertheless involve the inversion of a potentially ill-conditioned covariance matrix. In practice, the covariance matrix might become ill-conditioned when the assets are highly correlated or when the number of assets is close to the number of observations. In turn, the inversion of the ill-conditioned matrix often leads to an amplification of inherent estimation errors and ultimately results in a numerically unstable portfolio [6]. In fact, studies have been able to demonstrate that the estimation errors may be large enough to offset any benefits of diversification [27]. This phenomenon is normally referred to as Markowitz's curse, and naturally, the resulting portfolio is characterized by underperformance under these conditions [60].

As a consequence of the unreliability and underperformance of both Markowitz's mean-variance portfolio and risk-based optimization strategies, a significant amount of research has been devoted to enhancing the robustness and reducing the numerical instability of the covariance matrix. These efforts could be classified into two distinct directions of research. The first approach is concerned with the development of a more robust covariance estimation, with many proposed solutions, such as various shrinkage methods [50, 51], the inclusion of time-varying volatility and correlation structure [75], or solutions from the field of random matrix theory [96, 25]. The second strategy sidesteps the covariance matrix inversion by employing clustering techniques.

In 2016, Lopez de Prado developed a novel asset allocation approach in an attempt to mitigate the limitations of conventional optimization algorithms [60]. The suggested Hierarchical Risk Parity (HRP) method sought to address portfolio instability and reduce its concentration by combining hierarchical clustering of the covariance matrix with a heuristic risk-based allocation. Using Monte Carlo simulations, Lopez de Prado was able to demonstrate that HRP constructs portfolios with lower out-of-sample variance and higher return than both minimum-variance and inverse-variance portfolios. Burggraf and Vyas [14] tried to further improve HRP robustness by applying covariance matrix shrinkage. It was shown that HRP could achieve the most desirable diversification properties, while inverse volatility portfolios tended to be too static and mean-variance portfolios were too concentrated. Lohre et al. [57] have considered two different distance measures: the distance measure used by Lopez de Prado and a distance measure based on the lower tail dependence coefficient. This study confirmed that altering the distance measure might improve tail risk management. However, the results of both studies also suggested that improvements come at the cost of a higher turnover.

Raffinot [85] further extended HRP by proposing Hierarchical Clustering Asset Allocation (HCAA), which included several hierarchical clustering linkage criteria as well as estimation of the optimal number of clusters using the Gap index. The analysis of different linking criteria was unfortunately inconclusive. According to Raffinot, the estimation of the correlation matrix is the most crucial step, and the use of shrinkage should be further explored. In a later paper [86], Raffinot developed an improved approach, referred to as the Hierarchical Equal Risk Contribution Portfolio (HERC), which enabled the use of alternative risk metrics. Although HCAA was hard to beat, HERC performed better for some risk measures than others, and especially well for conditional drawdown at risk. Huang [43], however, showed that in the Chinese stock market, most HERC portfolios were not able to beat the equally weighted and inverse-variance portfolios in terms of several comparison measures. Moreover, the results did not indicate that any risk measure could outperform the others consistently.

In 2019, Lopez de Prado [61] introduced a novel way to apply clustering-based allocation in the general case and focused on the case of the mean-variance framework, with portfolios constructed using global minimum-variance and maximum Sharpe ratio methods. This framework, referred to as Nested Clustered Optimization (NCO), allowed for new possibilities in applying hierarchical clustering as a general tool in portfolio optimization instead of as a separate strategy.

Finally, there are a wide variety of alternative portfolio optimization approaches that, while significant, will not be covered in this thesis. Some of the most prominent research directions include utilizing Bayesian priors to incorporate exogenous insights [9, 71], multi-period portfolio optimization [54], and deep learning models [110].

1.2 Research Objectives

The purpose of this thesis is to make the following contributions to the theory of nested risk-based asset allocation strategies:

1. Utilize the following clustering algorithms in an effort to increase risk-adjusted returns:
   - partitioning around medoids with dynamic time warping distance;
   - discriminative functional mixture model.
2. Estimate the covariance matrix using Markov switching multifractal model with new distributional assumptions and dynamic conditional correlation type structure with nonlinear shrinkage.

The empirical study focuses on large-scale portfolio optimization and attempts to achieve the following:

1. Examine combinations of various intra- and inter-cluster optimization strategies, covariance matrix estimation methods, and clustering algorithms in terms of their contribution to portfolio risk and risk-adjusted returns;

2. Evaluate asset allocation method impact on portfolio concentration, turnover, and transaction costs;

3. Compare the out-of-sample performance of risk-based portfolio optimization techniques with practical weight constraints and transaction costs using empirical data.

2 Asset Allocation Methods

2.1 Equally Weighted Portfolio

The equally weighted (EW) portfolio is the portfolio where all assets are given the same weight, that is, a weighted inverse to the number of assets, N, in the portfolio, $w_i = \frac{1}{N}, \ i = 1, \ldots, N$. DeMiguel et al. [27] investigated the out-of-sample performance of the equally weighted portfolio relative to the mean-variance framework. The authors discovered that, despite its simplicity, the equally weighted portfolio regularly outperformed the other 14 models in terms of the Sharpe ratio, certainty-equivalent return, and turnover. They argue that the benefit of optimal diversification is negated by its estimation error, causing many optimization approaches to consistently underperform this naive allocation strategy. In fact, DeMiguel et al. estimate that, in order for the modern portfolio theory approach and its extensions to outperform the equally weighted benchmark, 500 years of data would be required to optimize a portfolio containing 50 assets. However, the question of whether or not portfolio optimization brings value has been hotly debated within the academic community. Several studies claim to demonstrate that optimized portfolios outperform naive diversification [49, 103].

2.2 Mean-Variance Portfolio

The mean-variance (MV) paradigm was proposed by Markowitz in 1952 as a novel way to think about portfolio optimization. The most significant part of his paper was the introduction of mean-variance efficient portfolio constructed as a convex optimization problem. This theory was elaborated upon with the release of his subsequent book [66], which provided a more generic model for portfolio selection known as Modern Portfolio Theory (MPT). Markowitz formulated the portfolio selection as a problem of finding a minimum variance portfolio of the assets in the investment universe that yields at least a target R of expected return. Mathematically, this formulation can be expressed as the following quadratic programming problem:

$$\begin{aligned} \min_{w}\ & w'\Sigma w \\ \text{s.t. } & w'\mathbf{1} = 1 \\ & w'\mu \geq R, \end{aligned} \tag{2.1}$$

where w' is the vector of asset weights, μ is the vector of expected returns, and Σ is the covariance matrix of the asset returns.

The formulation of the convex optimization problem is based on the assumption that the covariance matrix is positive-definite. When this assumption holds, the convex optimization problem can be defined as follows:

$$\max\ w'\mu - \frac{c}{2}w'\Sigma w, \tag{2.2}$$

where $c > 0$ is the risk aversion coefficient. This problem is usually solved using the Critical Line Algorithm (CLA).

The Markowitz paradigm yields two important economic insights. First, it demonstrates the effect of diversification: uncorrelated assets can be combined into portfolios with preferred expected risk-return characteristics. Second, it shows that, once a portfolio is fully diversified, higher expected returns can only be achieved through more extreme allocations and therefore by taking on more risk [12].

Despite the brilliance of Markowitz's theory, CLA solutions are rather unreliable due to various practical issues. A major caveat is that small deviations in the forecasted returns will cause CLA to produce very different portfolios [72]. Moreover, portfolios are based on variance rather than downside risk measures, which might better reflect investors' true preferences.

2.3 Risk-based Portfolios

The confounding effects of the uncertainty in the MV portfolio have led to the study of techniques that try to eliminate the need for estimated parameters, mainly expected returns and covariances. As shown by Chopra and Ziemba [21], the estimated covariance matrix causes less instability than the estimated expected returns, and it is suggested by Chopra [21] and Frahm and Wiechers [34] that simply removing the need for estimated expected returns from the optimization is possible and leads to primarily risk-based optimizations that are more stable.

2.3.1 Global Minimum Variance

Minimum variance investing has been inspired by early work from Haugen and Baker [42]. The authors discovered that a global minimum variance (GMV) portfolio outperformed the Wilshire 5000 at a lower risk from 1972 to 1989. A vast number of studies followed their original paper. For the US stock market, both higher returns and lower realized risks were found for the minimum variance portfolio versus a capitalization-weighted benchmark [19, 95, 44, 24].

Qualitatively similar results were also observed in global equity markets [35, 77, 82]. Scherer [94] showed that the GMV portfolio tends to hold a low beta and low residual-risk stocks. These results are known as low volatility anomalies.

On the other hand, the minimum variance portfolio is often highly concentrated in a small number of assets, since assets with low volatility are clearly favored [62]. It has also been shown that the minimum variance portfolio is very sensitive to errors in the estimated variance and correlation [5].

The GMV portfolio has the lowest risk of all portfolios based on the mean-variance method developed by Markowitz. Specifically, the optimization problem is to minimize the N-asset portfolio variance subject to short-sales constraints and a budget constraint, where the sum of the weights is 1:

$$\begin{aligned} &\min_{w} w'\Sigma w, \\ &\text{s.t. } \mathbf{1}'w = 1, \quad \mathbf{1} = (1, \dots, 1)', \\ &\quad w_i > 0, \qquad i = 1, \dots, N. \end{aligned} \tag{2.3}$$

In the absence of the short-sales constraints, the analytical solution is $w_{\text{opt}} = \frac{\Sigma^{-1}\mathbf{1}}{\mathbf{1}'\Sigma^{-1}\mathbf{1}}$. With additional constraints and bounds, the optimization problem is usually solved numerically.

2.3.2 Maximum Diversification

In contrast to the CAPM assumption that only systematic, non-diversifiable risk is priced in, Choueifaty and Coignard [22] claim that markets are risk-efficient in terms of total risk assessed in volatility. Authors define the diversification ratio as the ratio of the portfolio's weighted average volatility to its overall volatility:

$$\mathrm{DR}(w) = \frac{w'\sigma}{\sqrt{w'\Sigma w}}, \tag{2.4}$$

where σ is the vector of asset volatilities. The maximum diversification (MD) portfolio is obtained by solving the following optimization problem:

$$\begin{aligned} &\max_{w} \mathrm{DR}(w), \\ &\text{s.t. } \mathbf{1}'w = 1, \quad \mathbf{1} = (1, \dots, 1)', \\ &\quad w_i > 0, \qquad i = 1, \dots, N. \end{aligned} \tag{2.5}$$

It stands to reason that portfolios with a large concentration of uncorrelated assets would have a high $\mathrm{DR}(w)$. This insight can be formalized by deconstructing the portfolio's DR into its weighted correlation and weighted concentration measures [23]:

$$\mathrm{DR}(w) = \frac{1}{\sqrt{\rho(w)(1 - \mathrm{CR}(w)) + \mathrm{CR}(w)}}, \tag{2.6}$$

where $\rho(w)$ is the volatility-weighted average correlation of the assets in the portfolio,

$$\rho(w) = \frac{\sum_{i \neq j} \left(w_i \sigma_i w_j \sigma_j\right) \rho_{i,j}}{\sum_{i \neq j} (w_i \sigma_i w_j \sigma_j)}, \tag{2.7}$$

where $\rho_{i,j}$ is a Pearson correlation coefficient, and $\mathrm{CR}(w)$ is the volatility-weighted concentration ratio of the portfolio,

$$\mathrm{CR}(w) = \frac{\sum_{i=1}^{N} \left(w_i \sigma_i\right)^2}{\left(\sum_{i=1}^{N} w_i \sigma_i\right)^2}. \tag{2.8}$$

For a single asset portfolio CR = 1, while an equal volatility-weighted portfolio has the lowest CR, equal to the inverse of the number of assets. If correlations reach unity, the DR is also equal to 1 regardless of the CR, as the portfolio is no more diverse than a single asset.

Choueifaty et al. [23] define the core properties of the MD portfolio as follows:

1. Any stock not held by the MD portfolio is more correlated to the MD portfolio than any of the stocks that belong to it. In addition, the MD portfolio's constituents all have the same correlation with it;

2. The more diversified a given long-only portfolio, the greater its correlation with the MD portfolio.

2.3.3 Risk Parity

While revolutionary for its time, MPT came with some inherent problems, namely, allocating a large portion of the weight to a relatively small subset of assets within the portfolio. Qian [83], argues that risk is the true diversification factor, meaning that if a substantial portfolio loss can be attributed to a few assets, then the portfolio is not diversified. Furthermore, the author instead suggests an alternative investment approach, namely, risk parity (RP). Risk parity is an investment approach that allocates volatility instead of capital.

Theorem (Euler's Homogeneous Function Theorem)**.** *Let a continuous and differentiable function $f : \mathbb{R}^n \mapsto \mathbb{R}$ be a homogeneous function of degree one, i.e. $f(\lambda x) = \lambda f(x)$, $\lambda > 0$. Then*

$$f\left(w\right) = \sum_{i=1}^{n} w_i \frac{\partial f}{\partial w_i}. \tag{2.9}$$

Component $w_i \frac{\partial f}{\partial w_i}$ can be regarded as the risk contribution (RC) from asset to the total risk $f(w)$. Fortunately, most of the existing risk measures do satisfy the Euler's property. For example, the volatility of the portfolio $\sigma(w) = \sqrt{w' \Sigma w}$ can be decomposed as

$$\mathrm{RC}_i = \frac{w_i \left(\Sigma w\right)_i}{\sqrt{w' \Sigma w}}, \tag{2.10}$$

which satisfies $\sum_{i=1}^{N} \mathrm{RC}_i = \sigma(w)$. The relative risk contribution (RRC) is a normalized version:

$$\mathrm{RRC}_i = \frac{w_i\left(\Sigma w\right)_i}{w'\Sigma w}, \tag{2.11}$$

so that $\sum_{i=1}^{N} \mathrm{RRC}_i = 1$. The risk parity portfolio (RRP) attempts to equalize the risk contributions:

$$\mathrm{RC}_i = \frac{1}{N}\sigma\left(w\right) \quad \text{or} \quad \mathrm{RRC}_i = \frac{1}{N}. \tag{2.12}$$

More generally, the risk budgeting (RBP) attempts to allocate the risk according to the risk profile determined by the weights b:

$$\mathrm{RC}_i = b_i\sigma(w) \quad \textit{or} \quad \mathrm{RRC}_i = b_i. \tag{2.13}$$

In practice, one can express the condition $\mathrm{RC}_i = \frac{1}{N}\sigma(w)$ in different equivalent ways such as

$$w_i(\Sigma w)_i = w_j(\Sigma w)_j. \tag{2.14}$$

The budget condition $\mathrm{RC}_i = b_i\sigma(w)$ can also be expressed as

$$w_i(\Sigma w)_i = b_i w'\Sigma w. \tag{2.15}$$

Vanilla convex formulation Suppose we can only have constraints $\mathbf{1}'w = 1$ and $w \geq 0$. Then, after the change of variable $x = \frac{w}{\sqrt{w'\Sigma w}}$, the equations $w_i(\Sigma w)_i = b_i w'\Sigma w$ become $x_i(\Sigma x)_i = b_i$ or, more compactly in vector form, as

$$\Sigma x = \frac{b}{x} \tag{2.16}$$

with $x \geq 0$ and we can always recover the portfolio by normalizing: $w = \frac{x}{\mathbf{1}'x}$. With the goal of designing risk budgeting portfolios Spinu [98] proposed to solve the following convex optimization problem:

$$\min_{w\geq 0} \; \frac{1}{2}x'\Sigma x - \sum_{i=1}^{N} b_i \log x_i, \tag{2.17}$$

where the portfolio can be recovered as $w = \frac{x}{1'x}$.

Indeed, Spinu realized that the risk budgeting equation $\Sigma x = \frac{b}{x}$ correspond to the gradient of the convex function $f(x) = \frac{1}{2}x'\Sigma x - b'\log x$ set to zero:

$$\nabla f(x) = \Sigma x - \frac{b}{x} = 0. \tag{2.18}$$

Thus, a convenient way to solve the problem is by solving the following convex optimization problem:

$$\min_{w\geq 0} \; \frac{1}{2}x'\Sigma x - b'\log x, \tag{2.19}$$

which has optimality condition $\Sigma x = \frac{b}{x}$.

General nonconvex formulation The previous method is based on a convex reformulation of the problem, so it is guaranteed to converge to the optimal risk budgeting solution. However, it can only be employed for the simplest risk budgeting formulation. This method cannot be used if

- there are other constraints like allowing shortselling or box constraints, i.e. $l_i \leq w_i \leq u_i$;
- on top of the risk budgeting constraints $w_i(\Sigma w)_i = b_i w' \Sigma w$ we have other objectives like maximizing the expected return $w'\mu$ or minimizing the overall variance $w'\Sigma w$.

For those more general cases, we need more sophisticated formulations, which unfortunately are not convex. The idea is to try to achieve equal risk contributions $RC_i = \frac{w_i(\Sigma w)_i}{\sqrt{w'\Sigma w}}$ by penalizing the differences between the terms $w_i(\Sigma w)_i$. There are many possible reformulations, one such formulation is

$$\begin{aligned} \min_{w} \quad & \sum_{i,j=1}^{N} \left(w_i \left(\Sigma w\right)_i - w_j \left(\Sigma w\right)_j \right)^2 - F\left(w\right) \\ \text{s.t.} \quad & \mathbf{1}' w = 1, \qquad \mathbf{1} = (1, \ldots, 1)', \\ & w_i > 0, \qquad i = 1, \ldots, N, \\ & w \in \mathcal{W}. \end{aligned} \tag{2.20}$$

where $F(w)$ denotes some optional additional objective function and $\mathcal{W}$ denotes an arbitrary convex set of constraints. For example, expected return could be included as an additional objective by setting $F(w) = \lambda w'\mu$, where $w'\mu$ is the expected return and λ is a trade-off parameter. Similarly, the variance $w'\Sigma w$ could be added as an objective term by setting $F(w) = \lambda w'\Sigma w$.

In general, with different constraints and objective functions, exact parity cannot be achieved, and one needs to define a risk term to be minimized:

$$R\left(w\right) = \sum_{i=1}^{N} \left(g_i\left(w\right)\right)^2, \tag{2.21}$$

where g_i denote the different risk contribution errors. Risk term has many possible formulations [32]:

1. $R\left(w\right) = \sum_{i,j=1}^{N} \left(w_i \left(\Sigma w\right)_i - w_j \left(\Sigma w\right)_j \right)^2$;
2. $R\left(w\right) = \sum_{i=1}^{N} \left(\frac{w_i(\Sigma w)_i}{w'\Sigma w} - b_i \right)$;
3. $R\left(w\right) = \sum_{i=1}^{N} \left(w_i \left(\Sigma w\right)_i - b_i w'\Sigma w \right)^2$;
4. $R\left(w, \theta\right) = \sum_{i=1}^{N} \left(\frac{w_i(\Sigma w)_i}{b_i} - \theta \right)^2$;
5. $R\left(w, \theta\right) = \sum_{i=1}^{N} \left(w_i \left(\Sigma w\right)_i - \theta \right)^2$;
6. $R\left(w\right) = \sum_{i,j=1}^{N} \left(\frac{w_i(\Sigma w)_i}{b_i} - \frac{w_j(\Sigma w)_j}{b_j} \right)^2$;
7. $R\left(w\right) = \sum_{i=1}^{N} \left(\frac{w_i(\Sigma w)_i}{b_i\sqrt{w'\Sigma w}} \right)^2$;
8. $R\left(w\right) = \sum_{i=1}^{N} \left(\frac{w_i(\Sigma w)_i}{w'\Sigma w} \right)^2$.

An empirical study employs the first formulation. In practice, one would also like to set some general linear constraints. In such case, the optimization problem could be formulated as follows:

$$\begin{aligned} \min_{w} \; & R(w) + \lambda F(w) \\ \text{s.t.} \; & Cw = c, \\ & Dw \leq d. \end{aligned} \tag{2.22}$$

It was mentioned by Richard and Roncalli [89] that the problem of designing risk parity portfolios with general constraints is harder than it seems. Indeed, the authors show that, after imposing general linear constraints, the property of equal risk contributions is unlikely to be preserved among the assets affected by the constraints.

2.3.4 Hierarchical Risk Parity Portfolio

Originating from graph theory and machine learning, Hierarchical Risk Parity (HRP) approach provides a new contemporary prescription to the traditional challenges of asset allocation. HRP portfolios address three major concerns of quadratic optimizers in general and Markowitz's Critical Line Algorithm in particular: instability, concentration, and underperformance [60]. The key aspect is the introduction of hierarchical relationships among portfolio components. The algorithm takes advantage of the correlation structure without being dependent on the inversion of the covariance matrix. In fact, HRP can compute a portfolio on an ill-degenerated or even a singular covariance matrix, an impossible feat for quadratic optimizers. The main parts of the HRP algorithm are outlined below.

Clustering Let $\rho_{i,j}$ be the correlation coefficient between assets i and j, and ρ represents the correlation matrix. The hierarchical clustering is performed on a distance transformation of ρ, such that $d_{i,j} = \sqrt{2\left(1-\rho_{i,j}\right)}$. The Euclidean distance between columns in $d_{i,j}$ is utilized in the clustering algorithm. Lopez de Prado [60] apply single-linkage agglomerative nesting to $d_{i,j}$ which is described in section 4.1 of this thesis.

Quasi-Diagonalisation The quasi-diagonalisation step reorders the rows and columns in ρ such that the largest values lie close to the diagonal. This is achieved by rearranging the matrix based on the ordering generated by the clustering algorithm. A detailed discussion is provided in Lopez de Prado [60]. The loss function $\mathcal{L}$ is selected to measure the extent of diagonalization of the covariance matrix, with correlation coefficients far from the diagonal given higher weights:

$$\mathcal{L} = \sum_{i}^{N} \sum_{j}^{N} d_{i,j} \left(i-j\right)^{2}, \quad \forall i,j \in \left[1,\ldots,N\right]. \tag{2.23}$$

$\mathcal{L}$ takes on the value of zero, if ρ is a diagonal matrix. The use of $d_{i,j}$ in $\mathcal{L}$ sets the approach apart from Alipour et al. [3], who propose a version of $\mathcal{L}$, which uses $|\rho_{i,j}|$ instead of $d_{i,j}$.

Recursive Bisection The recursive algorithm consists of the following steps:

1. Initialize a list of assets in the portfolio, denoted L_0;
2. Initialize a vector of unit weights, $w_i - 1, \quad \forall i \in [1, \ldots, N]$;
3. Stop if $|L_i| = 1, \quad \forall L_i \in L$;
4. For each $L_i \in L$ such that $|L_i| > 1$:
   (a) Split L_i into two subsets $L_i^{(j)}$, preserving the order with $L_i^{(1)} \cup L_i^{(2)} = L_i$;
   (b) Calculate variance of $L_i^{(j)}$ such that $\tilde{V}_i^j$ is the covariance matrix of elements within cluster j, and $\tilde{w}_i^j = \frac{\mathrm{diag}\left(V_i^{(j)}\right)^{-1}}{\mathrm{tr}\left(\mathrm{diag}\left(V_i^j\right)^{-1}\right)}$, where diag(.) and tr(.) are the diagonal and trace operators respectively;
   (c) Compute the split factor: $\alpha_i = 1 - \frac{\tilde{V}_i^{(1)}}{\tilde{V}_i^{(1)} + \tilde{V}_i^{(2)}}$, so that $0 \leq \alpha_i \leq 1$;
   (d) rescale allocations w_n by a factor of $\alpha_i, \quad \forall n \in L_i^{(1)}$;
   (e) rescale allocations w_n by a factor of $(1 - \alpha_i), \quad \forall n \in L_i^{(2)}$;
5. Loop to step 3.

The original HRP is based on the single linkage (equivalent to the minimum spanning tree) algorithm, which suffers from the chaining effect: clusters are not dense enough and could span to very heterogeneous points since the algorithm merges clusters in a greedy fashion by considering only their two closest points. However, it is straightforward to replace the single linkage with another hierarchical clustering algorithm such as average or Ward linkage.

2.3.5 Hierarchical Equal Risk Contribution Portfolio

Raffinot [85] proposed the Hierarchical Clustering based Asset Allocation (HCAA), which agrees with the waterfall idea of HRP and, inspired by DeMiguel et al. [27], features dividing capital equally among hierarchical clusters and computing an equal-weighted allocation within each stock cluster. Yet, without incorporating sophisticated risk measures, HCAA's naive treatment of equal distribution suffers from oversimplification and subjectivity.

By integrating the HRP and HCAA methods, the Hierarchical Equal Risk Contribution (HERC) [86] algorithm adopts machine learning to allocate weights across and within asset clusters. HERC resembles HRP since both of them start by reorganizing the covariance matrix to group similar investments together. Nonetheless, HERC differs from HRP in that HRP makes no further use of clustering after an inverse-variance weighting allocation based on the number of assets. HERC benefits from HCAA's double-layer weighting scheme and alternative risk measures. Not limited to standard deviation, one can extend HERC to include downside risk measures such as Conditional Value at Risk (CVaR) (5.6) and Conditional Drawdown at Risk (CDaR) (5.14). This model consists of the following five steps:

1. **Hierarchical Tree Clustering**: use the relationships between financial assets to make a hierarchical structure that can be shown as a dendrogram.

2. **Selection of the Optimal Number of Clusters**: select the optimal number of clusters based on the hierarchical structure built in the first step. Raffinot suggests using the Gap Index.

3. **Matrix Seriation**: sort the assets in the dendrogram, minimizing the distance between leaves.

4. **Top-Down Recursive Division**: split the weights along the dendrogram in two parts using equal risk contribution allocation (risk contribution of each cluster is the sum of the inverse of the asset's risk) from the top of the tree to the clusters. The result is the weight for each cluster.

5. **Naive Risk Parity within Clusters**: use naive risk parity to calculate the weights within clusters and then multiply these weights by the cluster weight.

2.3.6 Nested Clustered Optimization

Nested Clustered Optimization (NCO) introduced by Lopez de Prado [61] is a machine learning-based approach to tackle the structural problems of covariance instability in modern portfolio theory. The goal is to contain the numerical instability at each clustering level, so that the instability within a subcluster does not extend to its parent cluster or the rest of the correlation matrix. The algorithm can be broken down into four steps:

1. Correlation clustering;

2. Estimation of optimal number of optimal number of clusters;

3. Intra-cluster weight allocation;

4. Inter-cluster weight allocation.

Because each security belongs to one cluster and one cluster only, the final allocation is the product of the intra- and inter-cluster weights. Lopez de Prado demonstrated that this dual clustering approach can significantly reduce Markowitz's estimation error. The author also argues that since similar assets are reduced to clusters, the reduced covariance matrix is, by construction, closer to a diagonal matrix and thus closer to the optimal solution of the original convex optimization proposed by Markowitz.

3 Covariance Estimation and Volatility Modeling

3.1 Shrinkage Methods

Many practical portfolio construction applications require an estimate of the covariance matrix and/or of its inverse when the matrix dimension N is large compared to the sample size T. It is well known that in such situations, the textbook estimator, the sample covariance matrix, performs poorly. It tends to be ill-conditioned and far from the population covariance matrix. The goal then becomes to find estimators that outperform the sample covariance matrix, both in finite samples and asymptotically. For the purposes of asymptotic analyses, to reflect the fact that N is large compared to T, one has to employ large-dimensional asymptotics where N is allowed to go to infinity together with T. In contrast, standard asymptotic theory would assume that N remains fixed while T tends to infinity. Ledoit and Wolf devoted two decades of their academic careers to shrinkage estimation of large-dimensional covariance matrices. In this thesis, we present only a tiny fraction of ideas related to shrinkage estimates, yet a much more comprehensive review is provided by Ledoit and Wolf themselves [53].

3.1.1 Linear Shrinkage

Ledoit and Wolf [50] demonstrate that the largest sample eigenvalues are systematically biased upwards, and the smallest ones downwards. It is then advantageous to correct this bias by pulling down the largest eigenvalues and pushing up the smallest ones, toward the grand mean of all sample eigenvalues. Working under large-dimensional asymptotics, Ledoit and Wolf derive the optimal linear shrinkage formula (when the loss is defined as the Frobenius norm of the difference between the estimator and the true covariance matrix). The same shrinkage intensity is applied to all sample eigenvalues, regardless of their positions. For example, if the linear shrinkage intensity is 0.5, then every sample eigenvalue is moved halfway toward the grand mean of all sample eigenvalues. Ledoit and Wolf both derive asymptotic optimality properties of the resulting estimator of the covariance matrix and demonstrate that it has desirable finite-sample properties via simulation studies.

Let us define the cross-sectional average of sample eigenvalues $(\lambda_i, \dots, \lambda_N)$ as

$$\overline{\lambda} = \frac{1}{N} \sum_{i=1}^{N} \lambda_i. \tag{3.1}$$

Linear shrinkage provides a consistent estimator $\rho \in [0, 1]$, which controls the amount by which the sample eigenvalues are dragged towards their cross-sectional average $\overline{\lambda}$. Let $(u_i, \dots, u_N)$ denote the set of eigenvectors. Then the linear shrinkage estimator is expressed as

$$\overline{C} = \sum_{i=1}^{N} \Big(\rho \overline{\lambda} + (1 - \rho)\, \lambda_i \Big)\, u_i u_i'. \tag{3.2}$$

Depending on the specifics, the gain over the sample covariance matrix can either be gigantic or negligible. Most of the potential improvement over the sample covariance matrix is captured when $\frac{N}{T}$ is large and the population eigenvalues are close to one another.

3.1.2 Nonlinear Shrinkage

The nonlinear shrinkage [51] method was proposed as an upgrade to the estimation of the covariance matrix when first-order approximation does not deliver a sufficient improvement. The intuition is as follows. Let Σ denote the population covariance matrix, S the sample covariance matrix, and u an eigenvector of S. Then, by basic linear algebra, the corresponding sample eigenvalue is equal to $u'Su$. It is the in-sample variance of a portfolio with weights given by the vector u. This is the quantity that needs to be rectified due to overfitting. Nonlinear shrinkage replaces it with a consistent estimator of $u'\Sigma u$, the out-of-sample variance of the same portfolio. Clearly, the objective is to allocate assets in the direction of the vector u based on its true out-of-sample risk $u'\Sigma u$, rather than its in-sample counterpart $u'\Sigma u$, which is heavily biased due to the curse of dimensionality. The nonlinear shrinkage formula depends on the unobservable population covariance matrix Σ, but thankfully it can approximated by an oracle shrinkage formula which depends only on the unobservable eigenvalues of Σ. Recovering the population eigenvalues from the sample eigenvalues requires inverting the Marčenko and Pastur equation, which governs their asymptotic relationship when the dimension is large. El Karoui [45] and Mestre [70] were the first to make an attempt in this direction. The solutions they proposed suffered from some limitations that made them unsuitable for general use. Subsequently, Ledoit and Wolf [52] introduced an effective method based on the numerical inversion of what they call the QuEST function; this acronym stands for Quantized Eigenvalues Sampling Transform. It is a deterministic N-dimensional function that discretizes the Marčenko-Pastur equation and lends itself to numerical inversion. By applying individualized shrinkage intensity to every sample eigenvalue Ledoit and Wolf derived a consistent estimator of population eigenvalues, which can then be used to find a consistent estimator of the oracle shrinkage. Asymptotically, when N and T grow large together, nonlinear shrinkage should perform better than linear shrinkage in the generic case.

3.2 Univariate Volatility Models

So far, we have used the assumption that the T observations are independent and identically distributed. Of course, such an assumption does not necessarily hold for financial return data, at least at shorter frequencies such as the daily frequency. It is thus advantageous to select a model that accommodates the time-varying nature of the conditional volatility and covariance matrix.

3.2.1 Generalized Autoregressive Conditional Heteroskedasticity

Particularly instrumental in the development of time varying higher order moments modeling techniques has been the autoregressive conditional heteroskedastic (ARCH) class of models introduced by Engle [29]. The key insight offered by the ARCH model lies in the distinction between the conditional and unconditional second order moments. While the unconditional covariance matrix for the variables of interest may be time invariant, the conditional variances and covariances often depend non-trivially on the past states of the world. In empirical applications of ARCH(q) models a long lag length and a large number of parameters are often called for. To circumvent this problem, Bollerslev [10] proposed the generalized ARCH model, GARCH(p, q).

Let r_t denote a discrete time stochastic process with conditional mean μ_t so that

$$r_t = \mu_t + \varepsilon_t, \tag{3.3}$$

$$\varepsilon_t = \sigma_t z_t, \tag{3.4}$$

where σ_t^2 is the conditional variance and z_t are independent identically distributed (i.i.d.) random variables with zero mean and unit variance.

The standard GARCH model may be written as:

$$\sigma_t^2 = \omega + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2 + \sum_{i=1}^{p} \beta_i \sigma_{t-i}^2 \tag{3.5}$$

with σ_t^2 denoting the conditional variance, ω denoting the intercept, and ε_t^2 denoting the residuals from the mean filtration process. Almost sure positivity of σ_t^2 requires that $\omega > 0$, $\alpha_i \geq 0$ and $\beta_j \geq 0$, $i = 1, \ldots, q$, $j = 1, \ldots, p$.

One of the shortcomings of the standard GARCH model is that conditional volatility only depends on the absolute values of previous observations. Thus, the model does not capture the asymmetry in financial return data. In the absence of a good theoretical model for this asymmetry, the GARCH literature has searched for econometric ways of capturing its effects. Models such as the EGARCH process introduced by Nelson [76], the GJR-GARCH process of Glosten, Jagannathan and Runkle [37], and the TGARCH model of Zakoian [108] are among the most popular solutions:

- EGARCH:

$$\ln\left(\sigma_t^2\right) = \omega + \sum_{i=1}^{q} \left(\alpha_i z_{t-i} + \gamma_i \left(\mid z_{t-i} \mid - \mathbb{E}|z_{t-i}|\right)\right) + \sum_{i=1}^{p} \beta_i \ln \sigma_{t-i}^2; \tag{3.6}$$

- GJR-GARCH:

$$\sigma_t^2 = \omega + \sum_{i=1}^{q} \left(\alpha_i \varepsilon_{t-i}^2 + \gamma_i \mathbb{1}_{(\varepsilon_{t-i}<0)} \varepsilon_{t-i}^2\right) + \sum_{i=1}^{p} \beta_i \sigma_{t-i}^2; \tag{3.7}$$

- TGARCH:

$$\sigma_t = \omega + \sum_{i=1}^{q} \left(\alpha_i \left(|\varepsilon_{t-1}| + \gamma_i \varepsilon_{t-i}\right)\right) + \sum_{i=1}^{p} \beta_i \sigma_{t-i}. \tag{3.8}$$

3.2.2 Markov Switching Multifractal

While the ARCH framework is a natural way to model changing volatility, it does not offer an integrated explanation of return phenomena at different frequencies. In contrast, stochastic regime-switching models permit the conditional mean and variance of financial returns to depend on an unobserved latent "state" that may change unpredictably. The Markov-switching multifractal (MSM) model designed by Calvet and Fisher [16] is a discrete-time Markov process with multi-frequency stochastic volatility. The return process is specified as

$$r_t = \sigma(M_t) z_t, \tag{3.9}$$

where the random variables z_t are i.i.d. standard Gaussians $\mathcal{N}(0,1)$ and σ_t is stochastic volatility with $\overline{k}$ volatility components $M_{1,t}, \ldots, M_{\overline{k},t}$ decaying at heterogeneous frequencies $\gamma_1, \ldots, \gamma_k$, such that

$$\sigma(M_t) = \overline{\sigma} \left(\prod_{k=1}^{\overline{k}} M_{k,t} \right)^{\frac{1}{2}}, \tag{3.10}$$

where $\overline{\sigma}$ is the unconditional standard deviation of the returns r_t under the assumption that the multipliers $M_{1,t}, \cdots, M_{k,t}$ are independent. The random volatility components $M_{k,i}$ are random multipliers that are persistent, non-negative and canonical such that $\mathrm{E}(M_{k,t}) = 1$, $\forall t$.

Volatility states are driven by a first-order Markov state vector

$$M_t = (M_{1,t}, \ldots, M_{\overline{k},t}) \in \mathbb{R}_+^{\overline{k}}. \tag{3.11}$$

When the state vector M_{t-1} is determined in period $t-1$, the t period multiplier $M_{k,t}$ for each $k \in 1, \ldots, \overline{k}$ is either drawn from a fixed distribution M with probability γ_k or remains unchanged at its current state: $M_{k,t} = M_{k,t-1}$. The multipliers differ in their transition probabilities γ_k but not in their marginal distribution M.

The transition probabilities $\gamma = (\gamma_1, \ldots, \gamma_{\overline{k}})$ are specified as

$$\gamma = 1 - (1 - \gamma_1)^{b^{k-1}}, \tag{3.12}$$

where $\gamma_1 \in (0,1)$ and $b \in (1, \infty)$. For small values of γ_1 and k, a Taylor polynomial approximation gives

$$\gamma_k \approx \gamma_1 b^{k-1}. \tag{3.13}$$

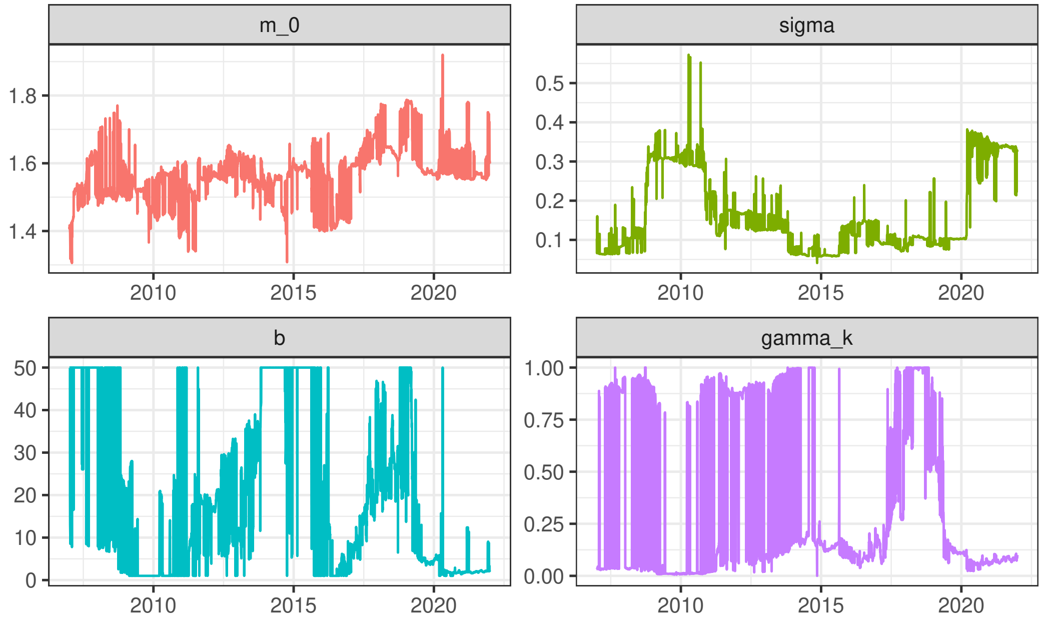
Hence, the transition probabilities of low-frequency components grow approximately at geometric rate b and the growth rate of the transition probabilities of high-frequency components slows down.

MSM imposes only minimal restrictions on the marginal distribution of the multipliers: $M \geq 0$ and $\mathrm{E}(M) = 1$, allowing flexible parametric or nonparametric specifications of M. A simple example is binomial MSM, in which the random variable M takes only two values, m_0 or m_1. For simplicity, it is often assumed that these two outcomes occur with equal probability, which implies that $m_1 = 2 - m_0$. The full parameter vector is then

$$\psi = (m_0, \overline{\sigma}, b, \gamma_{\overline{k}}) \in \mathbb{R}_+^4, \tag{3.14}$$

where m_0 characterizes the distribution of the multipliers, $\overline{\sigma}$ is the unconditional standard deviation of returns, b and $\gamma_{\overline{k}}$ define the set of switching probabilities.

Figure 1: MSM(4) parameter estimates over time for S&P500 index



Updating the State Vector Assuming that the distribution M is discrete, the Markov state vector M_t takes finitely many values $m^1, \ldots, m^d \in R_+^k$, and its dynamics are characterized by the transition matrix $A = (a_{i,j})_{1 \leq i,j \leq d}$ with components $a_{i,j} = \mathbb{P}(M_{t+1} = m^j \mid M_t = m^i)$. Conditional on the volatility state, the return r_t has Gaussian density $f_{r_t}(r \mid M_t = m^i) = n\left[r; \sigma^2\left(m^i\right)\right]$, where $n(.; \sigma^2)$ denotes the density of a centered normal with variance σ^2. The econometrician does not directly observe M_t but can compute the conditional probabilities

$$\Pi_t^j = \mathbb{P}\left(M_t = m^j \mid r_1, \ldots, r_t\right). \tag{3.15}$$

We can stack these probabilities in the row vector $\Pi_t = (\Pi_t^1, \ldots, \Pi_t^d) \in \mathbb{R}_+^d$.

The conditional probability vector is computed recursively. By Bayes' rule, Π_t can be expressed as a function of the previous belief Π_{t-1} and the innovation r_t:

$$\Pi_t = \frac{\omega\left(r_t\right) * \left(\Pi_{t-1} A\right)}{\left[\omega\left(r_t\right) * \left(\Pi_{t-1} A\right)\right] \mathbf{1}'}, \tag{3.16}$$

where $\mathbf{1} = (1, \ldots, 1) \in \mathbb{R}^d$, $x * y$ denotes the Hadamard product $(x_1 y_1, \ldots, x_d y_d)$ for any $x, y \in \mathbb{R}^d$, and

$$\omega\left(r_t\right)=\left(n\left[r_t;\sigma^2\left(m^1\right)\right],\ldots,n\left[r_t;\sigma^2\left(m^d\right)\right]\right). \tag{3.17}$$

In empirical applications, the initial vector Π_0 is chosen to be the ergodic distribution of the Markov process. Since the multipliers are mutually independent, the ergodic distribution is given by $\Pi_0^j=\Pi_{t=1}^{\bar{k}}\mathbb{P}(M=m_t^j)$ for all j.

Parameter estimates could be obtained by maximizing log-likelihood function

$$\ln L\left(r_q,\ldots,r_T;\psi\right)=\sum_{t=1}^{T}\ln\left(\omega\left(r_t\right)\cdot\left(\Pi_{t-1}A\right)\right), \tag{3.18}$$

where $x\cdot y$ denotes the inner product $x_1y_1+\cdots+x_dy_d$ for any $x,y\in\mathbb{R}^d$.

MSM with Various Innovation Distributions The modeling of financial assets as stochastic processes is determined by distributional assumptions on the increments and the dependence structure. It is well known that the returns of most financial assets have semi-heavy tails, i.e., the actual kurtosis is higher than the kurtosis of the normal distribution [63]. Although there are a lot of alternative distributions that have been used to model financial returns, to our knowledge, only Student distribution was considered in the context of the MSM model [58].

In this thesis, we try to replace the normality assumption by considering return r_t, which, conditional on the volatility state, has density $f\left[r;\sigma^2\left(m_i\right);\psi\right]$, where f denotes any parametric distribution, such as Student's t, hyperbolic, generalized hyperbolic skew Student's t, normal-inverse Gaussian or Johnson S_U, with finite variance and additional parameters denoted by ψ. Additional parameters are assumed to be independent of volatility state, therefore,

$$\omega\left(r_t\right)=\left(f\left[r_t;\sigma^2\left(m^1\right);\psi\right],\ldots,f\left[r_t;\sigma^2\left(m^d\right);\psi\right]\right). \tag{3.19}$$

Generalized Hyperbolic Distribution The one-dimensional generalized hyperbolic (GH) distribution is defined by the following Lebesgue density

$$f\left(x\right)=\frac{\left(\sqrt{\alpha^2-\beta^2}\right)^{\lambda}K_{\lambda-\frac{1}{2}}\left(\alpha\sqrt{\delta^2+\left(x-\mu\right)^2}\right)e^{\beta(x-\mu)}}{\sqrt{2\pi}\alpha^{\lambda-\frac{1}{2}}\delta^{\lambda}K_{\lambda}\left(\delta\sqrt{\alpha^2-\beta^2}\right)\left(\sqrt{\delta^2+\left(x-\mu\right)^2}\right)^{\frac{1}{2}-\lambda}} \tag{3.20}$$

where K_λ is a modified Bessel function and $x\in\mathbb{R}$. The domain of variation of the parameters is $\mu\in\mathbb{R}$ and

$$\begin{aligned}&\delta\geq 0,\ |\beta|<\alpha,\quad\text{if}\quad\lambda>0;\\&\delta>0,\ |\beta|<\alpha,\quad\text{if}\quad\lambda=0;\\&\delta>0,\ |\beta|\leq\alpha,\quad\text{if}\quad\lambda<0.\end{aligned}$$

An important aspect is, that GH distributions embrace many special cases, respectively limiting distributions:

- Student's t, for $\lambda = -\frac{\nu}{2}$, $\mu = 0$, $\delta = \sqrt{\nu}$, $\alpha = 0$ and $\beta = 0$;
- Generalized Hyperbolic Skew Student's t (GHST), for $\alpha \to |\beta|$, $\beta \neq 0$ and $\lambda = -\frac{\nu}{2}$;
- Hyperbolic (HYP), for $\lambda = 1$;
- Normal-inverse Gaussian (NIG), for $\lambda = -\frac{1}{2}$.

Detailed derivations of distribution densities from GH distribution, standartization, moments and tail asymptotics could be found in [99].

Johnson's S_U distribution Let U be a random variable that is uniformly distributed on the unit interval $[0, 1]$. Johnson's S_U random variables can be generated from U as follows:

$$x = \lambda \sinh\left(\frac{\Phi^{-1}(U) - \gamma}{\delta}\right) + \xi, \tag{3.21}$$

where Φ is the cumulative distribution function of the normal distribution. Johnson's S_U distribution has a probability density function given by

$$f(x) = \frac{\delta}{\lambda\sqrt{2\pi}} \frac{1}{\sqrt{1 + \left(\frac{x-\xi}{\lambda}\right)^2}} e^{-\frac{1}{2}\left(\gamma + \delta \sinh^{-1}\left(\frac{x-\xi}{\lambda}\right)\right)}. \tag{3.22}$$

The Johnson's S_U distribution can fit data that is leptokurtic and skewed [97], which makes it useful in a variety of areas, including modeling asset returns for portfolio management [101] and forecasting Value at Risk [18, 99].

3.3 Multivariate Volatility Models

3.3.1 Dynamic Conditional Correlation

The generalization of univariate GARCH models to the multivariate domain is conceptually simple. There exist many multivariate GARCH models such as Baba-Engle-Kraft-Kroner (BEKK), constant conditional correlation (CCC), and dynamic conditional correlation (DCC). In this thesis, we focus exclusively on the DCC framework. Consider the N-dimensional stochastic vector process $r_t,\ t = 1, \ldots, T$ and mean vector μ_t, given the information set $\mathcal{F}_{t-1}$:

$$r_t \mid \mathcal{F}_{t-1} = \mu_t + \varepsilon_t, \tag{3.23}$$

where the residuals of the process are modeled as:

$$\varepsilon_t = H_t^{\frac{1}{2}} z_t, \tag{3.24}$$

and $H_t^{\frac{1}{2}}$ is an $N \times N$ positive definite matrix such that H_t is the conditional covariance matrix of r_t, and z_t an N-dimensional vector of i.i.d. random variables with zero mean and unit variance.

Conditional correlation models are founded on a decomposition of the conditional covariance matrix into conditional standard deviations and correlations, so that it may be expressed in such a way that the univariate and multivariate dynamics may be separated, thus easing the estimation process. Engle [29] and Tse and Tsui [102] introduced the decomposition of the conditional correlation matrix H_t which allowed for the correlation matrix to be time varying with motion dynamics, such that

$$H_t = D_t R_t D_t, \tag{3.25}$$

where $D_t = \mathrm{diag}(\sqrt{h_{11,t}}, \sqrt{h_{22,t}}, \ldots \sqrt{h_{NN,t}})$ and R_t is the positive definite conditional correlation matrix with $\rho_{ij,t}$ and $\rho_{ii,t} = 1$ as its elements. The conditional variances, $h_{ii,t}$, can be estimated separately based on univariate GARCH-type models.

Apart from the fact that the time varying correlation matrix R_t must be inverted at every point in time, it is also important to constrain it to be positive definite. That could be achieved by modeling a proxy process, Q_t as:

$$Q_t = \Omega + A'(\varepsilon^*_{t-1}\varepsilon^*_{t-i}{}')A + B'Q_{t-1}B + G'(\eta_{t-i}\eta'_{t-i})G \tag{3.26}$$

$$\Omega = \overline{Q} - A'\overline{Q}A - B'\overline{Q}B - G'\overline{N}G \tag{3.27}$$

where A, B and G are the $N \times N$ parameter matrices, $\varepsilon^*_t = D_t^{-1}\varepsilon_t$, $\overline{Q} = \mathbb{E}[\varepsilon^*_t\varepsilon^{*\prime}_t]$ ir $\eta_t = \min[0, \varepsilon_t]$, $\overline{N} = \mathbb{E}[\eta_t\eta'_t]$. Because of its high dimensionality, restricted models have been used, including the standard (DCC), asymmetric (A-DCC), and generalized diagonal (GD-DCC) versions, with the specifications nested as follows [17]:

- DCC: $G = 0$, A and B are scalars;
- A-DCC: G, A, and B are scalars;
- GD-DCC: $G = 0$, A and B are diagonal matrices.

The conditional correlation matrix R_t is then obtained by rescaling Q_t such that

$$R_t = \mathrm{diag}(Q_t)^{-\frac{1}{2}} Q_t \mathrm{diag}(Q_t)^{-\frac{1}{2}}. \tag{3.28}$$

To ensure stationarity and positive definitness of Q_t, Ω and Q_0, the starting value of Q_t, has to be positive definite. In case of DCC(1,1), Ω is positive definite if $a \geq 0, b \geq 0$ and $a + b < 1$. In case of A-DCC, this constraint could be reformulated as $a + b + \delta g < 1$ where δ is the largest eigenvalue of $\overline{Q}^{-\frac{1}{2}}\overline{Q}^{-}\overline{Q}^{-\frac{1}{2}}$. More details about the existence and uniqueness of the stationary solution can be found in [33].

Remark. *Aielli [1] showed that* $\mathrm{E}[\varepsilon^*_t\varepsilon^{*\prime}_t] = \mathrm{E}[R_t] \neq \mathrm{E}[\overline{Q}]$, *therefore* $\overline{Q}$ *is not the unconditional covariance matrix of* ε^*_t. *He suggested a correction, named cDCC, which resolved this problem. Although there is a consistency issue related to the estimation of the intercept matrix in the* *standard DCC model described above, the correction seems to make very little difference in practice, if any [30].*

3.3.2 DCC with Nonlinear Shrinkage

It is well known that the number of assets in the investment universe generally poses a challenge to DCC-type models. One of the difficulties is inverting the conditional covariance matrix H_t for the log-likelihood computation. Pakel et al. [79] found a way to overcome this computational hurdle by summing up the log-likelihoods of pairs of assets instead of calculating the log-likelihood of all assets jointly, which they call the composite likelihood method. Authors showed that maximizing the composite likelihood yields consistent, if not efficient, estimators of the two correlation-dynamics parameters α and β. Later, Engle et al. [30] proposed a method to make the estimation of the DCC model computationally feasible by combining the composite likelihood with the nonlinear shrinkage of Ledoit and Wolf [51]. Following Hafner and Reznikova [40], authors applied the shrinkage to the intercept matrix rather than conditional covariance matrix itself. The resulting DCC-NLS estimator (where NLS stands for nonlinear shrinkage) unfolds in a three-stage process:

1. For each asset, fit a univariate GARCH model and use the fitted models to devolatilize the return series r_t to obtain the series $\hat{s}_t$.

2. Estimate the unconditional correlation matrix C by applying nonlinear shrinkage to the series $\hat{s}_t$ and use the resulting estimator $\hat{C}$ for correlation targeting.

3. Maximize the composite likelihood (over all neighboring pairs of assets) to estimate the two DCC parameters (α, β).

Nakagawa et al. [75] performed Monte Carlo simulations, which confirmed that DCC-NLS has the best estimation accuracy when compared to other DCC-type models. Although there were no statistically significant differences between the DCC and DCC-NLS or between the DCC and DCC-NLS in the MD and RP portfolios, the GMV portfolio had a significant improvement with a DCC-NLS model.

In this thesis, we also use the DCC-NLS model, but instead of combining univariate GARCH predictions, we obtain $\hat{s}_t$ through MSM-type models.

4 Clustering Methods

HRP, HERC, and NCO portfolio allocation strategies attempt to reduce estimation errors using clustering methods. As proposed by Papenbrock [80], clustering also presents a robust method to improve performance by finding groups of assets that show collective behavior, increasing the possibility to diversify by allocating according to the groupings. He further argues that, while there are numerous methods within finance to categorize groups of assets, an explorative machine learning method like clustering has the benefit of not requiring any economic, causal, institutional, or psychological explanation, which are all subject to potentially erroneous assumptions.

The impact of clustering on portfolio optimization has been studied in various studies. Both Leon et al. [55] and Duarte and De Castro [28] have considered partitional clustering methods such as k-means, k-medoids, and spectral clustering. Duarte and De Castro [28] found that when compared to the minimum variance portfolio, the Ibovespa Index, and the original HRP, partitional clustering based portfolio methods demonstrate the best average performance in terms of return and Sharpe ratio, but with slightly higher volatility, turnover, and drawdown. In contrast, Leon et al. [55] found that a particular hierarchical clustering method referred to as Ward's method outperformed all other clustering methods when evaluated using the Omega ratio.

This study considers alternative clustering methods, namely, partitioning around medoids with dynamic time warping distance and a discriminative functional mixture model. The aforementioned approaches are compared against hierarchical clustering with different linkages. Additionally, we explore various cluster validity indices to determine the optimal number of clusters.

4.1 Hierarchical Agglomerative Clustering

Hierarchical clustering requires a suitable distance measure. An appropriate function [64] is

$$D_{i,j} = \sqrt{2\left(1 - \rho_{i,j}\right)}, \tag{4.1}$$

where $D_{i,j}$ is the correlation-distance index between the ith and jth asset and $\rho_{i,j}$ is the respective Pearson's correlation coefficients. The distance $D_{i,j}$ is a linear multiple of the Euclidean distance between the vectors i, j after z-standardization, hence it inherits the true-metric properties of the Euclidean distance [60], to be specific:

1. $D_{i.j} \geq 0$;
2. $D_{i,j} = 0 \iff i = j$;
3. $D_{i,j} = D_{j,i}$;
4. $D_{i,j} \leq D_{i,k} + D_{k,j}$.

Hierarchical agglomerative clustering (HAC) starts with every observation representing a single cluster and then combines the clusters sequentially, reducing the number of clusters at each step until only one cluster is left. The least dissimilar clusters are merged into a single cluster at each $N-1$ step, resulting in one less cluster at the next higher level. It is important to define a measure of dissimilarity between two clusters, as different definitions of linkage might produce radically different dendrograms.

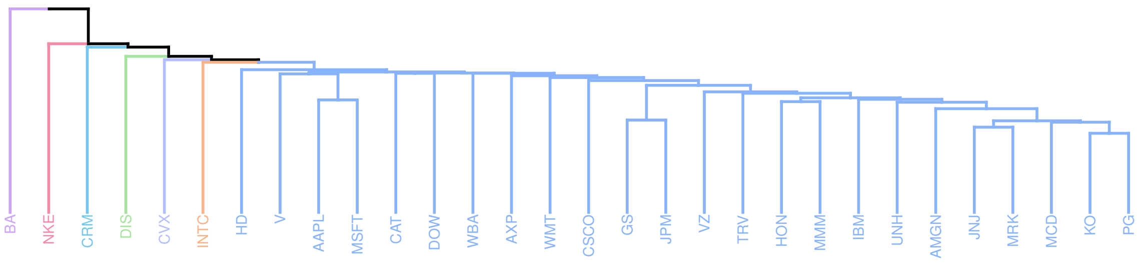
4.1.1 Linkage Methods

Single Linkage. The distance between two clusters is the minimum of the distance between any two points in the clusters. For clusters A, B:

$$d_{A,B} = \min_{a,b} \left[D\left(a,b\right) \mid a \in A,\ b \in B \right]. \tag{4.2}$$

This method is relatively simple and can handle non-elliptical shapes. Single linkage creates the $N-1$ edged hierarchical tree that minimizes the sum of the edge distances. This approach, however, is sensitive to outliers and can result in an issue known as “chaining” whereby clusters end up being long and straggly. The single linkage algorithm is intrinsically tied to the Minimum Spanning Tree (MST). However, the MST retains some information that the single linkage dendrogram throws away [86].

Figure 2: DJIA index constituents clustered using HAC with single linkage

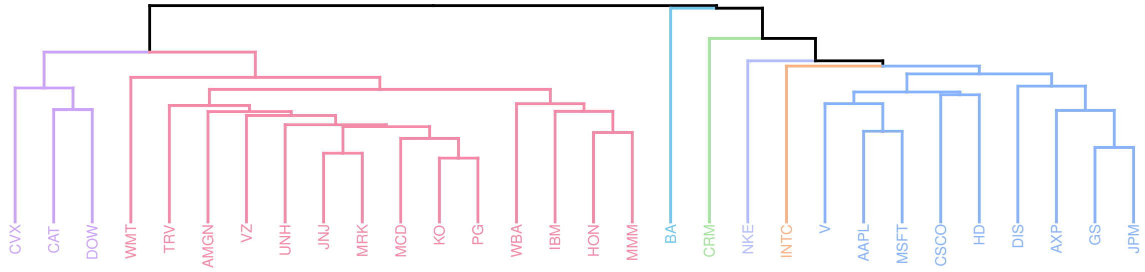


Complete Linkage. The distance between two clusters is the maximum of the distance between any two points in the clusters. For clusters A, B:

$$d_{A,B} = \max_{a,b} \left[D\left(a,b\right) \mid a \in A,\ b \in B \right]. \tag{4.3}$$

Complete linkage is at the opposite end of the spectrum from Single Linkage. In this case, two clusters are only considered close if all of the observations in their union are relatively similar. Therefore, complete linkage tends to produce compact clusters with small diameters. However, it can produce clusters that violate the “closeness” property. That is, observations assigned to a cluster can be much closer to members of other clusters than they are to some members of their own cluster [41].

Figure 3: DJIA index constituents clustered using HAC with complete linkage

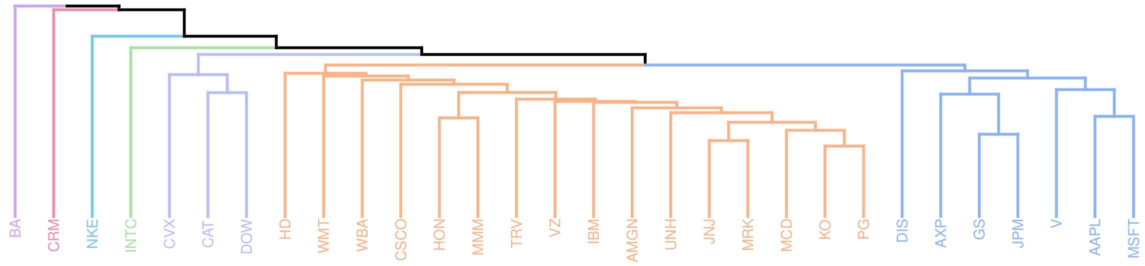


Average Linkage. The distance between two clusters is the average of the distance between any two points in the clusters. For clusters A, B:

$$d_{A,B} = \frac{1}{|A||B|} \sum_{a \in A} \sum_{b \in B} D(a,b)\,. \tag{4.4}$$

Average linkage represents a compromise between the two extremes of single and complete linkage. It attempts to produce relatively compact clusters that are relatively far apart and avoids the chaining effect as well as the closeness property [41].

Figure 4: DJIA index constituents clustered using HAC with average linkage

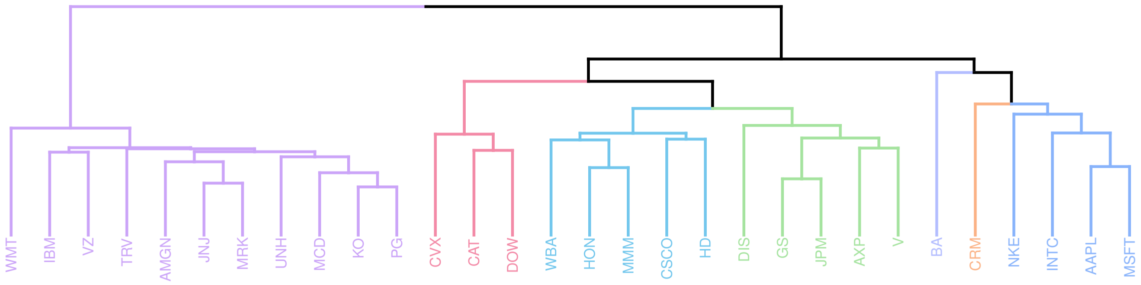


Ward's Method (Ward [1963]). The distance between two clusters is the increase of the squared error that results when two clusters are merged. For clusters A, B:

$$d_{A,B} = \frac{|A||B|}{|A|+|B|}\|a-b\|^2, \tag{4.5}$$

where a, b are the centroids for the clusters. This method is biased towards globular clusters, but less susceptible to noise and outliers. It is one of the most popular methods [86].

Figure 5: DJIA index constituents clustered using HAC with Ward's linkage



In the empirical results found by Papenbrock [80], single linkage and Ward's method are deemed superior in terms of performance among the explored portfolio allocation methods. Of these methods, single linkage produces a high concentration in weight allocation from the highly asymmetrical clusters but also results in a low correlation between clusters. On the contrary, Ward's method creates high-quality, balanced clusters with low weight concentrations, which is why it was chosen in the empirical study.

4.1.2 Optimal Number of Clusters

To determine the number of clusters, the Gap statistic [100] is employed. It compares the logarithm of the empirical within-cluster dissimilarity and the corresponding one for uniformly distributed data with no apparent clusters. Suppose that we have clustered the data into k clusters $C_1, \ldots, C_k$ with C_r denoting the indices of observations in cluster r. The within-cluster distance, W_k can be defined as the pooled within-cluster sum of quares around the cluster means.

$$W_k = \sum_{t=1}^{k} \frac{1}{2|C_r|} D_r, \tag{4.6}$$

where D_r is defined as a sum of pairwise distances for all points in cluster r:

$$D_r = \sum_{i,i' \in C_r} d_{ii'}. \tag{4.7}$$

The gap statistic measures the within-cluster dispersion around the cluster mean, which is used to investigate the relationship between $\log W_k$ for different values of k compared to a suitable null reference distribution. Hence, gaps are defined as

$$\mathrm{Gap}_n(k) = \mathrm{E}_n^* [\log (W_k)] - \log (W_k), \tag{4.8}$$

where E_n^* denotes expectation under a sample of size n from the reference distribution. Tibshirani et al. [100] consider two choices for the reference distribution:

1. generate each reference feature uniformly over the range of the observed values for that

feature;

2. generate the reference features from a uniform distribution over a box aligned with the principal components of the data.

Let $\sigma(k)$ denote the standard deviation of the B Monte Carlo replicates $\log(W_k^*)$. Accounting additionally for simulation error in $\mathrm{E}_n^*[\log(W_k)]$ results in the quantity

$$s_k = \sqrt{\left(1 + \frac{1}{B}\right)} \sigma\left(k\right). \tag{4.9}$$

Finally, the suggested number of clusters k can be inferred as

$$\begin{aligned} &\min_k \mathrm{Gap}(k), \\ &\text{s.t. } \mathrm{Gap}(k) \geq \mathrm{Gap}(k+1) + s_{k+1}. \end{aligned} \tag{4.10}$$

As the process of generating the reference distributions and running the clustering algorithm is computationally expensive, one has to weigh the reliability of the derived results with the computational costs. Yue et al. [107] proposed an alternative function, bypassing the computationally expensive method described above by utilizing second-order differencing to find the optimal number of clusters, using the maximisation function defined as

$$\begin{aligned} &\max_k W_k - 2W_{k+1} + W_{k+2}, \\ &\text{s.t. } 1 \leq k \leq \sqrt{n}. \end{aligned} \tag{4.11}$$

The results are argued by the author to be more stable, less dependent on random draws, and less computationally expensive as there is no need to compute the reference distributions.

4.2 Partitioning Around Medoids

Both partitioning around medoids (PAM) and k-means are partitional algorithms that aim to reduce the distance between points marked as belonging to a cluster and the center of that cluster. While the center of a cluster in k-means is not necessarily one of the input data points, PAM picks real data points as centers, allowing for enhanced interpretability of the cluster centers. Furthermore, PAM is more resilient to noise and outliers since it minimizes a sum of pairwise dissimilarities rather than a total of squared Euclidean distances.

PAM starts by selecting k random objects (medoids) as representative of the k clusters, and each of the remaining $n - k$ objects is assigned to the closest medoid according to the selected distance measure. The algorithm then attempts to improve on the initial clustering as follows: for each medoid object, swap each of the non-medoid objects with the medoid, and compute the cost, i.e., average dissimilarity, of the new clustering that results from this swap. If the cost increases, then the swap is undone.

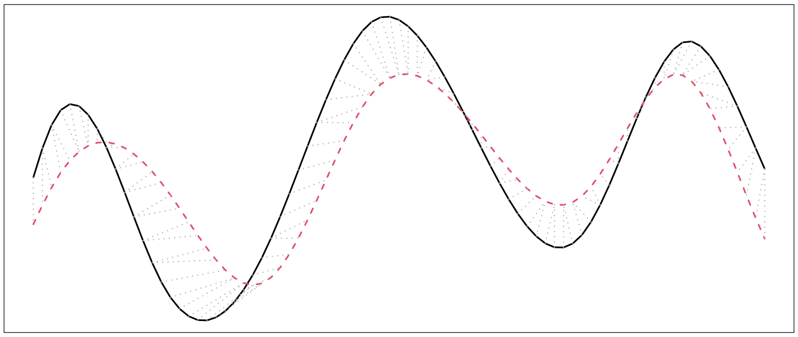
Unlike k-means, PAM may be implemented with arbitrary dissimilarity metrics. The selection of the distance measure has a significant influence on the algorithm's outputs as it defines how the similarity between clustering objects is calculated. The determination of dissimilarity has traditionally been done using the Euclidean and Manhattan distance measures. However, in this thesis, we use PAM with dynamic time warping distance as it allows matching of samples that are shifted in time.

4.2.1 Dynamic Time Warping

Dynamic time warping (DTW) is one of the most often employed distance metrics for time series clustering. This approach was initially created for voice recognition applications [104, 93] and is used to determine the optimal alignment between two time-dependent sequences. DTW started being used by the data mining community to overcome some of the limitations associated with the Euclidean distance [87, 7]. The main disadvantage of using Euclidean distance for time series data is that its results may appear very unintuitive. If two time series are identical but one is shifted slightly along the time axis, then Euclidean distance may consider them to be very different from each other. DTW overcomes this limitation and gives intuitive distance measurements between time series by ignoring both global and local shifts in the time dimension.

The simplest way to get an intuition of what DTW does is graphically. Figure 6 shows the alignment between two sample time series. In this instance, the initial and final points of the series must match, but other points may be time-warped.

Figure 6: DTW alignment



Given two sequences $X = (x_1, \ldots, x_N)$ and $Y = (y_1, \ldots, y_M)$, the DTW objective is to temporally align these two sequences in some optimal sense under certain constraints. This leads to the notion of a warping path. Let $[n] = \{1, \ldots, n\}$, where $n \in \mathbb{N}$. A warping path is a sequence $p = (p_1, \ldots, p_L)$ with $p_l = (n_l, m_l) \in [N] \times [M]$ for $l \in [L]$ satisfying the following conditions:

1. Boundary condition: $p_1 = (1, 1)$ and $p_L = (N, M)$;

2. Step size condition: $p_{l+1} - p_l \in (1, 0), (0, 1), (1, 1)$ for $l \in [L - 1]$.

A warping path defines an alignment between two sequences by assigning the element $x_{n_l} \in X$ to the element $y_{m_l} \in Y$. The boundary condition enforces that the first elements as well as the last elements of X and Y are aligned to each other. The step size condition expresses a kind of continuity requirement: no element in X and Y can be omitted, and there are no replications in the alignment. Note that the step condition also implies monotonicity: $n_1 \leq n_2 \leq \ldots \leq n_L$ and $m_1 \leq m_2 \leq \ldots \leq m_L$.

To determine an optimal warping path one needs to compare the elements of X and Y. Let $\mathcal{F}$ be a feature space and $x_n, y_m \in \mathcal{F},\ n \in [1 : N],\ m \in [1 : M]$. A local cost measure is a function $c : \mathcal{F} \times \mathcal{F} \mapsto \mathbb{R}^+$, the typical choice of c is Euclidean norm. Evaluating the local cost measure for each pair of elements of the sequences X and Y, we obtain a local cost matrix (LCM). The total cost $c_p(X, Y)$ of a warping path p between X and Y with respect to the local cost measure c is defined as

$$c_p\left(X, Y\right) = \sum_{l=1}^{L} c\left(x_{n_l}, y_{n_l}\right). \tag{4.12}$$

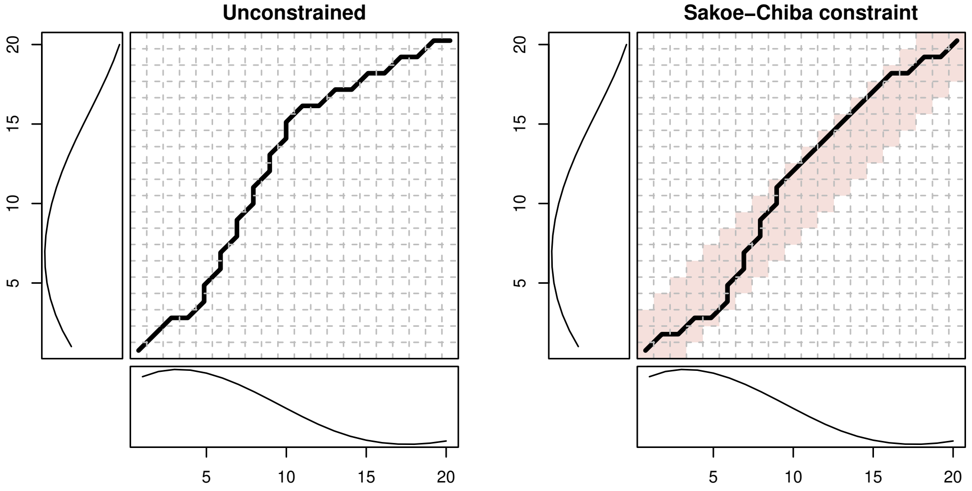
An optimal warping path between X and Y is a warping path p^* having minimal total cost among all possible warping paths. The DTW distance between X and Y is then defined as the total cost of p^*:

$$\mathrm{DTW}\left(X, Y\right) = c_{p^*}(X, Y) = \min\left[c_p\left(X, Y\right) \mid p \in P^{N \times M}\right], \tag{4.13}$$

where $P^{N \times M}$ is the set of all possible warping paths.

Global DTW constraints One of the possible modifications to DTW is the introduction of global constraints, also known as window constraints. These constraints restrict the LCM space that may be searched by the algorithm. There are numerous different types of windows [36], but one of the most prevalent ones is the Sakoe-Chiba window [93], which creates a permitted space along the diagonal of the LCM. These constraints can marginally accelerate the DTW computation, but their primary use is to prevent pathological warping. It is common to use a window 10% the length of the series. However, smaller windows can sometimes produce even better results [87].

Figure 7: Optimal warping path along the LCM



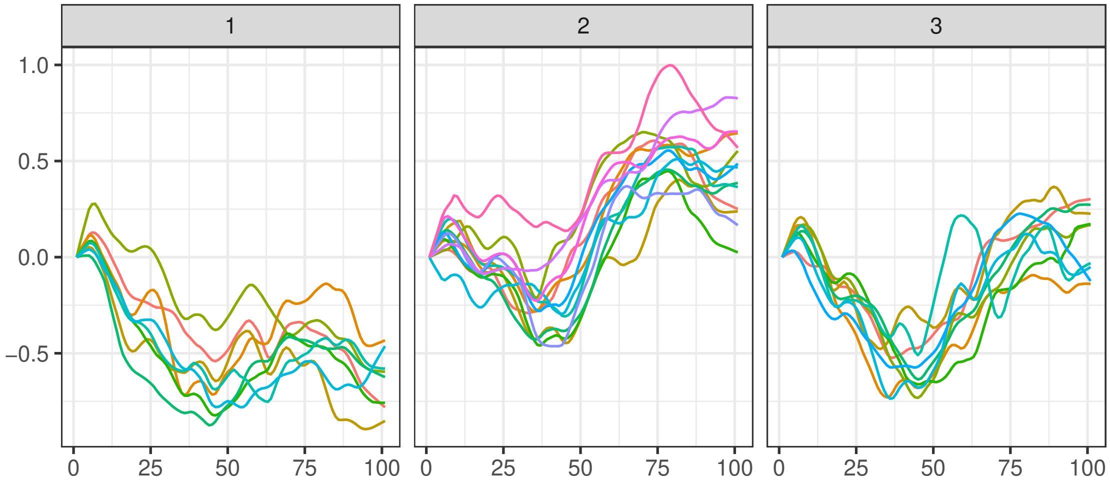
4.2.2 Noise Reduction

One of the disadvantages of DTW is that this algorithm is susceptible to noise, and the presence of noise may lead to singularities [78]. Singularities are defined as “unintuitive alignments where a single point on one time series maps onto a large subsection of another time series” [46]. For this reason, we consider data smoothing as the first stage of data preprocessing.

Data smoothing refers to methods of eliminating statistical noise from datasets to make the patterns more noticeable. One the most commonly used noise reduction algorithms is a moving average (MA) filter. The MA filter is a simple low pass finite impulse response filter commonly used for smoothing an array of sampled data. It takes a specified number of input samples at a time, calculates their average, and provides a single output point. The moving average, however, has two drawbacks. First, it distorts genuine peaks in the data. Since data are averaged over the window, peaks within the window are inevitably “pulled down” towards the mean value of the data within the window, which is undesirable if the peaks are not the result of noise. The second issue is the inverse of the first: big noise outliers affect smoothed data by dragging it in their direction.

In this thesis, we use the locally-weighted scatterplot smoothing (LOWESS) smoothing technique, which is a generalization of the moving average and polynomial regression, that mitigates both aforementioned problems. Using a window subset of the data, the weighted local polynomial fitting algorithm iteratively traverses the full dataset, moving through one data point at a time. Each step involves fitting a polynomial function (typically of first or second order) to the window’s data. The polynomial is fit using weighted least squares, giving more weight to points near the point whose response is being estimated and less weight to points further away. When the fit is complete, the central data point is replaced by the value of the polynomial at that point in the window.

Figure 8: DJIA index constituents clustered using PAM method



4.2.3 Cluster Validity Indices

Previous works have shown that there is no single CVI that outperforms the rest [73, 105, 67]. Arbelaitz et al. [4] compared 30 cluster validity indices in 720 synthetic and 20 real datasets. Comparison methodology included three clustering algorithms: k-means, Ward, and average linkage. The authors demonstrated that there are three main groups of indices and the indices in the first group — Calinski–Harabasz, Davies–Bouldin, Silhouette, Score, and COP — generally behave better.

Let us define a dataset X as a set of N objects represented as vectors in an F-dimensional space: $X = x_1, \ldots, x_N \subseteq \mathbb{R}^F$ A partition or clustering in X is a set of disjoint clusters that partitions X into K groups: $C = c_1, \ldots, c_K$ where $\cup_{c_k \in C} c_k = X$, $c_k \cap c_l = \varnothing$, $\forall k \neq l$ The centroid of a cluster c_k is its mean vector, $\overline{c_k} = \frac{1}{|c_k| \sum_{x_i \in c_k} x_i}$ and, similarly, the centroid of the dataset is the mean vector of the whole dataset, $\overline{X} = \frac{1}{N} \sum_{x_i \in X} x_i$. Finally, we will denote the distance between objects x_i and x_j as $\mathrm{d}(x_i, x_j)$.

Calinski–Harabasz (CH) [15] This index obtained the best results in the work of Milligan and Cooper [73]. It is a ratio-type index where the cohesion is estimated based on the distances from the points in a cluster to its centroid. The separation is based on the distance from the centroids to the global centroid. It can be defined as

$$\mathrm{CH}\left(C\right) = \frac{N-K}{K-1} \frac{\sum_{c_k \in C} |c_k| \mathrm{d}\left(\overline{c_k}, \overline{X}\right)}{\sum_{c_k \in C} \sum_{x_i \in c_k} \mathrm{d}\left(x_i, \overline{c_k}\right)}. \tag{4.14}$$

A higher value of the CH index means the clusters are dense and well separated, although there is no “acceptable” cut-off value.

Davies–Bouldin (DB) [26] This is probably one of the most used indices in CVI comparison studies. It estimates the cohesion based on the distance from the points in a cluster to its centroid and the separation based on the distance between centroids. It is defined as

$$\mathrm{DB}\left(C\right) = \frac{1}{K} \sum_{c_k \in C} \max_{c_i \in C \setminus c_k} \left[\frac{\mathrm{S}\left(c_k\right) + \mathrm{S}\left(c_l\right)}{\mathrm{d}\left(\overline{c_k}, \overline{c_l}\right)} \right], \tag{4.15}$$

where

$$\mathrm{S}\left(c_k\right) = \frac{1}{|c_k|} \sum_{x_i \in c_k} \mathrm{d}\left(x_i, \overline{c_k}\right). \tag{4.16}$$

Due to the way it is defined, as a function of the ratio of the within cluster scatter, to the between cluster separation, a lower value will mean that the clustering is better.

Davies–Bouldin* (DB*) [47] : This variation of the Davies–Bouldin index was proposed together with an interesting discussion about different types of CVIs. Its definition is

$$\mathrm{DB}^*\left(C\right) = \frac{1}{K}\sum_{c_k\in C}\frac{\max_{c_l\in C\setminus c_k}\left[\mathrm{S}\left(c_k\right)+\mathrm{S}\left(c_l\right)\right]}{\min_{c_l\in C\setminus c_k}\left[\mathrm{d}\left(\overline{c_k},\overline{c_l}\right)\right]}. \tag{4.17}$$

Silhouette index (Sil) [91] This index is a normalized summation-type index. The cohesion is measured based on the distance between all the points in the same cluster and the separation is based on the nearest neighbour distance. It is defined as

$$\mathrm{Sil}\left(C\right) = \frac{1}{N}\sum_{c_k\in C}\sum_{x_i\in c_k}\frac{\mathrm{b}\left(x_i,c_k\right)-\mathrm{a}\left(x_i,c_k\right)}{\max\left[\mathrm{a}\left(x_i,c_k\right),\mathrm{b}\left(x_i,c_k\right)\right]}, \tag{4.18}$$

where $\mathrm{a}\left(x_i,c_k\right)$ is the average distance of the point from the points in its own cluster:

$$\mathrm{a}\left(x_i,c_k\right) = \frac{1}{|c_k|}\sum_{x_j\in c_k}\mathrm{d}\left(x_i,x_j\right) \tag{4.19}$$

and $\mathrm{b}\left(x_i,c_k\right)$ is the average distance of the point from the points in the nearest cluster:

$$\mathrm{b}\left(x_i,c_k\right) = \min_{c_l\in C\setminus c_k}\left[\frac{1}{|c_l|}\sum_{x_j\in c_l}d\left(x_i,x_j\right)\right]. \tag{4.20}$$

The Silhouette score is bounded from -1 to 1 and higher score means more distinct clusters.

Score function (SF) [92] This is a summation-type index where the separation is measured based on the distance from the cluster centroids to the global centroid and the cohesion is based on the distance from the points in a cluster to its centroid. It is defined as

$$\mathrm{SF}\left(C\right) = 1-\frac{1}{e^{e^{\mathrm{bcd}(C)+\mathrm{wcd}(C)}}}, \tag{4.21}$$

where

$$\mathrm{bcd}\left(C\right) = \frac{\sum_{c_k\in C}|c_k|\mathrm{d}\left(\overline{c_k},\overline{X}\right)}{N\cdot K}, \tag{4.22}$$

$$\mathrm{wcd}\left(C\right) = \sum_{c_k\in C}\frac{1}{c_k}\sum_{x_i\in c_k}\mathrm{d}\left(x_i,\overline{c_k}\right). \tag{4.23}$$

The higher the value of the SF, the more suitable the number of clusters.

COP index [39] Although this index was first proposed to be used in conjunction with a cluster hierarchy postprocessing algorithm, it can also be used as an ordinary CVI. It is a ratio-type index where the cohesion is estimated by the distance from the points in a cluster to its centroid and the separation is based on the furthest neighbour distance. Its definition is

$$\mathrm{COP} = \frac{1}{N} \sum_{c_k \in C} \frac{\sum_{x_i \in c_k} \mathrm{d}\left(x_i, \overline{c_k}\right)}{\min_{x_i \notin c_k} \max_{x_j \in c_k} \mathrm{d}\left(x_i, x_j\right)}. \tag{4.24}$$

COP is bounded between 0 and 1 and takes its maximum value in the improbable case where the closest point not in the cluster is in the centroid of the cluster. Lower COP values signify more distinct clusters.

4.3 Discriminative Functional Mixture Model

The discriminative functional mixture (DFM) [11] model allows the clustering of the functional data in a unique and discriminative functional subspace. This model presents the advantage to be parsimonious and can therefore handle long time series. The first step of DFM is to smooth the functional data. Then, the functional data are fitted into a functional latent mixture model with lower dimensional subspaces. After specifying a cluster number K, inference of the latent mixture model is estimated by the modified expectation maximisation (EM) algorithm. The final clustering is obtained by estimating the probability that an observation belongs to a particular cluster.

Let $\{x_1, \ldots, x_n\}$ be the observed curves we want to cluster into K homogeneous groups. DFM introduces an unobserved random variable $Z = (Z_1, \ldots, Z_K) \in \{0, 1\}^K$ indicating the group membership of X: Z_k is equal to 1 if X belongs to the kth group and 0 otherwise. The clustering task aims to predict the value $z_i = (z_{i1}, \ldots, z_{iK})$ for each observed curve x_i.

In practice, the functional expressions of curves $\{x_1, \ldots, x_n\}$ are unknown, and only discrete observations $x_i(t_{is})$ at a finite set of ordered times $\{t_{is} : s = 1, \ldots, m_i\}$ are available. Therefore, functional data analysis typically begins with recovering the functional nature of the data. Generally, this is accomplished by assuming that the curves belong to a finite dimensional space spanned by a basis of functions. Let us therefore consider such a basis $\{\psi_1, \ldots, \psi_p\}$ and assume that the stochastic process X admits the following basis expansion:

$$X\left(t\right) = \sum_{j=1}^{p} \gamma_j\left(X\right) \psi_j\left(t\right), \tag{4.25}$$

where $\Gamma = \{\gamma_1(X), \ldots, \gamma_p(X)\}$ is a random vector in $\mathbb{R}^p$, and the number p of basis functions is assumed to be fixed and known.

Next, the functional data is represented in a lower dimension subspace in terms of basis functions $\{\varphi_j\}_{j=1,\ldots,d}$ with $d < K$ and $d < p$. The basis $\{\varphi_j\}_{j=1,\ldots,d}$ is obtained from $\{\psi\}_{j=1,\ldots,p}$ through a linear transformation $\varphi_j = \sum_{i=1}^{p} u_{ji}\psi_i$ such that the $p \times d$ matrix $U = (u_{ji})$ is orthogonal. Let $\{\lambda_1, \ldots, \lambda_n\}$ be the latent expansion coefficients of the curves $\{x_1, \ldots, x_n\}$ on the basis $\{\varphi_j\}_{j=1,\ldots,d}$. These coefficients are assumed to be independent realizations of a latent random vector $\Lambda \in \mathbb{R}^d$.

Then the observed curve x_i could be expressed using the following basis expansion:

$$x_i\left(t\right) = \sum_{j=1}^{d} \lambda_{ij} \varphi_j\left(t\right). \tag{4.26}$$

The relationship between random vectors Γ and Λ is then given by

$$\Gamma = U\Lambda + \varepsilon, \tag{4.27}$$

where $\varepsilon \in \mathbb{R}^p$ is an independent and random noise term, which is assumed to be distributed according to a multivariate Gaussian density:

$$\varepsilon \sim \mathcal{N}(0, \Omega). \tag{4.28}$$

Conditional distribution of Λ is also assumed to follow multivariate Gaussian density:

$$\Lambda_{|Z_k=1} = \mathcal{N}(\mu_k, \Sigma_k), \tag{4.29}$$

where μ_k and Σ_k are, respectively, the mean and the covariance matrix of the kth cluster. Under these distributional assumptions, the marginal distribution of Γ is a mixture of Gaussians:

$$p\left(\gamma\right) = \sum_{k=1}^{K} \pi_k \Phi\left(\gamma \mid U\mu_k,\ U'\Sigma_k U + \Omega\right), \tag{4.30}$$

where Φ is the standard Gaussian density function, and $\pi_k = P(Z_k = 1)$ is the prior probability of the kth group.

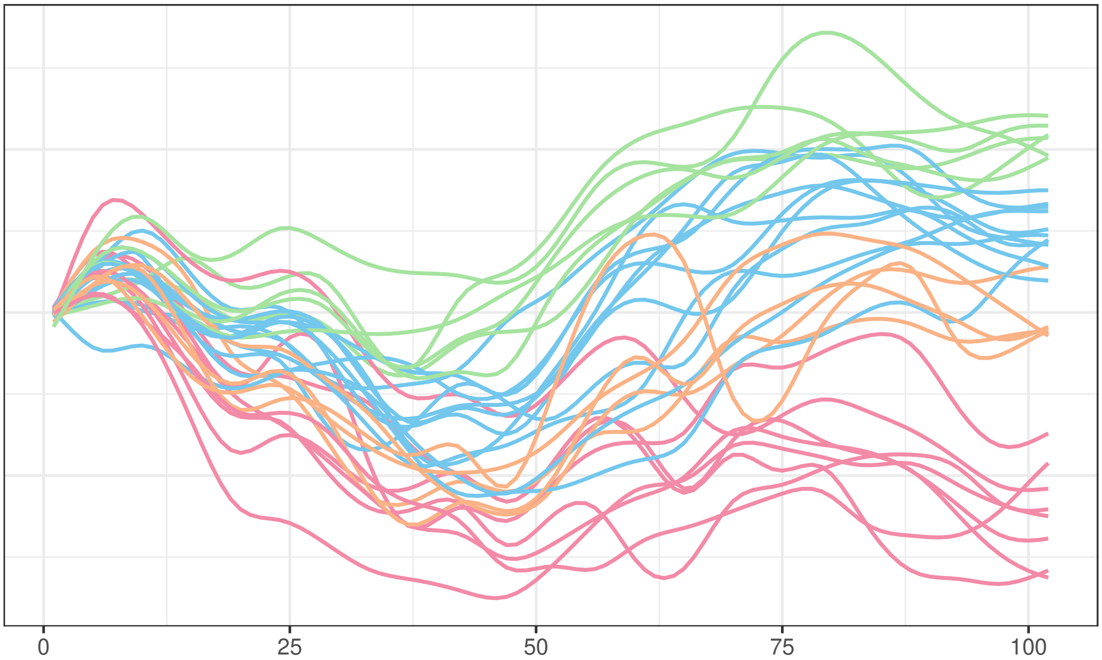
Finally, the conditional noise covariance matrix $\Omega_{Z_k=1} = \mathrm{Cov}[W'\Gamma \mid Z_k = 1] = W'\Sigma_k W$ is assumed to have the following form:

$$\begin{gathered}\left(\begin{array}{c:ccc} \Sigma_k & & 0 & \\ \hdashline & \beta & & 0 \\ 0 & & \ddots & \\ & 0 & & \beta \end{array}\right) \\ \underbrace{\hphantom{\Sigma_k}}_{d}\underbrace{\hphantom{\beta\quad\ddots\quad\beta}}_{p-d}\end{gathered}$$

with $W = [U, V]$, where V is the orthogonal complement of U. With this notation, and from a practical point of view, one can say that the variance of the actual data of the kth group is therefore modeled by Σ_k, whereas the parameter β models the variance of the noise outside the functional subspace.

Multiple submodels can be produced by putting restrictions to the $\Omega_{Z_k=1}$ parameters. For instance, it is possible to relax the constraint that the noise variance is common across clusters. It is also possible to constrain the model such that the covariance matrices $\Sigma_1, \dots, \Sigma_K$ in the latent space are common across groups. Similarly, in each group, Σ_k can be assumed to be diagonal. More details on various model specifications, complexity, inference, and model selection are provided in [11].

Figure 9: DJIA index constituents clustered using DFM model



5 Performance Measures

5.1 Risk Measures

Numerous criteria can be employed to quantify portfolio risk. The selection of acceptable risk measures remains a point of contention and research in financial mathematics, as all proposed risk measures have downsides and limited applications. Principally, the idea of risk is very subjective, as each market participant has their own perception of risk. Rachev et al. [84] tried to summarize the intrinsic properties of risk that all investors have to take into account. These properties relate to investment diversification, computational complexity, multi-parameter dependence, asymmetry, non-linearity, and incompleteness. Unfortunately, each proposed risk measure in the literature possesses only a subset of these characteristics. Consequently, proposed risk measures are insufficient and, based on this, authors conclude that a single measure cannot be relied upon to characterize uniquely investor choices. Therefore, in order to quantify and compare portfolio risk, we describe several popular risk measures used in empirical research.

5.1.1 Deviation Measures

Standard Deviation The standard deviation, also called volatility, is one of the most common tools for measuring risk in financial markets. This metric, which is defined as the square root of the variance, shows the extent to which the returns fluctuate around their mean. The more dispersion a return has around its mean, the more volatile it is, and thus, the riskier it is. In spite of its computation simplicity, standard deviation is not a satisfactory measure due to its symmetry property and inability to consider the risk of low-probability events.

Semi-Deviation The lower semi-deviation, commonly referred to in financial mathematics literature as semi-deviation or downside deviation, is an alternative to the standard deviation or variance. In contrast to those measures, semi-deviation exclusively considers negative price movements. Thus, semi-deviation is most often used to evaluate the downside risk of an investment. Mathematically, semi-deviation is defined as

$$\sigma_{-}(X) = \sqrt{\mathrm{E}\left[(X - \mathrm{E}[X])^2 \mathbb{1}_{X \leq \mathrm{E}[X]}\right]}, \tag{5.1}$$

where $\mathbb{1}_{X \leq \mathrm{E}[X]}$ is the indicator function.

5.1.2 Quantile-based Measures

Value at Risk Value at Risk (VaR) evaluates the downside risk as the possible maximum potential change in the value of a portfolio with a given probability over a particular horizon. VaR at a given confidence level $\alpha \in [0, 1]$ is the minimum loss such that higher losses will happen at most with probability $1 - \alpha$. Let X be a profit and loss distribution, then

$$\mathrm{VaR}_{\alpha}(X) = -\inf\{x \in \mathbb{R} : F_X(x) \geq \alpha\}. \tag{5.2}$$

In other words, $\mathrm{VaR}_{\alpha}(X)$ is the α-quantile of X.

Parametric VaR does a better job of accounting for the tails of the distribution by more precisely estimating the shape of the distribution tails. Let us assume that $X \sim \mathcal{N}(\mu, \sigma^2)$, then VaR is proportional to the standard deviation:

$$\mathrm{VaR}_{\alpha}(X) = -\mu - q_{\alpha}\sigma, \tag{5.3}$$

where q_{α} is the α-quantile of the standard normal distribution. In this study, μ is assumed to be 0.

To account for the possible fat-tailed nature of downside risk, Zangari [109] and Favre and Galeano [31] provided a modified VaR calculation that considers the higher moments through the use of a Cornish Fisher expansion and collapses to standard VaR if the return stream follows a

Gaussian distribution. They arrive at their modified VaR calculation in the following manner:

$$q_{\mathrm{cf},\alpha} = q_\alpha + \frac{(q_\alpha^2 - 1)S}{6} + \frac{(q_\alpha^3 - 3q_\alpha)K}{24} - \frac{(2q_\alpha^3 - 5q_\alpha)S^2}{36}, \tag{5.4}$$

$$\mathrm{mVaR}_\alpha(X) = \mu + q_{\mathrm{cf},\alpha}\sigma, \tag{5.5}$$

where S is the skewness of X and K is the excess kurtosis of X.

Conditional Value at Risk Conditional Value at Risk (CVaR), also known as Expected Shortfall (ES), corresponds to the average of all returns in the distribution that are worse than the VaR of the portfolio at a given level of confidence:

$$\mathrm{CVaR}_\alpha(X) = -\frac{1}{\alpha}\int_0^\alpha \mathrm{VaR}_\gamma\,(X)\,d\gamma. \tag{5.6}$$

In this thesis we use modified CVaR, which replaces VaR by modified VaR, defined in Equation(5.5).

CVaR was proposed in order to overcome some of the theoretical weaknesses of VaR. and satisfies coherent risk measure properties. Let X and Y be two loss random variables. The risk measure ρ is a coherent risk measure if it satisfies the following conditions:

- **Monotonicity**. If portfolio Y always has higher returns than portfolio X under almost all scenarios then the risk of X should be less than the risk of Y:

$$\text{If } X \leq Y \, \mathrm{withprobability1}, \mathrm{then} \; \rho(X) \leq \rho(Y). \tag{5.7}$$

- **Translation invariance** implies that the addition of a sure amount of capital reduces the risk by the same amount:

$$\rho(X + c) = \rho(X) - c, \quad \forall c > 0. \tag{5.8}$$

- **Positive homogeneity** implies that the risk of a position is proportional to its size:

$$\rho(\lambda X) = \lambda^k \rho(X), \quad \forall \lambda > 0, k \in \mathbb{R}. \tag{5.9}$$

- **Sub-additivity** states that the risk of the portfolio is not greater than the sum of the risks of the portfolio components:

$$\rho\,(X + Y) \leq \rho\,(X) + \rho\,(Y)\,. \tag{5.10}$$

  Compliance with this property tends to the diversification effect.

Value–at–Risk satisfies all conditions except subadditivity. In risk management, subadditivity violations might lead financial institutions to hold less capital than desired.

5.1.3 Drawdown Measures

A psychological issue in handling risk is the tendency of people to compare the current situation with the very best one from the past. Drawdowns measure the difference between two observable quantities – the local maximum and local minimum of the portfolio's wealth. Cheklov et al. [20] defined the drawdown function as the difference between the maximum of the total portfolio return up to time t and the portfolio value at t. More formally, if $X(t),\ t \geq 0$ is a stochastic process with $X(0) = 0$, the drawdown at time T, denoted $D(T)$, is defined as:

$$D(T) = \max\left[\max_{t\in(0,T)} X\left(t\right) - X\left(T\right), 0\right]. \tag{5.11}$$

Average Drawdown The average drawdown ADD up to time T is the time average of drawdowns that have occurred up to time T:

$$\mathrm{ADD}\left(T\right) = \frac{1}{T}\int_0^T D\left(t\right) dt. \tag{5.12}$$

Maximum Drawdown The maximum drawdown in an investment references the largest peak to valley loss during the course of any investment:

$$\mathrm{MDD}\left(T\right) = \max_{t\in(0,T)} D\left(t\right). \tag{5.13}$$

Conditional Drawdown at Risk Conditional drawdown-at-risk (CDaR) corresponds to the average $\alpha \cdot 100\%$ drawdowns. In the general case,

$$\mathrm{CDaR}_\alpha(T) = \min_\gamma \left[\gamma + \frac{1}{\alpha T}\int_0^T \left(D\left(t\right) - \gamma\right)_+ dt\right]. \tag{5.14}$$

There are two limiting cases:

- $\lim_{\alpha\to 1} \mathrm{CDaR}_\alpha = \mathrm{ADD}(T)$;
- $\lim_{\alpha\to 0} \mathrm{CDaR}_\alpha = \mathrm{MDD}(T)$.

5.2 Risk-Adjusted Measures

Sharpe Ratio The Sharpe ratio is probably the most commonly used risk adjusted ratio. It is a measure of risk-adjusted return, where risk is defined as the standard deviation. More precisely, the *ex-ante* Sharpe ratio is defined as:

$$\mathrm{S} = \frac{\mathrm{E}[R_p - R_b]}{\sigma_p} = \frac{\mathrm{E}[R_p - R_b]}{\sqrt{\mathrm{Var}[R_a - R_b]}}, \tag{5.15}$$

where R_p is the portfolio return, R_b is the risk-free return, which in this study is assumed to be 0. $\mathrm{E}[R_p - R_b]$ is the expected value of the excess of the asset return over the benchmark return, and σ_p is the standard deviation of the portfolio excess return.

In practice, the Sharpe ratio is calculated *ex-post* as the sample average of logarithmic returns divided by the sample standard deviation, both estimated over the same time period. The annualized Sharpe ratio is determined by multiplying Equation (5.15) by $\sqrt{252}$.

The Sharpe ratio is useful when assets are normally distributed, since the distribution of returns is then completely described by its mean and volatility. When the distribution of returns cannot be considered Gaussian, it becomes necessary to rely on performance measures that take non-normality into account. To solve the non-normality issue, Gregoriou [38] introduced a modification of the traditional Sharpe ratio, which defined as the ratio between the excess portfolio return and its modified Value at Risk, defined by (5.5):

$$\mathrm{S}_{\mathrm{VaR}} = \frac{\mathrm{E}[R_p - R_b]}{\mathrm{mVaR}_\alpha(R_p)}. \tag{5.16}$$

In this thesis, we also use another modification of the Sharpe ratio, which uses the CVaR instead of the sample standard deviation.

Sortino Ratio The Sortino ratio is used to score a portfolio's risk-adjusted returns relative to an investment target using downside risk:

$$\mathrm{SR} = \frac{\mathrm{E}[R_p - R_b]}{\sigma_-}, \tag{5.17}$$

where σ_- is the semi-deviation of the portfolio excess return.

This is analogous to the Sharpe ratio, which measures risk-adjusted returns relative to the risk-free rate using standard deviation. When return distributions are nearly symmetrical and the target return is close to the distribution median, these two measures will produce similar results. However, as skewness grows and targets deviate from the median, results can be expected to show substantial differences.

Upside Potential Ratio The Upside Potential Ratio is a further refinement developed by Frank A. Sortino that better addresses the risk preferences of investors. It is equal to the variation of the returns above a minimum acceptable return divided by the variation of the returns below a minimum acceptable return. This favours investments with stable growth above a minimum acceptable return. More formally,

$$\mathrm{U} = \frac{\sum_{\min}^{\infty} (R_r - R_{\min})\, P_r}{\sqrt{\sum_{-\infty}^{\min} (R_r - R_{\min})_r^P}} = \frac{\mathrm{E}\left[(R_r - R_{\min})_+\right]}{\sqrt{\mathrm{E}\left[(R_r - R_{\min})_-^2\right]}}, \tag{5.18}$$

where the returns R_r have been put into increasing order. Here P_r is the probability of the return R_r and $R_{\min}$, which occurs at $r = \min$ is the minimal acceptable return. In the secondary formula, $(X)_+ = \mathbb{1}_{X \geq 0} X$ and $(X)_- = (-X)_+$, where $\mathbb{1}$ is the indicator function. In this thesis, $R_{\min}$ is set to 0.

MAR Ratio A MAR ratio gets its name from the Managed Accounts Report newsletter. The MAR ratio is also similar to the Sharpe ratio but uses the maximum drawdown as a measure of risk:

$$\text{MAR} = \frac{\text{E}[R_p - R_b]}{\text{MDD}_p}. \tag{5.19}$$

The Calmar ratio, introduced by competitor newsletter California Managed Accounts Reports [106], is another popular ratio that measures the same metrics but instead only looks at the past 36 months. Despite its widespread usage in comparing the performance of commodity trading advisors, hedge funds, and trading techniques, the MAR and Calmar ratios' emphasis on drawdown provides a rather limited view of risk in comparison to other gauges.

6 Backtesting Methodology

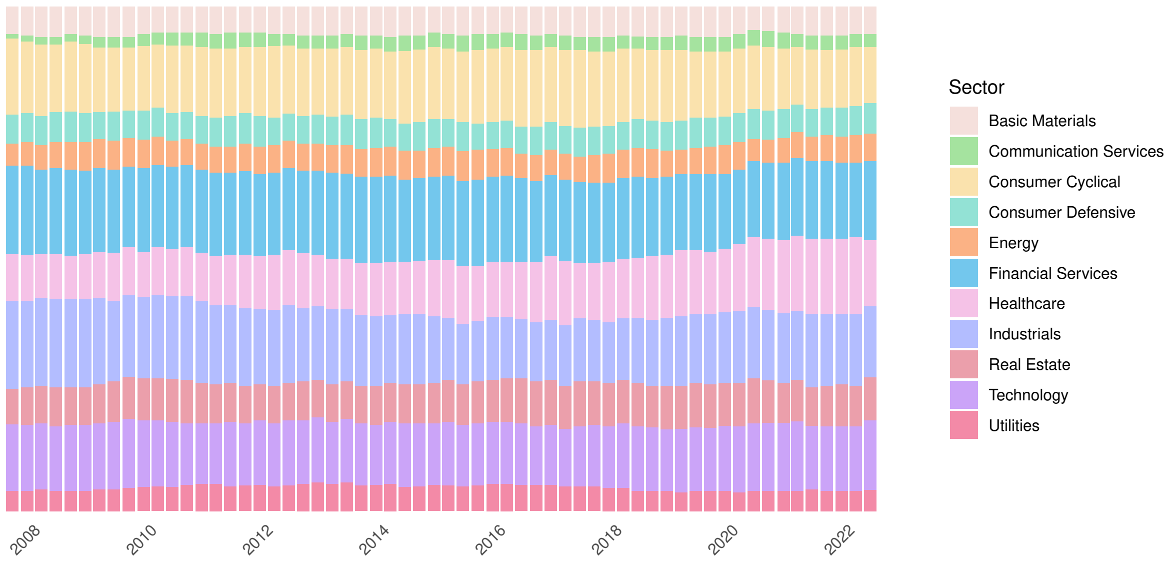
6.1 Data

The data used in the empirical study are the daily closing prices of the Russell 3000 historical constituents from 2005-10-01 to 2022-10-01, provided by Norgate Data. We select the first two years for initial optimization and leave an out-of-sample period of 15 years from 2007-10-01 for walk-forward analysis. The data include delisted stocks as well as actively trading ones and thus are survivorship bias-free. All dividend proceeds are assumed to be reinvested. Companies are also assigned to a specific economic sector using the Global Industry Classification Standard (GICS) methodology. GICS classification data is provided by Sharadar.

6.2 Sampling Methodology

In order to obtain relatively narrow confidence intervals for the medians of performance ratios described in Section 5, we create 100 portfolios using proportional stratified random sampling, where strata are defined by GICS sectors. First, 350 equities are sampled proportionally from the Russell 3000 investment universe. Then, 15% of the portfolio constituents are replaced each quarter using the same proportional stratified random sampling methodology. This way, we simulate well-diversified and realistic portfolios with a turnover ratio of around 60%, which is common across actively managed mutual funds. Figure 10 shows the proportions of GICS sectors at each quarter in one such portfolio.

Figure 10: GICS sector proportions



At each rebalancing period t equities must have at least 252 observations from $t-253$ to $t-1$ to be included in the portfolio so that enough observations are available for clustering and covariance estimation.

6.3 Walk-Forward Optimization

In order to measure postoptimization portfolio performance, all asset allocation techniques are tested using walk-forward optimization, which is a specific application of cross-validation technique for time-series data. The benefits of walk-forward analysis include [81]:

- Evaluation of the likelihood of an optimization algorithm performing well in real-time trading;
- Measurement of the robustness of the asset allocation strategy;
- Maintenance of superior performance through more effective adaptation to changing market conditions.

This thesis employs quarterly portfolio rebalancing with a rolling in-sample optimization window. Since we are mostly interested in practical applications in which the number of assets N is relatively large compared to the sample size T, the rolling window size is restricted to a maximum of 504 days, which also depends on the availability of price data.

6.4 Optimization Methods

Nested Clustered Optimization was selected as the primary comparison framework for clustering techniques and covariance estimation methods. Since NCO is split into intra- and inter-cluster optimizations, we investigate all combinations between GMV, MD, and RP portfolios.

All three clustering approaches described in Section 4 are applied in this study. Additional clustering details and hyperparameters are provided below:

- All clustering techniques search for the optimal number of clusters, ranging from 6 to 18. Assuming that the number of stocks in each cluster is appropriately balanced, this range is expected to provide a good compromise between theoretical clustering benefits, interpretability, and computing performance.
- Hierarchical clustering:
  - Ward linkage was selected as it tends to produce balanced clusters, which is a desirable property that might reduce weight concentration and improve the computational performance of covariance matrix estimation.
- Partitioning Around Medoids:
  - LOWESS technique applies weighted linear least squares regression over the smoothing span of 10% of return observations;
  - When calculating DTW distance, LCM space is restricted by Sakoe-Chiba window of 10% the series length;
  - Number of clusters is determined by voting between CH, DB, DB*, Sil, SF, and COP cluster validity indices. In case of a tie, the number of clusters is determined using the Silhouette index.
- Discriminative Functional Mixture Model:
  - DFM is applied on standardized cumulative logarithmic returns transformed using B-spline basis of order $n = 25$ and roughness penalty $\lambda = 16$. Order n was selected empirically, while the roughness penalty was selected using generalized cross-validation on a subset of data between 2006-10-01 and 2007-10-01;
  - $\text{DFM}_{[\alpha_j,\beta_k]}$ submodel is used, which constrains Σ_k to be diagonal and common across groups, whereas parameter β is left unconstrained.

Three covariance estimation methods, namely, sample covariance, the DCC(1,1)-EGARCH(1,1) model, and the DCC(1,1)-MSM(4) model with nonlinear shrinkage, are compared in this study. Distributional assumptions for MSM models are chosen from Normal, Student's t, Normal-inverse Gaussian, Johnson's S_U, and Generalized Hyperbolic Skew Student's t distributions based on Akaike information criterion for each equity separately.

To summarize, each portfolio is optimized using 81 combinations of clustering methods, covariance matrix estimation approaches, intra- and inter-cluster optimization strategies. Portfolios optimized using NCO are also benchmarked against equally weighted portfolios as well as the standard HRP and HERC algorithms.

6.5 Portfolio Constraints

Portfolio managers often impose limits on the portfolio weights of securities or groups of securities to avoid extreme weights that may emerge from model inaccuracies. Jagannathan and Ma [44] provide a theoretical justification for such practices. They demonstrate that the no short-selling constraints are equivalent to reducing the estimated covariances, whereas upper bounds have the reverse effect. For instance, equities with strong correlations with each other tend to receive negative portfolio weights. Therefore, when their covariance is decreased (which is equivalent to the effect of imposing no short-selling constraints), these negative weights diminish in magnitude. Similarly, stocks that have low covariances with other stocks tend to get overweighted, and the impact of these overweighted stocks might be reduced by increasing the corresponding covariances. Limiting single asset exposures also helps to control portfolio turnover. Nevertheless, while restrictions can help ensure robustness and stability, they must be used with caution. If the constraints are too tight, the portfolio allocation will be completely determined by the constraints rather than the forecasted expected returns and their covariances [48].

In this thesis, the asset weights within each cluster, $c_k \in \{c_1, \ldots, c_K\}$, are constrained using the following upper and lower bounds:

$$\frac{1}{10|c_k|} \leq w_{k,i}^{(\text{intra})} \leq \min\left\{\max\left\{\frac{4}{|c_k|},\ 0.25\right\},\ 1\right\}, \quad i = 1, \ldots, |c_k|. \tag{6.1}$$

The box constraints applied to inter-cluster weights are defined as follows:

$$\frac{1}{10K} \leq w_k^{(\text{inter})} \leq \min\left\{\max\left\{\frac{4}{K},\ 0.4\right\},\ 1\right\}, \quad k = 1, \ldots, K. \tag{6.2}$$

6.6 Transaction Costs

Fixed transaction costs (TC) of \$0.005 per share, with a minimum of \$1 and a maximum of 1% of trade value, are incorporated into the equity curve calculation *ex post*. In the event the calculated maximum per order is less than the minimum per order, the maximum per order is assessed. Given equity price $p > 0$ and trade quantity $q > 0$, transaction cost model could be formulated as follows:

$$\text{TC} = \min\left\{\max\left\{\frac{q}{200},\ 1\right\},\ \max\left\{\frac{pq}{100},\ 1\right\}\right\}. \tag{6.3}$$

Additional regulatory fees for selling shares include:

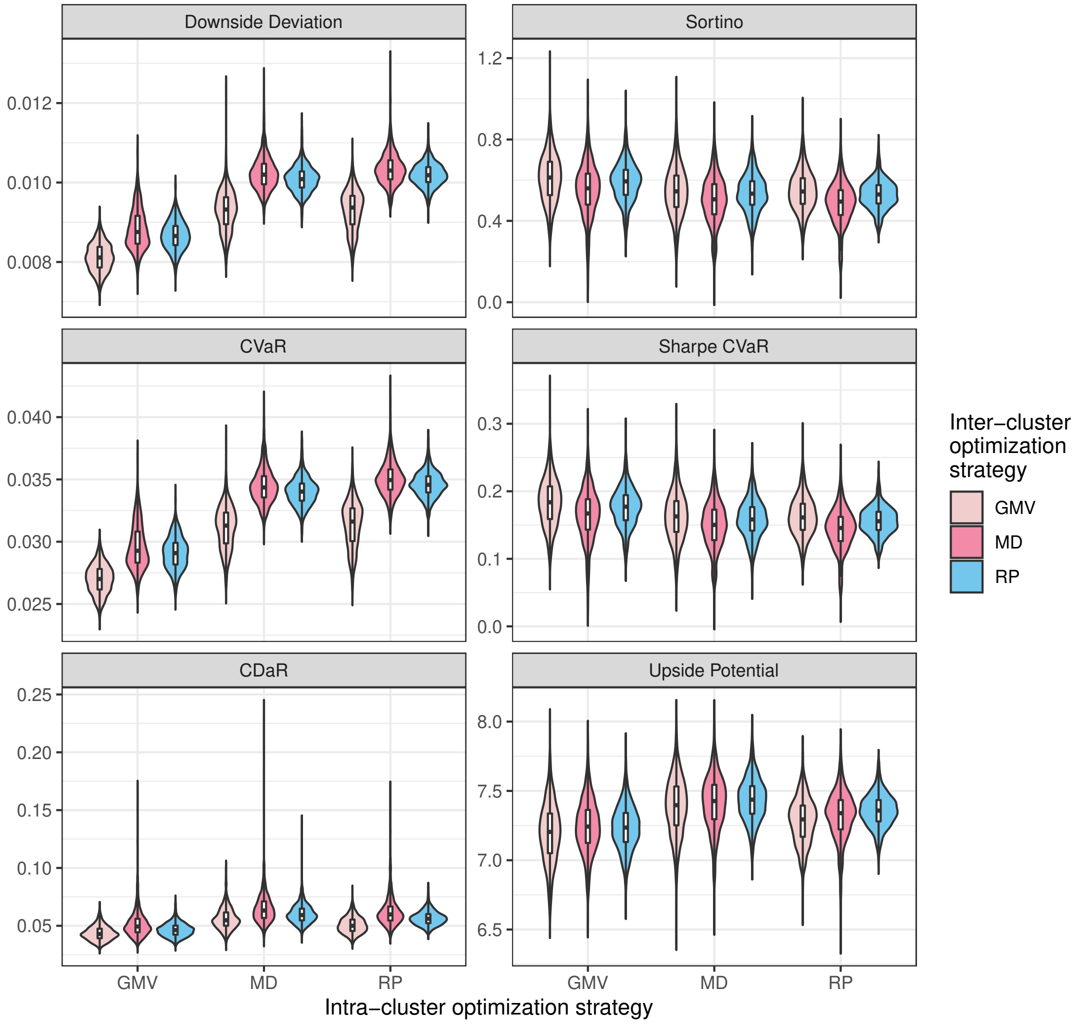
- SEC Transaction Fee: $0.0000229 \cdot p \cdot q$;
- FINRA Trading Activity Fee: $\min\{0.00013 \cdot q,\ 6.49\}$.

It is assumed that trading starts with the initial capital of \$100MM. Bid-ask spreads and indirect costs such as slippage, which are based on the liquidity of various securities, would require their own forecasting models and are thus disregarded in this study.

7 Results

The summary statistics and out-of-sample performance ratios are displayed using violin plots that feature the probability density of the data smoothed using a kernel density estimator. All metrics are calculated using daily price data. Figure 11 depicts the comparison between portfolios constructed using various combinations of intra- and inter-cluster optimization methodologies. It comes as no surprise that portfolio risk measured in terms of downside deviation, CVaR, and CDaR favor GMV optimization strategies. Moreover, risk-adjusted performance metrics such as Sortino and modified Sharpe ratios exhibit somewhat better values when utilizing GMV, which might be explained by the low-volatility anomaly.

Figure 11: Performance measure distributions



7.1 Risk-Based Performance Evaluation

This section presents a more in-depth view of the distributions of portfolio risk measures by contrasting the effects of clustering approaches and covariance estimation methods. Intra-cluster optimization strategies are represented by rows in a matrix of panels, whereas inter-cluster optimization strategies are represented by columns (same holds for other similarly structured violin plots).

Using hierarchical clustering in conjunction with MD inter-cluster optimization results in a greater variance and heavier distribution tails, whereas HAC with GMV inter-cluster optimization results in the lowest downside deviation and CVaR when compared to other combinations. Differences in covariance estimation approaches appear to be insignificant, with DCC-MSM and DCC-EGARCH yielding fairly comparable results in all circumstances and failing to consistently beat the sample covariance matrix.

Figure 12: Downside deviation distributions

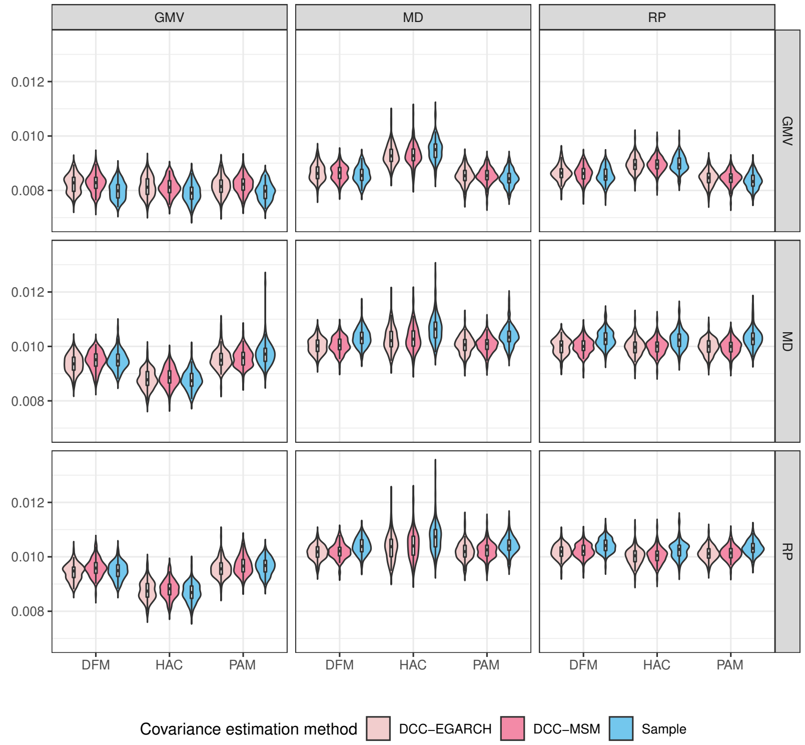


Figure 13: CVaR distributions

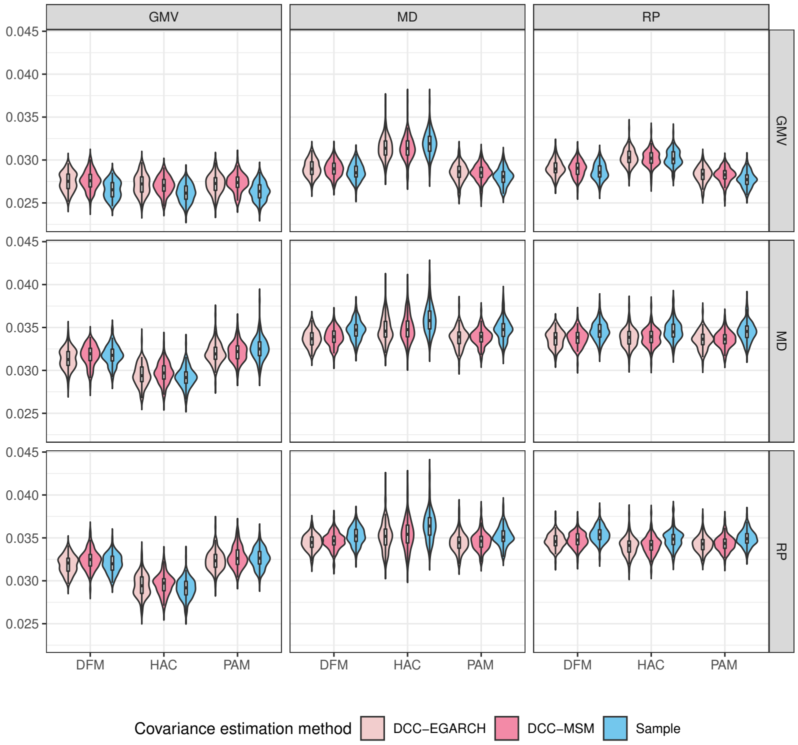
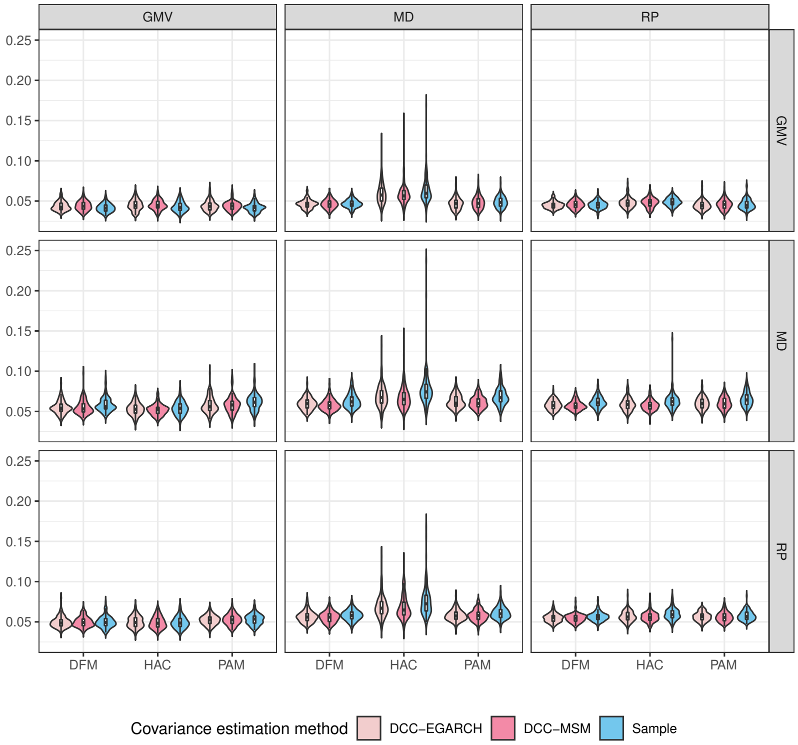


Figure 14: CDaR distributions



7.2 Risk-Adjusted Performance Evaluation

Figures 15, 16, and 17 provide a more comprehensive view in terms of risk-adjusted performance metrics. The most notable result is that hierarchical clustering with MD inter-cluster optimization has a relatively higher variance, whereas the strategies with RP intra- and inter-cluster optimization have a lower variance of out-of-sample performance ratios. On average, PAM and DFM clustering techniques produce similar distributions and sightly outperform hierarchical clustering. Interestingly, portfolios optimized with the MD intra-cluster approach provide greater upside potential ratios, indicating that the higher risk reported in figures 12, 13, and 14 may be balanced by a relatively strong upside performance.

Figure 15: Sortino ratio distributions

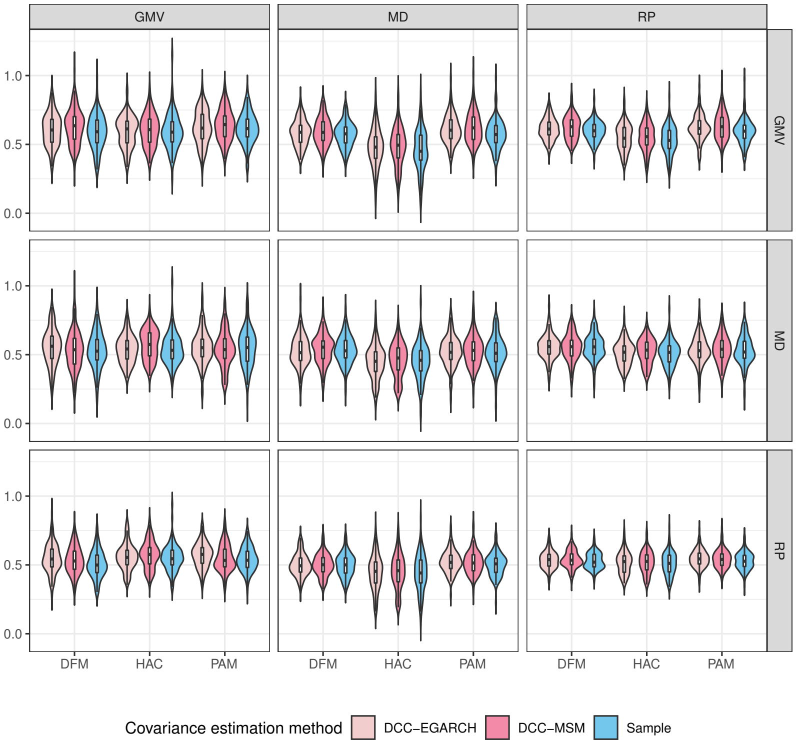


Figure 16: Sharpe$_{\text{CVaR}}$ ratio distributions

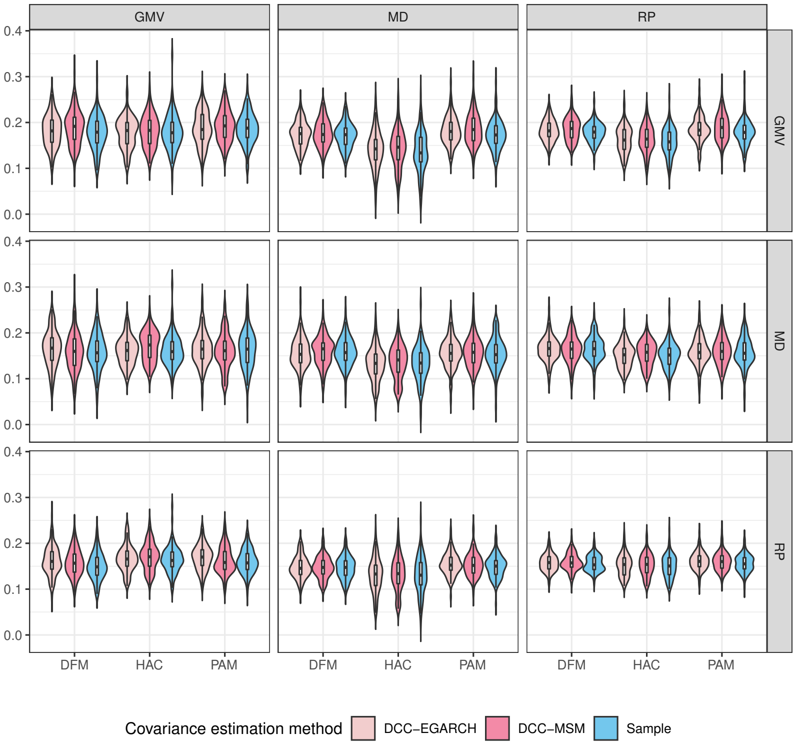
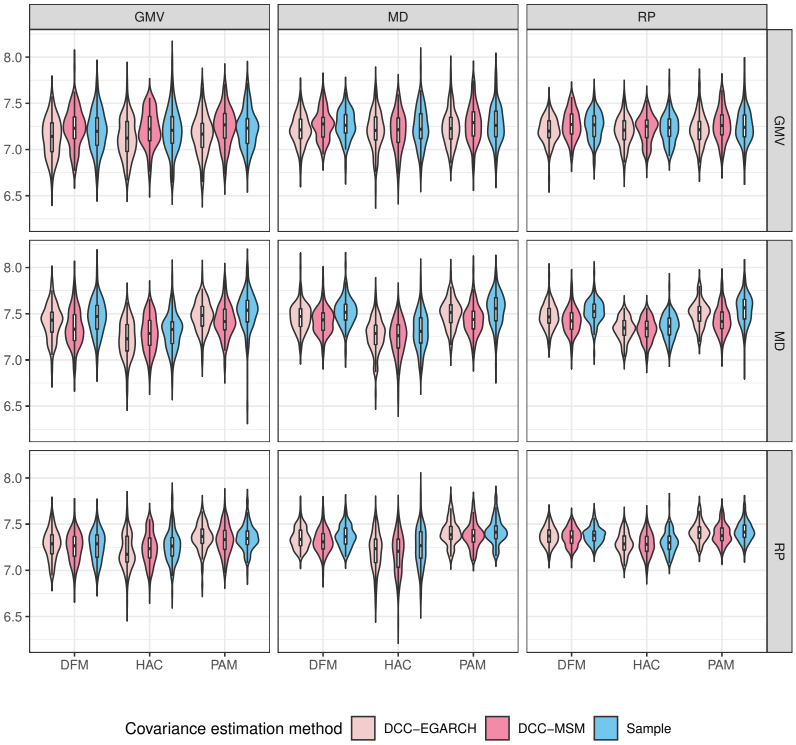


Figure 17: Upside potential distributions



7.3 Portfolio Turnover and Transaction Costs

This section investigates the stability of weight concentrations and the associated transaction costs. A large portfolio turnover might be a problem, as it indicates that the portfolio's maintenance costs are greater than those of comparable portfolios. Clearly, using the sample covariance matrix as the covariance estimation technique results in more stable portfolios, whereas the DCC-MSM method has the largest turnover and transaction costs, making it more suitable for active portfolio management strategies. More importantly, figures 18 and 19 imply that the similar risk characteristics and risk-adjusted performance of various covariance estimation approaches may be related to the fact that the benefits of a more precise estimate are negated by transaction costs. Furthermore, a relatively strong risk-based performance of GMV-type optimization is accompanied by the highest portfolio turnover. Although GMV-type optimization results in higher turnover, rebalancing MD-type portfolios is actually more expensive due to the non-linear relationship between turnover and transaction costs. Finally, using risk parity as an intra- and inter-cluster optimization technique results in portfolios with the lowest median turnover and transaction costs as well as the lowest variance of the aforementioned metrics.

Figure 18: Quarterly turnover ratios

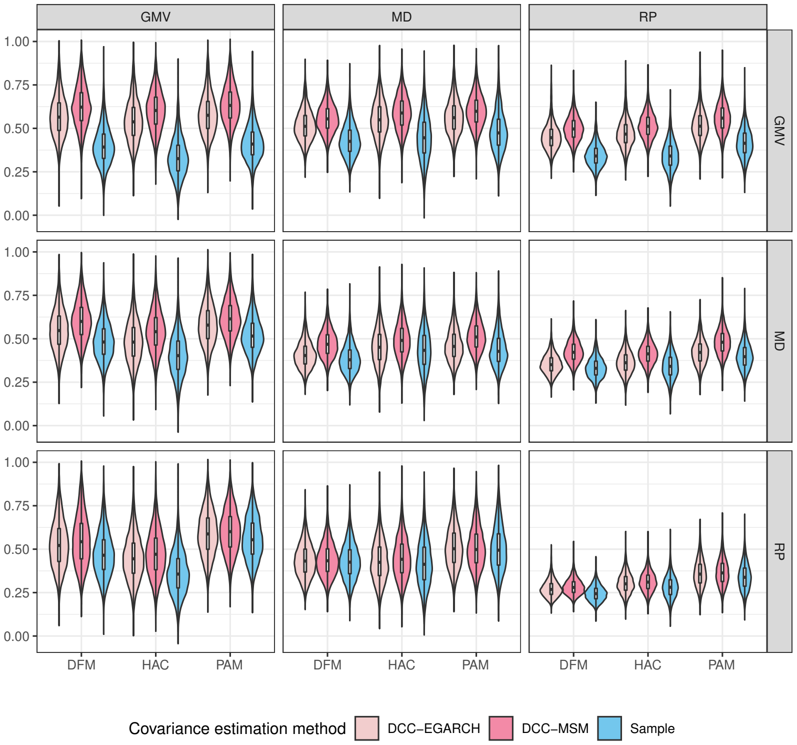
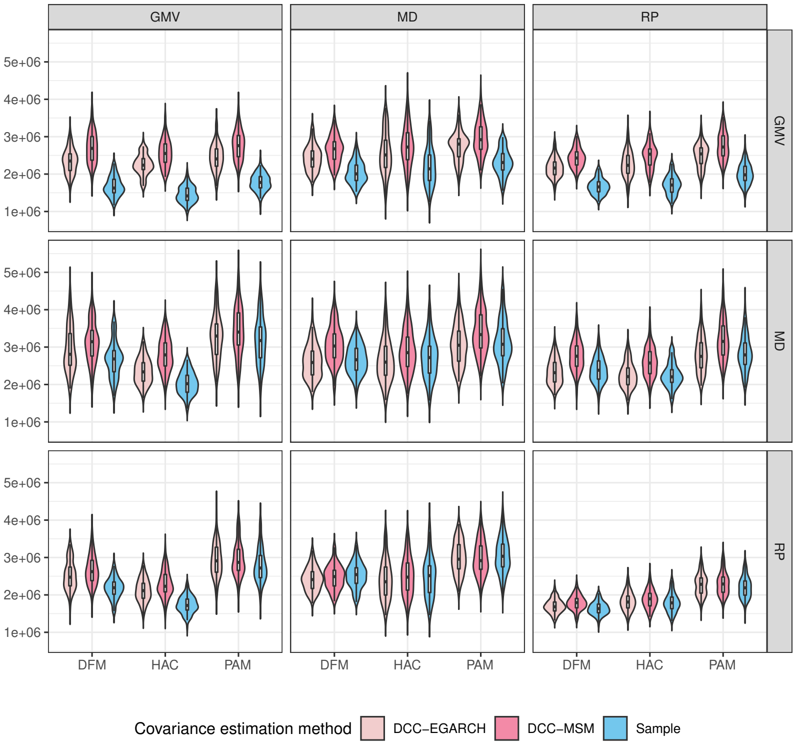


Figure 19: Transaction Costs



7.4 Comparison with Benchmark Portfolios

Based on overall risk characteristics and risk-adjusted performance ratios, we select the asset allocation strategy with GMV intra- and inter-cluster optimization, PAM clustering, and the DCC-MSM model as the optimal method. Nevertheless, the choice heavily depends on the investor's risk profile as well as desired portfolio turnover. In this section, we compare the selected method with the equally weighted portfolio, HRP, and HERC optimization strategies.

Table 1 displays 95% confidence intervals for the medians of all performance ratios. Median confidence intervals are calculated using the Clopper–Pearson interval method for calculating binomial confidence intervals. The Clopper–Pearson interval could be presented in a format that uses quantiles from the beta distribution:

$$B\left(\frac{\alpha}{2};\ x, n-x+1\right) < \theta < B\left(1-\frac{\alpha}{2};\ x+1, n-x\right), \tag{7.1}$$

where x is the number of successes, n is the number of trials, and $B(p; v, w)$ is the pth quantile from a beta distribution with shape parameters v and w. Hence, the 95% confidence interval for the median with a sample size $n = 100$ is between the 40th and the 60th observations in the ordered data.

Table 1: Confidence intervals for medians of performance ratios

	EW	**HERC**	**HRP**	**Proposed method**
Standard Deviation	[0.01534, 0.01547]	[0.01423, 0.01484]	[0.01296, 0.01318]	[0.01113, 0.01132]
Downside Deviation	[0.01119, 0.01129]	[0.01033, 0.01083]	[0.00942, 0.00964]	[0.00811, 0.00832]
VaR	[0.02477, 0.02504]	[0.02096, 0.02185]	[0.02022, 0.02063]	[0.01562, 0.01596]
CVaR	[0.05287, 0.05418]	[0.03380, 0.04062]	[0.04413, 0.04680]	[0.02714, 0.02767]
MDD	[0.56360, 0.58086]	[0.52411, 0.56398]	[0.50243, 0.51779]	[0.42900, 0.46217]
CDaR	[0.06245, 0.06645]	[0.05345, 0.05894]	[0.04919, 0.05123]	[0.04223, 0.04517]
Sharpe Ratio (σ)	[0.32194, 0.34522]	[0.30806, 0.35667]	[0.41704, 0.43584]	[0.44763, 0.48693]
Sharpe Ratio (VaR)	[0.19900, 0.21516]	[0.21139, 0.24937]	[0.26706, 0.27668]	[0.31933, 0.34629]
Sharpe Ratio (CVaR)	[0.09353, 0.09903]	[0.11493, 0.16137]	[0.11767, 0.12760]	[0.18360, 0.19813]
Sortino Ratio	[0.44209, 0.47597]	[0.42535, 0.49552]	[0.57335, 0.59460]	[0.61009, 0.66264]
Upside Potential	[7.48123, 7.52636]	[6.92660, 7.05538]	[7.27741, 7.37650]	[7.21092, 7.28881]

The results reveal that the proposed method outperforms all benchmark strategies across all performance metrics except upside potential, which is highest for the equally weighted portfolio but comes at a cost of a substantially higher tail risk.

Figure 20: Portfolio Turnover and Transaction Costs

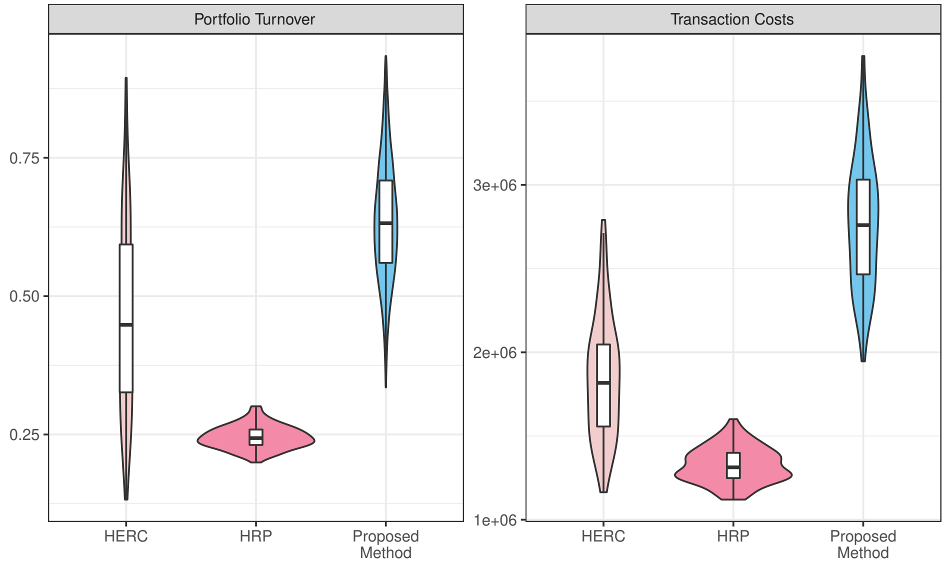
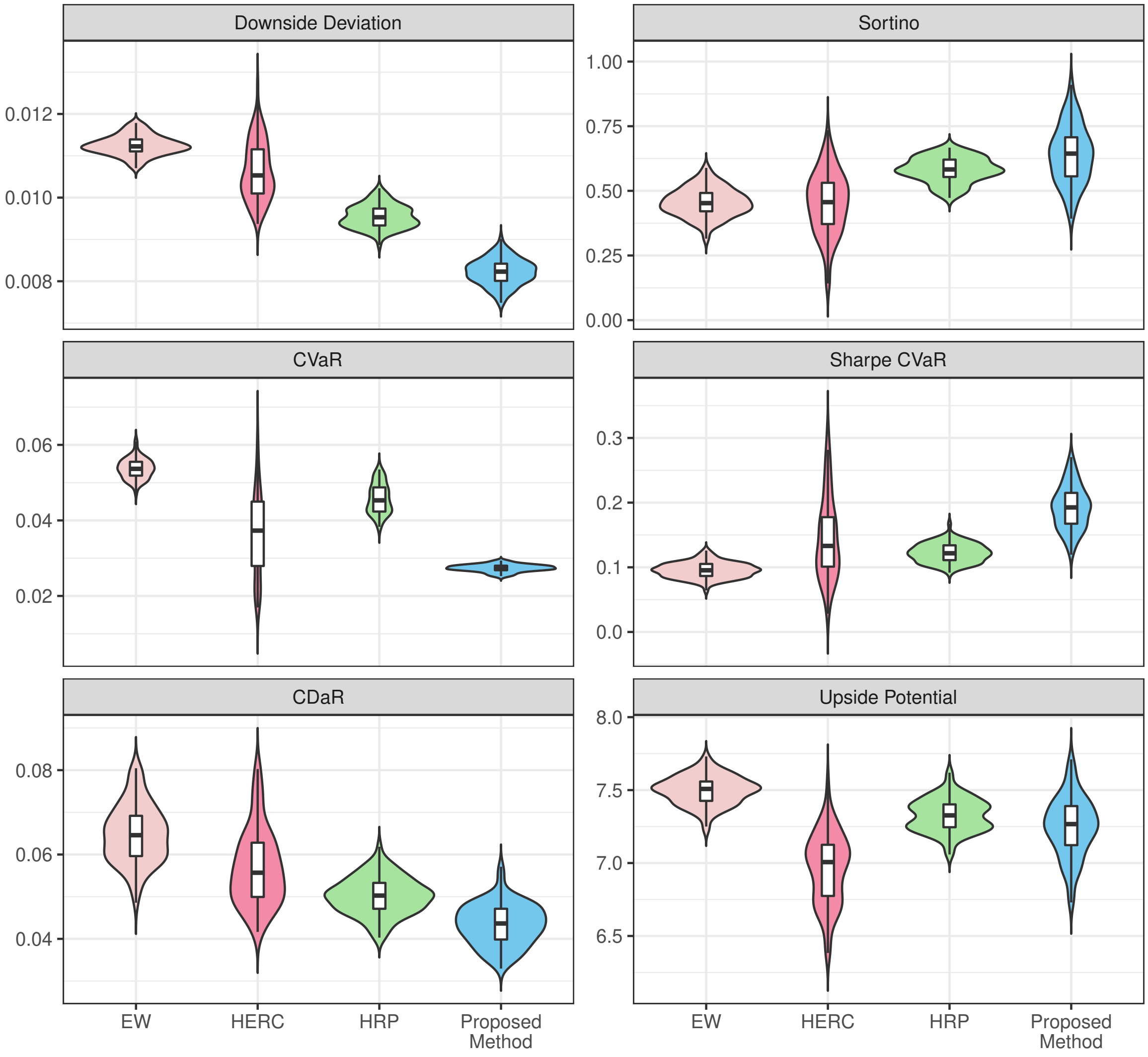


Figure 20 suggests that the proposed method results in a more active rebalancing approach that, despite higher turnover and transaction costs, still demonstrates superior risk characteristics as well as risk-adjusted performance. Selected performance measures are also displayed in Figure 21.

Figure 21: Benchmark performance measure distributions

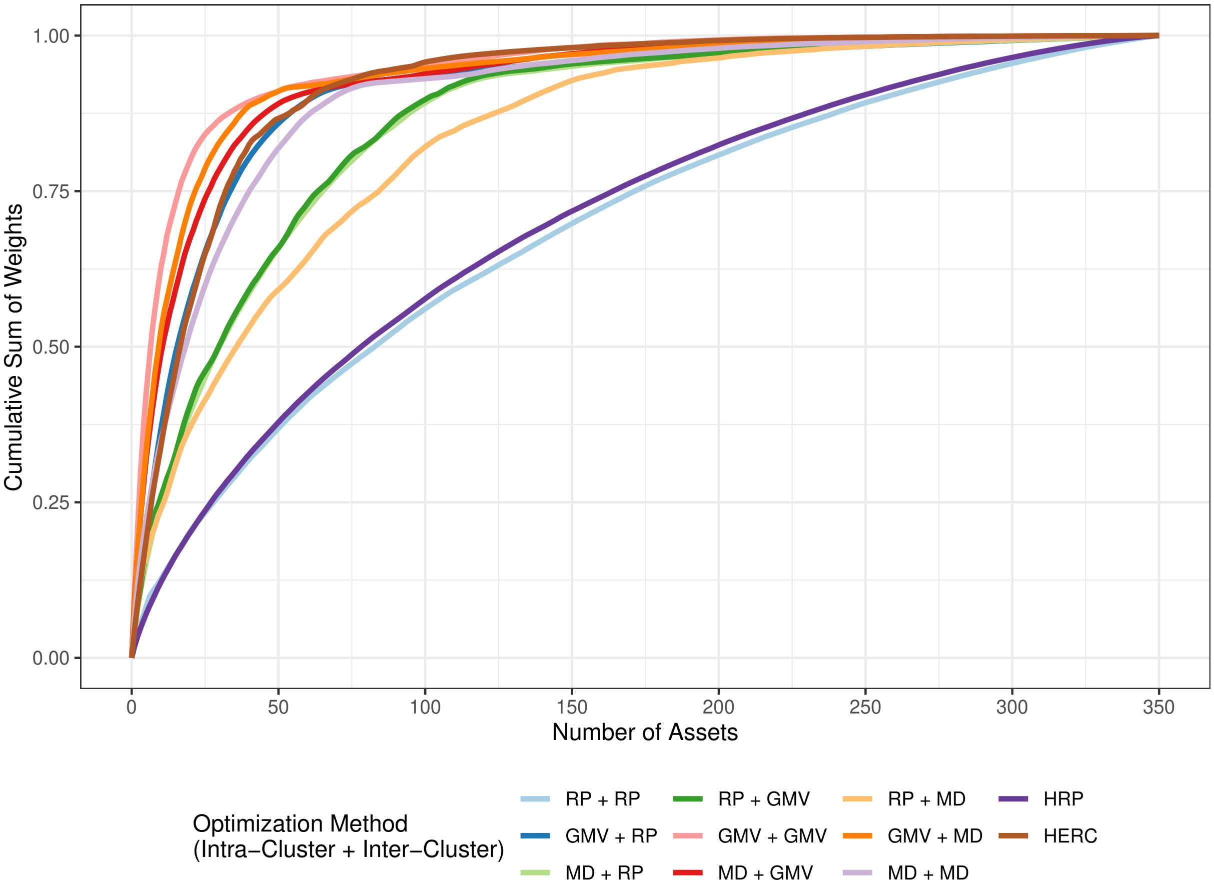


7.5 Portfolio Concentration

Financial advisors often tell their customers to have diverse portfolios to reduce unsystematic risk and optimize the average periodic return for a given amount of volatility. However, it is also possible to have too much diversification. Minimizing the concentration of portfolio weights leads to the well-known equally weighted portfolio as the optimal choice. The optimal portfolio diversification consists of holding a number of individual assets that is large enough to practically remove unsystematic risk but small enough to focus on the greatest opportunities. Over-diversification occurs when each incremental investment added to a portfolio lowers the expected return to a greater degree than the associated reduction in the risk profile. Frank Reilly and Keith Brown reported that in one set of studies for randomly selected stocks, “about 90% of the maximum benefit of diversification was derived from portfolios of 12 to 18 stocks” [88].

In this section, we measure the diversification in terms of weight concentration. Figure 22 depicts the least number of assets necessary to obtain a specified sum of weights for each combination of intra-cluster and inter-cluster methods, as well as benchmark portfolios. In this graph, median values are compared.

Figure 22: Weight Concentration



The findings show that risk concentrations are typically reduced in portfolios designed with the risk parity strategy. The weight concentrations of the NCO portfolio with risk parity as the intra- and inter-cluster optimization technique are comparable to those of the original HRP portfolio, which also explains the lower turnover rates and transaction costs associated with both portfolios. GMV-type optimization, on the other hand, has the highest weight concentration until 90% of the cumulative sum of weights is attained. With this method, around 60% of capital is allocated across 10 equities, 80% is allocated across 20 equities, and 90% is distributed across 45 equities. Nevertheless, as shown in Figure 11, a relatively high weight concentration does not necessarily correspond to a higher risk, which implies that other strategies might suffer from over-diversification.

8 Conclusion

In this thesis, we examined various risk-based optimization strategies, covariance matrix estimation methods, and clustering algorithms in terms of their contribution to portfolio risk and risk-adjusted returns. The empirical study centered on large-scale portfolio optimization with practical weight constraints and transaction costs. In order to model plausible scenarios and obtain relatively narrow confidence intervals for various performance ratios, optimization was performed on 100 randomly sampled 350-asset portfolios featuring realistic diversification across 11 GICS sectors.

It was demonstrated that among various combinations of intra- and inter-cluster optimization strategies, global minimum variance leads to higher weight concentration and portfolio turnover yet provides superior out-of-sample performance in terms of risk as well as risk-adjusted returns.

Two clustering algorithms were proposed: partitioning around medoids with dynamic time warping distance and the discriminative functional mixture model. To our knowledge, neither of these methods has been previously utilized for portfolio optimization. Clustering methodologies were compared against hierarchical agglomerative clustering with Ward's linkage. In general, the performance of the three clustering methods is comparable, with hierarchical clustering exhibiting somewhat better risk characteristics when the inter-cluster optimization approach is global minimum variance and slightly worse when it is maximum diversification.

This thesis also presents the Markov switching multifractal model with novel distributional assumptions and the dynamic conditional correlation structure with nonlinear shrinkage. This covariance estimation method was compared against the DCC-EGARCH model and the sample covariance matrix in the context of risk-based asset allocation. Since both the DCC-MSM and DCC-EGARCH methods are sensitive to the heteroscedasticity of equity returns, these models resulted in higher turnover rates. Consequently, the benefits of more precise estimates were offset by increased transaction costs. From a practitioner's point of view, this observation implies that there is little benefit to choosing computationally intensive methods such as DCC-GARCH when backtesting large-scale portfolios.

Overall, the choice of asset allocation strategy, clustering algorithm, and covariance estimation method heavily depends on the investor's risk profile as well as desired portfolio turnover and weight concentration. Based on risk characteristics and risk-adjusted performance ratios, this thesis suggests a combination of GMV intra- and inter-cluster optimization, PAM clustering with DTW distance, and DCC-MSM covariance matrix estimation techniques. The proposed methodology was able to consistently outperform the equally weighted portfolio as well as standard HRP and HERC optimization strategies.

Supplementary Material

The code used in the empirical study can be found at the following link: https://gitfront.io/r/mstagys/FtJ5iLgS6oms/thesis/.

References

[1] G. P. Aielli. Dynamic Conditional Correlation: On Properties and Estimation. *Journal of Business & Economic Statistics*, 31:282 – 299, 2011.

[2] I. E. Aldridge. Big Data in Portfolio Allocation: A New Approach to Successful Portfolio Optimization. In *The Journal of Financial Data Science*, 2019.

[3] E. Alipour, C. Adolphs, A. Zaribafiyan, and M. Rounds. Quantum-inspired hierarchical risk parity. *White paper, 1Qbit*, 2016.

[4] O. Arbelaitz, I. Gurrutxaga, J. Muguerza, J. M. Pérez, and I. Perona. An extensive comparative study of cluster validity indices. *Pattern Recognition*, 46(1):243–256, 2013.

[5] D. Ardia, G. Bolliger, K. Boudt, and J.-P. Gagnon-Fleury. The impact of covariance misspecification in risk-based portfolios. *Annals of Operations Research*, 254(1-2):1–16, 2017.

[6] D. Bailey and M. Lopez de Prado. Balanced baskets: a new approach to trading and hedging risks. *The Journal of Investment Strategies*, 1:21–62, 09 2012.

[7] D. J. Berndt and J. Clifford. Using Dynamic Time Warping to Find Patterns in Time Series. In *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining*, AAAIWS'94, page 359–370. AAAI Press, 1994.

[8] D. Bertsimas and A. Lo. Optimal control of execution costs. *Journal of Financial Markets*, 1(1):1–50, 1998.

[9] F. Black and R. B. Litterman. Asset Allocation. *The Journal of Fixed Income*, 1(2):7–18, 1991.

[10] T. Bollerslev. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3):307–327, 1986.

[11] C. Bouveyron, E. Côme, and J. Jacques. The discriminative functional mixture model for a comparative analysis of bike sharing systems. *The Annals of Applied Statistics*, 9(4):1726–1760, 2015.

[12] M. W. Brandt. Portfolio Choice Problems. In *Handbook of Financial Econometrics: Tools and Techniques*, volume 1 of *Handbooks in Finance*, pages 269–336. North-Holland, San Diego, 2010.

[13] B. Bruder and T. Roncalli. Managing risk exposures using the risk budgeting approach. MPRA Paper 37246, University Library of Munich, Germany, Jan. 2012.

[14] T. Burggraf. Beyond risk parity – A machine learning-based hierarchical risk parity approach on cryptocurrencies. *Finance Research Letters*, 38:101523, 2021.

[15] T. Caliński and J. Harabasz. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3:1–27, 1974.

[16] L. Calvet and A. Fisher. Forecasting multifractal volatility. *Journal of Econometrics*, 105(1):27–58, 2001.

[17] L. Cappiello, R. F. Engle, and K. Sheppard. Asymmetric Dynamics in the Correlations of Global Equity and Bond Returns. *Journal of Financial Econometrics*, 4(4):537–572, 09 2006.

[18] P. J. A. Cayton and D. S. Mapa. Time-varying conditional Johnson Su density in Value-at-Risk methodology. *The Philippine review of economics*, 52:23–44, 2015.

[19] L. K. Chan, J. Karceski, and J. Lakonishok. On Portfolio Optimization: Forecasting Covariances and Choosing the Risk Model. Working Paper 7039, National Bureau of Economic Research, March 1999.

[20] A. Chekhlov, S. Uryasev, and M. Zabarankin. Portfolio Optimization With Drawdown Constraints. In P. M. Pardalos, A. Migdalas, and G. Baourakis, editors, *Supply Chain And Finance*, World Scientific Book Chapters, chapter 13, pages 209–228. World Scientific Publishing Co. Pte. Ltd., September 2004.

[21] V. K. Chopra and W. T. Ziemba. The Effect of Errors in Means, Variances, and Covariances on Optimal Portfolio Choice. *The Journal of Portfolio Management*, 19(2):6–11, 1993.

[22] Y. Choueifaty and Y. Coignard. Toward Maximum Diversification. *The Journal Of Portfolio Management*, 35:40–51, 2008.

[23] Y. Choueifaty, T. Froidure, and J. Reynier. Properties of the Most Diversified Portfolio. *ERN: Econometric Modeling in Financial Economics (Topic)*, 2011.

[24] R. G. Clarke, H. de Silva, and S. Thorley. Minimum-Variance Portfolios in the U.S. Equity Market. *The Journal of Portfolio Management*, 33(1):10–24, 2006.

[25] J. Daly, M. Crane, and H. Ruskin. Random matrix theory filters in portfolio optimisation: A stability and risk assessment. *Physica A: Statistical Mechanics and its Applications*, 387(16):4248–4260, 2008.

[26] D. L. Davies and D. W. Bouldin. A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1:224–227, 1979.

[27] V. DeMiguel, L. Garlappi, and R. Uppal. Optimal Versus Naive Diversification: How Inefficient is the 1/N Portfolio Strategy? *The Review of Financial Studies*, 22(5):1915–1953, 12 2007.

[28] F. G. Duarte and L. N. De Castro. A Framework to Perform Asset Allocation Based on Partitional Clustering. *IEEE Access*, 8:110775–110788, 2020.

[29] R. Engle. Dynamic Conditional Correlation: A Simple Class of Multivariate Generalized Autoregressive Conditional Heteroskedasticity Models. *Journal of Business & Economic Statistics*, 20(3):339–350, 2002.

[30] R. F. Engle, O. Ledoit, and M. Wolf. Large Dynamic Covariance Matrices. *Journal of Business & Economic Statistics*, 37(2):363–375, 2019.

[31] L. Favre and J.-A. Galeano. Mean-Modified Value-at-Risk Optimization with Hedge Funds. *The Journal of Alternative Investments*, 5(2):21–25, 2002.

[32] Y. Feng and D. P. Palomar. SCRIP: Successive Convex Optimization Methods for Risk Parity Portfolio Design. *IEEE Transactions on Signal Processing*, 63(19):5285–5300, 2015.

[33] J.-D. Fermanian and H. Malongo. On the Stationarity of Dynamic Conditional Correlation Models. Working Papers 2013-26, Center for Research in Economics and Statistics, 2013.

[34] G. Frahm and C. Wiechers. On the diversification of portfolios of risky assets. Discussion Papers in Econometrics and Statistics 2/11, University of Cologne, Institute of Econometrics and Statistics, 2011.

[35] H. Geiger and J. Plagge. Minimum variance indices. *Deutsche Börse AG, Frankfurt*, 2007.

[36] T. Giorgino. Computing and Visualizing Dynamic Time Warping Alignments in R: The dtw Package. *Journal of Statistical Software*, 31(7):1–24, 2009.

[37] L. R. Glosten, R. Jagannathan, and D. E. Runkle. On the Relation between the Expected Value and the Volatility of the Nominal Excess Return on Stocks. *The Journal of Finance*, 48(5):1779–1801, 1993.

[38] G. N. Gregoriou and J.-P. Gueyie. Risk-Adjusted Performance of Funds of Hedge Funds Using a Modified Sharpe Ratio. *The Journal of Wealth Management*, 6(3):77–83, 2003.

[39] I. Gurrutxaga, I. Albisua, O. Arbelaitz, J. I. Martín, J. Muguerza, J. M. Pérez, and I. Perona. SEP/COP: An Efficient Method to Find the Best Partition in Hierarchical Clustering Based on a New Cluster Validity Index. *Pattern Recogn.*, 43(10):3364–3373, oct 2010.

[40] C. Hafner and O. Reznikova. On the estimation of dynamic conditional correlation models. *Computational Statistics & Data Analysis*, 56(11):3533–3545, 2012.

[41] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer series in statistics. Springer, 2009.

[42] R. A. Haugen and N. L. Baker. The efficient market inefficiency of capitalization–weighted stock portfolios. *The Journal of Portfolio Management*, 17(3):35–40, 1991.

[43] W. Huang. Performance of Hierarchical Equal Risk Contribution Algorithm in China Market. *ERN: Other Econometric Modeling: Capital Markets - Risk (Topic)*, 2020.

[44] R. Jagannathan and T. Ma. Risk Reduction in Large Portfolios: Why Imposing the Wrong Constraints Helps. *The Journal of Finance*, 58(4):1651–1683, 2003.

[45] N. E. Karoui. Spectrum Estimation for Large Dimensional Covariance Matrices Using Random Matrix Theory. *The Annals of Statistics*, 36(6):2757–2790, 2008.

[46] E. J. Keogh and M. J. Pazzani. Derivative dynamic time warping. In *In SIAM International Conference on Data Mining*, 2001.

[47] M. Kim and R. Ramakrishna. New indices for cluster validity assessment. *Pattern Recognition Letters*, 26(15):2353–2363, 2005.

[48] P. N. Kolm, R. Tütüncü, and F. J. Fabozzi. 60 Years of portfolio optimization: Practical challenges and current trends. *European Journal of Operational Research*, 234(2):356–371, 2014.

[49] M. Kritzman, S. Page, and D. Turkington. In Defense of Optimization: The Fallacy of 1/N. *Financial Analysts Journal*, 66(2):31–39, 2010.

[50] O. Ledoit and M. Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004.

[51] O. Ledoit and M. Wolf. Nonlinear shrinkage estimation of large-dimensional covariance matrices. *The Annals of Statistics*, 40(2):1024 – 1060, 2012.

[52] O. Ledoit and M. Wolf. Numerical implementation of the QuEST function. *Computational Statistics & Data Analysis*, 115:199–223, 2017.

[53] O. Ledoit and M. Wolf. The power of (non-)linear shrinking: A review and guide to covariance matrix estimation. Working Paper 323, University of Zurich, Department of Economics, Zurich, 2019.

[54] E. Lezmi, T. Roncalli, and J. Xu. Multi-Period Portfolio Optimization. *SSRN Electronic Journal*, 03 2022.

[55] D. León, A. Aragón, J. Sandoval, G. Hernández, A. Arévalo, and J. Niño. Clustering algorithms for Risk-Adjusted Portfolio Construction. *Procedia Computer Science*, 108:1334–1343, 2017. International Conference on Computational Science, ICCS 2017, 12-14 June 2017, Zurich, Switzerland.

[56] H. Liu. Optimal Consumption and Investment with Transaction Costs and Multiple Risky Assets, 2003.

[57] H. Lohre, C. Rother, and K. A. Schäfer. *Hierarchical Risk Parity: Accounting for Tail Dependencies in Multi-asset Multi-factor Allocations*, chapter 9, pages 329–368. John Wiley & Sons, Ltd, 2020.

[58] T. Lux and L. Morales-Arias. Forecasting volatility under fractality, regime-switching, long memory and student-t innovations. Kiel Working Papers 1532, Kiel Institute for the World Economy (IfW Kiel), 2009.

[59] A. W. Lynch and S. Tan. Multiple Risky Assets, Transaction Costs, and Return Predictability: Allocation Rules and Implications for U.S. Investors. *Journal of Financial and Quantitative Analysis*, 45(4):1015–1053, 2010.

[60] M. López de Prado. Building Diversified Portfolios that Outperform Out of Sample. *The Journal of Portfolio Management*, 42(4):59–69, 2016.

[61] M. López de Prado. A Robust Estimator of the Efficient Frontier. *SSRN Electronic Journal*, 01 2019.

[62] S. Maillard, T. Roncalli, and J. Teïletche. The Properties of Equally Weighted Risk Contribution Portfolios. *The Journal of Portfolio Management*, 36(4):60–70, 2010.

[63] B. Mandelbrot. The variation of certain speculative prices. *The Journal of Business*, 36, 1963.

[64] R. Mantegna. Hierarchical structure in financial markets. *The European Physical Journal B: Condensed Matter and Complex Systems*, 11(1):193–197, 1999.

[65] H. Markowitz. Portfolio Selection. *The Journal of Finance*, 7(1):77–91, 1952.

[66] H. Markowitz. *Portfolio Selection: Efficient Diversification of Investments.* Yale University Press, 1959.

[67] U. Maulik and S. Bandyopadhyay. Performance evaluation of some clustering algorithms and validity indices. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(12):1650–1654, 2002.

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

[69] R. C. Merton. On estimating the expected return on the market: An exploratory investigation. *Journal of Financial Economics*, 8(4):323–361, 1980.

[70] X. Mestre. On the asymptotic behavior of the sample estimates of eigenvalues and eigenvectors of covariance matrices. *IEEE Transactions on Signal Processing*, 56(11):5353–5368, 2008.

[71] A. Meucci. Fully Flexible Views: Theory and Practice. *Risk Magazine*, 21(10):97–102, 2008.

[72] R. O. Michaud. The Markowitz Optimization Enigma: Is 'Optimized' Optimal? *Financial Analysts Journal*, 45(1):31–42, 1989.

[73] G. Milligan and M. Cooper. An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50(2):159–179, June 1985.

[74] K. Muthuraman and S. Kumar. Multidimensional Portfolio Optimization with Proportional Transaction Costs. *Mathematical Finance*, 16, 2006.

[75] K. Nakagawa, M. Imamura, and K. Yoshida. Risk-Based Portfolios with Large Dynamic Covariance Matrices. *International Journal of Financial Studies*, 6(2), 2018.

[76] D. B. Nelson. Conditional Heteroskedasticity in Asset Returns: A New Approach. *Econometrica*, 59(2):347–370, 1991.

[77] F. Nielsen and R. Aylursubramanian. Far From the Madding Crowd – Volatility Efficient Indices, 2008.

[78] U. Ozertem and D. Erdogmus. Principal Curve Time Warping. *IEEE Transactions on Signal Processing*, 57(6):2041–2049, 2009.

[79] C. Pakel, N. Shephard, K. Sheppard, and R. F. Engle. Fitting Vast Dimensional Time-Varying Covariance Models. *Journal of Business & Economic Statistics*, 39(3):652–668, 2021.

[80] J. Papenbrock. *Asset Clusters and Asset Networks in Financial Risk Management and Portfolio Optimization.* PhD thesis, Karlsruher Institut für Technologie (KIT), 2011.

[81] R. Pardo. *The Evaluation and Optimization of Trading Strategies.* Wiley, 2008.

[82] T. Poullaouec. Things to consider when investing in minimum variance strategies. *State Street Global Advisor*, 2008.

[83] E. Qian. *Risk Parity Fundamentals.* Chapman and Hall/CRC., 2016.

[84] S. Rachev, S. Ortobelli, S. STOYANOV, F. Fabozzi, and A. Biglova. Desirable properties of an ideal risk measure in portfolio theory. *International Journal of Theoretical and Applied Finance (IJTAF)*, 11:19–54, 02 2008.

[85] T. Raffinot. Hierarchical Clustering-Based Asset Allocation. *The Journal of Portfolio Management*, 44(2):89–99, 2017.

[86] T. Raffinot. The Hierarchical Equal Risk Contribution Portfolio. *Risk Management eJournal*, 2018.

[87] C. Ratanamahatana and E. J. Keogh. Everything you know about Dynamic Time Warping is Wrong. In *Third Workshop on Mining Temporal and Sequential Data*, 2004.

[88] F. Reilly and K. Brown. *Investment Analysis and Portfolio Management*. Cengage Learning, 2011.

[89] J.-C. Richard and T. Roncalli. Constrained risk budgeting portfolios: Theory, algorithms, applications & puzzles. *Computation Theory eJournal*, 2019.

[90] T. Roncalli. Introduction to Risk Parity and Budgeting. MPRA Paper 47679, University Library of Munich, Germany, June 2013.

[91] P. J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.

[92] S. Saitta, B. Raphael, and I. F. C. Smith. A Bounded Index for Cluster Validity. In P. Perner, editor, *Machine Learning and Data Mining in Pattern Recognition*, pages 174–187, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg.

[93] H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1):43–49, 1978.

[94] B. Scherer. A note on the returns from minimum variance investing. *Journal of Empirical Finance*, 18(4):652–660, September 2011.

[95] T. Schwartz. How to beat the S&P 500 with portfolio optimization. *DePaul University*, pages 1–24, 2000.

[96] S. Sharifi, M. Crane, A. Shamaie, and H. Ruskin. Random matrix theory for portfolio optimization: a stability approach. *Physica A: Statistical Mechanics and its Applications*, 335(3):629–643, 2004.

[97] J.-G. Simonato. GARCH processes with skewed and leptokurtic innovations: Revisiting the Johnson Su case. *Finance Research Letters*, 9(4):213–219, 2012.

[98] F. Spinu. An Algorithm for Computing Risk Parity Weights. *Econometric Modeling: Capital Markets - Portfolio Theory eJournal*, 2013.

[99] M. K. Stagys. Kintamumo prognozavimas taikant skirtingus sąlyginio heteroskedastiškumo modelius, 2021.

[100] R. Tibshirani, G. Walther, and T. Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423, 2001.

[101] C. S.-Y. Tsai. The real world is not normal. In *Morningstar Alternative Investments Observer*, 2011.

[102] Y. K. Tse and A. K. C. Tsui. A Multivariate Generalized Autoregressive Conditional Heteroscedasticity Model With Time-Varying Correlations. *Journal of Business & Economic Statistics*, 20(3):351–362, 2002.

[103] J. Tu and G. Zhou. Markowitz meets talmud: A combination of sophisticated and naive diversification strategies. *Journal of Financial Economics*, 99(1):204–215, 2011.

[104] T. K. Vintsyuk. Speech discrimination by dynamic programming. *Cybernetics*, 4:52–57, 1968.

[105] A. Weingessel, E. Dimitriadou, and S. Dolnicar. An examination of indexes for determining the number of clusters in binary data sets. WorkingPaper 29, SFB Adaptive Information Systems and Modelling in Economics and Management Science, WU Vienna University of Economics and Business, 1999.

[106] T. W. Young. Calmar ratio: A smoother tool. *Futures*, 1991.

[107] S. Yue, X. Wang, and M. Wei. Application of two-order difference to gap statistic. *Transactions of Tianjin University*, 14(3):217–221, 2008.

[108] J.-M. Zakoian. Threshold heteroskedastic models. *Journal of Economic Dynamics and Control*, 18(5):931–955, 1994.

[109] P. Zangari. A VaR Methodology for Portfolios that include Options. Technical report, J.P. Morgan, 1996.

[110] Z. Zhang, S. Zohren, and S. Roberts. Deep Learning for Portfolio Optimization. *The Journal of Financial Data Science*, 2(4):8–20, aug 2020.