Aalborg University

10th semester

Mathematics-Economics

Master's thesis

# Index Tracking and Basket-Adjusted Integrated Covariance Estimation

28/5-2025

AALBORG
UNIVERSITY

**Institute of Mathematical Sciences**
Mathematics-Economics
Thomas Manns Vej 23
9220 Aalborg East
https://math.aau.dk

**Title:**

Index Tracking and Basket-Adjusted Integrated Covariance Estimation

**Project:**

Master's Thesis

**Project period:**

1st of February 2025 - 28th of May 2025

**Author:**

Jan Reiter Sørensen

**Supervisor:**

Orimar Sauri Arregui

**Pagecount: 50 + front page**
**Date of handin: 28/5-2025**

**Abstract:**

This thesis investigates the efficacy of using the ETF basket-adjusted covariance (BAC) estimator for improving the positive semidefinite covariation estimates of the *mrcCholCov*•-estimator. The estimators are applied and compared in an index tracking setup. For index tracking, a framework consisting of two parts is established. The first part picks an appropriate subset of component stocks from the index wished to be tracked, and the second part determines appropriate weights for the tracking portfolio. A simulation study investigates the rate, $G$, at which BAC-based tracking is beneficial to vanilla *mrcCholCov*•-based tracking across varying ETF and stock liquidities as well as for varying sizes of the tracking portfolio. The results suggest that $G$ is independent from the tracking portfolio size and the ETF-liquidity. The latter is, however, argued to be due to limitations of the simulation. Additionally, it is concluded that $G$ is higher for stock baskets, where many stocks are illiquid and few stocks have a high liquidity. Finally, an empirical study, using 55 S&P 500 component stocks, compares the tracking performance of the two estimators for index fund sizes of $q = 10, 20$, and 30. Overall, the BAC yields the best performance, and this effect is mostly profound for $q > 10$.

# Preface

The project period has elapsed from 1/2-2025 to 28/5-2025.

## Reader's guide

Literature and source code references are specified within the text. Literature references are given by author names and year, and refers to the bibliography at the end of the thesis. All source code used for this project is public and accessible in the following Github-repository:

- *https://github.com/janreiter793/MasterThesisIndexTracking.git*

All source code is free to use publicly without the permission of the author.

## Acknowledgements

This work has been supervised by Orimar Sauri Arregui. During the development of this project, he has contributed with a lot of good and useful advice. He should therefore have many thanks for his help, and additionally also for the idea for this project. I should also express my appreciation to Center for Clinical Data Science (CLINDA) for lending me office space to make my thesis. In that regard, I will also point a shout out to my office mates Rasmus Rask Jørgensen, Antheas Kapenekakis, and Jakob Bruhn Krøjgaard Skelmose for good advice and for many good productive and unproductive discussions. Last but not least, I will send my appreciation to my girlfriend Michele for always being very sympathetic, helpful, and a good listener, even at times when I go on to ramble about mathematics.

# Contents

# 1 Introduction

As it is explained by Boudt et al. (2023), the ability to accurately estimate the covariation between asset returns has a great importance in several fields within finance. These fields include, for instance, asset pricing, portfolio optimization, risk management, and index tracking. Consequently, as the availability of high-frequency data has become more common, the literature on covariation estimation has expanded significantly. However, as explained by Christensen et al. (2010), high-frequency data is prone to have microstructure noise, which induces autocorrelation in the returns and, thus, bias into the covariation estimates. Efforts to mitigate this effect have, among others, been made by Jacod et al. (2009) and Podolskij and Vetter (2009), who uses pre-averaging. Expanding their approach to a multivariate setting poses, however, a new problem, which is that of asynchronicity. As noted by Epps (1979), asynchronicity of high-frequency data tends to yield covariation estimates that are biased towards zero. Christensen et al. (2010) mitigate microstructure noise and asynchronicity by constructing a pre-averaged version of the estimator proposed by Hayashi and Yoshida (2005). This estimator does not discard observations, as is typically done with estimators that rely on synchronization procedures, however, it is not guaranteed to yield positive semidefinite estimates. On the other hand, Boudt et al. (2016) propose an estimator that exploits Cholesky decomposition to yield positive semidefinite estimates, but this estimator relies on refresh-time sampling for synchronizing observations.

In this thesis, we investigate the benefit of using ETF basket-adjusted covariance estimation for improving the covariation estimate of some pre-estimator. This is a novel method proposed by Boudt et al. (2023), which exploits the high-frequency availability of ETF-prices to improve covariation estimates. In their article, they take basis in the pre-averaged Hayashi-Yoshida estimator as the pre-estimator, however, in this thesis, we apply their method on the positive semidefinite estimator proposed by Boudt et al. (2016). We will conduct an investigation of the method through a simulation study and an empirical study, taking basis in stock-price data for components of the S&P 500 with index tracking as the use case. The problem can thus be boiled down into the following problem formulation:

*Does the ETF basket-adjusted covariance estimator pose a benefit, when comparing its index tracking performance with that of the Cholskey-based estimator of Boudt et al. (2016)? How do the estimators compare in a simulation study, and in a setting based on real S&P 500 stock-price data?*

The thesis is structured as follows. In Chapter 2, a framework for constructing index tracking funds is introduced. In Chapter 3, we present the BAC- and the *mrcCholCov*$^{\bullet}$-estimator for covariation estimation, along with the results of a simulation study comparing the index tracking performance of the two estimators. Chapter 4 contains the results of

an empirical evaluation of the index tracking performance for the two estimators. In Chapter 5, we discuss the obtained results and the methods, used in this thesis. Finally, Chapter 6 concludes with a summary of the findings.

# 2 | Index Tracking

When deciding how to invest, one will have to choose a strategy, and there are generally two approaches to choose from. These are active management and passive management. As described by Chen (2025), active management strategies relies on trying to gain a positive return by actively selling and buying assets on the basis of, for instance, technical analyses and forecasting. Passive investment strategies do not rely on forecasting, and are typically based on diversification. A passive investor will usually choose to either buy and hold or to do index tracking, of which the latter has gained a lot of popularity in recent decades. Cornuejols and Tütüncü (2007) point out the following three reasons for why index tracking has risen in popularity.

- **Market efficiency:** Under the assumption of the Efficient-market hypothesis (EMH), it is theoretically impossible to consistently obtain better returns than the market returns relative to the risk. It is therefore optimal to invest into a market tracking index.
- **Empirical evidence:** Studies generally show that active investment portfolios do not beat passive investment portfolios. See for instance Armour et al. (2024), which show that only 42% of active strategies beat their passive counterparts in 2024. There is, however, an ongoing discussion on whether the EMH holds in practice as there are evidence suggesting that the market can be beat, see Downey (2024).
- **Costs:** The performance of actively managed funds may be reduced by fees such as transaction or trading costs, and salaries for analysts and traders. These costs are avoided in passively managed index funds.

For pure index tracking one would have to buy all assets in the same proportions as in the index. Hence, tracking an index like the Standard and Poor's 500 (S&P 500) would involve buying all 500 different stocks in the index, which may be inconvenient, since the price of such a portfolio would be very expensive.[1] It is therefore ideal to find a subset of the target index consisting of $q$ stocks such that $q$ is smaller than the $d$ stocks of the target index, and such that a portfolio of these $q$ stocks closely replicates the returns of the target index. The construction of an index tracking fund involves two steps. First, an algorithm picks the optimal subset of stocks to represent the index. Next, an algorithm determines the optimal proportions of wealth to be invested into each position. Figure 2.1 illustrates this procedure.

[1] At the time of writing (18/03/2025 15:18 UCT+1) such a portfolio would cost around 1 497 000.00$, see `sp500_price.R`.

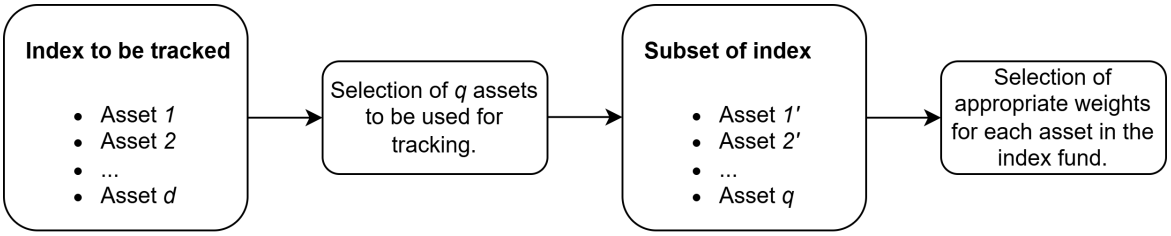



Figure 2.1: The procedure of creating an index fund. First, a subset of $q$ assets from the list of $d$ assets in the index is picked. Then the appropriate proportions of investments into each of the $q$ assets are selected.

In the following section we will delve into the procedure of selecting an appropriate subset of assets to track the index with.

## 2.1 Selecting a subset of stocks to use for index tracking

We will follow the approach described by Cornuejols and Tütüncü (2007). Suppose that we want to select $q$ stocks for an index fund that tracks a population of $d > q$ stocks. For each pair of stocks, we define a similarity index, denoted $\rho_{ij}$ for the pair of stocks $i$ and $j$, such that $\rho_{ij} \leq 1$ for all $i, j = 1, 2, \ldots, d$, and $i = j \Rightarrow \rho_{ij} = 1$. Specifically, $\rho_{ij}$ is larger for pairs of similar stocks than for pairs of less similar stocks. In this project, we will use the estimated correlation between the returns of each stock price process as the similarity index.[2] Let $y \coloneqq (y_1, y_2, \ldots, y_d)^\top$, where

$$y_j \coloneqq \begin{cases} 1, & \text{if stock } j \text{ was selected for the index fund,} \\ 0, & \text{otherwise} \end{cases}$$

for $j = 1, 2, \ldots, d$. Furthermore, define the matrix $x \in \{0, 1\}^{d \times d}$ with typical element $x_{ij}$ given by

$$x_{ij} \coloneqq \begin{cases} 1, & \text{if stock } j \text{ is the stock in the index fund that is most similar to stock } i, \\ 0, & \text{otherwise,} \end{cases}$$

for $i, j = 1, 2, \ldots, d$. To find the optimal set of stocks for the index fund, we should maximize the similarity between the stocks in the index fund, and the stocks in the target index. We can formulate this into the following maximization problem.

$$(M)$$
$$z \coloneqq \max_{x,y} \quad \sum_{i=1}^{d} \sum_{j=1}^{d} \rho_{ij} x_{ij},$$

[2]However not done in this project, it may be beneficial to use the absolute value of the estimated correlation instead, as a correlation close to $-1$ is, nevertheless, also a strong correlation.

$$
\begin{aligned}
\text{S.t.} \quad & \sum_{j=1}^{d} y_j = q, \\
& \sum_{j=1}^{d} x_{ij} = 1, && \text{for } i = 1, 2, \ldots, d \\
& x_{ij} \leq y_j, && \text{for } i = 1, 2, \ldots, d; j = 1, 2, \ldots, d \\
& x_{ij}, y_j \in \{0, 1\}, && \text{for } i = 1, 2, \ldots, d; j = 1, 2, \ldots, d.
\end{aligned}
$$

This is an integer programming problem, however obtaining a solution can be computationally heavy. Suppose for instance that we wanted to track the S&P 500 with an index fund consisting of 10 stocks. There would be $\binom{500}{10} \approx 2.46 \times 10^{20}$ ways of selecting $y$, such that

$$
\sum_{j=1}^{500} y_j = 10,
$$

and additionally we would have 250 000 constraints of $x_{ij} \leq y_j$, hence, searching through all possible solutions would be significantly time consuming. We will therefore instead obtain a heuristic solution to $(M)$ by following the proposed method of Cornuejols et al. (1977). In this article, they propose a solution to a variant of the Facility Location Problem (See chapter 2 of da Gama and Wang (2024)), which relies on a Lagrangian relaxation to turn an integer programming problem into a continuous problem. This is useful, since $(M)$ is a similar variant of the Facility Location Problem, hence, their solution can be used to find optimal subsets of stocks for an index fund. Rewriting $(M)$ as proposed by Cornuejols et al. (1977) yields the following maximization problem.

$$
\begin{aligned}
& (M') \\
\mathcal{L}(u) := \max_{x,y} \quad & \sum_{i=1}^{d} \sum_{j=1}^{d} \rho_{ij} x_{ij} \quad + \sum_{i=1}^{d} u_i \left( 1 - \sum_{j=1}^{d} x_{ij} \right), \\
\text{S.t.} \quad & \sum_{j=1}^{d} y_j = q, \\
& x_{ij} \leq y_j, && \text{for } i = 1, 2, \ldots, d; j = 1, 2, \ldots, d \\
& x_{ij}, y_j \in \{0, 1\}, && \text{for } i = 1, 2, \ldots, d; j = 1, 2, \ldots, d.
\end{aligned}
$$

Where $u = (u_1, u_2, \ldots, u_d)^\top$ is any vector in $\mathbb{R}^d$. The intuition behind reexpressing $(M)$ as $(M')$ is that the restriction of $\sum_{j=1}^{d} x_{ij} = 1$ for $i = 1, 2, \ldots, d$, has been written into the objective function, such that solutions are instead penalized for not satisfying the restriction. In other words, a solution to $(M')$ does not have to satisfy $\sum_{j=1}^{d} x_{ij} = 1$, but deviations from the restriction are still penalized. As a consequence of this, a solution to $(M')$ is not necessarily a solution to $(M)$, but according to the following proposition, we can use $(M')$ to estimate an upper bound to $z$.

**Proposition 2.1.1.**
Let $u \in \mathbb{R}^d$, let $z$ be given as in $(M)$, and let $\mathcal{L}(u)$ be given as in $(M')$. Then $\mathcal{L}(u) \geq z$.

**Proof.** We will prove Proposition 2.1.1 by contradiction, hence, let $u \in \mathbb{R}^d$ be given, and assume that $z$ and $\mathcal{L}(u)$ have been obtained such that $\mathcal{L}(u) < z$. Let the pairs $(x, y), (x', y') \in \mathbb{R}^{d \times d} \times \mathbb{R}^d$ be the pairs that are obtained by solving the maximization problems $(M)$ and $(M')$ respectively. The pair $(x, y)$ also satisfies the restrictions in $(M')$, but plugging $x$ into the objective function of $(M')$ yields $z$, which means that $(x, y)$ is a better solution to $(M')$ than $(x', y')$. This is a contradiction, and we can conlcude that Proposition 2.1.1 holds. ■

When finding an optimal subset of stocks for an index fund, we are only interested in $y$, since it specifies which stocks to include in the index fund. The following proposition states that $\mathcal{L}(u)$ can be calculated without needing to solve for $x$.

**Proposition 2.1.2.**
Let $u \in \mathbb{R}^d$, let $C_j \coloneqq \sum_{i=1}^{d} (\rho_{ij} - u_i)^+$ for $j = 1, 2, \ldots, d$, where

$$(\rho_{ij} - u_i)^+ \coloneqq \max\left\{0, \rho_{ij} - u_i\right\}.$$

Then

$$\begin{aligned} \mathcal{L}(u) = \max_{y} \quad & \sum_{j=1}^{d} C_j y_j + \sum_{i=1}^{d} u_i, \\ \text{S.t.} \quad & \sum_{j=1}^{d} y_j = q, \\ & y_j \in \{0, 1\}, \qquad \text{for } j = 1, 2, \ldots, d. \end{aligned}$$

**Proof.** We will prove Proposition 2.1.2 directly, and we will omit explicitly writing the constraints of $(M')$, but they are still implicitly present. Let $u \in \mathbb{R}^d$ be given, and consider the following reexpression of $\mathcal{L}(u)$

$$\begin{aligned} \mathcal{L}(u) &= \max_{x,y} \quad \sum_{i=1}^{d} \sum_{j=1}^{d} \rho_{ij} x_{ij} + \sum_{i=1}^{d} u_i \left(1 - \sum_{j=1}^{d} x_{ij}\right) \\ &= \max_{x,y} \quad \sum_{i=1}^{d} \sum_{j=1}^{d} (\rho_{ij} - u_i) x_{ij} + \sum_{i=1}^{d} u_i \end{aligned}$$

$$= \max_{x,y} \quad \sum_{i=1}^{d}\sum_{j=1}^{d}(\rho_{ij} - u_i)^+ x_{ij} + \sum_{i=1}^{d} u_i.$$

The last equality comes from considering that when $\rho_{ij} - u_i < 0$, we obtain a better solution by letting $x_{ij} = 0$ rather than $x_{ij} = 1$.We could maximize by setting all entries in $x$ to 1, but since $x_{ij} \leq y_j$, we can substitute $x_{ij}$ with $y_j$, hence

$$\max_{x,y} \quad \sum_{i=1}^{d}\sum_{j=1}^{d}(\rho_{ij} - u_i)^+ x_{ij} + \sum_{i=1}^{d} u_i = \max_{y} \quad \sum_{i=1}^{d}\sum_{j=1}^{d}(\rho_{ij} - u_i)^+ y_j + \sum_{i=1}^{d} u_i.$$

Finally, by turning the summation order around, we obtain

$$\begin{aligned}
\max_{y} \quad \sum_{i=1}^{d}\sum_{j=1}^{d}(\rho_{ij} - u_i)^+ y_j + \sum_{i=1}^{d} u_i &= \max_{y} \quad \sum_{j=1}^{d}\sum_{i=1}^{d}(\rho_{ij} - u_i)^+ y_j + \sum_{i=1}^{d} u_i \\
&= \max_{y} \quad \sum_{j=1}^{d} y_j \sum_{i=1}^{d}(\rho_{ij} - u_i)^+ + \sum_{i=1}^{d} u_i \\
&= \max_{y} \quad \sum_{j=1}^{d} C_j y_j + \sum_{i=1}^{d} u_i,
\end{aligned}$$

which was to be demonstrated. ■

Using Proposition 2.1.2 we see that we can quickly find solutions $y$ and calculate $\mathcal{L}(u)$ for any given $u \in \mathbb{R}^d$ by setting $y_j = 1$ for the $q$ largest values of $C_j$. In addition, we can also determine $x_{ij}$ by setting $x_{ij} = y_j$, when $\rho_{ij} - u_i > 0$, and $x_{ij} = 0$ otherwise.

As stated in Proposition 2.1.1, we can use $\mathcal{L}(u)$ as an upper bound for $z$. To improve this bound we are interested in approximating

$$\overline{z} \coloneqq \min_{u} \mathcal{L}(u). \tag{2.1}$$

Since $(M')$ is a relaxation of one of the constraints in $(M)$ the solution pair $(x, y)$ that solves $(M')$ is not necessarily an optimal solution to $(M)$ nor is it necessarily feasible, however, we can still use the solution to obtain a lower bound for $z$. To do this, we specify an $(M)$ feasible version of $x$ in the following way

$$x^*_{ij} \coloneqq \begin{cases} 1, & \text{if } j = \arg\max_j \rho_{ij} y_j, \\ 0, & \text{otherwise.} \end{cases} \tag{2.2}$$

We define this lower bound in the following way

$$\underline{z} \coloneqq \sum_{i=1}^{d}\sum_{j=1}^{d} \rho_{ij} x^*_{ij}.$$

When $\overline{z} = \underline{z}$ we know that the obtained solution $(x^*, y)$, where $x^*$ has typical element $x^*_{ij}$, is an optimal solution to $(M)$ according to Theorem 6.3 of Andréasson et al. (2020). Otherwise the solution will be suboptimal in $(M)$, however $y$ can still be regarded as a heuristic solution. In the following subsection, we will cover, how to approximate $\overline{z}$ and obtain a heuristic solution to $(M)$.

### 2.1.1 Approximating the least upper bound of *(M)*

In this section, we will investigate how to approximate $\overline{z}$. Recall that $\overline{z}$ was defined by (2.1) and recall as well that

$$\mathcal{L}(u) = \max_{y} \quad \sum_{j=1}^{d} C_j y_j + \sum_{i=1}^{d} u_i,$$

for any $u \in \mathbb{R}^d$ according to Proposition 2.1.2. Since $(\rho_{ij} - u_i)^+$ is convex with respect to $u_i$ we have that all $C_j$ are convex, hence the term $\sum_{j=1}^{d} C_j y_j$ must be convex. Additionally, the term $\sum_{i=1}^{d} u_i$ is convex as well, hence $\mathcal{L}(u)$ is convex. However, $(\rho_{ij} - u_i)^+$ is not differentiable in $u_i = \rho_{ij}$, when $\rho_{ij} \neq 0$, and $\mathcal{L}(u)$ is therefore not differentiable in all points of $\mathbb{R}^d$ unless $\rho_{ij} = 0$ for all $i, j = 1, 2, \ldots, d$. Thus, we cannot rely on gradient descent to approximate $\overline{z}$. Instead we use the subgradient optimization method, which is described by Andréasson et al. (2020). Subgradients are defined in the following way.

**Definition 2.1.3. (Subgradient)**
Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex function. A vetor $g \in \mathbb{R}^d$ is said to be a subgradient of $f$ at $x \in \mathbb{R}^d$ if

$$f(y) \geq f(x) + g^\top (y - x), \quad y \in \mathbb{R}^d.$$

The set of subgradients for $f$ at $x$ is called the subdifferential of $f$ at $x$, and it is denoted $\partial f(x)$.

Subgradient optimization can be used to obtain solutions for optimization problems, where we seek to minimize a convex function subject to a non-empty, closed, and convex subset of $\mathbb{R}^d$. The algorithm is defined in the following way.

**Algorithm 2.1.4. (Subgradient optimization)**
Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex function, and let $X \subseteq \mathbb{R}^d$ be a non-empty, closed, and convex subset. The subgradient optimization algorithm follows the following procedure:

1. Pick $x_0 \in X$,
2. For $k = 0, 1, \ldots$:

3. Generate $g_k \in \partial f(x_k)$, and $x_{k+1} = x_k - \alpha_k g_k$.
4. If $\|x_{k+1} - x_k\| < \varepsilon$ for some $\varepsilon > 0$ sufficiently close to 0:
5. Stop.

The step length sequence $(\alpha_k)_{k\in\mathbb{N}_0}$ is some chosen sequence where $\alpha_k > 0$ for $k \in \mathbb{N}_0$, $\lim_{k\to\infty} \alpha_k = 0$, and $\sum_{k=0}^{\infty} \alpha_k = \infty$.

Andréasson et al. (2020) proposes to pick sequences $(\alpha_k)_{k\in\mathbb{N}_0}$ that satisfy the additional criteria $\sum_{k\in\mathbb{N}_0} \alpha_k^2 < \infty$ for faster convergence, and they also prove that Algorithm 2.1.4 converges towards the global minimum, when it exists. An implementation of the subgradient optimization method can be found in this project's `Github` repo in the file `subgradient_optimization.R`. For the step length sequence we have used $\alpha_k \coloneqq \frac{\beta}{(k+1)}$ for $\beta > 0$ as is suggested by Andréasson et al. (2020).

In summary, we can apply subgradient optimization to (2.1) to obtain an approximation of $\overline{z}$ and a heuristic solution $(x^*, y)$ for $(M)$. Furthermore, we can assess, whether or not we obtained the optimal solution by checking if $|\overline{z} - \underline{z}| < \varepsilon$ for an $\varepsilon \geq 0$ sufficiently close to 0. Finally, the solution $y$ will specify which stocks to put into our index fund. In the next section, we will investigate how to select appropriate weights for each stock in an index fund.

## 2.2 Selecting weights for stocks in an index fund

In the former section, we investigated how to select a subset of stocks from an index in order to construct an index fund. In this section, we will cover how to determine the proportions of wealth to invest into each stock in the index fund. Fastrich et al. (2014) proposes the following optimization problem for determining both the weights and which set of stocks to use for tracking,

$$\begin{aligned} \arg\min_{\alpha} \quad & TE(\alpha;\Omega) \coloneqq \sqrt{\mathrm{Var}[r_b - r_p]} = \sqrt{(w_b - \alpha)^\top \Omega (w_b - \alpha)}, \\ \text{S.t.} \quad & \sum_{i\in\mathcal{J}} \alpha_i = 1, \\ & 0 \leq \alpha_i \leq 1, && \text{for } i = 1, 2, \ldots, d, \\ & |\mathcal{J}| \leq q. \end{aligned}$$

where $\alpha, w_b \in \mathbb{R}^d$ are the weights of investment into each stock for the tracking portfolio and the benchmark index respectively, $r_p, r_b \in \mathbb{R}$ are the tracking portfolio and the benchmark returns respectively, $\Omega \in \mathbb{R}^{d\times d}$ represents the covariance matrix for the $d$ stocks in the index, $\mathcal{J} \coloneqq \big\{i \in \{1, 2, \ldots, d\} \mid \alpha_i > 0\big\}$, and $|\mathcal{J}|$ denotes the cardinality of $\mathcal{J}$. The objective

function $TE$ is called the tracking error, and is defined as the standard deviation of the excess returns between the tracking portfolio and the index. The constraint of $|\mathcal{J}| \leq q$ ensures that only $q$ stocks are picked for the index fund, but since we already picked $q$ stocks for tracking in the former section, we can omit this constraint, and use the $q \times q$ submatrix of $\Omega$ with the chosen stocks. In addition, since we are only assessing the weights for the $q$ stocks, we can reduce the dimension of $\alpha$ to $q$, and the dimension of $\Omega$ to $q \times q$. In the empirical implementation of this project, we will follow the approach of Boudt et al. (2023). They introduce the use of ETF log-prices for an ETF that tracks the same index that we wish to track with the index fund. They suggest to include the stock-ETF covaration and ETF-variance in $\Omega$ in order to improve the estimate of $\alpha$. Hence,

$$\Omega \coloneqq \begin{bmatrix} \omega_E & \omega_{EK}^\top \\ \omega_{EK} & \Sigma \end{bmatrix},$$

where $\omega_E$ is the integrated variance of the ETF, $\omega_{EK}$ is the $q$-dimensional vector containing the stock-ETF covariation, and the $q \times q$-matrix $\Sigma$ is an estimate of the stock covariation. As a consequence of this, the dimension of $\Omega$ will be $(q+1) \times (q+1)$, and $\alpha$ will be $(q+1)$-dimensional. We will, however, constrain the entry associated with the ETF, that is the first entry in $\alpha$, to be zero, such that no investments are done into the ETF. Boudt et al. (2023) also propose to use a variant of the tracking error that is given by

$$TE(\alpha; \Omega) \coloneqq (\mathbf{1} - \alpha)^\top \Omega (\mathbf{1} - \alpha).$$

The explanation for using this tracking error is that for the minimization problem of finding $\alpha$, the weights $w_b$ are irrelevant, since they are constant and will be differentiated out in the first order conditions. We may therefore replace $w_b$ with a representation of an equal-weighted portfolio, which is the the vector given by

$$\mathbf{1} \coloneqq \begin{bmatrix} 1 & 1 & \dots & 1 \end{bmatrix}^\top.$$

In summary, the optimization problem that yields the optimal proportions of investments into each asset is given by

$$\begin{aligned} \arg\min_{\alpha} \quad & TE(\alpha; \Omega) \coloneqq (\mathbf{1} - \alpha)^\top \Omega (\mathbf{1} - \alpha), \\ \text{S.t.} \quad & \sum_{i=1}^{q+1} \alpha_i = 1, \\ & 0 \leq \alpha_i \leq 1, \qquad\qquad \text{for } i = 1, 2, \dots, q+1, \\ & \alpha_1 = 0, \end{aligned}$$

From the first order conditions, it is obtained that the optimal solution for $\alpha$ is given by

$$\alpha(\Omega) = \begin{bmatrix} 0 \\ \Sigma^{-1} \omega_{EK} \end{bmatrix}. \tag{2.3}$$

See Section A.4 for an argumnt for this result. To obtain $\alpha(\Omega)$ at some given point in time $t$, we will use an estimate of $\Omega_t$ denoted

$$\hat{\Omega}_t \coloneqq \begin{bmatrix} \hat{\omega}_{E,t} & \hat{\omega}_{EK,t}^\top \\ \hat{\omega}_{EK,t} & \hat{\Sigma}_t \end{bmatrix}.$$

To evaluate the performance of the index fund, we will insert $\alpha(\hat{\Omega}_t)$ into the tracking error and calculate $TE(\alpha(\hat{\Omega}_t); \hat{\Omega}_t)$. In the following chapter, we will delve into the estimation of $\Omega$.

# 3 Integrated Covariance Estimation

In this section, we will cover methods for estimating the integrated covariance for a $d$-dimensional Brownian semimartingale that represents the log-price processes of $d$ assets. We will take basis in the same theoretical setup that is used by Christensen et al. (2010) and Boudt et al. (2016). Let $(\Omega^0, \mathcal{F}^0, (\mathcal{F}_t^0)_{t\geq 0}, P^0)$ be a filtered probability space satisfying the usual conditions, let the $d$-dimensional log-price process, $X$, be defined on the probability space, let it be $(\mathcal{F}_t^0)_{t\geq 0}$-adapted, and assume that it is a solution to the following SDE,

$$X_t = X_0 + \int_0^t \mu_u du + \int_0^t \sigma_u dW_u, \quad t \geq 0, \tag{3.1}$$

where the $d$-dimensional drift process $\mu = (\mu_t)_{t\geq 0}$ is predictable and locally bounded, the $d \times d$ covolatility matrix $\sigma = (\sigma_t)_{t\geq 0}$ is adapted and càdlàg, and the $d$-dimensional process $W = (W_t)_{t\geq 0}$ is a Brownian motion. In this model, the indiviual log-price processes are Itô semimartingales of the form

$$X_t^i = X_0^i + \int_0^t \mu_u^i du + \sum_{j=1}^d \int_0^t \sigma_u^{ij} dW_u^j, \quad t \geq 0,$$

where $\mu^i = (\mu_t^i)_{t\geq 0}$ is the $i^{\text{th}}$ entry process in $\mu$, and $\sigma^{ij} = (\sigma_t^{ij})_{t\geq 0}$ is the $ij^{\text{th}}$ entry process in $\sigma$. The integrated covariance is also called the quadratic variation, and we define it in the following way. Let $\mathcal{P} = \{0 = t_0 < t_1 < \ldots < t_n = t\}$ be a partition of the interval $[0, t]$ for any $t \geq 0$, and let

$$|\mathcal{P}| \coloneqq \sup_i |t_i - t_{i-1}|.$$

The quadratic variation process of $X$ is defined as the process $\langle X \rangle_t$ that satisfies

$$\langle X \rangle_t \coloneqq P^0 \lim_{|\mathcal{P}|\to 0} \sum_{i=1}^n (X_{t_i} - X_{t_{i-1}})(X_{t_i} - X_{t_{i-1}})^\top, \quad t \geq 0.$$

Since $X$ is a semimartingale, the quadratic variation process is equal to the quadratic variation process of the local martingale component of $X$ which is given by

$$M_t = \int_0^t \sigma_u dW_u, \quad t \geq 0.$$

Which has the quadratic variation process

$$\langle M \rangle_t = \int_0^t \Sigma_u du,$$

where $\Sigma_t \coloneqq \sigma_t \sigma_t^\top$. We will also assume that $X$ is only observed in the time interval $[0,1]$, and we are, thus, interested in estimating

$$\Sigma \coloneqq \int_0^1 \Sigma_u du.$$

In this context, it is clear, why quadratic variation is also called the integrated covariance. In this thesis, we will use the terms integrated covariance and covariation to denote $\Sigma$. There are generally two elements that interfere with the estimation of the integrated covariance, and we will extend our model to incorporate both. These are

- Market microstructure noise,
- Asynchronous trading.

Market microstructure noise are perturbations of the observed log-price processes. According to Zhou (1996), microstructure noise produces autocorrelation in high-frequency return data, which causes bias in the integrated covariance estimates. In the context of our model, it means that instead of observing the log-price process $X$, we are observing the noisy process $Y$, given by

$$Y_t \coloneqq X_t + \varepsilon_t, \quad t \in [0,1],$$

where $(\varepsilon_t)_{t\in[0,1]}$ is a $d$-dimensional i.i.d zero mean process that is independent from $X$. Assume that $(\varepsilon_t)_{t\in[0,1]}$ is defined and adapted in the filtered probability space

$$(\Omega^1, \mathcal{F}^1, (\mathcal{F}_t^1)_{t\in[0,1]}, P^1),$$

satisfying the usual conditions, where $\Omega^1 \coloneqq \mathbb{R}^{[0,1]}$, and $\mathcal{F}^1$ is the Borel-$\sigma$-algebra generated by $\Omega^1$. The probability measure $P^1$ is defined by $P^1 \coloneqq \otimes_{t\in[0,1]} P_t^1$, where $P_t^1 \coloneqq Q$, and $Q$ is a probability measure on $\mathbb{R}$. Finally, let $\Omega \coloneqq \Omega^0 \times \Omega^1$, let $\mathcal{F} \coloneqq \mathcal{F}^0 \otimes \mathcal{F}^1$, let

$$\mathcal{F}_t \coloneqq \bigcap_{s>t} \mathcal{F}_s^0 \otimes \mathcal{F}_s^1, \quad t \in [0,1],$$

and let $P \coloneqq P^0 \otimes P^1$. Then we can define $Y$ on the filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\in[0,1]}, P)$.

Asynchronicity in the trading of assets involves observing log-prices of the assets at different times and different frequencies, and it is mainly caused by differences in the liquidity of each asset. As shown by Epps (1979), sampling asynchronous data at a high frequency can impose a bias towards zero for covariation estimates, hence asynchronicity is therefore problematic. Let $n_i$ be the number of observations for the log-price process $Y^i$ in the time interval $[0,1]$, then we have $n \coloneqq \sum_{i=1}^d n_i$ observations in total. The set of observation times for $Y^i$ will be denoted by

$$\mathcal{T}_i \coloneqq \left\{ 0 \le t_1^i < t_2^i < \ldots < t_{n_i}^i \le 1 \right\}.$$

To obtain useful integrated covariance estimates, we will use a method that is both robust to microstructure noise and asynchronicity. We will use the estimator proposed by Boudt et al. (2016), because it yields a positive semidefinite estimate. To account for microstructure noise, we will utilize the modulated realized covariance (MRC) approach described by Christensen et al. (2010), hence, we are using the *mrcCholCov*$^\bullet$ estimator of Boudt et al. (2016). The *mrcCholCov*$^\bullet$ estimator mitigates asynchronicity by doing refresh-time sampling, which will be described later. To improve the integrated covariance estimate we will use the method of ETF basket-adjusting proposed by Boudt et al. (2023). They propose a method of improving a pre-estimate of the integrated covariance of a set of log-price processes by exploiting the high frequency availability of ETF prices. In the following section we will cover the *mrcCholCov*$^\bullet$.

## 3.1 The *mrcCholCov* estimator

The *mrcCholCov*$^\bullet$ estimator of Boudt et al. (2016) is based on Cholesky factorization of the integrated covariance matrix, that is, it exploits that $\Sigma = HGH^\top$, where $H$ is a lower triangular matrix with ones in the diagonal and $G$ is a diagonal matrix. We can, thus, estimate $H$ and $G$, and calculate $\Sigma$, which ensures positive semidefiniteness, and since the entries of $H$ and $G$ can be calculated sequentially, it allows for efficient use of the log-price observations. This is because, when doing refresh-time sampling the number of observations that we have left will be determined by the least liquid asset, hence, by calculating the estimates for the most liquid assets first and the least liquid assets last we maximize the number of observations used for estimation. To rank the assets by liquidity, Boudt et al. (2016) proposes to use the squared duration criterion, which means sorting the assets by $\sum_{j=1}^{n_i-1}(\Delta_j^i)^2$, where $\Delta_j^i := t_{j+1}^i - t_j^i$. They show that this approach yields more observations after synchronization than sorting by the number of observations.

Refresh-time sampling is a way to synchronize observations by picking each refresh-time, such that all assets have been traded at least once since the last refresh-time. Suppose we want to generate a refresh-time grid of a subset of the assets $S \subseteq \{1, 2, \ldots, d\}$. The first refresh time is defined by

$$\tau_1^S := \max\left\{t_1^i \mid i \in S\right\}.$$

Let $N^i(t)$ be the counting process defined by

$$N^i(t) := \#\left\{\tau \in \mathcal{T}_i \mid \tau \leq t\right\},$$

then the succeeding refresh-times are defined by

$$\tau_{j+1}^S := \max\left\{t^1_{N^1(\tau_j)+1}, t^2_{N^2(\tau_j)+1}, \ldots, t^d_{N^d(\tau_j)+1}\right\},$$

for $j > 1$.[1] The set of refresh-times will be denoted

$$\mathcal{T}^S \coloneqq \left\{0 = \tau_0^S < \tau_1^S < \ldots < \tau_{\mathcal{N}^S}^S \leq 1\right\},$$

where $\mathcal{N}^S$ is the number of refresh-times. The returns of the grid will be denoted[2]

$$r_j(\mathcal{T}^S) \coloneqq Y_{\tau_j^S} - Y_{\tau_{j-1}^S}, \quad \text{for } j = 1, 2, \ldots, \mathcal{N}^S,$$

and the durations will be denoted

$$\Delta_j(\mathcal{T}^S) \coloneqq \tau_j^S - \tau_{j-1}^S.$$

Let $H$ and $G$ be specified by

$$H \coloneqq \begin{bmatrix} 1 & 0 & \cdots & 0 \\ h_{21} & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ h_{d1} & h_{d2} & \cdots & 1 \end{bmatrix}, \text{ and } G \coloneqq \begin{bmatrix} g_{11} & 0 & \cdots & 0 \\ 0 & g_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & g_{dd} \end{bmatrix}.$$

Define the factor $f_j(\mathcal{T}) \coloneqq H^{-1} r_j(\mathcal{T})$ for $j = 1, 2, \ldots, N$, where

$$\mathcal{T} \coloneqq \{0 = t_0 < t_1 < t_2 < \ldots < t_N = 1\},$$

is some partition of $[0, 1]$. Rewriting yields $r_j(\mathcal{T}) = H f_j(\mathcal{T})$, and since $H$ is a lower triangular matrix, we obtain that each entry in $r_j(\mathcal{T})$ is a function of the entries in $f_j(\mathcal{T})$ with a lower index, i.e, $f_j^1(\mathcal{T}) = r_j^1(\mathcal{T})$, and for $k = 2, \ldots, d$,

$$r_j^k(\mathcal{T}) = \sum_{i=1}^{k} h_{ki} f_j^i(\mathcal{T}),$$

where $h_{kk} \coloneqq 1$. In addition $f_j^k(\mathcal{T}) \sim N(0, \Delta_j(\mathcal{T}) g_{kk})$, where $\Delta_j(\mathcal{T}) \coloneqq t_j - t_{j-1}$. It is therefore proposed by Boudt et al. (2016) to estimate $g_{kk}$ as the integrated variance of $f^k(\mathcal{T})$, and $h_{kl}$ can be estimated as the realized beta between $r_j^k(\mathcal{T})$ and $f_j^k(\mathcal{T})$. Hence, to estimate $g_{kk}$ and $h_{kl}$ we will rely on the bias-corrected modulated realized covariance estimator of Christensen et al. (2010), which utilizes pre-averaging to mitigate microstructure noise.[3]

The following algorithm describes the *mrcCholCov*$^\bullet$ estimator.

[1] For an implementation of refresh-time sampling see `refresh-time_sampler.R`.

[2] In practice we will not necessarily have observations for all processes at exactly $\tau_j^S$, thus we will instead use last known price.

[3] See Section A.1 for an explanation of this estimator.

**Algorithm 1 (mrcCholCov$^\bullet$)**

Sort log-price processes from most liquid to least according to the squared duration criterion.
Set $f_j^1(\mathcal{T}_1) = r_j^1(\mathcal{T}_1)$, and obtain estimate $\hat{g}_{11}$ of $g_{11}$ as the integrated variance of $f^1(\mathcal{T}_1)$.
Let $S = \{1, 2\}$ and construct the refresh-time grid $\mathcal{T}^S$.
Obtain the estimate $\hat{h}_{21}$ of $h_{21}$ as the realized beta of $r_j^2(\mathcal{T}^S)$ and $f_j^1(\mathcal{T}^S)$.
Then estimate $\hat{g}_{22}$ as the integrated variance of $f^2(\mathcal{T}^S)$ using $f_j^2(\mathcal{T}^S) = r_j^2(\mathcal{T}^S) - \hat{h}_{21} f_j^1(\mathcal{T}^S)$.
Construct refresh-time grid $\mathcal{T}^S$ with $S = \{1, 3\}$.
Estimate $\hat{h}_{31}$ as the realized beta of $r_j^3(\mathcal{T}^S)$ on $f_j^1(\mathcal{T}^S)$.
Obtain a new refresh-time grid $\mathcal{T}^S$ with $S = \{1, 2, 3\}$.
Compute $f_j^2(\mathcal{T}^S) = r_j^2(\mathcal{T}^S) - \hat{h}_{21} f_j^1(\mathcal{T}^S)$.
Estimate $\hat{h}_{32}$ as the realized beta of $r_j^3(\mathcal{T}^S)$ on $f_j^2(\mathcal{T}^S)$.
Estimate $\hat{g}_{33}$ as the integrated variance of $f_j^3(\mathcal{T}^S) = r_j^3(\mathcal{T}^S) - \hat{h}_{32} f_j^2(\mathcal{T}^S) - \hat{h}_{31} f_j^1(\mathcal{T}^S)$.
**for** $k = 4, \ldots, d$ **do**
    **for** $l = 1, 2, \ldots, (k-1)$ **do**
        Construct refresh-time grid $\mathcal{T}^S$ with $S = \{1, 2, \ldots, l, k\}$.
        Calculate the factors $f_j^m(\mathcal{T}^S) = r_j^m - \sum_{n=1}^{m-1} \hat{h}_{mn} f_j^n(\mathcal{T}^S)$ for $m = 1, 2, \ldots, l$.
        Estimate $\hat{h}_{kl}$ as the realized beta of $r_j^k(\mathcal{T}^S)$ on $f_j^l(\mathcal{T}^S)$.
    **end for**
    Estimate $\hat{g}_{kk}$ using $f_j^k(\mathcal{T}^S) = r_j^k(\mathcal{T}^S) - \sum_{n=1}^{k-1} \hat{h}_{kn} f_j^n(\mathcal{T}^S)$.
**end for**
Estimate the diagonal entries of $\Sigma$ using all available observations, and construct correction matrix $\hat{D}$, such that we end up with the final estimate $\hat{\Sigma} = \hat{D}^{1/2} \hat{H} \hat{G} \hat{H}^\top \hat{D}^{-1/2}$.

Algorithm 1 describes how the *mrcCholCov*$^\bullet$-estimator works. It is robust to asynchronicity because of refresh-time sampling, and it is robust to microstructure noise because it uses the MRC estimator of Christensen et al. (2010), which uses pre-averaging. In the next section, we will investigate the procedure of adjusting the estimate provided by *mrcCholCov*$^\bullet$ using stock-ETF covariation. An implementation of the *mrcCholCov*$^\bullet$ estimator is available in the Github repo by in the file `mrcCholCov.R`.

## 3.2 ETF basket-adjusted covariance estimation

In the former section we covered the *mrcCholCov*$^\bullet$-estimator for the integrated covariation of a set of stock price processes. That estimator relies on refresh-time sampling to synchronize observations, however in the case of very illiquid stocks, many observations will be discarded, and as a consequence covariance estimates will be imprecise. Boudt et al. (2023) proposes to exploit the availability of high-frequency price data for ETFs that track the index in question to obtain a better estimate of the integrated covariance. They present a

way of estimating an adjustment matrix that can be added to a pre-estimate of the integrated covariance to improve the accuracy. The following definition is adapted from Boudt et al. (2023) and it specifies the basket-adjusted covariance estimator (BAC-estimator).

**Definition 3.2.1. (The BAC-estimator)**
Let the $d \times d$-matrix $\overline{\Sigma}$ denote a pre-estimator of the integrated covariance of $d$ stocks, let the $d$-dimensional vector $\overline{\beta}$ denote the stock-ETF covariation for each asset, and let $\overline{\beta}_\Delta$ be a given or estimated target. Then the BAC-estimator is given by

$$\overline{\Sigma}^{BAC} := \overline{\Sigma} - \overline{\Delta}^{BAC},$$

where $\overline{\Delta}^{BAC}$ is a $d \times d$ adjustment-matrix defined by

$$\text{vec}(\overline{\Delta}^{BAC}) := \overline{L}(\overline{\beta} - \overline{\beta}_\Delta),$$

and

$$\overline{L} := \left(I_{d^2} - \frac{1}{2}Q\right)\overline{W}^\top \left(I_d \left(\sum_{k=1}^{d} n_k^{-1} \sum_{m=1}^{n_k} \left(\overline{w}_{t_{m-1}^k}\right)^2\right) - \frac{\overline{W}Q\overline{W}^\top}{2}\right)^{-1},$$

$$\overline{W}^k := \left(0_{(k-1)d}^\top, \frac{1}{n_1}\sum_{m=1}^{n_1} \overline{w}^1_{t^1_{m-1}}, \ldots, \frac{1}{n_d}\sum_{m=1}^{n_d} \overline{w}^d_{t^d_{m-1}}, 0_{(d-k)d}^\top\right),$$

for $k = 1, 2, \ldots, d$, where $\overline{W}^k$ is the $k^{\text{th}}$ row of the $d \times d^2$-matrix $\overline{W}$,

$$\overline{w}^k_{t^k_m} := a^k_{t^k_m} \exp\left(\frac{1}{l_n}\sum_{j=0}^{l_n-1} Y^k_{t^k_{m+j}}\right), \tag{3.2}$$

for some given sequence $l_n \uparrow \infty$, the process $a_t := (a_t^1, a_t^2, \ldots, a_t^d)$ is an adapted càdlàg step function describing the amounts invested into each share, and finally, the rows of the $d^2 \times d^2$-matrix $Q$ is for $i, j = 1, 2, \ldots, d$ given by

$$Q^{(i-1)d+j} := (0^\top_{(i-1)d+j-1}, 1, 0^\top_{(d-i+1)d-j}) + (0^\top_{(j-1)d+i-1}, -1, 0^\top_{(d-j+1)d-i}),$$

for $i \neq j$, and $0^\top_{d^2}$ otherwise.

In this project, we use the *mrcCholCov*$^\bullet$-estimator as the pre-estimator $\overline{\Sigma}$. We will denote the observed log-price process of the ETF by $Y^{d+1}$, and assume that there is a latent

log-price process $X^{d+1}$, which is given by

$$X_t^{d+1} = \log\left(\sum_{k=1}^{d} a_t^k \exp(X_t^k)\right),$$

where the process $a_t \coloneqq (a_t^1, a_t^2, \ldots, a_t^d)$ is given as in Definition 3.2.1. Boudt et al. (2023) derive that the covariation between the $l^{\text{th}}$ asset and the ETF is given by

$$\beta^l = \sum_{k=1}^{d} \int_0^1 w_s^k \Sigma_s^{kl} ds,$$

where $\Sigma_t^{kl}$ is the spot covariation between assets $k$ and $l$, and $w_t^l \coloneqq a_t^l \exp(X_t^l)$. An estimator of $\beta^l$ is given by

$$\overline{\beta}^l \coloneqq \sum_{k=1}^{d} \sum_{m=0}^{[n_k/k_n]-1} \overline{w}^k_{t^k_{mk_n}} \hat{\Sigma}^{kl}_{t^k_{mk_n}} \left(t^k_{(m+1)k_n} - t^k_{mk_n}\right),$$

where $k_n \in \mathbb{N}$ is some local estimation window, $\overline{w}^k_{t^k_m}$ is given as in (3.2), and $\hat{\Sigma}_t^{kl}$ is an estimate of the spot covariation based on the pre-estimator. Let $\overline{\Sigma}_t$ denote the pre-estimate of the integrated covariance up to time $t \in [0,1]$, then the spot covariance is estimated by

$$\hat{\Sigma}_t^{kl} \coloneqq \frac{n_k}{k_n}\left(\overline{\Sigma}^{kl}_{t+k_n/n_k} - \overline{\Sigma}^{kl}_t\right), \tag{3.3}$$

for $t \in (0, 1 - k_n/n_{k]}$, and for $1 - k_n/n_k < t \leq 1$ set $\hat{\Sigma}_t^{kl} = \hat{\Sigma}^{kl}_{1-k_n/n_k}$. In the following section, we will showcase the finite sample performance of the BAC-estimator compared to the vanilla *mrcCholCov*$^\bullet$-estimator through a series of simulation experiments. For an implementation of the BAC-estimator see `BACestimator.R`.

## 3.3 Simulation study

In this section, we present the results of a simulation study that compares the performance of the basket-adjusted *mrcCholCov*$^\bullet$ to the performance of the vanilla *mrcCholCov*$^\bullet$. These simulations take basis in datasets of simulated finance data, and we will therefore start by covering, how this data is produced. We will follow the setup proposed by Barndorff-Nielsen et al. (2011). In essence, we want to simulate the log-price processes of $d$ stocks, and one ETF with the following stochastic volatility model. Let the $k^{\text{th}}$ latent stock log-price process be given by the following continuous Itô semimartingale,

$$dX_t^k = \mu^k dt + dV_t^k + dF_t^k,$$

where

$$dV_t^k \coloneqq \rho^k \sigma_t^k dB_t^k,$$

$B^k$ is a standard Brownian motion independent from $B^i$ for any $i \neq k$, and

$$dF_t^k := \sqrt{1 - \left(\rho^k\right)^2}\sigma_t^k dW_t,$$

where $W$ is a standard Brownian motion independent from $B^k$ for any $k = 1, 2, \ldots, d$. The process $F^k$ is, thus, driven by the Brownian motion $W$, which is common among all processes for $k = 1, 2, \ldots, d$, hence, the term $F^k$ is also called the common term. Let $\sigma_t^k$ be given by

$$\sigma_t^k := \exp(\beta_0 + \beta_1 g_t^k),$$

where $g_t^k$ is a process assumed to satisfy the Langevin equation

$$dg_t^k := \alpha^k g_t^k dt + dB_t^k.$$

We sample the initial values of $g^k$ according to $g_0^k \sim N(0, (-2\alpha^k)^{-1})$, which imposes stationarity on $g^k$. In this project, we will use the same values for the parameters as are suggested by Barndorff-Nielsen et al. (2011), which are

$$(\mu^k, \beta_0^k, \beta_1^k, \alpha^k, \rho^k) = (0.03, -5/16, 1/8, -1/40, -0.3). \tag{3.4}$$

As a consequence of this setup, we have that

$$\mathbb{E}\left[\int_0^1 \left(\sigma_s^k\right)^2 ds\right] = 1. \tag{3.5}$$

The covariation between two simulated log-price processes, say $X^k$ and $X^j$, is the covariation between their continuous martingale components, which are given by

$$\begin{aligned} M_t^k &:= \rho^k \int_0^t \sigma_s^k dB_s^k + \sqrt{1 - \left(\rho^k\right)^2} \int_0^t \sigma_s^k dW_s, \text{ and} \\ M_t^j &:= \rho^j \int_0^t \sigma_s^j dB_s^j + \sqrt{1 - (\rho^j)^2} \int_0^t \sigma_s^j dW_s. \end{aligned}$$

Using the bilinearity of covariation, we obtain that

$$\langle M^k, M^j \rangle_t = (\rho^k)^2 \int_0^t (\sigma_s^k)^2 ds \cdot \mathbb{1}_{\{k=j\}} + \sqrt{\left(1 - \left(\rho^k\right)^2\right)\left(1 - (\rho^j)^2\right)} \int_0^t \sigma_s^k \sigma_s^j ds.$$

Noise is simulated in the following way

$$U_t^k \mid \sigma, X \overset{i.i.d.}{\sim} N(0, \omega^2), \text{ where } \omega^2 := \xi^2 \sqrt{\frac{1}{n_k} \sum_{i=1}^{n_k} \left(\sigma_t^k\right)^4 \frac{i}{n_k}},$$

and the noise-to-signal ratio is given by $\xi^2 = 0.001$. The observed log-price processes are generated by $Y_t^k := X_t^k + U_t^k$. The ETF log-price process is generated by

$$X_t^{d+1} = \log\left(\sum_{k=1}^{d} a^k \exp\left(X_t^k\right)\right), \tag{3.6}$$

where the $d$-dimensional vector $a$, denoting the proportions of wealth invested into each asset, is sampled according to

$$a \sim \text{Dir}\underbrace{\left(\frac{1}{d}, \frac{1}{d}, \ldots, \frac{1}{d}\right)}_{d},$$

which is the Dirichlet distribution. Finally, the noise for the observed ETF log-price process is sampled according to

$$U_t^{d+1} \mid X^{d+1} \overset{i.i.d.}{\sim} N\left(0, \xi^2\left(\sigma_{ETF}^{IV}\right)^2\right),$$

where $\left(\sigma_{ETF}^{IV}\right)^2$ is the integrated variance of the unobserved ETF log-price process. The observed ETF log-price process is thus given by $Y_t^{d+1} = X_t^{d+1} + U_t^{d+1}$. The observations are assumed to arrive according to a Poisson process with intensity parameter $\lambda^k$ for $k = 1, 2, \ldots, d+1$. That is, for a simulated trading day, the expected number of trades for asset $k$ is $\lambda^k$. To generate sample paths of the log-price processes, we use the Euler-Maruyama scheme. For an implementation of the synthetic finance data generation process see `simulateFinanceData.R`. The adjustment matrix of the BAC-estimator relies on the information imposed by having high-frequency data available for ETF baskets. In the following subsection, we will present the results of a sensitivy analysis of the relationship between the ETF-liquidity and the compared performance of the BAC-estimator relative to the performance of the vanilla *mrcCholCov*$^\bullet$-estimator.

### 3.3.1 The performance sensitivity relative to the ETF-liquidity

In this subsection, we will show the results of a sensitivity analysis of the relationship between the ETF liquidity and the performance of the BAC-estimator with the *mrcCholCov*$^\bullet$ estimator as the pre-estimator versus the vanilla *mrcCholCov*$^\bullet$-estimator. We simulate data according to the scheme presented in the former section, and we will take basis in the parameters specified by (3.4). For the arrival intensities for the log-price observations, we will use the same approach as Boudt et al. (2023). We simulate data with minute resolution, which means that we have a maximum of $6.5 \cdot 60 = 390$ observations per day. Let the intensity for asset $k = 1$ be given by $\lambda^1 = 40$, let $\lambda^d = 390$ for asset $k = d$, and let

$$\lambda^k = \lambda^1 + \exp\left(\nu \frac{k-d}{d-1}\right)(\lambda^d - \lambda^1), \tag{3.7}$$

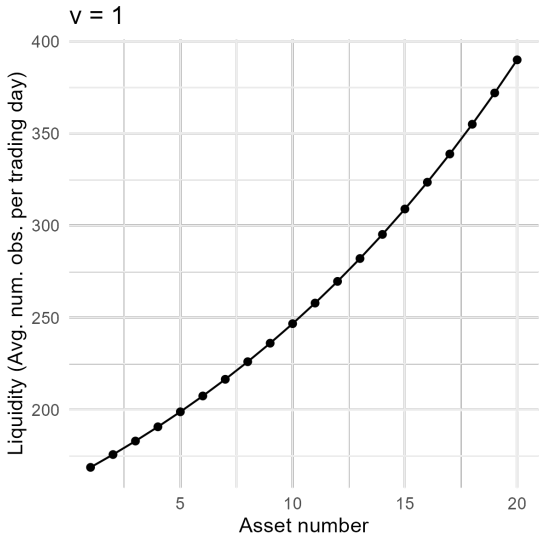
for $k = 1, 2, \ldots, d$. The parameter $\nu$ determines the steepness of the distribution of stock liquidities. For $\nu$ close to 0 the distribution of liquidities is more equally distributed, while higher values of $\nu$ causes a steeper growth in stock liquidities. Figure 3.1 demonstrates the distribution of stock liquidities for different values of $\nu$. For this sensitivity analysis, we use the value $\nu = 10$, which yields a set of stocks with low liquidity and a few high liquidity stocks. This is also the value of $\nu$ that is used in the simulation study by Boudt et al. (2023). We will let the liquidity of the ETF, $\lambda^{d+1}$, vary from 50 to 390 in steplengths of 10. For each value of $\lambda^{d+1}$, we conduct 1000 simulations, and estimate the rate at which the BAC-estimator had a better tracking error than the vanilla *mrcCholCov*$^\bullet$-estimator, with

$$G(\lambda^{d+1}) := \frac{1}{1000} \sum_{i=1}^{1000} \mathbb{1}_{\left\{TE\left(\alpha(\hat{\Omega}_t^{BAC});\hat{\Omega}_t^{BAC}\right) < TE\left(\alpha(\hat{\Omega}_t^{\text{CholCov}});\hat{\Omega}_t^{\text{CholCov}}\right)\right\}}. \tag{3.8}$$

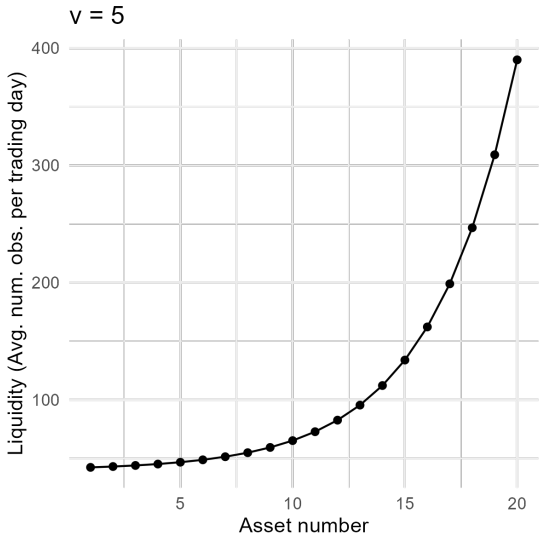
The standard deviation of $G$ is estimated as the standard deviation of a Bernoulli distribution, which is

$$\hat{\sigma}_G = \sqrt{\frac{G(\lambda^{t+1})(1 - G(\lambda^{t+1}))}{1000}}.$$

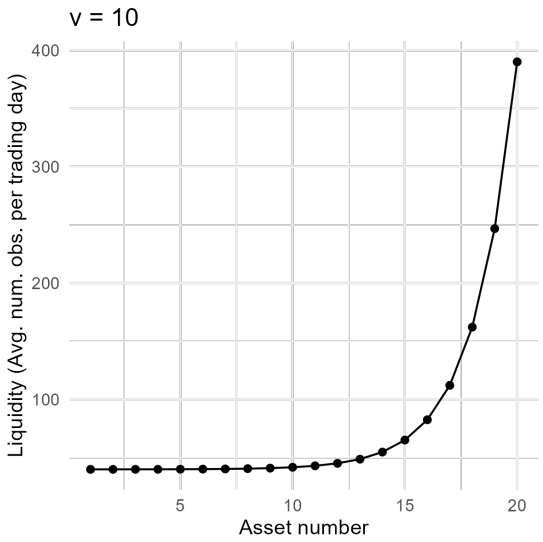
The results of the sensitivity analysis is illustrated in Figure 3.2. The blue lines illustrate the estimates of $G$, and the red dashed lines indicate the confidence intervals. None of the plots indicate a significant relationship between the ETF liquidity and the performance of the BAC-estimator with respect to the *mrcCholCov*$^\bullet$-estimator. This is also supported by Table 3.1, which shows that there is no statistical relationship between $G$ and $\lambda^{d+1}$. The idea behind the BAC-estimator is to use the high-frequency availability of ETF-prices to improve the pre-estimator, hence, the expected behavior should be that the higher the ETF liquidity the better the performance of the BAC-estimator relative to the vanilla *mrcCholCov*$^\bullet$. An explanation for the lack of this behavior in this experiment could lie in the way that the ETF prices are simulated. We simulate the ETF prices according to (3.6), and we generate observation times according a Poisson process with intensity $\lambda^{d+1}$. When we have an observation time for the ETF we simply calculate the ETF prices with the most recent observations of the stocks before that time. This excludes any supply and demand-effects that would be imposed on the ETF prices in a real market, hence, the information that the adjustment-matrix in the BAC-estimator yields, is not necessarily as profound as it would be in a real market scenario. The black dashed lines in Figure 3.2 indicates the mean of the estimates, $G$, across all ETF liquidity values. They show that on average for $d = 10$ the rate at which the BAC-estimator is better than the *mrcCholCov*$^\bullet$ is $\approx 0.61$ and for $d = 20$, and $d = 30$ the rate is on average $\approx 0.5$, which also conforms with the intercept estimates shown in Table 3.1. A rate of 0.5 likely indicates that the BAC-estimator performs similarly to the vanilla *mrcCholCov*$^\bullet$ for $d = 20$ and $d = 30$. However, the BAC-estimator was better on average 61% of times for $d = 10$. This behavior of having lower relative BAC-performance may be caused by the lack of information that the ETF prices relay, since the prices are simulated, and it may also be caused by the smaller cor-



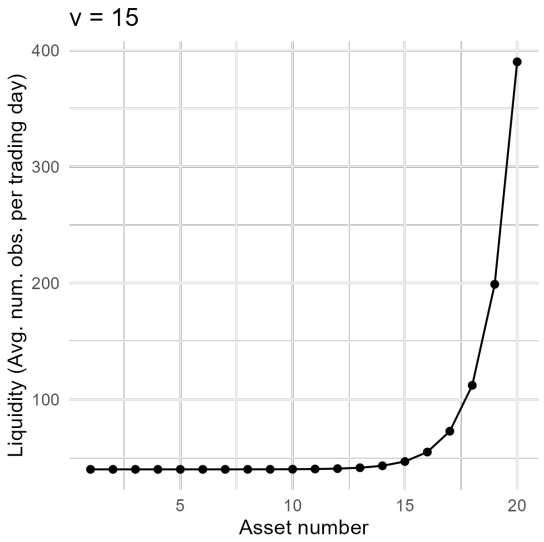

(a) Liquidity distribution for $\nu = 1$.




(b) Liquidity distribution for $\nu = 5$.



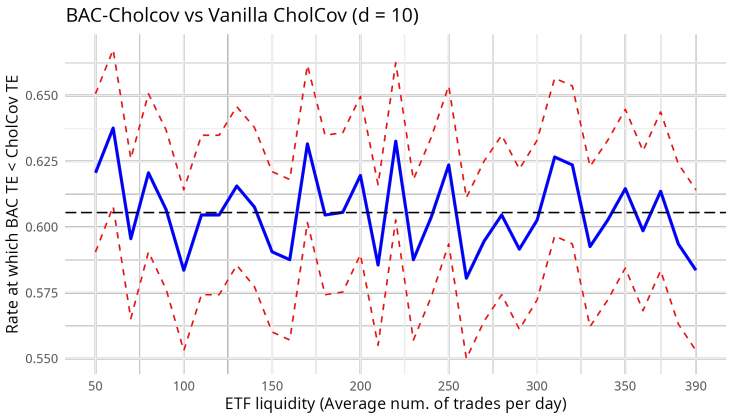

(c) Liquidity distribution for $\nu = 10$.



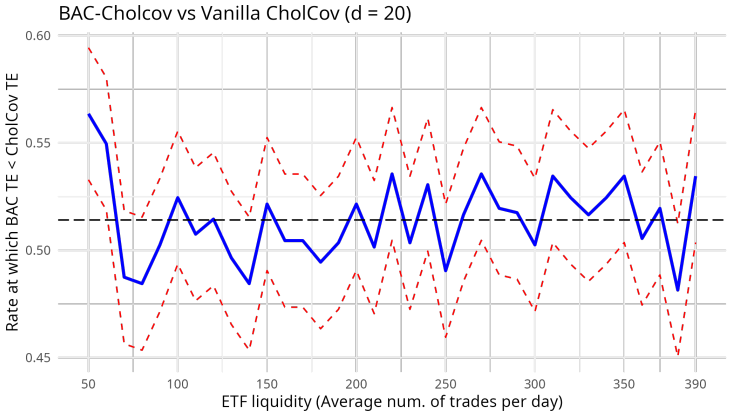

(d) Liquidity distribution for $\nu = 15$.

Figure 3.1: Distribution of the liquidity according to (3.7) with varying values of $\nu$.

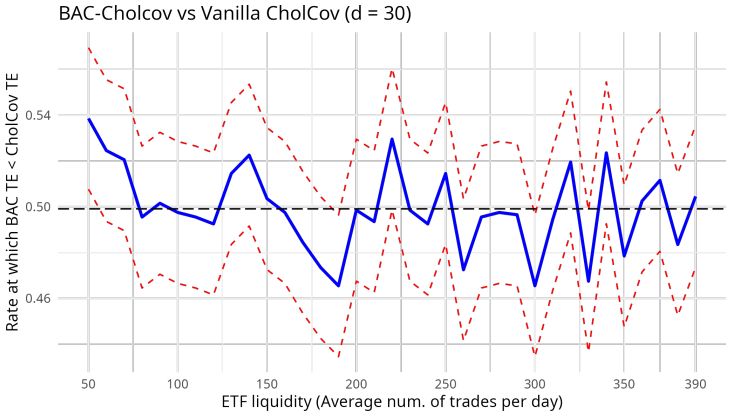
relation factor each individual stock will have on the ETF price, when there are more stocks present. These sources of error are likely less present in a scenario with real market data. We will continue our simulation study with a sensitivity analysis of the relationship between the parameter $\nu$ and the Monte Carlo estimate $G$.




(a) Results for $d = 10$ simulated stocks. Mean value is $\approx 0.6055$.




(b) Results for $d = 20$ simulated stocks. Mean value is $\approx 0.5141$.




(c) Results for $d = 30$ simulated stocks. Mean value is $\approx 0.4991$.

Figure 3.2: Results of running a sensitivity analysis of the relationship between the ETF-liquidity and the compared performance between the BAC-estimator and the $mrcCholCov^{\bullet}$-estimator. The blue lines are the Monte Carlo estimates of $G$, the red dashed lines are the confidence intervals, and the horizontal black dashed lines are the overall mean values of the $G$.

$d = 10$

| | Estimate | Std. Error | $t$-value | $p$-value |
|---|---|---|---|---|
| $\alpha$ | 0.6121 | 0.0062 | 98.4010 | $< 0.001$ |
| $\beta$ | 0.0000 | 0.0000 | $-1.1680$ | 0.2510 |

$d = 20$

| | Estimate | Std. Error | $t$-value | $p$-value |
|---|---|---|---|---|
| $\alpha$ | 0.5105 | 0.0078 | 65.2680 | $< 0.001$ |
| $\beta$ | 0.0000 | 0.0000 | 0.5030 | 0.6180 |

$d = 30$

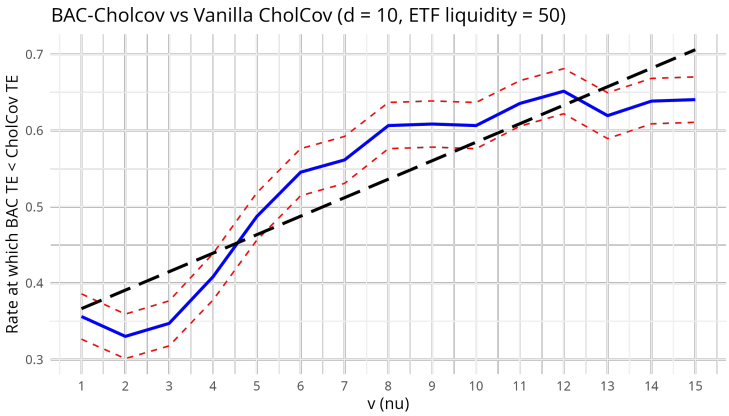
| | Estimate | Std. Error | $t$-value | $p$-value |
|---|---|---|---|---|
| $\alpha$ | 0.5101 | 0.0073 | 70.0850 | $< 0.001$ |
| $\beta$ | -0.0001 | 0.0000 | $-1.6690$ | 0.1050 |

Table 3.1: Summary of the linear regression $G(\lambda^{d+1}) = \alpha + \beta\lambda^{d+1}$ using the Monte Carlo estimates of the sensitivity analysis.

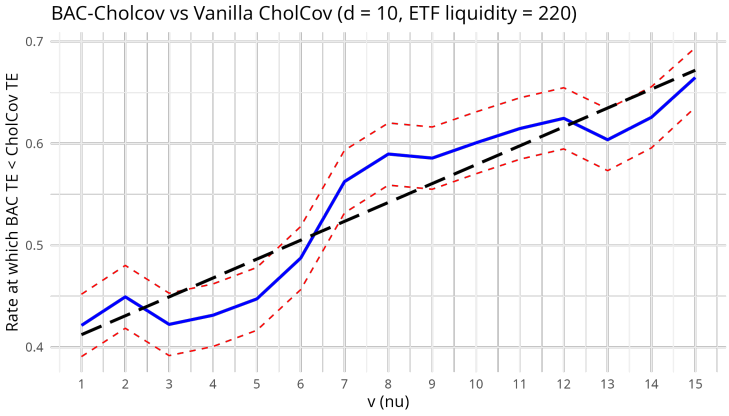
### 3.3.2 Sensitivity analysis of $\nu$

In this section, we present the results of a sensitivity analysis of the relationship between $\nu$ and the Monte Carlo estimate of the rate at which the BAC-estimator is better than the *mrcCholCov*$^\bullet$-estimator. This is done with a similar approach as in the former subsection, but with a fixed value for the ETF liquidity. We let $\nu$ vary from 1 to 15 with step lengths of 1, and for each value of $\nu$ we run 1000 simulations and calculate the rate $G$ according to (3.8). All simulations are run with $d = 10$, and we do this for ETF liquidities 50, 220, and 390. The results are shown in Figure 3.3. The plots generally show that for higher values of $\nu$ the better average performance of the BAC-estimator relative to the *mrcCholCov*$^\bullet$-estimator. This relationship is also supported by Table 3.2, which suggests that there is a statistical relationship between $\nu$ and $G$.[4] The improved relative performance of the BAC-estimator for high values of $\nu$ may be explained by the fact that high values of $\nu$ imply fewer simulated stocks with high liquidity and more stocks with low liquidity. As a consequence of the low liquidity there are less observations available for the *mrcCholCov*$^\bullet$-estimator, however the availability of ETF-prices even at low frequency aids to improve the integrated covariance estimates using the BAC-estimator. Likewise, for low values of $\nu$ the decreased relative performance of the BAC-estimator, is likely caused by having more liquid stocks, which decreases the contribution of the ETF-prices to the intergrated covariance estimate. In the next section, we will investigate, how well the BAC-estimator performs compared to the *mrcCholCov*$^\bullet$-estimator when tracking with subsets of stocks.

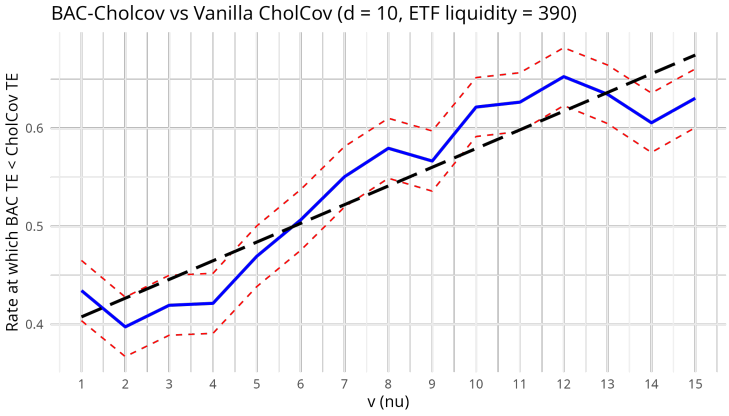
[4] A logistic regression was also fitted, but it suggested only insignificant estimates. This is likely caused by the narrow range of $G$. See Section A.2 for a summary of the logistic regression.




(a) Results based on an ETF liquidity of 50.




(b) Results based on an ETF liquidity of 220.




(c) Results based on an ETF liquidity of 390.

Figure 3.3: Results of a sensitivity analysis investigating the relationship between the parameter $\nu$ and the rate $G$. The blue lines illustrate the Monte Carlo estimates $G$, the red dashed lines are confidence intervals, and the black dashed lines are linear regression fits to the Monte Carlo estimates.

ETF liquidity of 50

| | Estimate | Std. Error | $t$-value | $p$-value |
|---|---|---|---|---|
| $\alpha$ | 0.3424 | 0.0270 | 12.6990 | $< 0.001$ |
| $\beta$ | 0.0242 | 0.0030 | 8.1750 | $< 0.001$ |

ETF liquidity of 220

| | Estimate | Std. Error | $t$-value | $p$-value |
|---|---|---|---|---|
| $\alpha$ | 0.3938 | 0.0160 | 24.5500 | $< 0.001$ |
| $\beta$ | 0.0185 | 0.0018 | 10.5100 | $< 0.001$ |

ETF liquidity of 390

| | Estimate | Std. Error | $t$-value | $p$-value |
|---|---|---|---|---|
| $\alpha$ | 0.3885 | 0.0184 | 21.0660 | $< 0.001$ |
| $\beta$ | 0.0191 | 0.0020 | 9.4040 | $< 0.001$ |

Table 3.2: Summary of the linear regression $G(\lambda^{d+1}) = \alpha + \beta\nu$ using the Monte Carlo estimates of the sensitivity analysis.

### 3.3.3 Tracking with subsets of stocks

Until now, we have used the entire set of component stocks for tracking and assess the performance of the BAC-estimator relative to the *mrcCholCov*$^\bullet$-estimator. In this section, we use the method of selecting an appropriate subset of stocks for index tracking, which was described in Section 2.1. The pipeline for each simulation consists of simulating a finance dataset, obtaining the integrated covariance estimates using the BAC-estimator and the *mrcCholCov*$^\bullet$-estimator, then finding an appropriate subset of stocks to include in the index fund using the method explained in Section 2.1, calculating the respective tracking errors, and finally comparing the tracking errors. This pipeline is also illustrated in Figure 3.4.

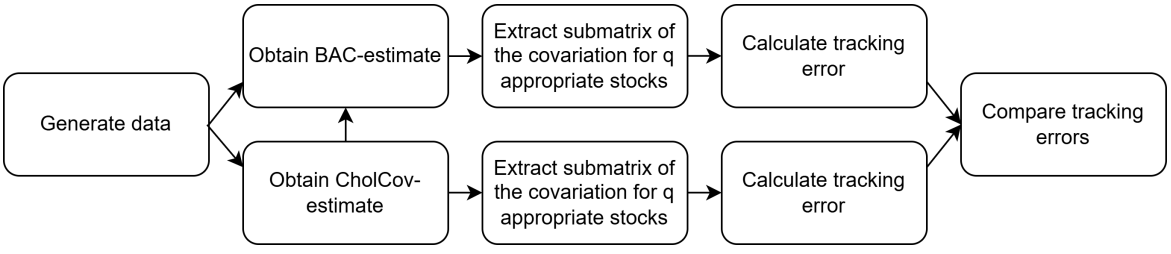



Figure 3.4: Pipeline for a single simulation run. First a dataset of finance data is simulated, then the BAC- and *mrcCholCov*$^\bullet$-estimates are obtained. Using these integrated covariance estimates, appropriate subsets of stocks are found, and their covariation matrices are used to calculate the tracking errors.

For the simulations, we let $d = 10$, fix $\nu = 10$, and let the ETF-liquidity be fixed at $\lambda^{d+1} = 390$. We let the number of stocks included in the subset, $q$, vary from 1 to 10 with steplengths of 1. Because of computational limitations, we only run 100 simulations per $q$ rather than 1000, which was used in the former analyses. This yields standard deviations that are larger by a factor of $\sqrt{10} \approx 3.16$ compared to the former analyses. The results of this simulation study can be seen in Figure 3.5.

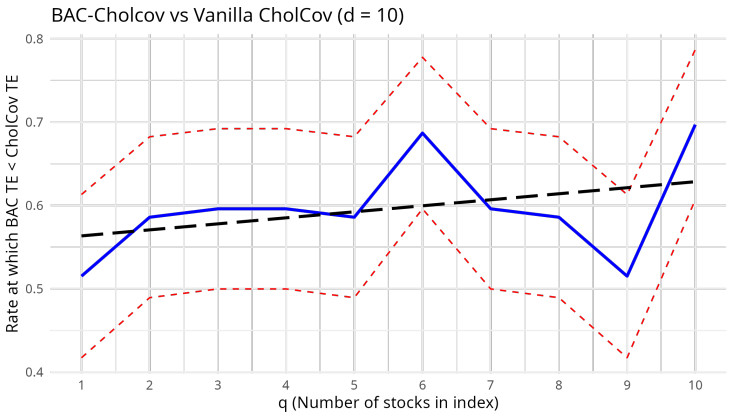



Figure 3.5: Results of a sensitivity analysis of the relationship between the relative BAC-performance and the number of stocks used for tracking. Blue represent the Monte Carlo estimates $G$, red dashed lines represent 95%-confidence intervals, and the black dashed line shows a linear regression of $G$ on $q$.

Figure 3.5 shows that overall the BAC-estimator performs better than the $mrcCholCov^{\bullet}$-estimator, and the black dashed lines also suggest that there is a slightly increasing trend in the relative performance as $q$ increases. Table 3.3 supports the claim that the BAC-estimator is better than the $mrcCholCov^{\bullet}$-estimator across all values of $q$, however, the slightly increasing relative performance relative to $q$ is not statistically significant.

| | Estimate | Std. Error | $t$-value | $p$-value |
|---|---|---|---|---|
| $\alpha$ | 0.5562 | 0.0401 | 13.8800 | $< 0.001$ |
| $\beta$ | 0.0072 | 0.0065 | 1.1190 | 0.2960 |

Table 3.3: Summary of the linear regression $G(\lambda^{d+1}) = \alpha + \beta q$ using the Monte Carlo estimates of the sensitivity analysis.

Since $G$ is the rate at which the BAC-estimator was better than the $mrcCholCov^{\bullet}$-estimator, the results do not specify the overall performance of the two integrated covariance estimators, however the overall tracking performance likely still improves significantly as $q$ increases. Therefore, it makes sense that there may not be an overall improvement in the rate $G$, when $q$ increases. In the next subsection, we will summarize the conclusions of the entire simulation study.

### 3.3.4 Summary of results

In the simulation study, we investigated the following three relationships:

- The relationship between the rate $G$ and the ETF-liquidity with $d = 10, 20$, and 30.
- The relationship between the rate $G$ and the distribution of the stock liquidities through the parameter $\nu$ all with $d = 10$ and ETF liquidity levels of $50, 220$, and 390.
- The relationship between the rate $G$ and the number of stocks included in the index fund.

We found that the ETF-liquidity did not seem to have a strong impact on the rate $G$, however there was an overall improvement with having fewer simulated stocks, that is $d = 10$, rather than $d = 20$ or 30. This might be caused by limitations in the simulation of the ETF log-prices. In practice, the adjustment-matrix of the BAC-estimator is likely more profound, since it conveys information from for instance supply-demand and other market-related effects. We continued the simulation study with $d = 10$, and investigated the sensitivity of the relationship between the rate $G$ and the parameter $\nu$ with $\nu$ varying from 1 to 15, and for ETF-liquidities of $50, 220$, and 390. The results suggested that the rate $G$ significantly improved for high values of $\nu$. An explanation for this could be that for high values of $\nu$, there are few low liquidity stocks, which results in few observations for covariation estimation. The inclusion of ETF log-prices through the BAC-estimator may therefore yield a significant improvement in the integrated covariance estimate. In the final simulation analysis, we found that the BAC-estimator consistently performed better than the $mrcCholCov^{\bullet}$-estimator.

In conclusion, we found that the BAC-estimator often has lower tracking errors than the $mrcCholCov^{\bullet}$-estimator, especially in the presence of low-liquidity stocks. In the next chapter, we investigate the tracking performance on real-world data.

# 4 Empirical study

In this section, the framework for estimating integrated covariance and index tracking is evaluated using real market data. We will in particular take basis in the S&P 500 index and a selection of 55 stocks from the index. For the ETF, used in the BAC-estimator, we utilize the $SPY$, since it is highly liquid and based on pure index tracking. We initiate this chapter with a section describing the data that we use.

## 4.1 Data description

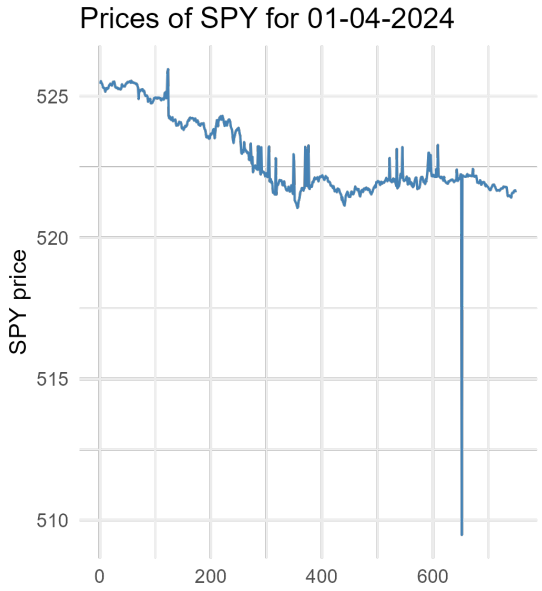
In this project, we take basis in the S&P 500, however because of computational limitations we will not conduct this empirical study with all component stocks. Instead, we pick 5 stocks from each of the 11 Global Industry Classification Standard (GICS) sectors that make up the S&P 500 index. Each sector represents its own index, and the 5 highest weighted stocks from each sector are chosen. Information about the stock weights within each sector index is available from `ssga.com`. We have added a complete overview of which stocks we picked from which sectors, as well as a short description of each sector in Section A.3. All data is retrieved from *polygon.io* through `R`. For each of the 55 stocks, their market capital are retrieved for the 1st of April 2024, and for each trading day in the period from the 1st of April 2024 to the 30th of June 2024 the volume adjusted log-prices are retrieved with a minute resolution. In this time interval, we retrieved data for 63 trading days. The market capital is used to assess the weights in the index, since the weights of the S&P 500 is determined by

$$a^i := \frac{\textit{market cap for asset } i}{\textit{sum of market caps for all stocks in S\&P 500}}.$$

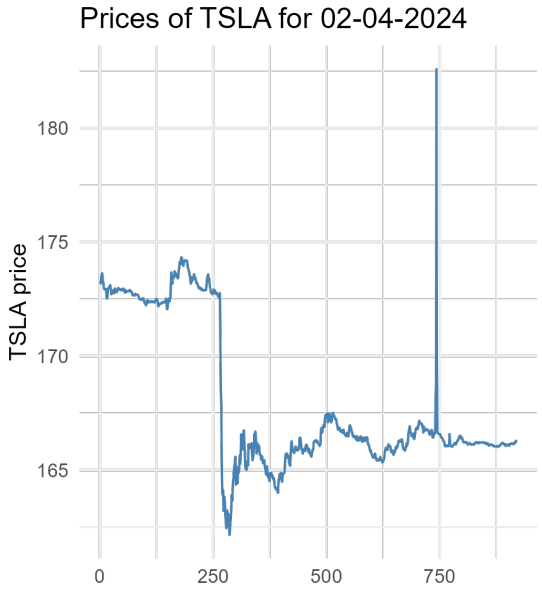
Recall that the weight processes for each stock are needed in order to estimate the invested wealth processes, (3.2), used in the BAC-estimator. We have chosen to only use data in the time interval from the 1st of April 2024 to the 30th of June 2024, since the S&P 500 is rebalanced every quarter (Hayes (2025)). As a consequence of this, we can let the càdlàg step function, $a_t$, of Definition 3.2.1 be constant.[1]

Figure 4.1 suggests that there are outliers present in the raw data. Furthermore, the data also contains observations outside the 9:30 EST to 16:00 EST time window for which the NYSE is open. We therefore use the following data cleaning procedures proposed by Barndorff-Nielsen et al. (2009):

[1] To obtain the data, use `dataRequest.R` and specify location of `stock_symbols.csv` in `WD`. Script can be run using a free *polygon.io* key.




(a) Raw price data for the *SPY* on the 1st of April 2024.




(b) Raw price data for the *TSLA* on the 2nd of April 2024.

Figure 4.1: Figures (a) and (b) suggest that there are outliers present in the raw data downloaded from *polygon.io*.

- **P1:** Delete entries with a time stamp outside of the 9:30 to 16:00 window when the exchange is open.
- **P2:** Delete entries with a bid, ask or transaction price equal to zero.
- **Q4:** Delete entries for which the mid-quote deviated by more than 10 mean absolute deviations from a rolling centred median (excluding the observation under consideration) of 50 observations (25 observations before and 25 after).
- **T3:** If multiple transactions have the same time stamp use the median price. (In this project, we use the volume weighted average instead of the median).[2]

Barndorff-Nielsen et al. (2009) suggest a few more steps in the data cleaning procedure, but these are not included in this project, since we are not retrieving bid-ask-spread or letter codes. Figure 4.2 shows how many observations we have on average for each of the 63 trading days in the time interval after cleaning. The maximum liquidity is 390 trades per day, since we obtain the data with minute resolution, and we see that most of the stocks have above 300 trades per day on average. Only three assets fall below the 300 trades per day on average, and the least traded asset is `BKNG` with an average of 115 trades per day. If we sort the assets by their squared duration, and conduct refresh time-sampling with first the most liquid asset, then the most liquid and the next most liquid asset, and so on,

[2]The data cleaning procedures are implemented in `dataCleaning.R`.

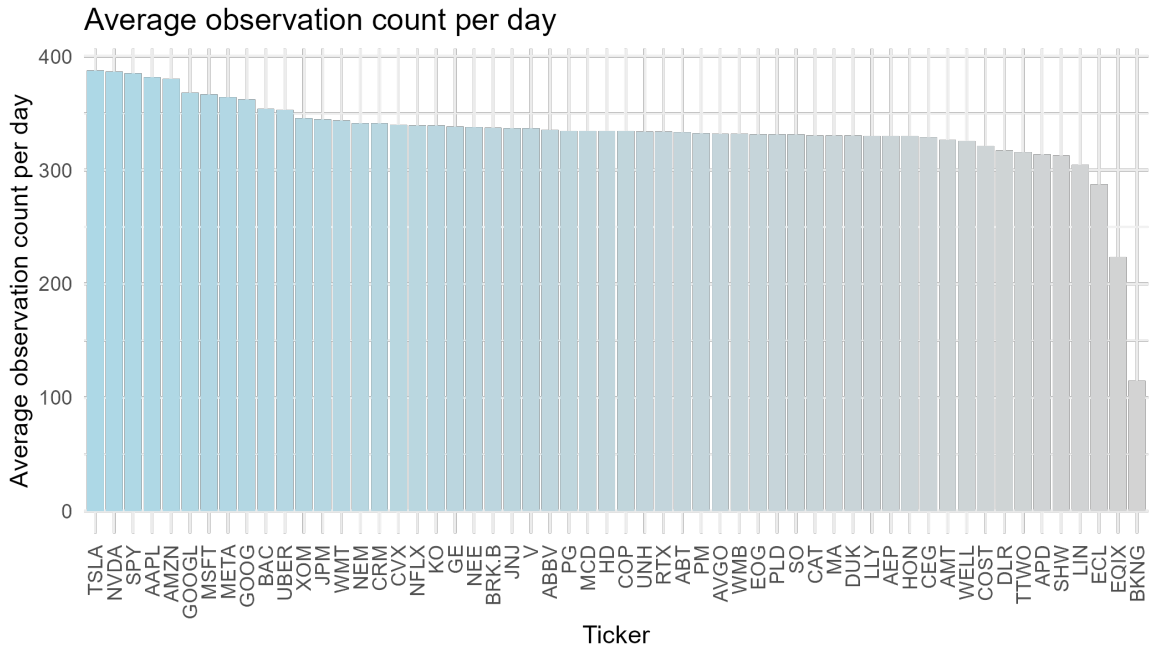



Figure 4.2: Average observation count per day for the $SPY$ and all the selected stocks.

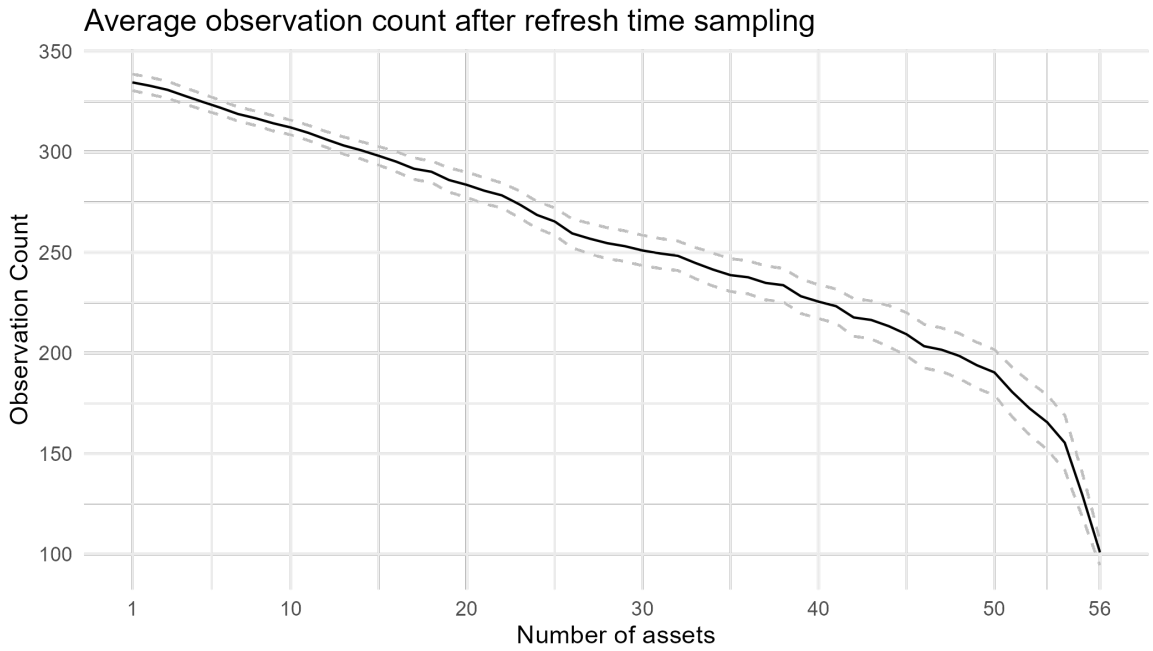


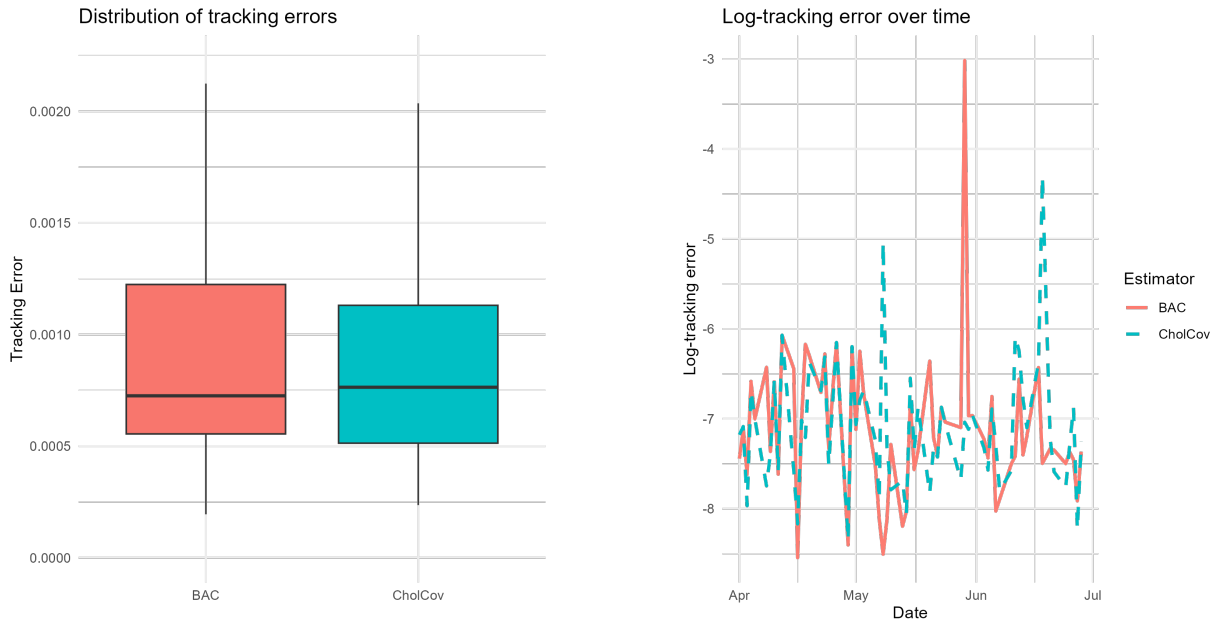

Figure 4.3: Average observation count after refresh time sampling relative to the number of included stocks. Dashed lines indicate the 95%-confidence interval.

we obtain Figure 4.3. It shows how many observations are available, when we estimate

the entries of the Cholesky-decomposition sequentially in the *mrcCholCov*•-estimator. The number of observations decreases steadily as more assets are included in the refresh time-sampling, and from 50 to 56 assets, it rapidly decreases towards 100 observations. In summary, we have obtained the price data for a subset of the S&P 500, cleaned it, and assessed the liquidity. In the next section, we evaluate the tracking performance based on the *mrcCholColv*•- and the BAC-estimator.

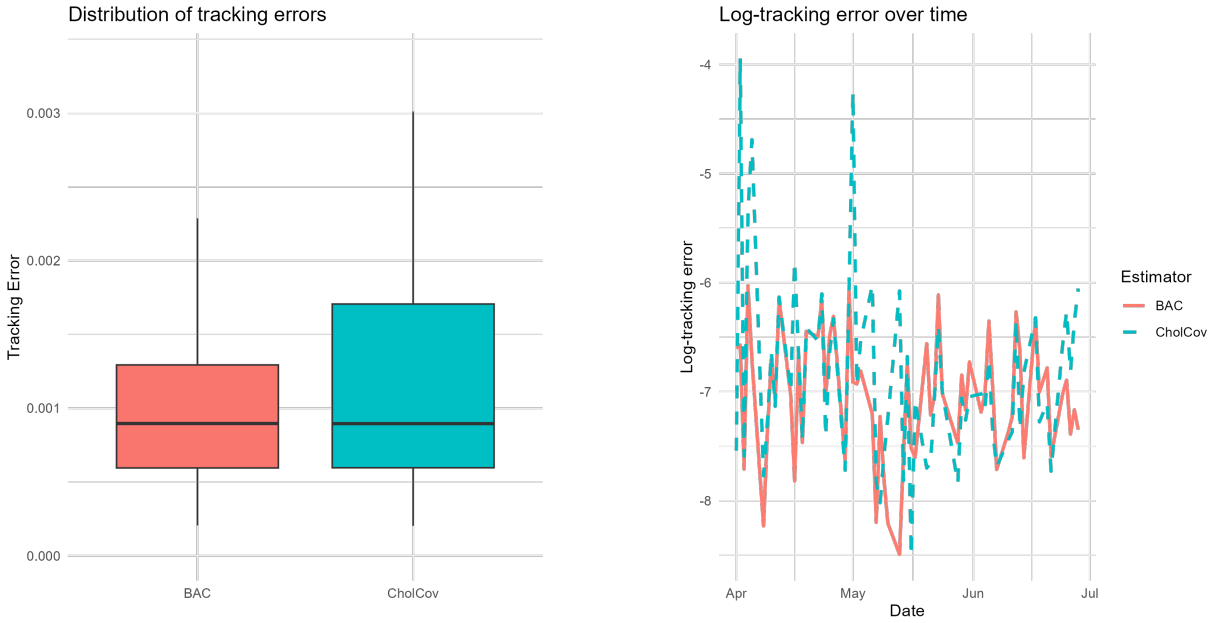
## 4.2 Tracking performance of the CholCov and the BAC

In this section, we present an evaluation of the tracking performances of the BAC-estimator and *mrcCholCov*•-estimator. To obtain the results, we follow the procedure outlined in Figure 3.4 for each trading day, but instead of generating data in the first step, we use the cleaned data of the 55 stocks. This yields tracking errors for index tracking based on the BAC-estimator and the *mrcCholCov*•-estimator for each day, and this experiment is conducted for $q = 10$, 20, and 30. That is, we evaluate the tracking performance using first a subset consisting of 10 of the 55 selected stocks, next a subset of 20 of the selected stocks, and finally a subset consisting of 30 of the selected stocks. Figure 4.4, Figure 4.5, and Figure 4.6 show the results of running this procedure.




(a) (b)

Figure 4.4: Plot (a) illustrates the distribution of the tracking errors for the BAC-estimator and the *mrcCholCov*•-estimator across all 63 trading days. Plot (b) shows the log-tracking errors for each day in the data retrieving window. ($q = 10$)

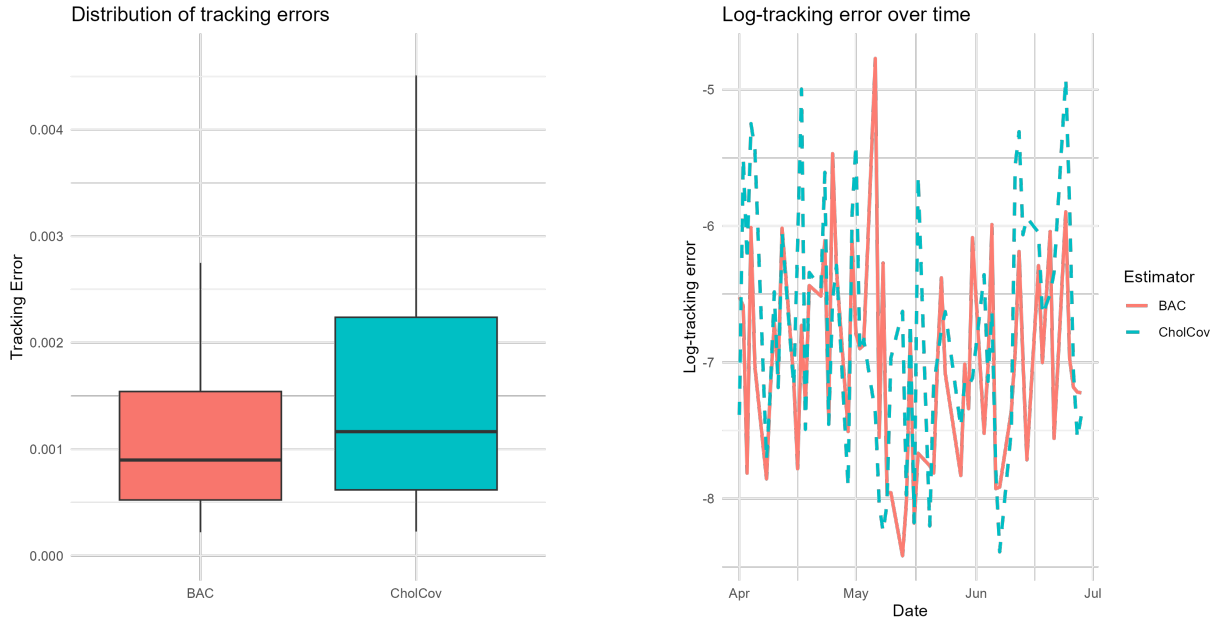



(a) (b)

Figure 4.5: Plot (a) illustrates the distribution of the tracking errors for the BAC-estimator and the *mrcCholCov*•-estimator across all 63 trading days. Plot (b) shows the log-tracking errors for each day in the data retrieving window. ($q = 20$)

In each figure, we have included boxplots that show the distribution of tracking errors for index tracking using either the BAC-estimator or the *mrcCholCov*•-estimator, and a line plot that shows the log-tracking errors for each trading day, likewise for index tracking using either of the two estimators. For $q = 10$ the boxplots show that the quantiles for the tracking errors are slightly higher, when using the BAC-estimator rather than when the *mrcCholCov*•-estimator is used, however, the median is sligthly lower for the BAC-estimator. For $q = 20$ the quantiles are more similar, however, the upper quartile and the maximum are largest for the *mrcCholCov*•-estimator, suggesting that index tracking based on the BAC-estimator is the better choice. For $q = 30$ all quantiles are lower for the BAC-tracking error than for the *mrcCholCov*•-tracking error. All in all, the results suggest that the BAC-estimator tracks better than the *mrcCholCov*•-estimator, however, the difference is mostly profound for $q > 10$.

Figure 4.7, Figure 4.8, and Figure 4.9 show the ratio of which each of the sectors are represented in the tracking portfolios. Both estimators tend to prioritize Communication Services and Consumer Staples for $q = 10$. Additionally, Consumer Discretionary and Energy stocks are prioritized along with Communication Services and Consumer Staples for $q = 20$. For $q = 30$ Communication, Consumer Disc., Consumer Staples, Energy, Financials, and Health Care stocks are prioritized. It generally appears that Information Technology stocks are prioritized slightly higher when tracking with the BAC-estimator

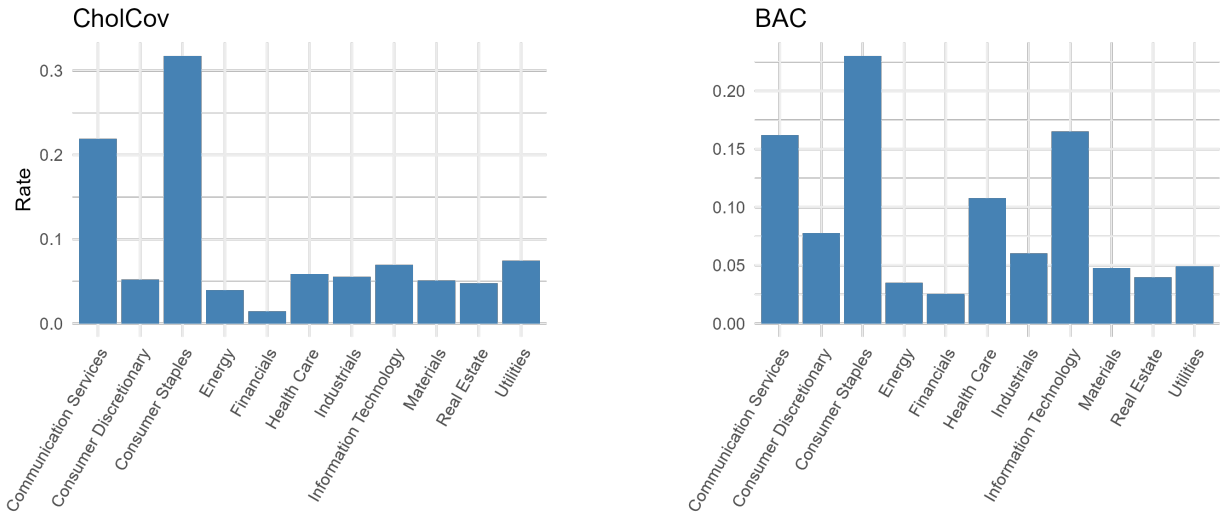



(a) (b)

Figure 4.6: Plot (a) illustrates the distribution of the tracking errors for the BAC-estimator and the *mrcCholCov*•-estimator across all 63 trading days. Plot (b) shows the log-tracking errors for each day in the data retrieving window. ($q = 30$)

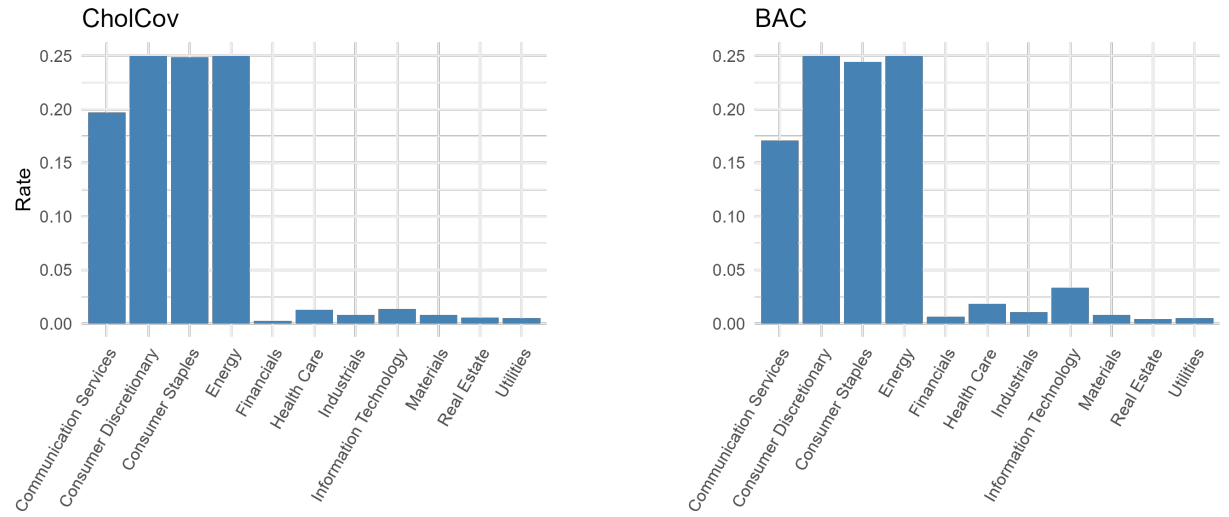
than with the *mrcCholCov*•-estimator. This means that the BAC-adjusment increases the overall covariation between tech-stocks and all other stocks, which may be due to illiquidity of the tech-related stocks. However, Figure 4.2 shows that the five tech stocks, `AAPL`, `MSFT`, `NVDA`, `AVGO`, and `CRM` each have more than 300 observations per day on average. Another explanation may instead be that a significant amount of observations are discarded for these stocks after the refresh-time sampling procedure. This can happen if trades comes in clusters rather than being evenly distributed across time, causing the *mrcCholCov*•-estimator to yield inaccurate estimates.

The method described in Section 2.1 for picking which stocks to go into the index fund, tries to pick the set of stocks that maximizes the correlation with all stocks in the index. Hence, if some sectors generally have a higher correlation across all the stocks in the index, the stocks of these sectors will be prioritized to stocks of other sectors. This may explain why some sectors are often and equally represented, while other sectors are almost never used. In addition, a source of error in the portfolio selection may stem from the subgradient descent algorithm not getting sufficiently close to the correct solution before terminating due to a limit of 200 iterations. This limit was determined because of computational limitations, however, a higher iteration limit may yield other results. Also, the method of selecting the stocks is based on maximizing correlation, however, this means that hypothetically stocks with a correlation of 0 with all other stocks are more likely to be included in the index



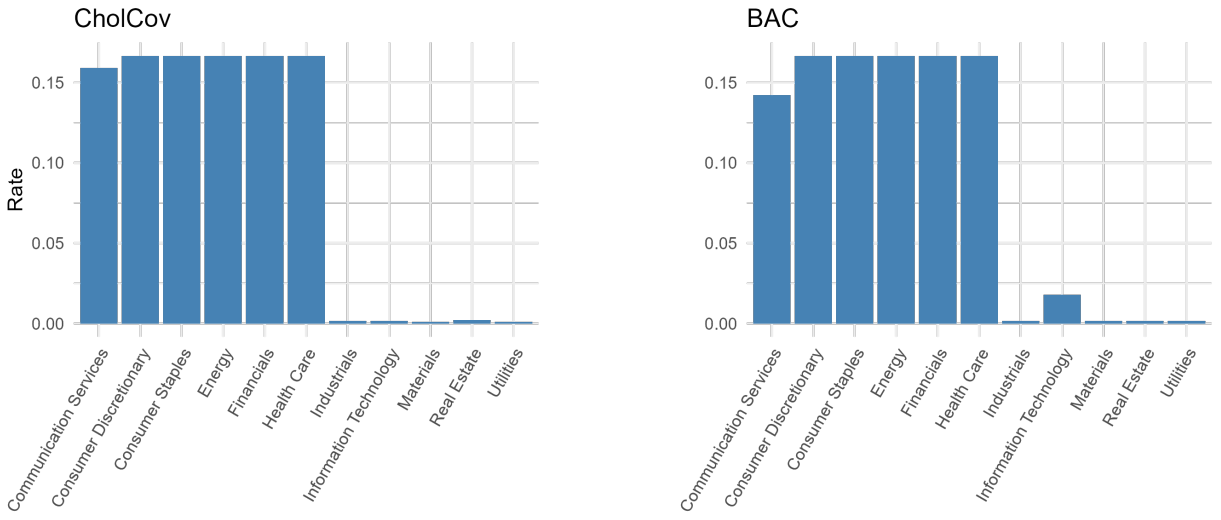

(a) (b)

Figure 4.7: Rate at which the sectors are picked for the index fund portfolio positions, when $q = 10$ stocks are used for tracking.




(a) (b)

Figure 4.8: Rate at which the sectors are picked for the index fund portfolio positions, when $q = 20$ stocks are used for tracking.




(a) (b)

Figure 4.9: Rate at which the sectors are picked for the index fund portfolio positions, when $q = 30$ stocks are used for tracking.

fund than stocks with a correlation close to $-1$. This may not be optimal, since a correlation close to $-1$ is a strong negative correlation, and could contribute more to the tracking performance than the uncorrelated stock, which would contribute only with noise. The results may therefore be improved by maximizing the absolute value of the correlations rather than just using the raw correlation estimates.

# 5 Discussion

This thesis can be divided into two parts. The first part revolves around the methodology of index tracking and covariation estimation, and the second part showcases the methods covered in the first part through a simulation study and an investigation based on real data. The central aim is to compare the BAC-estimator to the *mrcCholCov*$^\bullet$-estimator with index tracking as a use case.

Index tracking, as it is proposed by Fastrich et al. (2014), is done by finding the appropriate set of weights that minimizes the tracking error and has constrained the number of strictly positive weight entries. The approach of this thesis is to separate this minimization problem into two parts, where the first part finds an optimal subset of stocks to use for tracking, and the second part finds the optimal set of weights due to a closed form formula. This allows the use of subgradient descent to efficiently construct optimal index funds. A downside to this approach is, however, that the subgradient descent converges slowly compared to regular gradient descent. Finding appropriate subsets for index fund creation is therefore computationally heavy with this approach. Further investigation could look into the use of integer programming methods such as, for instance, the branch-and-bound algorithm.

The covariation estimation takes basis in the *mrcCholCov*$^\bullet$-estimator and the basket-adjusted covariance estimator using the *mrcCholCov*$^\bullet$-estimator as the pre-estimator. This framework is robust to both market microstructure noise and asynchronicity, as the *mrcCholCov*$^\bullet$-estimator uses pre-averaging and refresh-time sampling. The refresh-time sampling procedure does discard some observations, but due to the sequential estimation of the covariation, the number of observations can be optimized by queueing the stock-price processes from most to least liquid. Instability was encountered in the estimation of spot covariation, (3.3), for $t$ close to 0, however, this may be mitigated by using higher frequency data, than what is used in this thesis.

The simulation study investigates several relationships. First, we look into the relationship between the ETF-liquidity and the rate at which the BAC-estimator is better than the *mrcCholCov*$^\bullet$-estimator. No significant relationship is found, however, this may be due to limitations in the data simulation. In this project the ETF prices are generated at points in time, according to a Poisson process, as a weighted sum of the simulated component stock prices. However, if the stocks do not have simulated prices for the exact observation times of the ETF, the most recent simulated stock price is used instead. As a consequence of this, the ETF prices are entirely based on known stock prices, hence, the simulated ETF prices may likely not contribute with much information. A reasonable way to mitigate this would be to simulate all component stocks for each minute, then calculate the ETF prices for each minute, and then finally filter out observations for all processes, such that we end

up with the daily expected number of observations for each stock.

The simulation study also looks into the relationship between the distribution of stock liquidity and the performance of the BAC-estimator and the *mrcCholCov*$^{\bullet}$-estimator. For small $\nu$ the liquidity is evenly distributed and generally higher, while for high values of $\nu$ many stocks will have low liquidity and only few will have a high liquidity. The simulation results show that the BAC-estimator performs better than the *mrcCholCov*$^{\bullet}$-estimator at an increasingly higher rate as $\nu$ increases. This behavior is expected, since a high value for $\nu$ implies many stocks with low liquidity and few stocks with high liquidity, and, hence, the ETF-adjustment will yield more information relative to if all stocks had a high liquidity. This suggests that the benefit of using the BAC-estimator rather than the pre-estimator in the real world is much greater in cases with many low liquidity stocks and few high liquidity stocks.

It is also found that the rate at which the BAC-based tracking error is lower than the *mrcCholCov*$^{\bullet}$-based tracking does not show a significant dependence on the number of stocks included in the index fund. This suggests that it is equally beneficial to use the BAC-estimator rather than the *mrcCholCov*$^{\bullet}$-estimator regardless of the number of stocks picked for index tracking. This result does, however, not reveal the objective performance of the estimators for each level of $q$. The overall tracking performance is likely highly related to $q$, hence the result could be improved by looking at the distribution of tracking errors for both estimators for each level of $q$. This would reveal to which degree the BAC-based tracking would be better than the *mrcCholCov*$^{\bullet}$-based tracking.

Finally, the results of the empirical study relies only on a subset of the components of the S&P 500 due to computational limitations. The performance of the BAC-estimator and the *mrcCholCov*$^{\bullet}$-estimator are tested in index tracking scenarios, where the index fund is constructed from $10, 20$, and 30 stocks. The results show that in all cases the BAC-based tracking performs better than *mrcCholCov*$^{\bullet}$-based tracking, but the effect is mostly significant for $q > 10$. The two estimators also seemed to disagree about the importance of including Information Technology stocks into the index fund, as the BAC-based index tracking portfolio tended to include tech stocks at a slightly higher rate than the *mrcCholCov*$^{\bullet}$-based index fund. This suggests that the *mrcCholCov*$^{\bullet}$-estimator may underestimate the covariation for tech-related stocks.

The window in time for which data is retrieved spans one quarter, such that the index is not rebalanced within the trading window. This could easily be mitigated for data windows spanning more than a quarter by simply retrieving the market caps for the stocks each quarter and taking the new weights into account. More data would yield more reliable results. Another limitation to the approach of this thesis is that the subgradient descent algorithm is limited to a maximum of 200 iterations. For $q = 30$ the algorithm has a significantly lower likelihood of reaching the optimal solutions than for $q = 10$ or $q = 20$. For further investigation, an increased maximum iteration count is likely to yield more accurate results.

# 6 Conclusion

The purpose of this thesis was to assess whether the BAC-estimator posed an improvement relative to its pre-estimator, when using the *mrcCholCov*$^\bullet$-estimator as the pre-estimator. The use case for this project was index tracking, and the tracking performance was used to evaluate the estimators.

A framework for constructing index funds was created in two parts. The first part was an algorithm for choosing which of the component stocks of the index to go into the index fund. This part relied on maximizing the correlation between the index stocks and the index fund stocks. The second part was a solution to a minimization problem in which the optimal set of weights for the index was calculated. The covariation estimates of the stock-price processes, for which the BAC-estimator and the *mrcCholCov*$^\bullet$-estimator were used, were needed for running the framework. Hence, the estimators were introduced following the specification of the framework.

A simulation study delved into the comparative performance between BAC-based tracking and *mrcCholCov*$^\bullet$-based tracking. In general the BAC-based tracking seemed to outperform the *mrcCholCov*$^\bullet$-based tracking, but in particular the following relationships where investigated.

- It was found that the liquidity of the ETF did not pose a significant effect on the relative tracking performance of the BAC-estimator and the *mrcCholCov*$^\bullet$-estimator. However, this relationship was likely due to the simulation scheme used for generating the ETF prices, and the conclusion may have been different if a more realistic simulation scheme was adopted for the ETF prices.
- The BAC-based tracking had a significantly greater performance compared to the *mrcCholCov*$^\bullet$-based tracking as the distribution parameter $\nu$ increased. This finding suggests that the BAC-estimator may yield significantly better results in scenarios, where the component stocks of the index have low or uneven liquidity.
- For different sizes of tracking portfolios, that is, different values for $q$, the relative performance between the BAC-based tracking and the *mrcCholCov*$^\bullet$-based tracking stayed constant.

An empirical analysis compared the tracking performance for the two estimators for tracking portfolio sizes of $q = 10, 20$, and 30. Stock-price processes were retrieved for the five largest weighted S&P 500 stocks within each of the 11 GICS sectors for a total of 55 stocks, and the time interval spanned from the 1$^{\text{st}}$ of April 2024 to the 30$^{\text{th}}$ of June 2024. The overall result was that the BAC-based tracking performance was better than the *mrcChol-*

$Cov^{\bullet}$-based tracking, however the improvement was mostly profound for $q > 10$.

Further investigation could improve on the results of this thesis by applying an improved simulation scheme, higher frequency data, and more computationally intensive reruns of the analyses. Investigations could also delve into the use of for instance branch-and-bound or other optimization algorithms for improving the framework for index fund creation.

In summary, this thesis provides results which suggest that the BAC-estimator offers a significant improvement in index tracking performance relative to the $mrcCholCov^{\bullet}$-estimator.

# A | Appendix

## A.1 The MRC-estimator

In this section, we will cover the modulated realized covariance estimator described by Christensen et al. (2010). Their estimator utilizes pre-averaging, which is a method for modifying a process in order to mitigate noise - in our case, microstructure noise. Let $V \coloneqq (V_t)_{t\geq 0}$ be any $d$-dimensional process, and define the difference process in the following way, $\Delta_i^n V \coloneqq V_{i/n} - V_{(i-1)/n}$ for $i = 1, 2, \ldots, n$, assuming that we have observations of $V$ in the time interval $[0, 1]$ at times $t = 0, \frac{1}{n}, \frac{2}{n}, \ldots, 1$. Hence, we have $n + 1$ observations of $V$. The pre-average of $V$ is given by

$$\overline{V}_i^n \coloneqq \sum_{j=1}^{k_n-1} g\left(\frac{j}{k_n}\right) \Delta_{i+j}^n V, \quad \text{for } i = 0, 1, \ldots, n - k_n + 1,$$

where $k_n$ is the window of averaging satisfying

$$\frac{k_n}{\sqrt{n}} = \theta + o(n^{-1/4}), \tag{A.1}$$

and the weight function $g : [0, 1] \to \mathbb{R}$ is continuous, piecewise continuously differentiable, has a piecewise Lipschitz continuous first derivative, $g(0) = g(1) = 0$, and

$$\int_0^1 g^2(s)ds > 0.$$

Let the following two constants $\psi_1$ and $\psi_2$ be given by

$$\psi_1 \coloneqq \int_0^1 g'(u)^2 du, \text{ and } \psi_2 \coloneqq \int_0^1 g(u)^2 du,$$

Then the modulated realized covariance estimator is given by

$$\text{MRC}[V]_n' \coloneqq \left(\frac{n}{n - k_n + 2}\right)\left(\frac{1}{\psi_2 k_n}\right) \sum_{i=0}^{n-k_n+1} \overline{V}_i^n \left(\overline{V}_i^n\right)^\top.$$

The following theorem of Christensen et al. (2010) states, however, that this estimator is inconsistent.

**Theorem A.1.1. (Inconsistency of the MRC-estimator)**
Let $Y = (Y_t)_{t\geq 0}$ be a noisy log-price process as defined in Chapter 3, let $\varepsilon = (\varepsilon_t)_{t\geq 0}$ be the associated noise process with $\mathbb{E}[|\varepsilon_j|^4] < \infty$ for all $j = 1, 2, \ldots, d$, and let $(k_n, \theta)$ satisfy (A.1). Then

$$\text{MRC}[Y]'_n \xrightarrow{P} \int_0^1 \Sigma_s ds + \frac{\psi_1}{\theta^2 \psi_2} \Psi, \text{ as } n \to \infty, \tag{A.2}$$

where $\Psi \coloneqq \mathbb{E}[\varepsilon_t \varepsilon_t^\top]$, and $\Sigma_s$ is the spot covariation of $Y$ at time $s$.

**Proof.** *Omitted, but it can be found in the appendix of Christensen et al. (2010).* ■

To obtain a consistent estimate of the integrated covariance we need to estimate the bias. We obtain estimates of $\psi_1$ and $\psi_2$ by Riemann approximation, hence

$$\psi_1^{k_n} \coloneqq k_n \sum_{i=1}^{k_n} \left( g\left(\frac{i}{k_n}\right) - g\left(\frac{i-1}{k_n}\right) \right)^2, \text{ and } \psi_2^{k_n} \coloneqq \frac{1}{k_n} \sum_{i=1}^{k_n - 1} g^2\left(\frac{i}{k_n}\right).$$

For the estimate of $\Psi$ Christensen et al. (2010) use the following estimator

$$\hat{\Psi}_n \coloneqq \frac{1}{2n} \sum_{i=1}^{n} \Delta_i^n Y \left(\Delta_i^n Y\right)^\top.$$

They state, however, that this estimator has the following finite sample property

$$2n\hat{\Psi}_n = 2n\Psi + \int_0^1 \Sigma_s ds + o_P(n^{-1}),$$

where $\mathbb{E}[o_p(n^{-1})] = 0$. As a consequence of this, we have for finite $n$ that the bias-adjusted version version of (A.2) estimates

$$\text{MRC}[Y]'_n - \frac{\psi_1^{k_n}}{\theta^2 \psi_2^{k_n}} \hat{\Psi}_n \approx \left(1 - \frac{\psi_1^{k_n}}{\theta^2 \psi_2^{k_n}} \cdot \frac{1}{2n}\right) \int_0^1 \Sigma_s ds.$$

Therefore, for finite samples we will use the following rescaled bias-adjusted MRC-estimator,

$$\text{MRC}[Y]_n \coloneqq \left(1 - \frac{\psi_1^{k_n}}{\theta^2 \psi_2^{k_n}} \cdot \frac{1}{2n}\right)^{-1} \left(\text{MRC}[Y]'_n - \frac{\psi_1^{k_n}}{\theta^2 \psi_2^{k_n}} \hat{\Psi}_n\right).$$

Because of the bias-correction, the estimator $\text{MRC}[Y]_n$ is not necessarily positive semidefinite. There are several solutions to this. Christensen et al. (2010) proposes to choose $\theta$ in a way that ensures positive semidefiniteness with the trade-off of having slower convergence.

For this project, we have made an implementation that utilizes the approach described by Fan et al. (2012) for dealing with non-positive semidefiniteness. They conduct a singular value decomposition of the covariation matrix estimate, replace negative eigenvalues in the diagonal matrix with 0, and multiply back in order to obtain a positive semidefinite covariation matrix. The implementation can be found in the `Github-repo` in the file `Pre-averaged_Bias-adjusted_MRC(Christensen et al 2010).R`. For the weight function $g$ we use the proposed weight function of Podolskij and Vetter (2009), which is given by

$$g(x) \coloneqq \min(x, 1-x),$$

and for $k_n$ and $\theta$, we use the $k_n = \sqrt{n}$ and $\theta = 1$ as is done by Christensen et al. (2010).

## A.2 Logistic regression results

In this section, we present the results of fitting a logisitc regression between $\nu$ and $G$ in Subsection 3.3.2, where $\nu$ is the independent variable, and $G$ is the dependent variable. The results are shown in the following table.

ETF liquidity of 50

| | Estimate | Std. Error | $t$-value | $p$-value |
|---|---|---|---|---|
| $\alpha$ | $-0.6491$ | 1.1158 | $-0.5820$ | 0.5610 |
| $\beta$ | 0.1002 | 0.1249 | 0.8020 | 0.4220 |

ETF liquidity of 220

| | Estimate | Std. Error | $t$-value | $p$-value |
|---|---|---|---|---|
| $\alpha$ | $-0.4336$ | 1.1014 | $-0.3940$ | 0.6940 |
| $\beta$ | 0.0759 | 0.1228 | 0.6180 | 0.5370 |

ETF liquidity of 390

| | Estimate | Std. Error | $t$-value | $p$-value |
|---|---|---|---|---|
| $\alpha$ | $-0.4554$ | 1.1025 | $-0.4130$ | 0.6800 |
| $\beta$ | 0.0781 | 0.1230 | 0.6350 | 0.5250 |

Table A.1: Summary of the logistic regression logit $\left(G(\lambda^{d+1})\right) = \alpha + \beta\nu$ using the Monte Carlo estimates of the sensitivity analysis in Subsection 3.3.2.

## A.3 Selected sector stocks

In this section, we present an overview of the selected sector stocks used in the empirical study of this project. The overview can be seen in Table A.3. A short description for each of the GICS sectors is also given below. All information is available from `ssga.com`.

| GICS Sector | Ticker | Name |
|---|---|---|
| Communication Services | META | Meta Platforms Inc Class A |
| Communication Services | GOOGL | Alphabet Inc. Class A |
| Communication Services | NFLX | Netflix Inc. |
| Communication Services | GOOG | Alphabet Inc. Class C |
| Communication Services | TTWO | Take- Two Interactive Software |
| Consumer Staples | COST | Costco Wholesale Corporation |
| Consumer Staples | WMT | Walmart Inc. |
| Consumer Staples | PG | Procter & Gamble Company |
| Consumer Staples | KO | Coca-Cola Company |
| Consumer Staples | PM | Philip Morris International Inc. |
| Consumer Discretionary | AMZN | Amazon.com Inc. |
| Consumer Discretionary | TSLA | Tesla Inc. |
| Consumer Discretionary | HD | Home Depot Inc. |
| Consumer Discretionary | MCD | McDonald's Corporation |
| Consumer Discretionary | BKNG | Booking Holdings Inc. |
| Energy | XOM | Exxon Mobil Corporation |
| Energy | CVX | Chevron Coroporation |
| Energy | COP | ConocoPhillips |
| Energy | WMB | Williams Companies Inc. |
| Energy | EOG | EOG Resources Inc. |
| Financials | BRK.B | Berkshire Hathaway Inc. Class |
| Financials | JPM | JPMorgan Chase & Co. |
| Financials | V | Visa Inc. Class A |
| Financials | MA | Mastercard Incorporated Class A |
| Financials | BAC | Bank of America Corp |
| Health Care | LLY | Eli Lilly and Company |
| Health Care | UNH | UnitedHealth Group Incorporated |
| Health Care | JNJ | Johnson & Johnson |
| Health Care | ABBV | AbbVie Inc. |
| Health Care | ABT | Abbott Laboratories |
| Industrials | GE | GE Aerospace |
| Industrials | RTX | RTX Corporation |
| Industrials | UBER | Uber Techonologies Inc. |
| Industrials | CAT | Caterpillar Inc. |
| Industrials | HON | Honeywell Internation Inc. |
| Materials | LIN | Linde plc |
| Materials | SHW | Sherwin-Williams Company |
| Materials | NEM | Newmont Corporation |
| Materials | ECL | Ecolab Inc. |
| Materials | APD | Air Products and Chemicals Inc. |
| Real Estate | AMT | American Tower Corporation |

| | | |
|---|---|---|
| Real Estate | PLD | Prologis Inc. |
| Real Estate | WELL | Welltower Inc. |
| Real Estate | EQIX | Equinix Inc. |
| Real Estate | DLR | Digital Realty Trust Inc. |
| Information Technology | AAPL | Apple Inc. |
| Information Technology | MSFT | Microsoft Corporation |
| Information Technology | NVDA | NVIDIA Corporation |
| Information Technology | AVGO | Broadcom Inc. |
| Information Technology | CRM | Salesforce Inc. |
| Utilities | NEE | NextEra Energy Inc. |
| Utilities | SO | Southern Company |
| Utilities | DUK | Duke Energy Corporation |
| Utilities | CEG | Constellation Energy Corporation |
| Utilities | AEP | American Electric Power Company Inc. |

Table A.2: Overview of which stocks were selected for the empirical study. The five highest weighted stocks from each SPDR sector index of the S&P 500 was selected.

Here is a plot illustrating the distribution of market capital for each of the 55 selected stocks. A list with the market capital for all 55 companies is retrievable from

- `MarketCapsSP500_Sectors_2024-04-01.csv`.

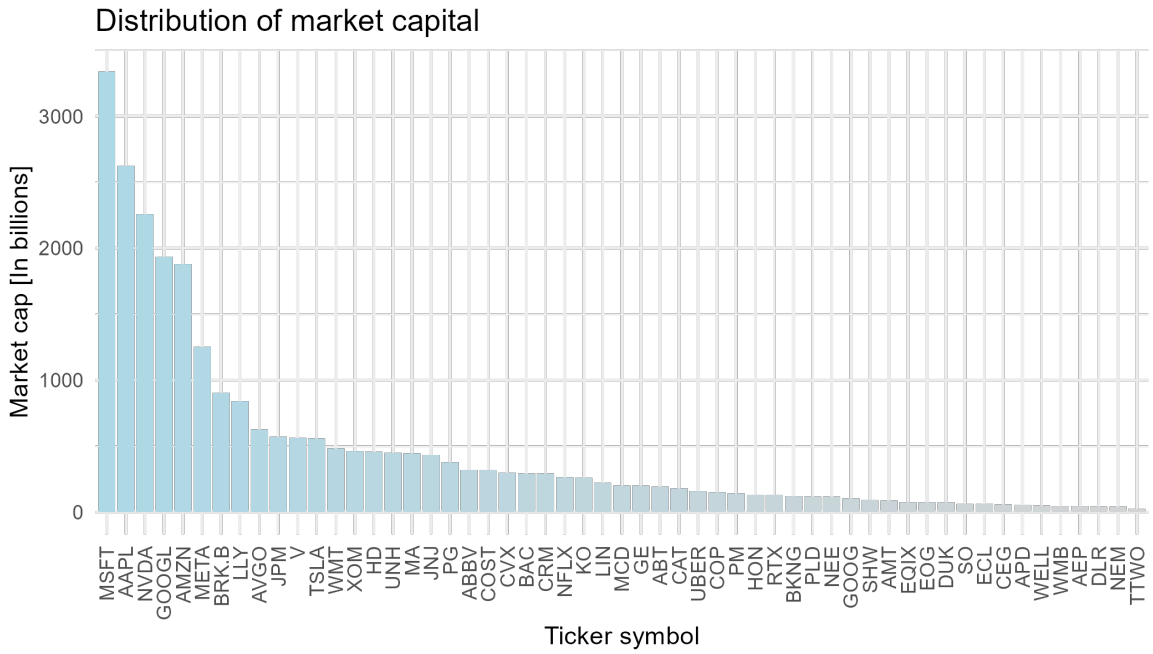



Figure A.1: Distribution of market capital for each selected sector stock on the 1st of April 2024.

Here is a list with short descriptions for each of the GICS sectors.

- **Communication Services:** Consists of companies from industries, such as: diversified telecommunication services, wireless telecommunication services, media, entertainment, and interactive media and services.
- **Consumer Staples:** Includes companies dealing with consumer staples distribution and retail, household products, food products, beverages, tobacco, and personal care products.
- **Consumer Discretionary:** Comprises of companies within the industries of specialty retail, broadline retail, hotels, restaurants and leisure, textiles, apparel and luxury goods, household durables, automobiles, automobile components, distributors, leisure products, and diversified consumer services.
- **Energy:** Includes companies from the following industries: oil, gas and consumable fuels, and energy equipment and services.
- **Financials:** Consists of firms from the industries: financial services, insurance, banks, capital markets, mortgage real estate investment trusts, and consumer finance.
- **Health Care:** Includes companies from the industries: pharmaceuticals, health care equipment and supplies, health care providers and services, biotechnology, life sciences tools and services, and health care technology.
- **Industrials:** Includes companies within: aerospace and defense, industrial congolmerates, marine transportation, transportation infrastructure, machinery, ground transportation, air freight and logistics, commercial services and supplies, professional services, electrical equipment, construction and engineering, trading companies and distributors, passenger airlines, and building products.
- **Materials:** Comprises of companies from the following industries: chemicals, metals and mining, paper and forest products, containers and packaging, and construction materials.
- **Real Estate:** Consists of companies dealing with the following industries: real estate management and development and REITs, excluding mortgage REITs.
- **Information Technology:** Includes companies from the following industries: technology hardware, storage, and peripherals, software, communications equipment, semiconductors and semiconductor equipment, IT services, and electronic equipment, instruments and components.
- **Utilities:** Comprises of companies dealing with electric utilities, water utilities, multi-utilities, independent power and renewable electricity producers, and gas utilities.

## A.4 Derivation of a solution by the first order conditions

In this section, we will derive (2.3) based on the first order conditions. Suppose we have an ETF that tracks the index that we want to track and $d$ component stocks to choose from. Let $\alpha \in \mathbb{R}^d$ denote the weights of investment into each component stock in the portfolio used for tracking. Consider the portfolio consisting of 1 share of the ETF and short positions of the component stocks, i.e.

$$w = \begin{bmatrix} 1 \\ -\alpha \end{bmatrix}.$$

Letting $R = [r_b \ r^\top]^\top$ denote the return vector, where $r_b$ is the return of the ETF/benchmark index, and $r \in \mathbb{R}^d$ is the vector of returns for all the component stocks in the ETF, the returns of $w$ are given by

$$w^\top R = \begin{bmatrix} 1 & -\alpha^\top \end{bmatrix} \begin{bmatrix} r_b \\ r \end{bmatrix} = r_b - \alpha^\top r = r_b - r_p,$$

where $r_p$ is the return of the tracking portfolio. Since the returns of $w$ corresponds to the excess returns between the benchmark and the tracking portfolio, the tracking error of the tracking portfolio is the portfolio standard deviation of $w$, hence

$$TE(\alpha; \Omega) = \sqrt{V(w)} = \sqrt{\begin{bmatrix} 1 & -\alpha^\top \end{bmatrix} \Omega \begin{bmatrix} 1 & -\alpha^\top \end{bmatrix}^\top},$$

where $V(w)$ is the portfolio variance of $w$. The solution $\alpha$ is the same regardless if we are using $\sqrt{V(w)}$ or $V(w)$ as the objective function for minimizing tracking error, hence, we will use the latter. By the first order condition, we have for $i = 1, 2, \ldots, d$ that

$$\frac{\partial}{\partial \alpha_i} V(w) = 0.$$

We will reexpress the left hand side:

$$\begin{aligned}
\frac{\partial}{\partial \alpha_i} V(w) &= \frac{\partial}{\partial \alpha_i} \begin{bmatrix} 1 & -\alpha^\top \end{bmatrix} \Omega \begin{bmatrix} 1 & -\alpha^\top \end{bmatrix}^\top \\
&= \frac{\partial}{\partial \alpha_i} \begin{bmatrix} 1 & -\alpha^\top \end{bmatrix} \begin{bmatrix} \omega_E & \omega_{EK}^\top \\ \omega_{EK} & \Sigma \end{bmatrix} \begin{bmatrix} 1 & -\alpha^\top \end{bmatrix}^\top \\
&= \frac{\partial}{\partial \alpha_i} \left( \omega_E - \omega_{EK}^\top \alpha - \alpha^\top \omega_{EK} + \alpha^\top \Sigma \alpha \right) \\
&= \frac{\partial}{\partial \alpha_i} \alpha^\top \Sigma \alpha - 2 \frac{\partial}{\partial \alpha_i} \omega_{EK}^\top \alpha \\
&= \frac{\partial}{\partial \alpha_i} \sum_{j=1}^{d} \sum_{k=1}^{d} \Sigma_{j,k} \alpha_j \alpha_k - 2 \frac{\partial}{\partial \alpha_i} \sum_{j=1}^{d} \omega_{EK,j} \alpha_j
\end{aligned}$$

$$= 2\sum_{j=1}^{d} \Sigma_{i,j}\alpha_j - 2\omega_{EK,i}.$$

Thus for $i = 1, 2, \ldots, d$ we have that

$$2\sum_{j=1}^{d} \Sigma_{i,j}\alpha_j - 2\omega_{EK,i} = 0 \Leftrightarrow \sum_{j=1}^{d} \Sigma_{i,j}\alpha_j = \omega_{EK,i},$$

which expressed in vector form yields that

$$\Sigma\alpha = \omega_{EK} \Leftrightarrow \alpha = \Sigma^{-1}\omega_{EK},$$

which was to be demonstrated.

# Bibliography

[1] Niclas Andréasson, Anton Evgrafov, Michael Patriksson, Emil Gustavsson, Zuzana Nedelková, Kin Cheong Sou, and Magnus Önnheim. *An Introduction to Continuous Optimization Foundations & Fundamental Algorithms*. Dover Publications Inc., 2020. ISBN 0-486-80287-6.

[2] Bryan Armour, Ryan Jackson, Eugene Gorbatikov, and Hyunmin Kim. Morningstar's us active/passive barometer year-end 2024. *Morningstar*, 2024. URL `https://www.morningstar.com/lp/active-passive-barometer`.

[3] Ole E. Barndorff-Nielsen, Peter Reinhard Hansen, Asger Lunde, and Neil Shephard. Realized kernels in practice: trades and quotes. *The Econometrics Journal*, 2009. URL `https://www.jstor.org/stable/23116045`.

[4] Ole E. Barndorff-Nielsen, Peter Reinhard Hansen, Asger Lunde, and Neil Shephard. Multivariate realised kernels: Consistent positive semi-definite estimators of the covariation of quity prices with noise and non-synchronous trading. *Journal of Econometrics*, 2011. URL `https://doi.org/10.1016/j.jeconom.2010.07.009`.

[5] Kris Boudt, Sébastien Laurent, Asger Lunde, Rogier Quaedvlieg, and Orimar Sauri. Positive semidefinite integrated covariance estimation, factorizations and asynchronicity. *Journal of Econometrics*, 2016. URL `https://doi.org/10.1016/j.jeconom.2016.09.016`.

[6] Kris Boudt, Kirill Dragun, Orimar Sauri, and Steven Vanduffel. Etf basket-adjusted covariance estimation. *Journal of Econometrics*, 2023. URL `https://doi.org/10.1016/j.jeconom.2022.10.002`.

[7] James Chen. *Passive Investing: Definition, Pros and Cons, vs. Activce Investing*, 2025. URL `https://www.investopedia.com/terms/p/passiveinvesting.asp`.

[8] Kim Christensen, Silja Kinnebrock, and Mark Podolskij. Pre-averaging estimators of the ex-post covariance matrix in noisy diffusion models with non-synchronous data. *Journal of Econometrics*, 2010. URL `https://doi.org/10.1016/j.jeconom.2010.05.001`.

[9] Gerard Cornuejols and Reha Tütüncü. *Optimization Methods in Finance*. Cambridge University Press, 2007. ISBN 978-0-521-86170-0.

[10] Gerard Cornuejols, Marshall L. Fisher, and George L. Nemhauser. Location of bank accounts to optimize float: An analytic study of exact and approximate algorithms. *Management Science*, 1977. URL `https://www.jstor.org/stable/2630709`.

[11] Francisco Saldanha da Gama and Shuming Wang. *Facility Location under Uncertainty: Models, Algorithms and Applications*. Springer, 2024. ISBN 978-3-031-55927-3.

[12] Lucas Downey. *Efficient Market Hypothesis (EMH): Definition and Critique*, 2024. URL https://www.investopedia.com/terms/e/efficientmarkethypothesis.asp.

[13] Thomas W. Epps. Comovements in stock prices in the very short run. *Journal of the American Statistical Association*, Vol. 74:pp. 291–298, 1979. URL https://doi.org/10.2307/2286325.

[14] Jianqing Fan, Yingying Li, and Ke Yu. Vast volatility matrix estimation using high-frequency data for portfolio selection. *Journal of the American Statistical Association*, 2012. URL https://doi.org/10.1080/01621459.2012.656041.

[15] Björn Fastrich, Sandra Paterlini, and Peter Winker. Cardinality versus q-norm constraints for index tracking. *Quantitative Finance*, 2014. URL https://doi.org/10.1080/14697688.2012.691986.

[16] Takaki Hayashi and Nakahiro Yoshida. On covariance estimation of non-synchronously observed diffusion processes. *Bernoulli*, 2005. URL https://www.jstor.org/stable/3318933.

[17] Adam Hayes. *Index Rebalancing: What Every Investor Should Know*, 2025. URL https://www.investopedia.com/index-rebalancing-7972596.

[18] Jean Jacod, Yingying Li, Per A. Mykland, Mark Podolskij, and Mathias Vetter. Microstructure noise in the continuous case: The pre-averagin approach. *Stochastic Processes and their Applications*, 2009. URL https://doi.org/10.1016/j.spa.2008.11.004.

[19] Mark Podolskij and Mathias Vetter. Estimation of volatility functionals in the simultaneous presence of microstructure noise and jumps. *Bernoulli*, 2009. URL https://www.jstor.org/stable/20680171.

[20] Bin Zhou. High-frequency data and volatility in foreign-exchange rates. *Journal of Business & Economic Statistics*, 1996. URL https://doi.org/10.2307/1392098.