LUND UNIVERSITY
School of Economics and Management

An alternative approach for solving the problem of close to singular covariance matrices in modern portfolio theory

Martin Claeson

STAM01: Master's thesis in Statistics (15 ECTS)

Lund university
School of Economics and Managment
Supervisor: Krzysztof Podgórski
January 11, 2018

Abstract

In this thesis the effects of utilizing the sample covariance matrix in the estimation of the global minimum variance (GMV) portfolio are presented. When the number of assets, N, are close to the number of observations, T, the sample covariance matrix approaches singularity, leading to a lot of uncertainties in form of estimation error. Due to that the computations of the sample GMV portfolio are linear transformations of the sample covariance matrix inverse, the error is transferred to the portfolio weights. As a consequence, the portfolio performance out-of-sample is misleading and inadequate.

To solve this shortcoming an alternative approach is presented. By grouping similar stocks together, utilizing sector indices, the dimension of the sample covariance matrix is decreased and consequently the estimation error. As a result, the sample sector index portfolio displays a more stable structure than the sample GMV portfolio counterpart as $N\,/\,T \to 1$. This leads to more accurate parameter estimates and less volatile portfolio performances out-of-sample.

Keywords: global minimum variance portfolio, singular covariance matrix, sector index portfolio

1 Introduction

The global minimum variance (GMV) portfolio is an important part of modern portfolio theory, introduced by Markowitz (1952, 1959). Due to basic computations, completely based on the covariance matrix, it is a well-established approach for portfolio risk minimization purposes. If the covariance matrix is known one cannot find a portfolio with lower variance. However, the covariance matrix is unknown in practice and needs to be estimated. This can either be done by subjective judgement or by a statistical approach, using the sample covariance matrix based on historical data. In this thesis, the latter of these methods is utilized. Compared to future mean returns, the estimation of variance and covariance parameters is less problematic, see Merton (1980), Chopra and Ziemba (1993). The reason for this is that the volatility of financial data tends to be non-random and it exhibits a positive serial correlation, see Engle (1982) and Bollerslev (1986), i.e. the future variance and covariance parameters will approximately show a similar pattern as it has historically. Consequently, the usage of historical data seems to be an appropriate procedure for this task.

However, some major flaws of the sample covariance matrix appear as the number of assets increase and approach the number of observations. Besides that more variance and covariance parameters need to be estimated, the sample covariance matrix approaches singularity. Both problems lead to uncertainties in form of estimation error. Due to that the GMV portfolio weights are completely based on the sample covariance matrix inverse, the obtained weights are affected, becoming unstable and sensitive, see Best and Grauer (1991). As a consequence, the out-of-sample portfolio variance will be higher than the in-sample counterpart, see Basak (2005). Hence, the investor will be gravely disappointed in the portfolios future performance.

Due to these problems a lot of research have been made in this field. Frankfurter et al. (1971) concludes that the GMV portfolio might not outperform a randomly selected portfolio. This is something that DeMiguel et al. (2009) acknowledge as well, after comparing a variety of risk estimating strategies to the naive approach of an equally weighted portfolio. Another important finding, see Jobson and Korkie (1980), is that the most influential estimates of the covariance matrix tend to be extreme due to the high amount of error rather than the reality. As a consequence, one will obtain a misleading result due to unrealistic and unstable portfolio weights. Bengtsson and Holst (2002) endorse this with the addition that the sample covariance matrix requires estimation of too many parameters and that it overfits the sample data due to that it imposes to little structure. To solve the above declared problems, regarding singularity and high dimensional covariance matrices, some different shrinkage approaches have been suggested, see Ledoit and Wolf (2003, 2004) and Bodnar et al. (2014) among others.

The author of this paper introduced a different approach in an attempt to solve some of these problems, see Claeson (2016). The presented method attempts to decrease the amount of estimation error contained in the weights by utilizing sector indices. However, due to that the weights obtained by this approach imposes to little structure, it needs to be developed further, which is the main purpose of this thesis. Rather than completely compute the weights based on sector indices, the structure within each index can be taken into consideration as well. Hence, the weights are computed by the usage of the GMV theory between the sector indices as well as between the stocks within each sector.

There are a number of advantages of this approach that are presented and further

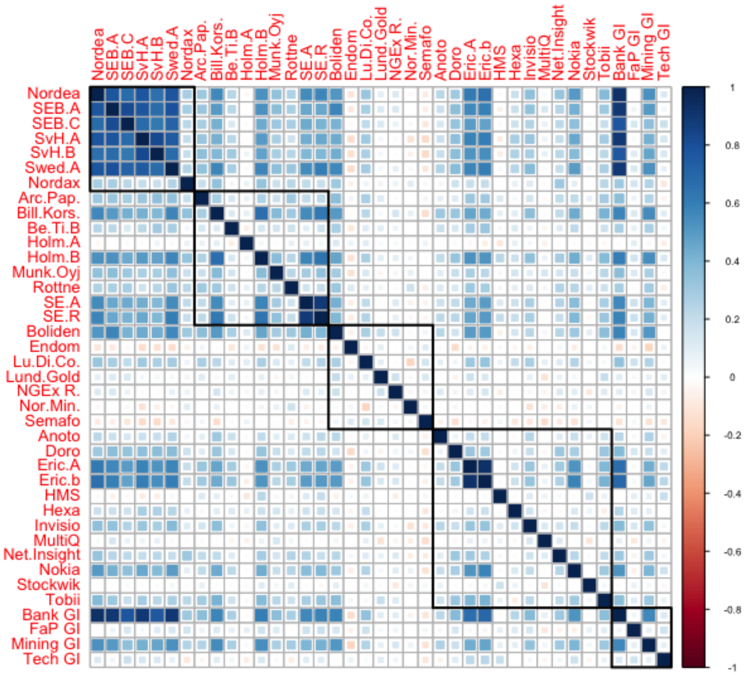


Figure 1: Correlation plot representing the structure of the empirical data. The first 35 assets are stocks and the last four are sector indices, all 39 covering 182 observations between 18 of June 2015 to 7 of April 2016. The first four black squares represent the clusters of stocks within each sector index. These four sector indices are presented in the fifth black square.

analyzed in this paper. First, by utilizing sector indices, the number of estimated parameters decrease. Secondly, due to that the data set are divided into smaller subsamples, the sample covariance matrices will get further away from singularity. Third, when comparing the sector indices with self-composed average weighted sector indices as underlying structure in the proposed model, we conclude that the former outperform the latter. Hence, we suspect that the sector indices contain some structural sector information not covered by simply averaging the stocks. However, the structure of the indices is unknown, due to a special and quite complex calculation methodology, which leads to difficulties in fully evaluating the indices construction. To illustrate, a correlation plot based on the empirical data used in this thesis are presented in Figure 1. On the diagonal axis of the plot, five black squares divide the assets into clusters. The first four squares represent the stocks within each sector index and the last square contains the sector indices in question. Before we draw any deeper conclusions from this plot, keep in mind that the there are some drawbacks with this data set, see Section four, which leads to this quite unrealistic correlation structure. When evaluating the correlation plot, it is clear that the structure between the clusters of stocks, the first 35 assets in Figure 1, does not conform with the structure between the sector indices, within the last black square. However, when looking closer into the correlation between the sector indices and the underlying stocks one acknowledge that these time series not necessarily imitate each other. After evaluating the stocks of largest market capital, it seems that these have a higher influence on the corresponding sector index compared to the smaller counterparts. Due to that the purpose of the indices is to give the investor knowledge of the economic climate of a whole branch, it seems reasonable to weight the stocks depending on some sort of market capital. To summarize, by utilizing these sector indices instead of computing our own indices, one might gain some information concerning the sector that are not covered when we look into the average weighted indices. Finally, the model might be able to utilize the properties of the law of total covariance. Hence, by grouping similar stocks together while aiming for low correlation between the sectors, the portfolio variance might be reduced.

It is important to state that one cannot find a less risky portfolio than the global minimum variance portfolio if the covariance matrix is known. However, this is not the aim of the suggested approach. The idea is to present a model that contain less amount of error in the weights when the GMV portfolio is estimated with the sample covariance matrix. Such a model is especially useful when a limited amount of data is available and the sample covariance matrix is close to singularity. As a consequence of more stable weights, the in-sample as well as out-of-sample portfolio performance will be more accurate estimates of the reality.

The thesis is structured as follows. In Section two the theoretical and mathematical framework of the sample GMV portfolio is presented. This is followed by the methodology of the proposed sample sector index portfolio in Section three. Furthermore, the problems of the sample GMV portfolio are more in deep discussed with propositions of how the sample sector index portfolio might be able to solve these shortcomings. In the fourth Section, four Monte-Carlo studies are performed to get a deeper knowledge of the uncertainties associated with respectively approach and to fully understand the both portfolios strengths in a variety of situations. Finally, in Section five conclusions and final remarks of the findings are made.

2 The global minimum variance portfolio

Let $\mathbf{x}_t$ be a N dimensional column vector of daily logarithmic returns, covering N assets at time $t = 1, ..., T$. The expected values and the covariance matrix of $\mathbf{x}_t$ are defined as,

$$\underset{(N\times 1)}{E[\mathbf{x}_t]} = \boldsymbol{\mu} \text{ and } \underset{(N\times N)}{\boldsymbol{\Sigma}_{\mathbf{x}}} = E\Big[(\mathbf{x}_t - \boldsymbol{\mu})(\mathbf{x}_t - \boldsymbol{\mu})'\Big]. \tag{1}$$

Given a selection of assets, the global minimum variance portfolio is defined as the portfolio with the lowest variance possible. By utilizing the covariance matrix of the selected holdings, the GMV portfolio is obtained by,

$$\underset{(N\times 1)}{\mathbf{w}_{GMV}} = \text{argmin}\{\mathbf{w}'\boldsymbol{\Sigma}_{\mathbf{x}}\mathbf{w}; \mathbf{w}'\mathbf{1}_N = 1\}, \tag{2}$$

where $\mathbf{w} = [\mathrm{w}_1, \mathrm{w}_2, ..., \mathrm{w}_N]'$ is a column vector of portfolio weights and $\mathbf{1}_N$ is a N dimensional column vector of only ones. Due to that the only constraint is that the weights summarize to 1, short sales are allowed. Given that $\boldsymbol{\Sigma}_{\mathbf{x}}$ is positive definite the following holds,

$$\underset{(N\times 1)}{\mathbf{w}_{GMV}} = \frac{\mathbf{\Sigma}_{\mathbf{x}}^{-1}\mathbf{1}_N}{\mathbf{1}_N'\mathbf{\Sigma}_{\mathbf{x}}^{-1}\mathbf{1}_N}. \tag{3}$$

$\mathbf{\Sigma}_{\mathbf{x}}^{-1}$ is defined as the inverse of the covariance matrix $\mathbf{\Sigma}_{\mathbf{x}}$. The GMV portfolio mean, variance and standard deviation are given by,

$$\mu_{GMV} = \mathbf{w}_{GMV}'\boldsymbol{\mu},\ \sigma_{GMV}^2 = \mathbf{w}_{GMV}'\mathbf{\Sigma}_{\mathbf{x}}\mathbf{w}_{GMV} \quad \text{and} \quad \sigma_{GMV} = \sqrt{\sigma_{GMV}^2}. \tag{4}$$

However, as mentioned in the introduction, $\boldsymbol{\mu}$ and $\mathbf{\Sigma}_{\mathbf{x}}$ are in reality unknown and need to be estimated. The sample mean vector and sample covariance matrix of $\mathbf{x}_t$ are defined as,

$$\underset{(N\times 1)}{\bar{\mathbf{x}}} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t \quad \text{and} \quad \underset{(N\times N)}{\widehat{\mathbf{\Sigma}}_{\mathbf{x}}} = \frac{1}{T-1}\sum_{t=1}^{T}(\mathbf{x}_t - \bar{\mathbf{x}})(\mathbf{x}_t - \bar{\mathbf{x}})'. \tag{5}$$

Wishart (1928) concludes that $\widehat{\mathbf{\Sigma}}_{\mathbf{x}}$ is Wishart distributed with $T-1$ degrees of freedom,

$$(T-1)\widehat{\mathbf{\Sigma}}_{\mathbf{x}} \quad \sim \quad \mathcal{W}_N(T-1, \mathbf{\Sigma}_{\mathbf{x}}). \tag{6}$$

$\widehat{\mathbf{\Sigma}}_{\mathbf{x}}$ has the property of being an unbiased estimator of $\mathbf{\Sigma}_{\mathbf{x}}$ and due to the law of large numbers, for large T relative to N, $\widehat{\mathbf{\Sigma}}_{\mathbf{x}} \xrightarrow{P} \mathbf{\Sigma}_{\mathbf{x}}$, see Johnson and Wichern (2007). The sample GMV portfolio is obtained by replacing $\mathbf{\Sigma}_{\mathbf{x}}^{-1}$ with $\widehat{\mathbf{\Sigma}}_{\mathbf{x}}^{-1}$ in (3),

$$\underset{(N\times 1)}{\widehat{\mathbf{w}}_{\mathbf{x}}} = \frac{\widehat{\mathbf{\Sigma}}_{\mathbf{x}}^{-1}\mathbf{1}_N}{\mathbf{1}_N'\widehat{\mathbf{\Sigma}}_{\mathbf{x}}^{-1}\mathbf{1}_N}, \tag{7}$$

where $\widehat{\mathbf{\Sigma}}_{\mathbf{x}}^{-1}$ is the inverse of the sample covariance matrix and is distributed as,

$$(T-1)\widehat{\mathbf{\Sigma}}_{\mathbf{x}}^{-1} \quad \sim \quad \mathcal{W}_N^{-1}(T-1, \mathbf{\Sigma}_{\mathbf{x}}^{-1}). \tag{8}$$

Here $W_N^{-1}(T-1, \mathbf{\Sigma}_{\mathbf{x}}^{-1})$ denotes the inverse Wishart distribution with $T-1$ degrees of freedom. The estimated in-sample GMV portfolio mean, variance and standard deviation are obtained by replacing the weights in (4) with the ones obtained in (7),

$$\widehat{\mu}_{\mathbf{x}} = \widehat{\mathbf{w}}_{\mathbf{x}}'\bar{\boldsymbol{x}}, \quad \widehat{\sigma}_{\mathbf{x}}^2 = \widehat{\mathbf{w}}_{\mathbf{x}}'\widehat{\mathbf{\Sigma}}_{\mathbf{x}}\widehat{\mathbf{w}}_{\mathbf{x}} \quad \text{and} \quad \widehat{\sigma}_{\mathbf{x}} = \sqrt{\widehat{\sigma}_{\mathbf{x}}^2}. \tag{9}$$

Due to that $\bar{\boldsymbol{x}}$, $\widehat{\mathbf{\Sigma}}_{\mathbf{x}}$ and $\widehat{\mathbf{\Sigma}}_{\mathbf{x}}^{-1}$ are maximum likelihood estimates of $\boldsymbol{\mu}$, $\mathbf{\Sigma}_{\mathbf{x}}$ and $\mathbf{\Sigma}_{\mathbf{x}}^{-1}$, $\widehat{\mathbf{w}}_{\mathbf{x}}$ is a maximum likelihood estimate of $\mathbf{w}_{GMV}$. Hence, in-sample, $\widehat{\mathbf{w}}_{\mathbf{x}}$ is the most efficient estimator of $\mathbf{w}_{GMV}$ asymptotically as $T \to \infty$. However, since $\widehat{\mathbf{w}}_{\mathbf{x}}$, $\widehat{\mathbf{\Sigma}}_{\mathbf{x}}$ and $\widehat{\mathbf{\Sigma}}_{\mathbf{x}}^{-1}$ are random parameters, $\widehat{\sigma}_{\mathbf{x}}$ is random as well. Consequently, despite these asymptotic properties, $\widehat{\sigma}_{\mathbf{x}}$ might differ considerably from σ_{GMV}, even for large T relative to N, see Kan and Smith (2008).

However, investors are primarily interested in the out-of-sample performance of a portfolio $\widehat{\mathbf{w}}_{\mathbf{x}}$ rather than the in-sample counterpart. More precisely, if one were to invest according to the sample GMV portfolio based on historical data, how would this portfolio perform in the unknown future? As Kan and Zhou (2007) concludes, in these situations,

$\widehat{\mathbf{w}}_{\mathbf{x}}$ is no longer optimal in terms of maximizing the performance. The estimated mean, variance and standard deviation of holding such a portfolio out-of-sample is defined as,

$$\tilde{\mu}_{\mathbf{x}} = \widehat{\mathbf{w}}_{\mathbf{x}}'\boldsymbol{\mu}, \quad \tilde{\sigma}_{\mathbf{x}}^2 = \widehat{\mathbf{w}}_{\mathbf{x}}'\boldsymbol{\Sigma}_{\mathbf{x}}\widehat{\mathbf{w}}_{\mathbf{x}} \quad \text{and} \quad \tilde{\sigma}_{\mathbf{x}} = \sqrt{\tilde{\sigma}_{\mathbf{x}}^2}. \tag{10}$$

Besides that $\widehat{\mathbf{w}}_{\mathbf{x}}$ is random, the unknown future mean returns $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}_{\mathbf{x}}$ are random, hence, the out-of-sample portfolio parameters $\tilde{\mu}_{\mathbf{x}}$, $\tilde{\sigma}_{\mathbf{x}}^2$ and $\tilde{\sigma}_{\mathbf{x}}$ are also random. For an investor, these estimates are of higher importance than the in-sample counterparts, computed in (9). As Kan and Smith (2008) remarks, the estimated portfolio variance $\widehat{\sigma}_{\mathbf{x}}$ is on average much smaller than the out-of-sample portfolio variance, $\tilde{\sigma}_{\mathbf{x}}$; especially as the ratio of N/T approaches one. If one relies on the former measurement as a good indicator of the latter, one will be grossly disappointed. The reason for this is that the sample GMV portfolio tends to weight heavily on assets with low variance and covariance. These might be low by chance, especially when the number of observations are close to the number of holdings. When N / T is close to one, $\widehat{\Sigma}_{\mathbf{x}}$ consists of a major amount of error due to that the distribution of the sample covariance matrix is more heavy-tailed, which leads to much uncertainties in $\widehat{\mathbf{w}}_{\mathbf{x}}$ and consequently in $\tilde{\sigma}_{\mathbf{x}}$, see Kan and Zhou (2007). Due to these out-of-sample uncertainties, a less optimistic and more efficient model needs to be taken into account, especially as N / T approaches one.

3 Model based on sector indices

Let $\mathbf{y}_t$ be a M dimensional column vector of daily logarithmic *quasi* returns, covering M sector indices at time $t = 1, ..., T$. We call them *quasi* returns since one cannot invest in these indices per se, but in the underlying stocks. In similarity to the previous section, the expected values and the covariance matrix are unknown and need to be estimated with the sample counterparts, these are defined as,

$$\underset{(M\times 1)}{\bar{\mathbf{y}}} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{y}_t \quad \text{and} \quad \underset{(M\times M)}{\widehat{\boldsymbol{\Sigma}}_{\mathbf{y}}} = \frac{1}{T-1}\sum_{t=1}^{T}(\mathbf{y}_t - \bar{\mathbf{y}})(\mathbf{y}_t - \bar{\mathbf{y}})'. \tag{11}$$

The sample GMV portfolio of sector indices is obtained by replacing $\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}}^{-1}$ with $\widehat{\boldsymbol{\Sigma}}_{\mathbf{y}}^{-1}$ in (6),

$$\underset{(M\times 1)}{\widehat{\mathbf{w}}_{\mathbf{y}}} = \frac{\widehat{\boldsymbol{\Sigma}}_{\mathbf{y}}^{-1}\mathbf{1}_M}{\mathbf{1}_M'\widehat{\boldsymbol{\Sigma}}_{\mathbf{y}}^{-1}\mathbf{1}_M}. \tag{12}$$

Hence, $\widehat{\mathbf{w}}_{\mathbf{y}}$ is a column vector of portfolio weights between the sector indices. The vector of stocks $\mathbf{x}_t$, presented in Section 2, is separated into M smaller vectors, containing the stocks within each sector,

$$\underset{(N\times 1)}{\mathbf{x}_t} = \left[\begin{array}{c} \underset{(n_1\times 1)}{\mathbf{x}_t{}^{(1)}} \\ \hline \underset{(n_2\times 1)}{\mathbf{x}_t{}^{(2)}} \\ \hline \vdots \\ \hline \underset{(n_M\times 1)}{\mathbf{x}_t{}^{(M)}} \end{array}\right],$$

where n_j is the number of stocks within the j:th sector, $j = \{1, 2, .., M\}$ and $N = \sum_{j=1}^{M} n_j$. For respectively sector, the sample mean and sample covariance matrix are defined as,

$$\underset{(n_j\times 1)}{\bar{\mathbf{x}}^{(j)}} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t^{(j)} \quad \text{and} \quad \underset{(n_j\times n_j)}{\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}^{(j)}}} = \frac{1}{T-1}\sum_{t=1}^{T}(\mathbf{x}_t^{(j)} - \bar{\mathbf{x}}^{(j)})(\mathbf{x}_t^{(j)} - \bar{\mathbf{x}}^{(j)})'. \tag{13}$$

The sample GMV portfolio within a sector is computed for each of these subsamples of stocks,

$$\underset{(n_j\times 1)}{\widehat{\mathbf{w}}_{\mathbf{x}^{(j)}}} = \frac{\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}^{(j)}}^{-1}\mathbf{1}_{n_j}}{\mathbf{1}_{n_j}'\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}^{(j)}}^{-1}\mathbf{1}_{n_j}}. \tag{14}$$

The definitive sample sector index portfolio, $\widehat{\mathbf{w}}_z$, is the product between the sample GMV portfolio of sector indices, $\widehat{\mathbf{w}}_{\mathbf{y}}$, and the sample GMV portfolio within a sector, $\widehat{\mathbf{w}}_{\mathbf{x}^{(j)}}$,

$$\underset{(N\times 1)}{\widehat{\mathbf{w}}_z} = \left[\begin{array}{c} \widehat{\mathbf{w}}_{\mathbf{y}_1}\cdot\widehat{\mathbf{w}}_{\mathbf{x}^{(1)}} \\ \hline \widehat{\mathbf{w}}_{\mathbf{y}_2}\cdot\widehat{\mathbf{w}}_{\mathbf{x}^{(2)}} \\ \hline \vdots \\ \hline \widehat{\mathbf{w}}_{\mathbf{y}_M}\cdot\widehat{\mathbf{w}}_{\mathbf{x}^{(M)}} \end{array}\right], \tag{15}$$

where $\widehat{\mathbf{w}}_{\mathbf{y}_k}$ denotes the k:th row of the vector $\widehat{\mathbf{w}}_{\mathbf{y}}$. The estimated in-sample sector index portfolio mean, variance and standard deviation are obtained by replacing $\widehat{\mathbf{w}}_{\mathbf{x}}$ with $\widehat{\mathbf{w}}_z$ in (9),

$$\widehat{\mu}_z = \widehat{\mathbf{w}}_z'\bar{\boldsymbol{x}}, \quad \widehat{\sigma}_z^2 = \widehat{\mathbf{w}}_z'\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}}\widehat{\mathbf{w}}_z \quad \text{and} \quad \widehat{\sigma}_z = \sqrt{\widehat{\sigma}_z^2}. \tag{16}$$

As mentioned in the previous section, the investors primarily interest is the out-of-sample performance. Hence the definitions in (10) are redefined as,

$$\tilde{\mu}_z = \widehat{\mathbf{w}}_z'\boldsymbol{\mu}, \quad \tilde{\sigma}_z^2 = \widehat{\mathbf{w}}_z'\boldsymbol{\Sigma}_{\mathbf{x}}\widehat{\mathbf{w}}_z \quad \text{and} \quad \tilde{\sigma}_z = \sqrt{\tilde{\sigma}_z^2}, \tag{17}$$

where $\widehat{\mathbf{w}}_z$, $\tilde{\mu}_z$ and $\tilde{\sigma}_z$ all are random. One has to keep in mind that due to the sector index model computations, the obtained portfolio $\widehat{\mathbf{w}}_z$ is not minimizing the variance of $\boldsymbol{\Sigma}_{\mathbf{x}}$. However, due to a variety of reasons, presented in *Remarks* 1-4 below, we suspect that the estimated in-sample sector index portfolio is more realistic and that the estimated out-of-sample sector index portfolio contain less error than the sample GMV portfolio counterpart, especially as $N / T \to 1$.

Remark 1. Due to the computations in (12) and (14), the GMV portfolio methodology is utilized twice in the computations of the weights $\widehat{\mathbf{w}}_z$. First between the sector indices and thereafter between the stocks within the sectors. As is partly noted in the previous section, the sample GMV portfolio shows some distinct properties in its weightings. First, less volatile stocks are weighted more heavily than riskier assets. Secondly, the increased risk associated with a pair of highly correlated assets is solved by selling one of these holdings short. Due to this, the sample GMV portfolio based on sector indices, $\widehat{\mathbf{w}}_{\mathbf{y}}$, weight more heavily into stable sectors and less into more volatile. In similarity, the sample GMV portfolio within a sector, $\widehat{\mathbf{w}}_{\mathbf{x}^{(j)}}$, weight more heavily into into stable stocks and less into more volatile assets. Due to the computations in (15), the weights of respectively stock in the sample sector index portfolio are relative to the risk of the sectors and are therefore less extreme. More precisely, compared to the sample GMV portfolio there are a smaller risk that the sample sector index portfolio weight heavily on a couple of assets. As a consequence, $\widehat{\mathbf{w}}_z$ has a similar structure as $\widehat{\mathbf{w}}_{\mathbf{x}}$, but more stable.

An observant reader might notice some potential shortcomings of this approach. If a couple of sector indices are strongly correlated the model will "sell" one of these short. Due to the computations in (15), this leads to short positions within the sector being long and vice versa in $\widehat{\mathbf{w}}_z$. This is one of multiple reasons to aim for low correlation between the sector indices. One other is more closely discussed in *Remark* 4. However, due to the concept and construction of sector indices, these might be less correlated than stocks. As explained further in Section 4, the data presented in Figure 1 is not representative of the reality and over a longer time horizon the strong correlation between the sectors should decline.

One solution of the problem concerning short sector index position is to use the constraint that short sales are not allowed. However, to keep the model as simple as possible and to get a better understanding of the models behavior this is not further discussed in this thesis.

Remark 2. One important property of the suggested sample sector index portfolio is that less parameters need to be estimated compared to the sample GMV portfolio. In the former, based on $\widehat{\boldsymbol{\Sigma}}_{\mathbf{y}}$ and $\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}^{(j)}}$, $M(M+1)/2+\sum_{j=1}^{M}\Big(n_j(n_j+1)/2\Big)$ variance and covariance parameters are estimated. In contrary, the latter, based on $\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}}$, contains $N(N+1) / 2$ estimates. Hence, the sector index model will always utilize sample covariance matrices of smaller dimension than the sample GMV portfolio. Consequently, as $N \to \infty$, one might be able to decrease the number of required estimates drastically by utilizing the sector index portfolio. Due to this, the estimation error in $\widehat{\mathbf{w}}_z$ might be decreased in comparison to $\widehat{\mathbf{w}}_{\mathbf{x}}$, which leads to more stable and realistic estimates in $\tilde{\sigma}_z$ than in $\tilde{\sigma}_{\mathbf{x}}$.

Remark 3. The problems of singularity are of major interest due to that the sample GMV portfolio weights are linear transformations of the sample covariance matrix inverse, see (6). This singularity problem arises for several reasons. First, it occurs when the rank of the covariance matrix is less than the dimension of data, i.e. there is linear dependence between at least two of the holdings. Secondly, it happens as $N/T \geq 1$, i.e. the number of assets is equal or larger than the number of observations. The risk of singularity increases when the number of assets in the portfolio arises, for both reasons. However, if one is able to reduce the dimension of $\widehat{\mathbf{\Sigma}}_{\mathbf{x}}$, utilized in (6), one will solve the latter of these problems to some degree. As mentioned in *Remark* 2, $\widehat{\mathbf{\Sigma}}_{\mathbf{y}}$ and $\widehat{\mathbf{\Sigma}}_{\mathbf{x}(j)}$ are always of smaller dimension than $\widehat{\mathbf{\Sigma}}_{\mathbf{x}}$ and therefore further away from singularity. Hence, an alternative approach in solving the problem of a singular covariance matrix is presented.

Besides singularity, close to singularity is a source of problems as well, which arise as $N\,/\,T$ approaches 1. Pappas et al. (2010) states that close to singular and numerically ill-conditioned covariance matrix estimates are a source of estimation error. The reason for this is that small values of the covariance matrix leads to large values in its inverse. Hence, small errors get amplified by the large inverse values which lead to unstable weights, see Michaud (1989).

In similarity to $\widehat{\mathbf{\Sigma}}_{\mathbf{x}}$, the distribution of $\widehat{\mathbf{\Sigma}}_{\mathbf{y}}$ and $\widehat{\mathbf{\Sigma}}_{\mathbf{x}(j)}$ are,

$$(T-1)\widehat{\mathbf{\Sigma}}_{\mathbf{y}} \quad \sim \quad \mathcal{W}_M(T-1, \mathbf{\Sigma}_{\mathbf{y}}) \quad \text{and} \quad (T-1)\widehat{\mathbf{\Sigma}}_{\mathbf{x}(j)} \quad \sim \quad \mathcal{W}_{n_j}(T-1, \mathbf{\Sigma}_{\mathbf{x}(j)}) \tag{18}$$

respectively. By utilizing sector indices, the ratios $M\,/\,T$ and $n_j\,/\,T$ are always smaller than N/T. Hence, the distributions of $\widehat{\mathbf{\Sigma}}_{\mathbf{y}}$ and $\widehat{\mathbf{\Sigma}}_{\mathbf{x}(j)}$ will be less heavy tailed than $\widehat{\mathbf{\Sigma}}_{\mathbf{x}}$. As a consequence, $\widehat{\mathbf{w}}_z$ is more stable than $\widehat{\mathbf{w}}_{\mathbf{x}}$, having a more realistic out-of-sample portfolio performance estimate $\tilde{\sigma}_z$ compared to $\tilde{\sigma}_{\mathbf{x}}$.

Remark 4. As mentioned in the introduction, one might gain some informational advantages by grouping stocks together and utilize sector indices. The law of total covariance are defined as follows,

$$\mathbf{\Sigma}_{\mathbf{x}} = E\big[\mathrm{Cov}[\mathbf{x}_t|\mathbf{y}_t]\big] + \mathrm{Cov}\big[E[\mathbf{x}_t|\mathbf{y}_t], E[\mathbf{x}_t'|\mathbf{y}_t]\big], \tag{19}$$

where $\mathbf{x}_t$ and $\mathbf{y}_t$ have the same definition as in the previous sections. The first term on the right side of the equality sign, $E\big[\mathrm{Cov}[\mathbf{x}_t|\mathbf{y}_t]\big]$, can be interpreted as the expected value of the covariance of the stocks $\mathbf{x}_t$ conditional on the indices $\mathbf{y}_t$, i.e. the expected value of the covariance matrices within the sectors. The last term in (19), $\mathrm{Cov}\big[E[\mathbf{x}_t|\mathbf{y}_t], E[\mathbf{x}_t'|\mathbf{y}_t]\big]$, can be interpreted as the covariance matrix of the expected values of the stocks conditional on the indices, i.e. the covariance between the sectors.

The above formula is useful in the context of this thesis for multiple reasons. First, the interpretation of the conditioning is that one better explains the covariance of $\mathbf{x}_t$ by adding more information in form of $\mathbf{y}_t$. Hence, the investor might experience some informational gain by utilizing sector indices and as a consequence better explain and reduce $\mathbf{\Sigma}_{\mathbf{x}}$. The second utility of (19) arises from the above interpretation of respectively term. Due to that sector indices are utilized and stocks within a branch tend to imitate each other, the first term is hard to decrease. More precisely, even though the market climate over a longer time horizon might reduce the correlation between the stocks of different sectors, the stocks within the sector might still show signs of strong correlation. The second term however, might be reduced by low correlation between the sectors. As a consequence, $\mathbf{\Sigma}_{\mathbf{x}}$

contains less covariation, which leads to more stable portfolio weights.

4 Empirical illustration

To substantiate the above remarks, four Monte Carlo-studies are performed on a variety of simulated data sets. The main reasons for using simulation rather than actual data for this procedure are primarily due to the small number of observations in the obtained empirical data set and furthermore to be able to control some of the underlying structure. Due to that the distribution of data is unknown, the assumptions of the multivariate normal (MVN) distribution are utilized. Even though financial data are far from Gaussian distributed, the solid properties help us to understand, explain and compare the different portfolios behavior more transparently.

The empirical sector index and stock data are collected from Nasdaq OMX Nordic and Handelsbanken. The data set contains of 182 daily observations between 18 of June 2015 to 7 and April 2016. Due to comparative reasons, some compromises had to be done, which lead to this short span of data. Beside that Nordax was IPOed at the 18 of June 2015 and a variety of stocks have been added to the indices after the 15 of April 2016, a dozen of daily observations are removed due to missing data. Keep in mind that this small sample of data might be inadequate in explaining the structure of reality. When examining Figure 1, a variety of the stocks and indices tend to be highly correlated, even between the sectors, due to the strong economic market effect during this time period. However, as an underlying structure in the multivariate normal distribution, the data set covers the intentions of this thesis.

Four sectors are selected, consisting of a total of 35 stocks. The sectors are presented in Table 1 below and the underlying stocks can be found in Table 2 in Appendix A, all 39 assets are presented in Figure 1. For a more realistic analyzes, gross indices and price adjusted stock price data are utilized. Meaning that dividends and splits are included to avoid extreme jumps in the time series.

Table 1: Sector indices

Index	Abbreviation	Symbol
OMX Stockholm Banks GI	Banks GI	SX8300GI
OMX Stockholm Forestry & Paper GI	FaP GI	SX1730GI
OMX Stockholm Mining GI	Mining GI	SX1770GI
OMX Stockholm Technology Hardware & Equipment GI	Tech GI	SX9570GI

Four MVN distributed data sets of different size and structure are simulated to give us prerequisites for further analyzes of the portfolios and to substantiate the remarks presented in Section 3. The first study represents a sort of optimal situation for the sample GMV portfolio, where the in-sample as well as out-of-sample estimates are based on large MVN distributed sample sizes with similar covariance structure. To appraise the statements presented in *Remarks* 1 and 4, the covariance structure between the sectors is put to zero in the second study. Hence, the covariance structure between the sector indices is removed as well as the covariance structure between the first four black squares in Figure 1. This newly constructed empirical data set is then used as underlying data for the simulation of a MVN distributed data set of same size as in the previous study. To illustrate, a correlation plot of the simulated data set is presented in Figure 2. The labels have

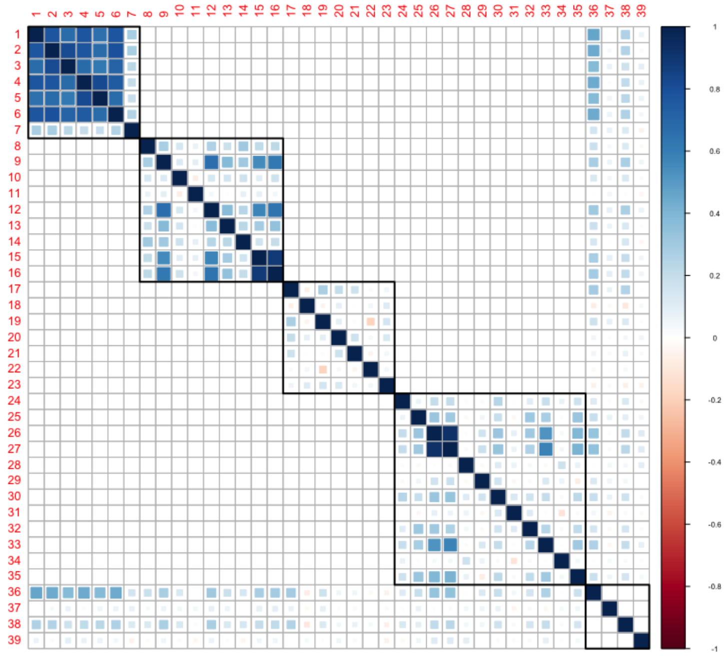


Figure 2: Correlation plot representing the structure of the empirical data with the correlation between the sectors removed. The first 35 time series are simulated stocks and the last four are simulated sector indices. The labels are removed due to that these time series are no longer the assets presented in Figure 1 but simulations of similar structure. The first four black squares represents the clusters of simulated stocks within each simulated sector index. These four sector indices are presented in the fifth black square.

been removed to avoid misunderstandings. To clarify, the underlying time series are now simulated assets and no longer empirical data. Due to that the sample covariance matrix is adjusted, some computational problems in R arise as the sample covariance matrix is no longer positive definite. This is solved by reducing the structure between the stocks and the sector indices by half. Hence the difference between the correlation in the last four assets when comparing Figure 1 and Figure 2. The third study have the original empirical data set, used in the first study, as underlying structure. However, this time a smaller amount of observations are simulated to illustrate the problems of singularity discussed in *Remark* 3. Finally, the fourth study is of same size as the third, but now based on the underlying empirical structure obtained in the second study. Hence, with similar structure as data presented in Figure 2. In this situation we are able to fully utilize the strength of the sector index portfolio in comparison to the sample GMV portfolio.

The procedure of the Monte-Carlo studies has the following structure,

1. Simulate a multivariate normally distributed data set of T observations. As input variables are a vector of the sample means and the sample covariance matrix covering the empirical data, stocks as well as indices.

2. The obtained data set are then divided into two subsamples, one containing simulated stocks and the other simulated sector indices. From respectively subsample the first 80 % of the observations are used as in-sample data and the remaining 20 % as out-of-sample data.

3. Obtain the bootstrap sample data set by drawing observations uniformly, with replacement, from the in-sample data set. Notice that the sampling order is the same for the simulated stocks as well as simulated sector indices. Furthermore, the out-of-sample data are kept constant.

4. Compute $\widehat{\mathbf{w}}_{\mathbf{x}}$ using formula (7) presented in Section 2 and $\widehat{\mathbf{w}}_z$ using formulas (12), (14) and (15) presented in Section 3 on the bootstrapped in-sample data.

5. Evaluate respectively models out-of-sample portfolio performance by computing $\tilde{\mu}_{\mathbf{x}}$, $\tilde{\sigma}_{\mathbf{x}}$, $\tilde{\mu}_z$ and $\tilde{\sigma}_z$, found in equations (10) and (17), on the out-of-sample data set.

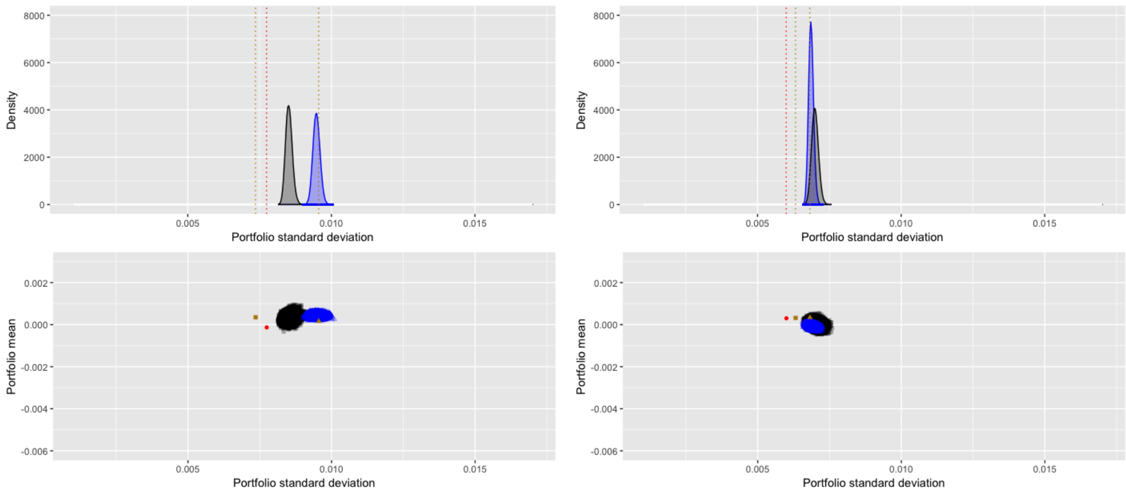
6. Redo step 3 to 5 until an arbitrary amount of bootstrap samples, B, are obtained.

7. For comparison reasons, the average in-sample bootstrap GMV portfolio as well as the average in-sample bootstrap sector index portfolio are computed. Furthermore, the out-of-sample GMV portfolio are computed to illustrate the estimated smallest portfolio variance possible out-of-sample.

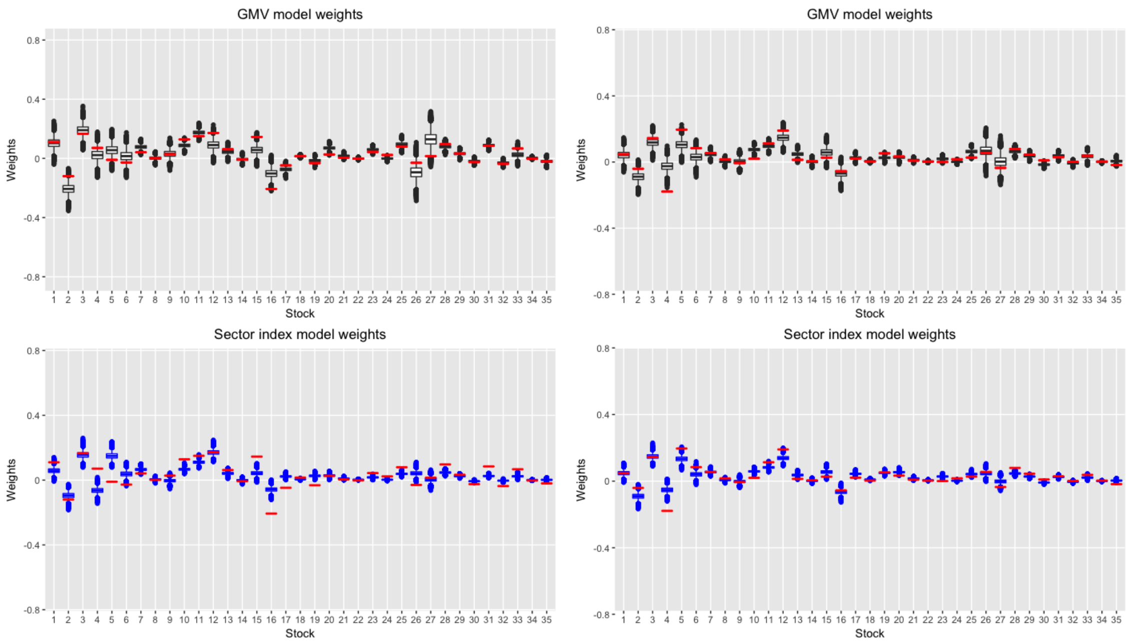
Study 1. As presented above, the first study represents a sort of optimal situation for the sample GMV portfolio, where $N \,/\, T$ is small. 1000 MVN distributed observations are simulated, the first 800 of these observations are used as an in-sample data set and the remaining 200 as an out-of-sample counterpart. B = 100 000 bootstrap samples are simulated and steps 4 and 5 are performed on each of these samples, the result is presented in Figure 3a. The black squares represent the out-of-sample estimates of the GMV portfolios with mean $\tilde{\mu}_{\mathbf{x}}$ and standard deviation $\tilde{\sigma}_{\mathbf{x}}$. The blue triangles represent the out-of-sample estimates of the sector index portfolios with mean $\tilde{\mu}_z$ and standard deviation $\tilde{\sigma}_z$. For comparison reasons the average in-sample bootstrap portfolios for respectively model are presented as brown geometric shapes. More precisely, these represents the average of the bootstrapped in-sample GMV portfolios means $\widehat{\mu}_{\mathbf{x}}$ and $\widehat{\mu}_z$ and the average of the bootstrapped in-sample GMV portfolios standard deviations $\widehat{\sigma}_{\mathbf{x}}$ and $\widehat{\sigma}_z$. The estimated out-of-sample GMV portfolio is presented in red, this is the portfolio with the estimated smallest standard deviation possible given the out-of-sample data. Respectively models bootstrap weights, $\widehat{\mathbf{w}}_{\mathbf{x}}$ and $\widehat{\mathbf{w}}_z$, are presented in boxplots in Figure 3c to illustrate the stability of the portfolios weight. For simplicity, the colors in the boxplots conform with the colors in Figure 3a. To be able to compare the different studies, the limits are kept constant for all plots in Figures 3 and 4.

Due to the large number of observations and the stability of the multivariate Gaussian distribution, the simulated assets contain much structure and few extreme observations. The potential problems discussed in previous sections are far from being an issue. The sample covariance matrix is Wishart distributed with a high amount of degrees of freedom and the sample covariance matrix inverse is therefore far from singularity. Furthermore, the higher amount of estimates in the sample GMV portfolio, mentioned in *Remark* 2, does not have much of an impact in this situation due to the stable underlying structure. In this quite optimal situation the out-of-sample GMV portfolio (black squares) outperform the out-of-sample sector index portfolio (blue triangles), presented in Figure 3a. This is not surprising however, under such circumstances one expects the sample GMV portfolio to be superior the sample sector index counterpart in reducing risk. Furthermore, as mentioned in Section 2, the in-sample GMV portfolio tends to be optimistic compared to its out-of-sample counterpart. This is clear when comparing the brown square with the cluster of black squares in Figure 3a. In contrast, the average of the sector index portfolios in-sample standard deviations (the brown triangle) is close to the average of the cluster of bootstrap samples (blue triangles). Hence, even though the estimated GMV portfolios contain less standard deviation out-of-sample, the sector index portfolios propose more realistic estimates. The estimated GMV portfolio out-of-sample, presented in red, is far from being covered by the bootstrap samples. This is not surprising either, due to that it is based on future unknown data. This portfolio is presented to give the reader an idea of where the portfolio with the estimated lowest standard deviation possible is situated in comparison to respectively models bootstrap portfolios.

In Figure 3c the weights of the bootstrap samples, $\widehat{\mathbf{w}}_{\mathbf{x}}$ and $\widehat{\mathbf{w}}_z$, are presented together with the out-of-sample GMV portfolio weights (in red). After comparison, it is clear that the estimated in-sample sector index portfolios have more stable weights compared to the estimated in-sample GMV portfolios. However, as evaluated when analyzing Figure 3a, this leads to higher portfolio standard deviation, i.e. the portfolios based on sector indices impose to little structure compared to its GMV counterpart. Even though we are not particularly interested in specific stocks the behavior between them helps us understand how the portfolios are weighted. As mentioned in Sections 2 and 3, the sample GMV portfolio tend to weigh heavily on stable assets. This is confirmed when evaluating the weights in Figure 3c. Large company stocks, which tend to be less volatile than smaller counterparts, have medians of more extreme weights and a higher variation between the size of the position. It is important to state that there are a handful of situations where the same company have multiple stocks, i.e. where the company issued stocks of different voting rights. Consequently, these stocks will often be highly correlated, see Figure 1, which obviously is problematic when concerning risk. However, as commented in *Remark* 1, the GMV theory solves this by selling one of the positions short. When analyzing Figure 3c it is clear that a large weighting in one of the companies' stocks often is hedged with a large short position in the companies' other stock. Hence, the movement between the time series will be smoothed out, which leads to less portfolio volatility. When comparing the boxplots to the sample GMV portfolio computed on the out-of-sample data, presented in red, it is clear that the GMV model is closer to these weights than the sector index model, which is confirmed when evaluating Figure 3a.



(a) Empirical data structure (b) Zero correlation between sectors



(c) Empirical data structure (d) Zero correlation between sectors

Figure 3: Density plots, scatter plots and box plots covering 100 000 bootstrap samples. The black squares represent the out-of-sample GMV portfolios mean $\tilde{\mu}_{\mathbf{x}}$ and standard deviation $\tilde{\sigma}_{\mathbf{x}}$ based on the in-sample GMV portfolio weights $\widehat{\mathbf{w}}_{\mathbf{x}}$. The blue triangles represent the out-of-sample sector index portfolios mean $\tilde{\mu}_z$ and standard deviation $\tilde{\sigma}_z$ based on the the in-sample sector index portfolio weights $\widehat{\mathbf{w}}_z$. The brown geometric symbols are respectively models average in-sample performance and the red dot represents the estimated GMV portfolio computed on the out-of-sample data. Figure (a) displays the first study, based on 1000 MVN simulated observations with the empirical covariance structure presented in Figure 1. Figure (b) covers the second study, based on 1000 MVN distributed observations with empirical structure within the sectors and zero correlation between, presented in Figure 2. In Figures (c) and (d) the weights from respectively study are presented, the colors agree with the colors in Figures (a) and (b).

Study 2. To investigate the ideas presented partly in *Remark* 1 and mainly in *Remark* 4 above, a new multivariate normally distributed data sets of 1000 observations are simulated. The correlation plot of the simulated data is presented in Figure 2. The difference from the previous study is that the correlation between the sectors is put to zero. Even though this is an unrealistic situation, the idea is to utilize this extreme condition to substantiate our intentions. Besides this the bootstrap procedure follows the same steps as in the previous example and $B = 100\,000$ samples are obtained. Respectively models out-of-sample portfolio performances are presented in Figure 3b, where the geometric shapes in the plot have the same denotation as in Figure 3a. In similarity with the previous sample, the in-sample bootstrap portfolio weights are obtained and presented in Figure 3d.

Some interesting attributes of the sector index model arise in this situation. In *Remark* 4 two important properties of the law of total covariance are presented. Besides some informational gains by introducing sector indices to the model, the idea of decreasing the portfolio variance by aiming for zero correlation between the sectors is suggested. In Figure 3b one observe that the cluster of bootstrapped out-of-sample sector index portfolios is moved within the cluster of the out-of-sample GMV portfolios counterparts. When comparing the density plot in Figure 3b with the corresponding density plot in Figure 3a one acknowledge that the mean as well as standard deviation of the sector index portfolios cluster is reduced. When evaluating the average in-sample sector index portfolio it is clear that this is within the width of the out-of-sample sector index cluster. Hence, the risk estimation of the average in-sample portfolio is not particularly optimistic in comparison of the out-of-sample counterparts. The average in-sample GMV portfolio as well as the estimated out-of-sample GMV portfolio are optimistic compared to the cluster of out-of-sample bootstrap GMV portfolios.

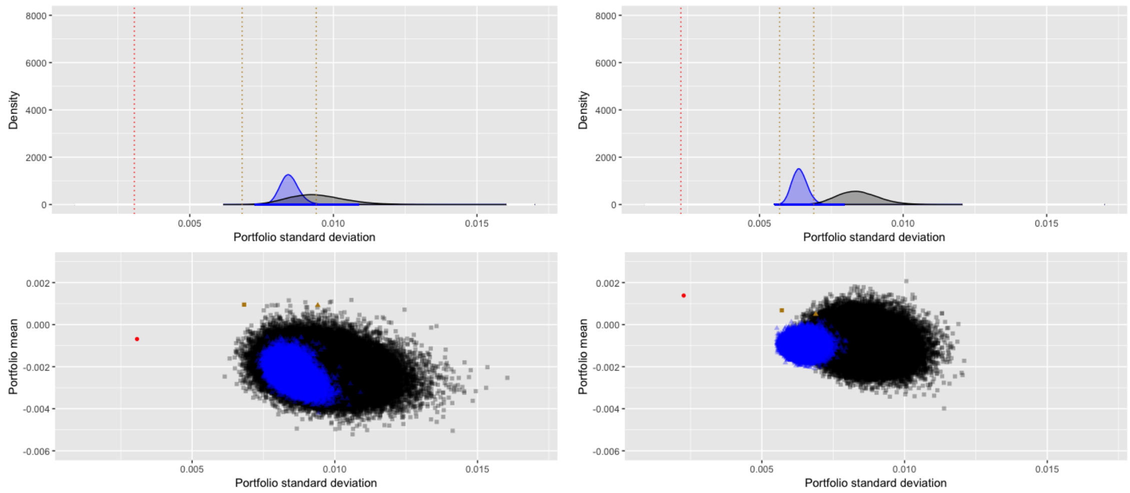
To further investigate the reason for these behaviors one need to evaluate the weights presented in Figure 3d. There are still some differences between the two models, however, not as extreme as in the previous study. Most interesting is the comparison with the boxplots in Figure 3c. The boxplots of the sample GMV portfolios show signs of more stable weights compared to the previous study. The reason for this is probably due to that less covariation need to be taken into account. The sample GMV portfolio, in red, computed on the out-of-sample data confirms this, as it indicates more stable weights compared to the previous case. The sector index portfolios denote close to identical weights as in the previous study. Hence, even though the correlation between the sectors is removed, the model is not particularly affected. A reason for this might arise from the comments in *Remark* 1. The weighting within the sectors are not affected by the fact that the covariance between the sectors are removed. Hence, in this case, it seems that the covariance between the sector indices does not affect the weights particularly.

After evaluating respectively out-of-sample portfolio performances one acknowledge that the cluster of sector index portfolios is of smaller width then the sample GMV portfolio cluster. Hence, in this study it seems like the sector index portfolio is superior to the sample GMV portfolio in reducing risk as well as having less optimistic in-sample estimates. However, even though the underlying structure in this example is unrealistic, the comments in *Remarks* 1 and 4 have proven to be of relevance.

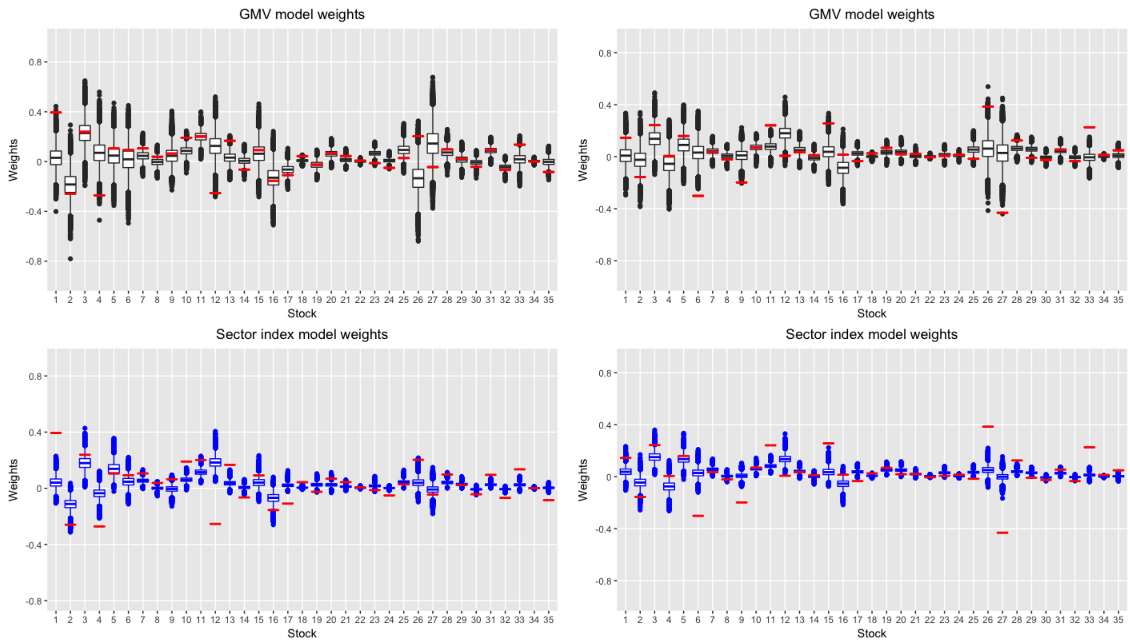
Study 3. To illustrate the problem with close to singular covariance matrices, discussed in *Remark* 3, a smaller data set of 200 MVN distributed observations is simulated. The first 160 of these are used as an in-sample data set and the remaining 40 as an out-of-sample counterpart. Hence, compared to the two previous studies there are a major decrease in observations which leads to that $N\,/\,T$ is closer to one. Consequently, the properties commented in *Remark* 2 might have a bigger influence on the results than in the previous studies due to less stable underlying structure. The bootstrap procedure is similar to previous studies and $B = 100\,000$ bootstrap samples are simulated. The result of the out-of-sample portfolios performance and the in-sample portfolios weight are presented in Figures 4a and 4c. The limits, denotation of geometric shapes and colors conform with Figure 3.

Some shortcomings followed by the usage of the sample GMV portfolio, discussed in Section 3, reveals itself in Figure 4a. There are a tangible increase of the width and height of the cluster of bootstrapped out-of-sample GMV portfolios compared to previous studies. By evaluating the weights in Figure 4c some of the suspicions discussed in Section 2 are confirmed. The weights are extremely unstable and much of the structure from previous studies have disappeared. The width and height of the out-of-sample sector index portfolios cluster have increased compared to previous studies as well. However, it is now located in the lower part of the out-of-sample GMV portfolios cluster. In contrast to the latter, the sector index cluster show signs of smaller variation between the portfolios, hence the more peaked density plot. The sector index portfolios weight are more unstable compared to the ones obtained in the larger data sets, found in Figures 3c and 3d. However, in relation to the sample GMV portfolios, the weights are less volatile and reveal more structure. The result in Figure 4c is a good illustration of the main purpose of this thesis. Our intention is to decrease the error in the sample GMV portfolio to obtain more accurate out-of-sample portfolio performances as $N\,/\,T$ approaches one.

Furthermore, the average sector index in-sample portfolio, the brown triangle, is within the width of the out-of-sample sector index cluster. Hence, the estimates of the sector index portfolio showing signs of being more realistic than the sample GMV portfolio counterpart, i.e. the in-sample portfolios are within range of the out-of-sample cluster. It is important to understand the properties of the sample GMV portfolio computed on the out-of-sample data, displayed in red. As noticed in Section 2 formula (6), the sample covariance matrix is Wishart distributed and an unbiased estimator of the covariance matrix which converge in probability to the covariance matrix for large T relative to N. Due to that the underlying sample covariance matrix in the computations of this portfolio is close to singular, $N = 35$ and $T = 40$, this estimate contains a lot of error. As a consequence, the out-of-sample GMV portfolio weights, in red, are quite extreme. Hence, the uncertainties in the sample covariance matrix have a spillover effect on the weights and finally on the estimation of the portfolio performance.



(a) Empirical data structure (b) Zero correlation between sectors



(c) Empirical data structure (d) Zero correlation between sectors

Figure 4: Density plots, scatter plots and box plots covering 100 000 bootstrap samples. The black squares represent the out-of-sample GMV portfolios mean $\tilde{\mu}_{\mathbf{x}}$ and standard deviation $\tilde{\sigma}_{\mathbf{x}}$ based on the in-sample GMV portfolio weights $\widehat{\mathbf{w}}_{\mathbf{x}}$. The blue triangles represent the out-of-sample sector index portfolios mean $\tilde{\mu}_z$ and standard deviation $\tilde{\sigma}_z$ based on the the in-sample sector index portfolio weights $\widehat{\mathbf{w}}_z$. The brown geometric symbols are respectively models average in-sample performance and the red dot represents the estimated GMV portfolio computed on the out-of-sample data. Figure (a) displays the first study, based on 200 MVN simulated observations with the empirical data presented in Figure 1 as covariance structure. Figure (b) covers the second study, based on 200 MVN distributed observations with empirical covariance structure within the sectors and zero correlation between, presented in Figure 2. In Figures (c) and (d) the weights from respectively study are presented, the colors agree with the colors in Figures (a) and (b).

Study 4. The fourth and final study is specially interesting in regards to all the above declared *Remarks* (1-4). The simulated sample is of same size as the previous example, more precisely, 200 observations. The first 160 are used as in-sample data set and the remaining 40 as out-of-sample counterpart. In contrast of the previous study, but in similarity with *Study* 2, the correlation between the sectors are put to zero. Hence, we are able to evaluate the two models on a data structure with the properties of both being close to singularity as well as having reduced correlation between the sectors. In similarity to previous studies $B = 100\,000$ bootstrap samples are simulated. The limits, geometric shapes and colors in Figure 4b and 4d are in line with the figures presented in *Studies* 1-3.

When evaluating the results in Figure 4b it is clear that the reduction of covariance between the sectors does affect the out-of-sample bootstrap GMV portfolio as well as the sector index counterpart. The width of the cluster of the former of these methods is reduced and less portfolios of high standard deviation is obtained. The out-of-sample sector index portfolios show similar characteristics as in *Study* 2. This cluster is situated in the lower area of the out-of-sample GMV portfolio cluster. Hence, average standard deviation of the out-of-sample sector index portfolios are lower than in the previous study. This is reasonable due to that much of the covariance are removed. In similarity to previous studies, the average in-sample GMV portfolio is optimistic in comparison to the bootstrapped out-of-sample GMV portfolio cluster. While on the other hand, the average in-sample sector index portfolio is within reach of the standard deviations of the out-of-sample sector index cluster. The estimated out-of-sample GMV portfolio is quite extreme and have a similar interpretation as in the previous study.

The in-sample GMV portfolio weights show signs of being a little more stable and structured in comparison to the in-sample GMV portfolio weights in *Study* 3, but less stable than the sector index weights. The letter of these might be slighter more stable than in the previous study and the structure is recognizable. The out-of-sample GMV portfolio, in red, are within the reach of the outliers of the bootstrapped in-sample GMV portfolio weights and out of reach of the bootstrapped in-sample sector index weights.

In summation, in this study we are able to illustrate and strengthen the properties of the suggested sector index portfolio. By reducing the correlation between the sectors while simulating a small data set close to singularity, a precarious situation is analyzed with results that strongly supports our hypothesized advantages of the proposed method.

5 Conclusion

After evaluating the results of the four studies a lot of the concerns are straightened out and some of the suspicions are confirmed. For large data sets, having a stable underlying structure of the multivariate normal distribution, the sample GMV portfolio outperforms the sector index portfolio. As concluded above, this is not surprising due to the optimal conditions for the estimation of the GMV portfolio. Furthermore, this is a situation that we are not particularly interested in. However, to completely understand the concept of the sample GMV portfolio it is of importance to evaluate this sort of situation as well.

The positive effects associated with low correlation between the sector indices are presented in *Remarks* 1 and 4. After evaluating the results in *Studies* 2 and 4 these properties are confirmed and we conclude that aiming for low correlation between the sector indices have a positive effect on the results of both models, but particularly the sector index model. From the above studies it is difficult to evaluate the effect that the reduced amount of variability of estimates, signaled in *Remark* 2, has on the weights and consequently on the portfolio performance. For larger data sets, with MVN distributed assets, the effect seems negligible. The underlying assets are stable and does not contain much error due to outliers. As the number of observations in relation to the number of assets decrease, with more heavy tailed distributions as consequence, the importance of this effect is probably more evident. However, whether the more stable in-sample weights of the sector index portfolios in contrast to the estimated in-sample GMV portfolios in *Studies* 3 and 4 are primarily due to less variance and covariance estimates or that the sample covariance matrices are further away from singularity is unknown. Likely there are a combination of the two.

After evaluating the results in the last two studies, which are of particular interest in this thesis, the strength of the sample sector index portfolios is acknowledged. In situations where the sample covariance matrices are close to singular, the sample sector index portfolio is superior the sample GMV portfolio in reducing risk. Due to its more stable and structured weights, the out-of-sample performance is less risky and more accurate in comparison to the the in-sample counterpart, especially when the correlation between the sector indices are low.

To fully evaluate the results, some further thoughts are in order. The assumptions of MVN distributed stock data should be questioned. In reality financial data are more heavy tailed than the Gaussian distribution. However, based on the properties of the sample covariance matrix and the concept of indices, one might suspect that the sample sector index portfolio performs even better in comparison to the sample GMV portfolio in situations where the empirical data are more heavy tailed.

During the research and writing of this thesis some interesting ideas appeared that would be of interest for future research. First, in reality an investor scarcely invests in all the stocks within a sector. He or she will rather invest in a couple of stocks that are of most interest and diversify over different sectors. Therefore it would be exciting to evaluate how the sector index portfolio would perform in such a situation.

Finally it would be of interest to evaluate the structure behind the sector indices to fully understand the properties of the sector index portfolio. With such knowledge one would be able to construct the indices oneself and utilize the benefits of grouping similar assets together further.

References

Basak, G. K. (2005), Estimating the Risk in Sample Efficient Portfolios1, PhD thesis, Citeseer.

Bengtsson, C. and Holst, J. (2002), *On portfolio selection: Improved covariance matrix estimation for Swedish asset returns*, Univ.

Best, M. J. and Grauer, R. R. (1991), 'On the sensitivity of mean-variance-efficient portfolios to changes in asset means: some analytical and computational results', *Review of Financial Studies* **4**(2), 315–342.

Bodnar, T., Parolya, N. and Schmid, W. (2014), 'Estimation of the global minimum variance portfolio in high dimensions', *arXiv preprint arXiv:1406.0437* .

Bollerslev, T. (1986), 'Generalized autoregressive conditional heteroskedasticity', *Journal of econometrics* **31**(3), 307–327.

Chopra, V. K. and Ziemba, W. T. (1993), 'The effect of errors in means, variances and covariances on optimal portfolio choice', *Journal of Portfolio Management* **19**(2), 6–11.

Claeson, M. (2016), A critical review of the global minimum variance theory.

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

Engle, R. F. (1982), 'Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation', *Econometrica: Journal of the Econometric Society* pp. 987–1007.

Frankfurter, G. M., Phillips, H. E. and Seagle, J. P. (1971), 'Portfolio selection: the effects of uncertain means, variances, and covariances', *Journal of Financial and Quantitative Analysis* pp. 1251–1262.

Jobson, J. and Korkie, B. (1980), 'Estimation for markowitz efficient portfolios', *Journal of the American Statistical Association* pp. 544–554.

Johnson, R. and Wichern, D. (2007), *Applied Multivariate Statistical Analysis*, Applied Multivariate Statistical Analysis, Pearson Prentice Hall.

Kan, R. and Smith, D. R. (2008), 'The distribution of the sample minimum-variance frontier', *Management Science* **54**(7), 1364–1380.

Kan, R. and Zhou, G. (2007), 'Optimal portfolio choice with parameter uncertainty', *Journal of Financial and Quantitative Analysis* **42**(03), 621–656.

Ledoit, O. and Wolf, M. (2003), 'Improved estimation of the covariance matrix of stock returns with an application to portfolio selection', *Journal of empirical finance* **10**(5), 603–621.

Ledoit, O. and Wolf, M. (2004), 'Honey, i shrunk the sample covariance matrix', *The Journal of Portfolio Management* **30**(4), 110–119.

Markowitz, H. (1952), 'Portfolio selection', *The journal of finance* **7**(1), 77–91.

Markowitz, H. (1959), 'Portfolio selection: Efficient diversification of investments', *New York: Wiley* .

Merton, R. C. (1980), 'On estimating the expected return on the market: An exploratory investigation', *Journal of financial economics* **8**(4), 323–361.

Michaud, R. O. (1989), 'The markowitz optimization enigma: Is optimized optimal?', *ICFA Continuing Education Series* **1989**(4), 43–54.

Pappas, D., Kiriakopoulos, K. and Kaimakamis, G. (2010), Optimal portfolio selection with singular covariance matrix, *in* 'International Mathematical Forum', Vol. 5, pp. 2305–2318.

Wishart, J. (1928), 'The generalised product moment distribution in samples from a normal multivariate population', *Biometrika* pp. 32–52.

A Appendix

Table 2: Stocks

	Bank GI		FaP GI		Mining GI		Tech GI
1	Nordea	8	Arctic paper	17	Boliden	24	Anoto
2	SEB A	9	Billerud Korsns	18	Endomies	25	Doro
3	SEB C	10	Bergs Timber B	19	Lucara Diamond Corp.	26	Ericsson A
4	Sv.H A	11	Holmen A	20	Lundin Gold	27	Ericsson B
5	Sv.H B	12	Holmen B	21	NGEx Resourses	28	HMS
6	Swed. A	13	Munksj Oyj	22	Nordic Mines	29	Hexatronic
7	Nordax	14	Rottneros	23	Semafo	30	Invisio
		15	Stora Enso A			31	MultiQ
		16	Stora Enso R			32	Net Insight
						33	Nokia
						34	Stockwik
						35	Tobii