A Cluster Analysis of Stocks to Define an Investment Strategy

SARAH BJÄRKBY

SOFIA GRÄGG
A Cluster Analysis of Stocks to Define an Investment Strategy

How can investment strategies be formulated based on calculating similarities in the volatilities of stock indices?

SARAH BJÄRKBY
SOFIA GRÄGG
This thesis investigates the possibilities of creating an investment strategy by performing a cluster analysis on stock returns. This to provide a diversified portfolio, which has multiple advantages, for instance that the risk of the investment decreases. The cluster analysis was performed using various methods – Average linkage, Centroid and Ward’s method, for the purpose of determining a preferable method. Ward’s method was the most appropriate method to use according to the results, since it was the only method providing an analysable result. The investment strategy was therefore based on the result of Ward’s method. This resulted in a portfolio consisting of eight stocks from four different clusters, with the eight stocks representing four sectors. Most of the results were not interpretable and some of the decision making regarding the number of clusters and the appropriate portfolio composition was not entirely scientific. Therefore, this thesis should be considered as a first indication of the adequacy of using cluster analysis for the purpose of creating an investment strategy.
Sammanfattning

Acknowledgement

This thesis was conducted in cooperation with Nordea. We would like to thank Krister Alvelius and Peter Seippel at Nordea for providing us with this exciting project and for the expertise, time and effort contributed. We also want to thank our supervisors at KTH – Jörgen Säve-Söderbergh and Julia Liljegren, for providing us with literature suggestions, guidance and quick responses when problems occurred.
List of Tables

4.1 Type of data obtained ................................................. 14
5.1 Calinski-Harabasz index ................................. 17
5.2 Depicts the cluster stock division – Ward’s method ........ 22
5.3 Depicts the best stock in each cluster – Ward’s method ... 23
5.4 Depicts the second best stock in the largest clusters – Ward’s method 23
5.5 Stocks composing the final portfolio .......................... 23
# List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Example of a dendrogram</td>
<td>8</td>
</tr>
<tr>
<td>5.1</td>
<td>Stock index overview</td>
<td>16</td>
</tr>
<tr>
<td>5.2</td>
<td>Stock return overview</td>
<td>17</td>
</tr>
<tr>
<td>5.3</td>
<td>Calinski-Harabasz index plot</td>
<td>18</td>
</tr>
<tr>
<td>5.4</td>
<td>Dendrogram Ward’s method</td>
<td>19</td>
</tr>
<tr>
<td>5.5</td>
<td>Dendrogram Ward’s method divided in nine clusters</td>
<td>20</td>
</tr>
<tr>
<td>5.6</td>
<td>Dendrogram Ward’s method with a cutoff at nine clusters</td>
<td>21</td>
</tr>
<tr>
<td>5.7</td>
<td>Dendrogram Centroid method</td>
<td>24</td>
</tr>
<tr>
<td>5.8</td>
<td>Dendrogram Average linkage method – Euclidean distance</td>
<td>25</td>
</tr>
<tr>
<td>5.9</td>
<td>Dendrogram Average linkage method – Correlation measure</td>
<td>26</td>
</tr>
</tbody>
</table>
6 Discussion and Analysis
   6.1 Mathematical discussion
   6.2 Economic discussion

7 Further Research

8 Conclusion

References

Appendix
Introduction

1.1 Background

Financial investments are used all around the world by both companies and individuals. Investments are performed for various purposes, for instance by a financial investment firm that administers investor’s capital. Consequently the actors have different prerequisites, aims and methods for performing their investments. Investment strategies can be obtained by both primitive methods and more advanced algorithms. However, all investors have one ambition in common, to achieve a high return combined with a low risk. To actually achieve such a portfolio is more complicated.

The first step when creating an investment strategy is data analysis. Financial data, often collected from daily operations, usually needs to be processed and converted into manageable data, which is easier to analyse. Further, the data have to be analysed using one or several methods and multiple aspects should be considered. One aspect that is important to consider is risk. Some actors have a high tolerance for it, while others have lower. As stated previously, finding the right balance between risk and return is sought after. There are a number of different approaches for assessing and handling risk. For portfolio optimisation, diversification is a commonly used and well renowned approach for dealing with it. The risk can thus be reduced by creating more diversified portfolios. The ultimate goal of diversification is to reduce the volatility of the portfolio by including data prone to risk from different sources in a portfolio.

A diversified portfolio can in theory be obtained by performing a cluster analysis of financial assets, such as stocks, futures and bonds. Cluster analysis is used in various areas and previous research indicates that it might be appropriate for financial purposes too. A cluster analysis systematises disorganised data into subsets of groups, called clusters. The principle of cluster analysis is that the difference, specifically the distance between stocks within a cluster, should be as small as possible, whereas the distance between clusters should be large. Cluster analysis therefore intends to find correlations within data. The most similar data is grouped together, creating clusters of similar data sets. These clusters can then be used to create investment strategies. A number of different methods of cluster analysis will be compared and analysed in hope of finding a reliable method. The cluster analysis will be be evaluated based on data sets provided by Nordea and the investment strategy evaluated and assessed theoretically.

---

1 Berry and Linoff 2004.
2 Chen 2018.
3 Education 2019.
1.2 Project aim

The aim of this thesis is to create an investment strategy with low risk and steady return. This is accomplished by investigating the possibility of creating investment strategies based on finding similarities in the volatilities of stock indices. Subsequently analysing how different methods for clustering impact an investment strategy. The goal is to calculate how the volatilities of the stock indices correlate. This allows grouping of stocks into distinct clusters, which further allows for cluster analysis using different methods. The clustering enables the possibility of drawing conclusions about the return of grouped stocks. An additional aim of this thesis is to generate an investment strategy that can act as a complement to other investment strategies.

1.3 Research question

The research question of the thesis is as follows: How can investment strategies be formulated based on calculating similarities in the volatilities of stock indices? The research question can be divided into two sections. The first part aims to find and examine the similarities in the volatilities of the stock indices. Afterwards, the obtained results about the similarities have to be analysed for the purpose to formulate an investment strategy.

1.4 Scope

Cluster analysis is a broad topic that includes a wide range of methods. This thesis will apply three different methods for grouping stocks and two for retrieving correlations. The Centroid-, Average linkage- and Ward’s method will be used for the clustering and the Euclidean distance and the Correlation matrix to obtain correlations. All methods will be discussed further in the theory section. The data used is stock indices for 305 stocks from the time period 2017-10-02 to 2018-10-02. The stock data is provided by Nordea and concerns the sectors of Communication Services, Consumer Discretionary, Consumer Staples, Energy, Financial, Health Care, Industrial, Information Technology, Materials, Real Estate and Utilities.
Previous Research

Cluster analysis is used in various areas in order to find similarities in data. However, cluster analyses based on stock indices or financial data in general, have not been performed in a great extent.

One paper written on the subject aims to examine the optimal method of measuring the difference between stocks. The report performs cluster analysis with a self-composed distance measure, with the aim of decreasing the disadvantages with other well known methods. An example is the Correlation method, which often changes during financial distress and therefore could provide faulty results. The report concludes that diversification is favourable but that it does not exist one clear procedure on how to compose and maintain a diversified portfolio.\(^1\)

Another paper investigates the possibility of creating a classification of the OMXS 30 stocks based on the correlation of its co-movements using cluster analysis. The aim of the paper is to investigate whether ultrametric space is appropriate for this type of classification. The author concludes the possibility to obtain a classification based on the co-movements of the stocks and the ultrametric assumption.\(^2\)

A further paper brands cluster analysis as the new approach to solving problems related to estimation errors in portfolios composed by stocks. The paper compares the estimation errors obtained from both cluster analysis and a resampling method, which was more commonly used in the past. The results indicate that cluster analysis is useful for improving the robustness and performance of a portfolio. It also addresses the fact that few studies have been performed on the subject. Furthermore, a number of questions linked to the usage and interpretation of the cluster analysis method is left unanswered.\(^3\)

There is one report concerning cluster analysis of a similar data set as used in this thesis. It concludes that investments, according to investment strategies obtained from the cluster analysis of stock data, can result in both profit and loss.\(^4\)

There is, as shown, some research conducted on cluster analysis of data sets of stocks. The previous research does not have the exact same approach as this thesis but indicates that the method of cluster analysis is worth examining further for this purpose.

\(^1\)Marvin 2015.  
\(^2\)Rosén 2006.  
\(^3\)Ren 2005.  
Theory

3.1 Economic theory

Financial markets are remarkably complex systems. They are composed of many actors interacting in non-linear ways. Financial markets are monitored closely, and unlike other complex systems there exists a massive amount of data, making the financial system well defined.\(^1\)

A financial portfolio is a collection of financial assets such as stocks, futures and bonds. This thesis will be restricted to stocks, however it is possible that the method employed is applicable also on other forms of financial assets. The composition of a portfolio requires the determination of the strengths, weaknesses, opportunities and threats of the probable assets, making it a difficult task to manage successful portfolios. Human decision making is sometimes biased and illogical and there is also some randomness in the market, along with other factors that can affect the outcome. This makes the process of creating a portfolio more difficult. Achieving high returns combined with low risk is ideal, but a difficult task. Nowadays, portfolio theory states that diversification of assets is the most prominent way of obtaining low risk-reward ratios.\(^2\)

3.1.1 Stocks and their properties

A stock is a financial asset that represents ownership in a company or corporation and symbolizes a claim on its assets and earnings. Depending on the company or corporation, the stocks are influenced by various factors. One factor that influence the majority of stocks, independent of the company, is the change in prices of other stocks. It is important to understand the co-movements of stocks to be able to compose a well performing portfolio.\(^3\)

3.1.2 Relative return

It is commonly known that the financial market is volatile over time and so is each individual stock. In order to compare stocks to each other, the stocks’ volatilities can be computed. The comparison can be performed through various methods, this thesis using the relative return (see equation 3.1). The relative return describes an asset’s performance by computing the price quotient from one day compared to the day before.\(^4\)

---

\(^1\)Bonanno, Lilloa, and Mantegna 2001, p. 16.
\(^2\)Marvin 2015, p. 1.
\(^3\)C. Sharma, Habib, and Bowry 2018, p. 2.
\(^4\)Chen 2017.
\[
Relative\ return = \frac{I_t}{I_{t-1}} - 1
\]  

A cluster analysis concerning financial data, with the goal of creating an investment strategy, should process the return of the assets as input data. An investor’s goal when investing in a portfolio is to maximise the profit while minimising the risk. This equals minimising the risk for a specific amount of return. By this definition, the input data for the cluster analysis in this thesis will be the relative return.\(^5\)

### 3.1.3 Investment strategy

An investment strategy consists of for instance the rules and tactics that constitutes the base for the investments. The investors are mainly dealing with balancing risk and return, independently of the purpose of the investments. A commonly used method for dealing with the risk versus return problem is diversification.\(^6\) Diversification spreads the risk in the portfolio by merging together assets that are different from each other according to some chosen aspects. These aspects can for instance be related to the industry, country, type of asset or size of corporation. It is more likely that two stocks from, for example, the same industry or the same country are more closely related in their movements. A portfolio composed of stocks from one single industry is generally more exposed to risk, than a portfolio composed of stocks from different industries.\(^7\)

Making a diversified portfolio is by definition not difficult since finding objectively different stocks is relatively simple. To spread the risk and maximise the return requires more effort though. There is a need to know which stocks to choose among, in for example the different industries and countries. The human behaviour biases mentioned above could be minimised if an automated method for classification of stocks is developed. Cluster analysis is a method of classification to find combinations of stocks that would compose a well diversified portfolio, also taking the return into account.\(^8\) The clusters will have high inter-correlation and low intra-correlation, making the risk diversification simple. The high inter-correlation can be used to find stocks with higher returns within the cluster. As the stocks in each cluster are supposed to follow the same market movements, it can be assumed that the stocks with high returns in the past, will continue to have high return in the future and vice versa when considering lower returns.

To minimise the human interference when composing a diversified portfolios does not guarantee success. Sometimes events affect the market in unexpected ways or cause unforeseen movements. A diversified portfolio would be affected if, for instance, an exogen event, affecting the whole market, would occur. An example is the 2008 financial crisis, which affected the whole market during a period of time.

---

\(^5\)Marvin 2015.  
\(^6\)Education 2019.  
\(^7\)Marvin 2015, p. 3.  
\(^8\)Ibid., p. 3.
### 3.1.4 Long and short term investments

An important aspect to consider when dealing with investments is the time horizon of the investments. If the portfolio is based on data from a short period of time, for example six months, it is unlikely that the assumptions and expectations of the returns will be reliable in six years or even the six following months. Longer history of the data results in greater duration for the investment strategy. The state of the economy at the time of making the investment strategy also affects the outcome. A more unstable economic situation at the time of making the investment strategy, increases the necessity of a quicker change of the investment strategy.
3.2 Mathematical theory

Investment strategies are often based on mathematical models since they can predict future outcomes of a financial market, based on history. The mathematical model bases decisions on several factors that a human being would not be able to take into consideration. Furthermore, it is easy to successively make small changes in a mathematical model due to unexpected events in the market. However, a mathematical model cannot always predict the future of the market since all factors cannot be predicted and included in the model. Mathematical models do still need the human factor for correcting the model due to unexpected disruptions.

There exists a great amount of mathematical models with the purpose of creating investment strategies. Every method has different pros and cons. Some methods are for example favourable when predicting an investment strategy in a specific line of business, while others are better at dealing with outliers.

This thesis concerns the mathematical method cluster analysis for the creation of an investment strategy. This method is used due to its perks of creating diversified portfolios. There exists different methods for conducting cluster analysis, which in turn have various advantages. The choice of methods used in this thesis is based on, among other things, the type of data, the purpose of the analysis and the definition of distances between the stock returns.

3.2.1 Cluster analysis

Cluster analysis is a technique used for grouping observations into clusters. The obtained clusters are homogeneous with respect to certain characteristics chosen. Each cluster will be different from the others according to the chosen criteria. The chosen characteristics differ depending on the data and the purpose of performing the cluster analysis. Cluster analysis is a common method for genomics, marketing, financial purposes and mapping groups of illness types - for example types of breast cancer.

Since the cluster analysis can be used in various fields, a distinct number of methods has been evolved, beside the choice of characteristics. There are two main methods for performing a cluster analysis – hierarchical and non-hierarchical clustering. The two methods have various advantages, often depending on the type of data and the reason for performing the clustering.

Hierarchical clustering

Hierarchical clustering merges smaller clusters, of one or more data points, into larger clusters or splits larger clusters into smaller clusters. The more common, to merge smaller clusters into larger, is called agglomerative clustering. The result of hierarchical clustering is usually depicted as a tree, called a dendrogram (see Figure 3.1). The lowest linked nodes (in Figure 3.1 the lowest link nodes correspond to data points 1 and 4) consist of the data points which are the most similar. Moving further up the tree, it is depicted that the data points are linked together at a greater

---

10James et al. 2013, pp. 385-386.
height. There are a number of different linkage methods for performing hierarchical clustering, depending on the data used. Hierarchical clustering does not have the need for a pre-specified number of clusters and can therefore be used on a wide range of problems.\textsuperscript{11}

![Dendrogram example](image)

Figure 3.1: Example of a dendrogram

**Non-hierarchical clustering**

Non-hierarchical clustering on the other hand, directly partitions the data into a set of disjoint clusters. The K-means method is the most popular method in the non-hierarchical clustering. K-means clustering is an elegant and simple method for creating K distinct clusters. To perform K-means clustering, the number of clusters, K, must be chosen beforehand. Deciding the number of clusters is often a difficult task and a major drawback of the K-means clustering method.\textsuperscript{12} Another disadvantage with this method is that it is not possible to depict the result in a figure, as viable for the hierarchical method.

### 3.2.2 Choice of main method

The main method of choice in this thesis is hierarchical clustering. This thesis concerns data of stock indices therefore the hierarchical clustering is more appropriate than non-hierarchical since there is no previous knowledge on how the stock indices are linked together. Further, it is difficult to determine the number of clusters in advance, as needed for non-hierarchical method.

\textsuperscript{11}James et al. 2013, pp. 389-390.
\textsuperscript{12}Marvin 2015, p. 7.
3.2.3 Methods for performing hierarchical clustering

There exists a wide range of linkage methods and distance measures for performing a hierarchical cluster analysis. This thesis examines three different linkage methods and two distance measures with the aim of drawing a reliable conclusion. The linkage methods of choice are Ward’s method, the Centroid method and the Average linkage method. The methods are chosen since they are commonly used and have various advantages which increase the probability of finding a reliable result. The distance measures used are the Euclidean distance measure and the Correlation measure. The Euclidean distance measure is chosen since it is by far the most common measure further it can be applied to every linkage method. The Correlation measure will be used as a complement to investigate the differences between the Euclidean- and the Correlation distances. The Correlation measure can only be combined with the Average linkage method of the three linkage methods chosen. The properties of the linkage methods and measures will be discussed in detail below.

**Linkage methods**

**Ward**

The Ward’s method is a hierarchical linkage method. It calculates the distance between two observations, stocks or clusters, through merging the sum of squares and analyse how much the within-cluster sum of squares increases.

\[
  d_{\text{Ward}}(A, B) = \sqrt{\frac{2n_A n_B}{n_A + n_B} \| \bar{m}_A - \bar{m}_B \|}
\]  

Equation 3.2 shows the equation for the merging cost, the increase in sum of squares, when merging clusters A and B. The \( m_j \) is the centre of cluster \( j \) and \( n_j \) is the number of data points in the cluster. The \( \| \| \) represents the Euclidean distance which is explained in detail below. The sum of squares starts at zero, since every point is its own cluster and increases as the clusters merge. By using Ward’s method the increase aims to be as small as possible. If two clusters have the same merging cost, Ward’s method will merge the cluster containing the least number of data points.

A disadvantage with the method is that it is greedy and constrained by previous choices of clustering. It is not possible for a data point to change cluster after being assigned to one.

Ward’s method can only be combined with one of the methods that calculates the distance between stocks, namely the Euclidean distance. This because the algorithm requires the Euclidean calculation for the initial set up, when all data points are individual clusters.

**Centroid**

The Centroid method measures the distance between the centroids of the clusters. The centroid is the most representative point of the cluster, the average element. The value is obtained by calculating the average value of the data points within the

---

13 James et al. 2013, p. 395.
14 Cosma 14 September 2009, p. 3.
15 Ibid., p. 3.
The centroid is the centre of mass of a cluster and will be the point of comparison when determining the distance between clusters.

\[ d_{\text{centroid}}(A, B) = \| \overline{m}_A - \overline{m}_B \|, \text{where } \overline{m}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} m_{ji}, i = 1, ..., n \] (3.3)

\( \overline{m}_j \) = centroid of cluster \( j \). Further \( i \) portraits the number of data points in the cluster.\(^{17}\)

The Centroid method is a commonly used method, especially in genomics. It does however have a disadvantage. When two clusters are grouped at a height, distance less than the height of one already existing cluster in the dendrogram, an inversion occurs. The inversion can bring difficulties in visualisation and interpretation of the dendrogram.\(^{18}\)

**Average linkage**

The Average linkage method is a method that calculates the average distance between all of the cluster pairs to provide an accurate evaluation of the distance between clusters. For Average linkage, the distances between each data point in one cluster and all of the data points in the other cluster are compared and afterwards averaged.\(^{19}\)

\[ d_{\text{average}}(A, B) = \frac{1}{n_An_B} \sum_{i \in A, j \in B} d_{ij} \] (3.4)

where \( n_i \) is the number of data points in the cluster.

The formula for the Average linkage calculation is shown in equation 3.4. The distance between clusters A and B is equal to the average distance between all of the data points in the cluster.

The Average linkage method can, unlike the other methods, use both Euclidean distance measure and other methods like Correlation measure.

**Distance measures**

**Euclidean**

The Euclidean distance measure is a common method for calculating distance between observations to detect similarities. The method origins from classical geometry and calculates the shortest distance between observations through a straight line distance based on the Pythagorean theorem. The Euclidean distance measure calculates the square root of the sum of the squared differences between two clusters, \( p \) and \( q \), in the \( i \)th dimension.\(^{20,21}\)

---

\(^{16}\)Berry and Linoff 2004, p. 369.

\(^{17}\)S. Sharma 1996, pp. 188-191.

\(^{18}\)James et al. 2013, p. 395.

\(^{19}\)Yim and Ramdeen 2015, p. 4.

\(^{20}\)Gonçalves et al. 2015.

\(^{21}\)Rosén 2006.
Euclidean distance \[ = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2} \] (3.5)

**Correlation**

The Correlation measure can be used as a distance measure for the Average linkage method. The Correlation matrix measures how data points fluctuate compared to each other. However the Correlation matrix is not a measure strictly speaking. The coefficients can easily be converted into distance measures, which enables conducting a cluster analysis based on the correlations. The Correlation measure is a favourable approach for cluster analysis when analysing data based on stocks. If the stock price of one stock in a cluster decreases, it is expected that the other stocks in the cluster will decrease as well. This results in an efficient way of creating and optimising a portfolio.\(^{22,23}\)

The Correlation matrix includes comparisons between the stocks and is portrayed in Matrix 3.1. The diagonal elements are equal to one and \( S_m \) denotes stock \( m \). \( \sigma^2 \) denotes the variance of the stocks. Further every element provides the correlation between stock \( m \) and stock \( n \).

\[
\begin{bmatrix}
1 & \frac{\sigma^2(S_1, S_2)}{\sqrt{\sigma^2(S_1) * \sigma^2(S_2)}} & \frac{\sigma^2(S_1, S_3)}{\sqrt{\sigma^2(S_1) * \sigma^2(S_3)}} & \cdots & \frac{\sigma^2(S_1, S_n)}{\sqrt{\sigma^2(S_1) * \sigma^2(S_n)}} \\
\frac{\sigma^2(S_2, S_1)}{\sqrt{\sigma^2(S_2) * \sigma^2(S_1)}} & 1 & \frac{\sigma^2(S_2, S_3)}{\sqrt{\sigma^2(S_2) * \sigma^2(S_3)}} & \cdots & \frac{\sigma^2(S_2, S_n)}{\sqrt{\sigma^2(S_2) * \sigma^2(S_n)}} \\
\frac{\sigma^2(S_3, S_1)}{\sqrt{\sigma^2(S_3) * \sigma^2(S_1)}} & \frac{\sigma^2(S_3, S_2)}{\sqrt{\sigma^2(S_3) * \sigma^2(S_2)}} & 1 & \cdots & \frac{\sigma^2(S_3, S_n)}{\sqrt{\sigma^2(S_3) * \sigma^2(S_n)}} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\frac{\sigma^2(S_m, S_1)}{\sqrt{\sigma^2(S_m) * \sigma^2(S_1)}} & \frac{\sigma^2(S_m, S_2)}{\sqrt{\sigma^2(S_m) * \sigma^2(S_2)}} & \frac{\sigma^2(S_m, S_3)}{\sqrt{\sigma^2(S_m) * \sigma^2(S_3)}} & \cdots & 1 \\
\end{bmatrix}
\]

Matrix 3.1: Correlation matrix

### 3.2.4 Number of clusters

Hierarchical clustering does not require the specification of the number of clusters before the analysis is performed. However some limitations of the quantity are preferably made to get a reliable result. The number of clusters are in general determined by looking at the obtained dendrogram. Further the sensible number of clusters is selected by eye. The quantity is chosen based on the number of desired clusters and the heights of fusion. However this method is not scientific and which level to cut the dendrogram is not always clear.\(^{24}\) Therefore, an algorithm that indicates where to cut the dendrogram could preferably be used. The Calinski-Harabasz criterion is a method used to evaluate an optimal number of clusters. The highest

---

\(^{22}\) Sharma 1996, p. 220.

\(^{23}\) Marvin 2015.

\(^{24}\) James et al. 2013, pp. 393-394.
Calinski-Harabasz value indicates the optimal number of clusters.\textsuperscript{25}

The Calinski-Harabasz index is defined as
\[
VRC_k = \frac{SS_b}{SS_w} \times \frac{N - k}{k - 1}
\] (3.6)

\(SS_b = \sum_{i=1}^{k} n_i \|m_i - m\|^2\) is the overall between-cluster variance and \(SS_w = \sum_{i=1}^{k} \sum_{x \in c_j} \|x - m_i\|^2\) is the overall within-cluster variance. \(k\) is the number of clusters and \(N\) is the number of observations.\textsuperscript{26}

### 3.2.5 Best stock in cluster

When performing the cluster analysis, the observations, stocks, are divided in groups because of their similarities. In order to formulate the investment strategy, the best stocks in each cluster has to be detected. Afterwards, the best performing stocks, in some combination, depict the recommended portfolio for investing.

The Sharpe ratio (see equation 3.7) is used to determine the best performing stock in each cluster by calculating the risk of an investment compared to its return.\textsuperscript{27} The ratio calculates the excess return of an asset divided by the asset’s standard deviation of returns. A higher Sharpe ratio presents a stock with higher return over time. Calculating the Sharpe ratio for every stock in a cluster gives an indication on the stocks with the highest ratios and are the best performing according to the data obtained.\textsuperscript{28}

\[
S_a = \frac{E[R_a - R_b]}{\sigma_a}
\] (3.7)

\(R_a\) is the asset return and \(R_b\) is the return on a benchmark asset. \(E[R_a - R_b]\) is the expected value of the excess of the asset return and \(\sigma\) is the standard deviation of the assets excess return.

The Sharpe ratio compares the return to the risk of an investment which generates a way to isolate profits associated with risk taking activities. A greater Sharpe ratio value is therefore in general more attractive. Sharpe ratio is a well used risk versus return measure, mostly because of its simplicity. Further it has a high credibility which makes it a trustworthy method. However, to use past data to predict the future might not accurately replicate the future, which should be taken into consideration.\textsuperscript{29}

### 3.2.6 Limitations in clustering

The data used in this thesis is obtained from Nordea and is composed by stock indices. The analysis includes 305 stocks and their associated indices over a time period of one year. The number of stocks in the analysis is 305 and the time period is limited to trading days. A year consists of approximately 252 trading days,

\textsuperscript{25}MathWorks 2019a.
\textsuperscript{26}Ibid.
\textsuperscript{27}Hargrave 2019.
\textsuperscript{28}MathWorks 2019b.
\textsuperscript{29}Hargrave 2019.
however the number of trading days varies each year. The one year period analysed 2017-10-02 – 2018-10-02, includes 257 days.

Apart from the data used, some further specific and important aspects might have an impact on the result, when conducting the cluster analysis. Some aspects to consider are; do the observations have to be standardised before the calculations, what methods will be used to calculate the distance between the data observations, what type of linkage methods will be used and where should the dendrogram be cut off in order to obtain an appropriate amount of clusters? The decision for each aspect is based on the most appropriate solution for the cause and what gives the most interpretable result.\footnote{James et al. 2013, pp. 399-400.}
Method and Model

4.1 Data collection

The data used in this thesis is a time series of index documentations of stock prices provided by Nordea. The time period used for the analysis is 2017-10-02 – 2018-10-02, 257 trading days. Each trading day correspond to a specific index for each stock.

The data is on the form:

<table>
<thead>
<tr>
<th>Day/Stock</th>
<th>Stock_1</th>
<th>Stock_2</th>
<th>Stock_3</th>
<th>...</th>
<th>Stock_n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Day_1</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>...</td>
<td>100</td>
</tr>
<tr>
<td>Day_2</td>
<td>index_{S_12}</td>
<td>index_{S_22}</td>
<td>index_{S_32}</td>
<td>...</td>
<td>index_{S_n2}</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td></td>
<td>...</td>
</tr>
<tr>
<td>Day_m</td>
<td>index_{S_1m}</td>
<td>index_{S_2m}</td>
<td>index_{S_3m}</td>
<td>...</td>
<td>index_{S_nm}</td>
</tr>
</tbody>
</table>

where \( n \) is the number of stocks and \( m \) is the number of trading days.

The state of the economy during the time period that the data concerns includes one period (2017 to early 2018) of strong growth. During the second half of 2018, the growth slowed down as a result of various factors affecting major economies. An example is US–China market tensions and changes in the automotive industry in Germany. The beginning of 2019 was a more beneficial period for the economy partly due to that the US signalled a more accommodative monetary policy.\(^1\) The current unstable economic environment makes investing more difficult since it becomes more uncertain to predict how the market will change. The current state of the economy will be important to take into consideration as composing an investment strategy based on the data from an unstable time.

The data contains stocks from various sectors, the distribution between the sectors is depicted in the Appendix.

4.2 Computations

All calculations in this thesis are performed in the programming platform MATLAB, using its built-in functions. The specifics of the code for this thesis are depicted in the Appendix. Below, a detailed description of the calculations are found.

The clustering is produced by the built-in function \textit{linkage} in MATLAB, using various inputs of the categories methods and metrics. The function

\(^1\)Fund April 2019, p. 13.
$Z = \text{linkage}(\text{data}, \text{method}, \text{metric})$ performs agglomerative hierarchical clustering by passing the metric to a function called $\text{pdist}$, which computes the distance between the rows of the data. This implies that a transpose of the data used in this thesis is necessary, since the data used is portrayed with the stocks in the columns.\textsuperscript{2}

The function $\text{dendrogram}(Z)$ creates dendrograms for the different linkage method-measure combinations.\textsuperscript{3} The nodes at the bottom of the tree, dendrogram, are called leaf nodes and are numbered from one to $m$, where $m$ is the number of stocks. The leaf nodes are the singleton clusters that all higher clusters are built from. It is possible to chose the number of clusters depicted in the tree and if doing so the leaf nodes will consist of more than one stock each.\textsuperscript{4}

The Calinski-Harabasz index is, as stated in 3.2.4, calculated to find the appropriate number of clusters for the data present. The convenient number of clusters is calculated by MATLAB’s function $\text{eva} = \text{evalclusters}(\text{data},'\text{linkage}',\text{CalinskiHarabasz}')$. The function creates a clustering evaluation object containing data used to evaluate the optimal number of clusters. To be able to use this function the data has to be on the form $N \times P$, where $N$ equals the number of observations and $P$ corresponds to the variables, leading to a need to transform the data.\textsuperscript{5} When the appropriate number of clusters is determined, the dendrogram has to be adapted to the constraint. This is done by limiting the dendrogram function $\text{dendrogram}(Z, \text{NumberOfClusters})$, where $\text{NumberOfClusters}$ define the appropriate number of clusters. This function provides a solution with a chosen number of clusters.

When the dendrogram with the appropriate number of clusters is produced for each combination of linkage methods and distance measures, the formulation of the investment strategy is composed. To be able determine which stocks to include in the portfolio, it is necessary to investigate which of the stocks that are clustered in each of the clusters. Thereafter calculate which of these stocks, cluster-wise, that are the best performing. In order to find out which stocks that are clustered in each cluster, the MATLAB function $\text{find}(\text{T} == \text{k})$, where $k$ is cluster $1, \ldots, n$ decided by the Calinski-Harabasz and $\text{T}$ is a vector containing the leaf node number for each object in the original data set, is used for each cluster\textsuperscript{6}. Secondly, the Sharpe ratio is used for finding the stocks with the highest return and lowest risk in each cluster. In MATLAB the Shape ratio is calculated by the function $\text{sharpe}(\text{x})$, where $\text{x}$ is the input data for the Sharpe ratio\textsuperscript{7}. Afterwards the best performing stocks form the portfolio.

The results provided are used for the composition of an investment strategy and analysed based on expectations from the theory.

\textsuperscript{2}MathWorks 2019c.
\textsuperscript{3}Ibid.
\textsuperscript{4}MathWorks 2019d.
\textsuperscript{5}MathWorks 2019d.
\textsuperscript{6}MathWorks 2019c.
\textsuperscript{7}MathWorks 2019b.
Results

5.1 Stocks

Figure 5.1 depicts an overview of the development of each stock index (a total of 305 stocks) in the data set during time period 2017-10-02 to 2018-10-02, which equals 257 trading days. All stock indices was normalised to start at an index value of 100 at day 0 (by 2017-10-02). Thereafter they developed differently. Some stocks do not seem to follow the development of the other stocks. The most extreme example is the light blue graph at the top of the diagram, represented by stock number 300.

Figure 5.1: Stock index overview
The following figure illustrates the return of each stock index in the data set. The return describes the development of the stock from day to day. A large increase in the stock index from one day to the day after is described as a large y-value, return. In Figure 5.2, it is shown that the light blue stock index increased in value from around day 50 to 51. Similar for the purple stock, which increased in stock index value from approximately day 180 to the day after. A stock’s return with a negative value, imply that the stock index has decreased from one day to another. This is for instance depicted in the figure of the dark blue stock at around day 30 with a return value of -0.2.

![Figure 5.2: Stock return overview](image)

### 5.2 Calinski-Harabasz index

The Calinski-Harabasz index calculates an indication on the optimal number of clusters for the data chosen. With 305 observations and 257 trading days, the Calinski-Harabasz index proposed the optimal number of clusters to be set to two. The result of the Calinski-Harabasz index calculation is shown, both in text and in visual form (Table 5.1 and Figure 5.3). Table 5.1 presents data of both the optimal number of clusters and information about the input data.

<table>
<thead>
<tr>
<th>Number of Observations</th>
<th>305</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inspected K</td>
<td>[1x100 double]</td>
</tr>
<tr>
<td>Criterion Values</td>
<td>[1x100 double]</td>
</tr>
<tr>
<td>Optimal K</td>
<td>2</td>
</tr>
</tbody>
</table>
Figure 5.3 visualises the result of the Calinski-Harabasz index in a graph. The largest value of the graph represents the optimal number of clusters proposed by the Calinski-Harabasz index. The optimal number of clusters is therefore proposed to be two.

Figure 5.3: Calinski-Harabasz index plot
5.3 Results of methods

5.3.1 Ward’s method

Euclidean distance measure

A dendrogram of the clustering result obtained by Ward’s method is shown in Figure 5.4. All of the data observations were clustered with Ward’s method using Euclidean distance measure. At height 0 the 305 stocks are represented individually. The greater the height, the more stocks are clustered together. At the height of 1.4, all of the stocks are clustered in one cluster. At approximately height 0.6, a number of quite distinct clusters can be distinguished.

Figure 5.4: Dendrogram Ward’s method
The dendrogram obtained with Ward’s method, combined with the Euclidean distance measure is divided in nine clusters (chosen by eye), by a cutoff at approximately height 0.6. Each cluster is depicted by a colour shown in Figure 5.5. In Figure 5.6 a more clear picture of the nine clusters is presented.

Figure 5.5: Dendrogram Ward’s method divided in nine clusters
The following figure depicts the Ward’s method’s dendrogram with a cutoff at nine clusters. Cluster number 3 corresponds to the green cluster in Figure 5.5. Cluster number 4 correlates to the purple section, cluster number 1 equals the light blue segment and cluster number 2 correlates to the yellow cluster. Cluster number 8 correlates to the black cluster with one stock between the yellow and red cluster. Cluster number 7 corresponds to the red section and cluster number 5 is equal to the dark blue section. Further, clusters 6 and 9 correspond to the two black clusters to the far right in Figure 5.5, with one stock in each cluster.

Figure 5.6: Dendrogram Ward’s method with a cutoff at nine clusters
The clusters and their corresponding stocks along with the sectors they represent.

Table 5.2: Depicts the cluster stock division – Ward’s method

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Stocks in cluster</th>
<th>Sector present in cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster 1</td>
<td>1 2 4 5 6 9 10 11 14 15</td>
<td>5 Communication Services</td>
</tr>
<tr>
<td></td>
<td>16 17 19 28 29 31 32 34 43 45</td>
<td>11 Consumer Discretionary</td>
</tr>
<tr>
<td></td>
<td>47 48 49 58 66 76 79 96 97 101</td>
<td>1 Consumer Staples</td>
</tr>
<tr>
<td></td>
<td>104 108 110 122 126 127 128 134 138 140</td>
<td>1 Energy</td>
</tr>
<tr>
<td></td>
<td>143 149 153 154 167 171 172 173 176 177</td>
<td>10 Financial</td>
</tr>
<tr>
<td></td>
<td>180 181 189 190 200 202 207 210 212 216</td>
<td>8 Health Care</td>
</tr>
<tr>
<td></td>
<td>218 220 223 225 226 235 236 242 244 247</td>
<td>23 Industrial</td>
</tr>
<tr>
<td></td>
<td>250 257 260 263 265 266 272 274 275 280</td>
<td>11 Information Technology</td>
</tr>
<tr>
<td></td>
<td>281 282 285 286 292 293 298 303 304</td>
<td>19 Materials</td>
</tr>
<tr>
<td>Cluster 2</td>
<td>12 22 25 33 36 40 42 50 57 62</td>
<td>3 Communication Services</td>
</tr>
<tr>
<td></td>
<td>71 81 83 98 102 105 114 120 121 123</td>
<td>16 Consumer Discretionary</td>
</tr>
<tr>
<td></td>
<td>124 145 152 156 157 165 178 179 193 194</td>
<td>1 Financial</td>
</tr>
<tr>
<td></td>
<td>203 205 213 221 222 228 231 243 245 251</td>
<td>1 Health Care</td>
</tr>
<tr>
<td></td>
<td>270 271 276 284 287 288 289 295 301</td>
<td>12 Industrial</td>
</tr>
<tr>
<td></td>
<td>7 Information Technology</td>
<td></td>
</tr>
<tr>
<td></td>
<td>9 Materials</td>
<td></td>
</tr>
<tr>
<td>Cluster 3</td>
<td>7 8 18 20 21 24 26 27 38 41</td>
<td>14 Communication Services</td>
</tr>
<tr>
<td></td>
<td>44 51 52 55 56 60 61 63 64 68</td>
<td>7 Consumer Discretionary</td>
</tr>
<tr>
<td></td>
<td>72 73 74 75 78 80 86 87 89 90</td>
<td>19 Consumer Staples</td>
</tr>
<tr>
<td></td>
<td>91 92 93 95 99 106 111 113 116 119</td>
<td>2 Energy</td>
</tr>
<tr>
<td></td>
<td>129 131 132 133 135 136 137 139 141 142</td>
<td>14 Health Care</td>
</tr>
<tr>
<td></td>
<td>147 148 151 155 158 160 161 162 168 169</td>
<td>20 Industrial</td>
</tr>
<tr>
<td></td>
<td>170 175 183 185 186 187 188 191 192 195</td>
<td>1 Information Technology</td>
</tr>
<tr>
<td></td>
<td>196 197 198 201 208 209 211 214 215 217</td>
<td>13 Real Estate</td>
</tr>
<tr>
<td></td>
<td>219 227 229 230 232 233 234 237 239 240</td>
<td>19 Utilities</td>
</tr>
<tr>
<td></td>
<td>241 246 249 252 253 254 258 259 267 268</td>
<td></td>
</tr>
<tr>
<td></td>
<td>269 277 278 279 290 291 296 297 299</td>
<td></td>
</tr>
<tr>
<td>Cluster 4</td>
<td>164 174</td>
<td>2 Communication Services</td>
</tr>
<tr>
<td>Cluster 5</td>
<td>30 35 163 283</td>
<td>4 Energy</td>
</tr>
<tr>
<td>Cluster 6</td>
<td>300</td>
<td>1 Health Care</td>
</tr>
<tr>
<td>Cluster 7</td>
<td>3 13 23 37 39 46 53 54 59 65</td>
<td>1 Communication Services</td>
</tr>
<tr>
<td></td>
<td>67 69 70 77 82 84 85 88 94 100</td>
<td>6 Energy</td>
</tr>
<tr>
<td></td>
<td>103 107 109 112 115 117 118 125 130 144</td>
<td>41 Financial</td>
</tr>
<tr>
<td></td>
<td>146 150 159 166 182 184 199 204 206 224</td>
<td>1 Industrial</td>
</tr>
<tr>
<td></td>
<td>238 248 255 256 261 262 264 273 294</td>
<td></td>
</tr>
<tr>
<td>Cluster 8</td>
<td>302</td>
<td>1 Health Care</td>
</tr>
<tr>
<td>Cluster 9</td>
<td>305</td>
<td>1 Information Technology</td>
</tr>
</tbody>
</table>
The following tables list the stocks with the highest Sharpe ratio, the best performing stocks in each cluster. These stocks provide support for creating the portfolio. In Table 5.3 the best stock from each cluster is depicted, combined with the sector they belong to.

Table 5.3: Depicts the best stock in each cluster – Ward’s method

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Sharpe Ratio</th>
<th>Stock</th>
<th>Sector</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster 1</td>
<td>47.7501</td>
<td>79</td>
<td>Financial</td>
</tr>
<tr>
<td>Cluster 2</td>
<td>28.9230</td>
<td>114</td>
<td>Consumer Discretionary</td>
</tr>
<tr>
<td>Cluster 3</td>
<td>69.6630</td>
<td>60</td>
<td>Industrial</td>
</tr>
<tr>
<td>Cluster 4</td>
<td>11.6460</td>
<td>174</td>
<td>Communication Services</td>
</tr>
<tr>
<td>Cluster 5</td>
<td>14.6958</td>
<td>35</td>
<td>Energy</td>
</tr>
<tr>
<td>Cluster 6</td>
<td>2.8356</td>
<td>300</td>
<td>Health Care</td>
</tr>
<tr>
<td>Cluster 7</td>
<td>45.8393</td>
<td>3</td>
<td>Financial</td>
</tr>
<tr>
<td>Cluster 8</td>
<td>5.6951</td>
<td>302</td>
<td>Health Care</td>
</tr>
<tr>
<td>Cluster 9</td>
<td>2.1771</td>
<td>305</td>
<td>Information Technology</td>
</tr>
</tbody>
</table>

Table 5.4 depicts the second best stock in the four largest clusters, combined with the sector they belong to.

Table 5.4: Depicts the second best stock in the largest clusters – Ward’s method

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Sharpe Ratio</th>
<th>Stock</th>
<th>Sector</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster 1</td>
<td>43.8727</td>
<td>173</td>
<td>Industrial</td>
</tr>
<tr>
<td>Cluster 2</td>
<td>24.2175</td>
<td>12</td>
<td>Consumer Discretionary</td>
</tr>
<tr>
<td>Cluster 3</td>
<td>37.3467</td>
<td>75</td>
<td>Consumer Staples</td>
</tr>
<tr>
<td>Cluster 7</td>
<td>35.6003</td>
<td>82</td>
<td>Financial</td>
</tr>
</tbody>
</table>

Table 5.5 depicts the stocks included in the final portfolio. The portfolio consists of two stocks, the best and second best stock, from each of the four largest clusters, namely cluster 1, 2, 3 and 7.

Table 5.5: Stocks composing the final portfolio

<table>
<thead>
<tr>
<th>Stock</th>
<th>Sector</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>Financial</td>
</tr>
<tr>
<td>12</td>
<td>Consumer Discretionary</td>
</tr>
<tr>
<td>60</td>
<td>Industrial</td>
</tr>
<tr>
<td>75</td>
<td>Consumer Staples</td>
</tr>
<tr>
<td>79</td>
<td>Financial</td>
</tr>
<tr>
<td>82</td>
<td>Financial</td>
</tr>
<tr>
<td>114</td>
<td>Consumer Discretionary</td>
</tr>
<tr>
<td>173</td>
<td>Industrial</td>
</tr>
</tbody>
</table>
5.3.2 Centroid method

Euclidean distance measure

A dendrogram of the clustering result, obtained by the Centroid method, is shown in Figure 5.7. All of the data observations are clustered with the Centroid method using Euclidean distance measure. Depicted in the graph are the clusters obtained, which are not distinctive.

Figure 5.7: Dendrogram Centroid method
5.3.3 Average linkage method

Euclidean distance measure

A dendrogram of the clustering result, obtained by the Average linkage method, is shown in Figure 5.8. All of the data observations are clustered with the Average linkage method using Euclidean distance measure. Depicted in the graph are the clusters obtained, which are not distinctive.

Figure 5.8: Dendrogram Average linkage method – Euclidean distance
Correlation measure

A dendrogram of the clustering result, obtained by the Correlation measure, is shown in Figure 5.9. All of the data observations are clustered with the Average linkage method using Correlation. The result is a figure with various clusters obtained. To the right in the figure, several small clusters can be distinguished. To the left in the figure, some distinct clusters are depicted.

Figure 5.9: Dendrogram Average linkage method – Correlation measure
Discussion and Analysis

6.1 Mathematical discussion

Comparison of methods

By looking at the dendrograms obtained by the different models (see Figures 5.4, 5.7, 5.8 and 5.9), the conclusion is drawn that one of the dendrograms is significantly different. The dendrogram that is unlike the others is also the one that has the most interpretable result and is obtained by Ward’s method, see Figure 5.4. Further it is the only analysable result since it provides distinct clusters, which the other methods do not do. In the other methods (Figure 5.7, 5.8 and 5.9), several stocks are one by one merged to the first cluster, resulting in only one large cluster and no distinct groups.

The Centroid method combined with the Euclidean distance and the Average linkage method, combined with both Correlation and Euclidean distance, fit poorly with the data used. The Average linkage method with Correlation distance measure (see Figure 5.9) does provide some distinct clusters but the stocks are clustered one by one to one large cluster at a large height. Therefore, the dendrogram cannot be divided into distinct clusters. This due to that a cutoff will result in some decent clusters (shown to the left in Figure 5.9) and several clusters with only one stock in each cluster (shown to the right in Figure 5.9). The amount of data points in each cluster will not be enough to make a sufficient analysis. The Centroid method combined with the Euclidean distance and the Average linkage method combined with the Euclidean distance does not provide distinct clusters (see Figures 5.7 and 5.8). A cutoff of the dendrograms will result in several clusters with only one stock in each cluster, and the rest of the stocks in one large cluster.

Ward’s method generates a dendrogram (see Figure 5.4) that is easily interpreted compared to the other methods. The dendrogram depicts distinct clusters, which is necessary for performing an analysis. The data obtained from Ward’s method is referred to as actionable data, which is data that offers strategic information possible for the user to act on.

According to theory, there are no indications that a cluster analysis on a data set of stock indices would not be possible to execute. A number of explanations on why all of the methods used did not provide analysable results could be considered. However, it is difficult to explicitly determine why it did not work this time. The answer might be found in the combination of the data, the methods, the amount of data or something completely different. Because of the different approaches of the methods, they do have different segments where they are preferably used. The reason for obtaining different results from each method could as such be because...
of the different focus of and calculations used within each method. Further, the calculations result in each method having different main aims and are not equally qualified for all types of data. This indicating that the methods might be suitable for other types of financial data. Other sectors or other time periods could for instance give other analysable results.

The investment strategy is based on finding stocks with minimum risk and the maximum return in each cluster. Each of the clusters has to contain several stocks to be able to perform this analysis. Therefore, the conclusion is that Ward’s method is most fitted for the thesis purpose. It is not obvious though that Ward’s method is the most suitable clustering method for the data used or the problem stated overall, but it is the only method that generates a result valid for analysis in this thesis. The result should be evaluated with some restraint regarding the validity of the results obtained. It is also possible that some other form of cluster analysis, such as non-hierarchical clustering, is a more suitable approach to this problem. However, it is essential to remember that there is no specific practise to follow on how to conduct a cluster analysis. Therefore the "right" answer can vary from time to time depending on several factors like the type of data or method used.

Choice of number of clusters

The appropriate number of clusters for Ward’s method is, in this case, chosen to nine (see Figure 5.5). The choice of dividing the data set into nine clusters is based on both the height of fusion and an estimate of a reasonable number of clusters by looking at the graph. The Calinski-Harabasz index indicates that the optimal number of clusters is supposed to be two (see Table 5.1 and Figure 5.3). The graph in Figure 5.3 is decreasing exponentially, meaning that it cannot have a maximum value that is not the first value (in this case 2). This indicates that the Calinski-Harabasz index is not an appropriate method for the data used.

Since most of the methods was not appropriate for the data used, and by the look of the Ward’s method dendrogram (see Figure 5.4), the result of Calinski-Harabasz index was not considered appropriate for the aim of this thesis. The decision then followed to choose the number of nine clusters. The choice of the number of clusters is not entirely scientific and should not be seen as the only, or definitely correct partition. By evaluating the clusters, there may be indications that the choice of nine clusters is not the optimal number. This opens up to critic of the analysis since there does not exist a scientific way of determining the appropriate number of clusters as further confirmed by the results of the different methods in this thesis.

Cluster stock division

In Table 5.2 the cluster stock division is presented. Clusters 4, 5, 6, 8 and 9 are made up by only one to four stocks. Those clusters will therefore be difficult to consider when composing the investment strategy. This because it is difficult to know how to analyse or use such small groups of stocks for the formulation of an investment strategy. Looking at the Sharpe ratios (see Table 5.3), the values are not very high for the clusters with only a few stocks (cluster 4, 5, 6, 8 and 9). Therefore, they might not be appropriate to include in the portfolio. It is though, important to examine these stocks further to see, if it regardless of the size of the cluster, would be profitable to include the stocks in the portfolio.
Clusters 6, 8 and 9 are only composed of one stock each. The reason that these clusters only include one stock, is that the stocks in those clusters (stock number 300, 302 and 305) do not follow any other stock’s movements during the days examined. The stocks 300, 302 and 305 could be viewed as outliers. Stock 300 in cluster 6 has a significant growth compared to the other cluster’s stocks. Stock 305 in cluster 9 has a steady zero return for the first $\sim 175$ days, indicating that the stock was introduced on the market at approximately day 175, which explains the lack of similarity with the other stocks. The late introduction also indicates that the stock should not be taken into account for the formulation of the investment strategy since the duration of the data is too short to provide a reliable result. Stock 302 in cluster 8 is the stock with the most decreasing index during the analysed period. It significantly decreases at multiple times compared to the other stocks, making it clustered to itself.

In cluster 4 and 5 there are more than one but no more than four stock per cluster. Interesting in these cases are that the stocks in the different clusters origin from the same sector. As for the single stock clusters, it is difficult to make a valid analysis based on these few stocks. For cluster 4 and 5 the largest Sharpe ratios are somewhat higher than for the single stock clusters (see Table 5.3), indicating that the stocks in cluster 4 and 5 have better risk versus return. However, cluster 4 and 5 need to be evaluated further before deciding if they should be considered when composing the portfolio.

Cluster 3 and 4 are fused at a height, not much higher than the nine cluster division (see Figure 5.6), indicating that cluster 3 and 4 might be better of as one combined cluster. The sector of the stocks in cluster 4 (only from sector Communication Services), is also represented in a relatively large proportion in cluster 3, adding to the indication (see Table 5.2). In cluster 5, it is more difficult to conduct why the stocks are clustered the way they are. It can however be noted that cluster 5 contains four stocks from the energy sector, a proportion making the clustering seem a bit more understandable since it in total are 13 stocks from the energy sector in the data set.

Cluster 1, 2, 3 and 7 contain the majority of the stocks. As the theory states, the probability that stocks from the same sectors, countries, etc. would be clustered together is high. In cluster 7 there are 41 stocks from the financial sector and only 8 from other sectors. In cluster 1, the sectors are dominated by the industrial and material sectors, but there are some other sectors present as well. Cluster 2 is dominated by the consumer discretionary sector as well as the industrial sector. Cluster 3 is probably the most diverse cluster as there are six sectors with more than 10 stocks each present in the cluster.

The distribution of stocks in the clusters somewhat coincides with theory. Even the clusters that have multiple sectors present, do not have every sector present and some of the sectors can nearly all be found in one or two different clusters. These four large clusters will be central when formulating the investment strategy.
6.2 Economic discussion

Investment strategy

It is worth taken into consideration that the mathematical result can be complemented by an economical analysis to create a more elaborate conclusion. This because of that some uncertainties in the financial data are not taken into consideration in the mathematical computations. The economical aspect includes a parallel aspect of the problem, focusing more on the combination of the data concerning social and economical aspects. When composing an investment strategy based on the nine clusters obtained (see Figure 5.6), an important factor to consider is if the size of the cluster has an impact on the cluster’s role in the investment strategy.

An investment strategy can be assembled in several ways. To get a diversified portfolio, that is not exposed to a known high risk, it is preferable not to include the stocks in the smaller clusters in the investment strategy (namely cluster 4, 5, 6, 8 and 9). The investment strategy is therefore based on cluster 1, 2, 3 and 7. This decision is mainly due to the fact that it is impossible to, in advance, determine how the market will evolve. Including the single stocks expose the portfolio to a higher risk. The stocks with the highest Sharpe ratio in each cluster (see Table 5.3), are preferred to invest in. Is it enough to have a portfolio consisting of one stock from each cluster? There are various aspects for deciding how to do this – the portfolio could be created of the two best performing stocks in each cluster (based on the Sharpe ratio), include a percentage of stocks from the clusters dependent on the size of the clusters or something else. The more stocks in a portfolio indicate a lower unsystematic risk exposure. A portfolio obtains higher diversity if various sectors are included in the portfolio. On the other hand transaction costs can add up if an investor holds several stocks, which argues that the number of stocks should not be too large.¹

A preferable strategy for the results in this thesis, is to use the four largest clusters and use the best performing stock in each cluster. Since these clusters include a large amount of stocks from various sectors it will not be possible, or even wished for, to include a stock from every sector. Furthermore, it might be preferable to ensure that a stock from the largest sector in each cluster is represented in the portfolio. The best performing stock in cluster 2, 3 and 7 is also the sector that is most represented in the cluster (see Table 5.3). However in cluster 1, the best performing stock is not in the largest sector in that cluster (see Table 5.3). Since all of the largest sectors in all clusters are not covered and the portfolio only consists of four stocks, it was decided to also include the second best stock in each of the largest clusters (see Table 5.4). When including the second best performing stock in each cluster, the most represented sector in cluster 1 will as well be included in the investment strategy. This results in that the most represented sector in each of the largest clusters is included in the portfolio. Including the second best performing stock in the portfolio, is also beneficial as a portfolio composed of only four stocks is risky and does not include enough diversity to cover the risk factor. The portfolio and thereby the investment strategy will finally be composed by stock 3, 12, 60, 75, 79, 82, 114 and 173 which contain three stocks from the financial sector, two stocks from the consumer discretionary sector, two stocks from the industrial sector and

¹Staff 2019.
one stock from the consumer staple sector (see Table 5.5).

The created investment strategy will, if used, hopefully generate a steady positive return, although not very high since the risk is minimised simultaneously. As the strategy is based on one year’s data, this investment strategy is short–lived. The obtained strategy cannot be used for a long time before new calculations and assessments have to be performed. This is done in order to see if the current investment strategy is still optimal. If data for a longer period of time would have been used for the analysis, the investment strategy would probably be valid longer. However, as markets change new stocks are introduced, removed and other external factors impact the result which lead to that the investment strategy needs to be modified continuously. The financial market fluctuates and the conditions often change, also creating a need for a continuous change in the investment strategies. The data used in the analysis is approximately six months old and the portfolio derived is therefore already outdated. Since the world economy was unstable for a large period of time, when obtaining the data, and has now stabilised in some extent, it might be even more difficult to use the investment strategy for a longer duration. The methods used could though be assessed continuously for recently obtained data and the investment strategy could be altered thereafter.

According the discussion, that the best and second best performing stocks in the four largest clusters should be included in the investment strategy is strengthened by theory. It is possible that the investment strategy could give indications on which stocks to invest in. However, since there are some uncertain factors the model should be tested further, in order to provide a more precise investment strategy.
Further Research

Of the four method-measure combinations tested, only one resulted in an interpretable result namely Ward’s method. However, that does not necessarily mean that it is the best method for performing cluster analysis on the type of data used in this thesis. Further studies should be conducted in order to determine the method that provides the most reliable result. It is also interesting to further investigate the reasons why other methods do not provide analysable results.

The choice of the appropriate number of clusters could be evaluated further. The method of estimation of the number of clusters by intuition and by looking at the dendrogram is not scientific. It does indeed let the human bias influence the creation of the investment strategies in a way that the clustering aims to avoid. There might be another method to determine the appropriate number of clusters, contrary to the Calinski-Harabasz index, which is more suitable for the type of data provided.

To make sure that the investment strategy obtained is reliable, a suggestion might be to consider different and modified data sets both considering the stocks used and the dates included in the analysis. An example is to use the data but modify it by excluding the last or first day. It might provide a totally different result and should then be taken into consideration when composing the investment strategy.

In order to consider taking a risk and investing in a portfolio obtained from this method of cluster analysis, some testing on previous data should be performed. Portfolios obtained should be tested on a time period after the period used for composing the portfolio. For this thesis the test could, for instance be on the time period 2018-10-02 to 2019-05-02. This to assure that the methods for composing the investment strategies will indeed perform as the analyses indicate.
Conclusion

The method of cluster analysis cannot provide only one correct answer since there are many ways of conducting the analysis and interpreting the results. To be able to determine if cluster analysis is a suitable method for composing investment strategies, further analysis has to be conducted.

The produced investment strategy is composed by the stocks 3, 12, 60, 75, 79, 82, 114 and 173. The portfolio contains three stocks from the financial sector, two stocks from the consumer discretionary sector, two stocks from the industrial sector and one stock from the consumer staple sector.

This investment strategy is already outdated since the financial market is in continuous change, creating a need for constant modification of the strategy. Due to the fact that only one method gave an analysable result, this investment strategy could not be verified and should be regarded with caution. Further, deciding the number of clusters was not scientific, making the result even more unreliable. The investment strategy is affected by the state of the economy, which can be difficult to take into consideration when performing the computations. The more data used, the more reliable results are obtained. If further studies are conducted in this area, an investment strategy created by a cluster analysis, will probably be functioning and provide a small but positive return.
References


[22] Investopedia Staff. 2019. URL: [https://www.investopedia.com/ask/answers/05/optimalportfoliosize.asp](https://www.investopedia.com/ask/answers/05/optimalportfoliosize.asp) (visited on 04/26/2019).

Appendix

MATLAB code

close all; clear all; clc;

data1 = xlsread('inputdata.xlsx);
data1(:, 1) = []; % remove the first column containing the dates of the days.

numberDays = length(data1(:,1)); % number of rows
returnVector = [];
data = [];

for i = 1:length(data1) % number of stocks
    returnVector = [];
    for t = 1:numberDays-1;
        ret = data1(t+1,i)/data1(t,i) - 1; % equation for the return
        returnVector = [returnVector ; ret]; % a vector of each stock’s return
    end
    data = [data, returnVector]; % a matrix of all stock’s and their returns.
end

figure;
plot(data)
title('Stock return overview')
xlabel('Days')
ylabel('Index')

figure;
plot(data1)
title('Stock index overview')
xlabel('Days')
ylabel('Index')
hold on

%%
ClusterNumbers = evalclusters(data,'linkage','calinskiharabasz',
'KList',[1:100])
% kList states the range of the number of clusters reasonable.
figure;
plot(ClusterNumbers)
Ward's method & Euclidean

rng('default')
Z = linkage(data', 'ward', 'euclidean');

cutoff = 0.605; % The height that limit the dendrogram at the appropriate numbers of clusters
dendrogram(Z, 0, 'ColorThreshold', cutoff) % Creates the dendrogram
title('Clustering Ward\'s method')
xlabel('Stock number')
ylabel('Height')
hold on

[H, T, outperm] = dendrogram(Z, 9, 'ColorThreshold', cutoff);
title('Cutoff at nine clusters')
xlabel('Cluster number')
ylabel('Height')

for i = 1:9
    StocksInCluster = find(T == i);
    SharpeRatio = [];
    SharpeRatio = sharpe(data1(:, StocksInCluster)); % Calculates all sharpe ratios for cluster i
    [Max, Index] = max(SharpeRatio); % Find the highest sharp ratio in cluster i and its corresponding index.
    Max
    StockNummer = StocksInCluster(Index)
end

Centroid & Euclidean

rng('default')
Z = linkage(data', 'centroid', 'euclidean');

cutoff = 200; % Limit the dendrogram
dendrogram(Z, 0, 'ColorThreshold', cutoff) % Creates the dendrogram
title('Clustering centroid method')
xlabel('Stock number')
ylabel('Height')
hold on

Average Linkage & Euclidean

rng('default')
Z = linkage(data', 'average', 'euclidean');
cutoff = 200; % Limit the dendrogram
dendrogram(Z,0,'ColorThreshold',cutoff) % Creates the dendrogram
title('Clustering average linkage method, euclidean')
xlabel('Stock number')
ylabel('Height')
hold on

%%
Average linkage & Correlation

rng('default')
Z = linkage(data','average','correlation');

cutoff = 200; % Limit the dendrogram
dendrogram(Z,0,'ColorThreshold',cutoff) % Creates the dendrogram
title('Clustering average linkage method, correlation')
xlabel('Stock number')
ylabel('Height')
hold on
## Table of stock division

<table>
<thead>
<tr>
<th>Sector</th>
<th>Number of stocks</th>
<th>Stocks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Communication Services</td>
<td>25</td>
<td>8 19 26 27 48 51 56 64 95 143 148</td>
</tr>
<tr>
<td></td>
<td></td>
<td>162 164 174 179 180 196 206 215 228</td>
</tr>
<tr>
<td></td>
<td></td>
<td>232 258 269 285 288</td>
</tr>
<tr>
<td>Consumer Discretionary</td>
<td>34</td>
<td>9 12 22 28 42 43 50 57 58 61</td>
</tr>
<tr>
<td></td>
<td></td>
<td>66 68 83 92 102 105 108 114 132 134</td>
</tr>
<tr>
<td></td>
<td></td>
<td>205 207 208 222 230 243 247 251 253</td>
</tr>
<tr>
<td></td>
<td></td>
<td>266 271 276 289 292</td>
</tr>
<tr>
<td>Consumer Staples</td>
<td>20</td>
<td>10 21 44 52 63 74 75 106 111 113</td>
</tr>
<tr>
<td></td>
<td></td>
<td>116 129 135 195 209 211 214 217 234 291</td>
</tr>
<tr>
<td>Energy</td>
<td>13</td>
<td>13 30 35 46 53 84 88 137 142</td>
</tr>
<tr>
<td></td>
<td></td>
<td>163 184 189 283</td>
</tr>
<tr>
<td>Financial</td>
<td>52</td>
<td>3 23 37 39 45 54 59 65 67 69 70</td>
</tr>
<tr>
<td></td>
<td></td>
<td>77 79 82 94 100 101 103 107 109 112 115</td>
</tr>
<tr>
<td></td>
<td></td>
<td>117 118 125 130 144 146 149 150 159 166</td>
</tr>
<tr>
<td></td>
<td></td>
<td>176 182 190 199 202 204 221 224 238 248</td>
</tr>
<tr>
<td></td>
<td></td>
<td>250 255 256 260 261 262 264 273 282 294</td>
</tr>
<tr>
<td>Health Care</td>
<td>25</td>
<td>16 24 38 49 55 86 120 139 183 201</td>
</tr>
<tr>
<td></td>
<td></td>
<td>219 237 240 242 244 246 259 263 268</td>
</tr>
<tr>
<td></td>
<td></td>
<td>272 280 297 300 302 304</td>
</tr>
<tr>
<td>Industrial</td>
<td>56</td>
<td>5 11 14 15 17 18 29 33 34 40 47 60</td>
</tr>
<tr>
<td></td>
<td></td>
<td>62 73 81 85 90 98 99 119 127 131</td>
</tr>
<tr>
<td></td>
<td></td>
<td>133 138 151 152 158 160 161 165 168 170</td>
</tr>
<tr>
<td></td>
<td></td>
<td>171 172 173 186 187 188 192 193 198 200</td>
</tr>
<tr>
<td></td>
<td></td>
<td>210 216 223 225 231 236 245 249 257 270</td>
</tr>
<tr>
<td></td>
<td></td>
<td>274 281 284 293</td>
</tr>
<tr>
<td>Information Technology</td>
<td>21</td>
<td>4 31 36 71 96 110 122 123 124 156 212</td>
</tr>
<tr>
<td></td>
<td></td>
<td>213 218 233 265 286 295 298 301 303 305</td>
</tr>
<tr>
<td>Materials</td>
<td>27</td>
<td>1 2 6 25 32 76 97 104 121 126</td>
</tr>
<tr>
<td></td>
<td></td>
<td>128 140 145 153 154 157 167 177 178</td>
</tr>
<tr>
<td></td>
<td></td>
<td>181 194 203 220 226 235 275 287</td>
</tr>
<tr>
<td>Real Estate</td>
<td>13</td>
<td>78 141 185 197 229 239</td>
</tr>
<tr>
<td></td>
<td></td>
<td>241 252 254 267 290 296 299</td>
</tr>
<tr>
<td>Utilities</td>
<td>19</td>
<td>7 20 41 72 80 87 89 91 93 136</td>
</tr>
<tr>
<td></td>
<td></td>
<td>147 155 169 175 191 227 277 278 279</td>
</tr>
</tbody>
</table>