On Cluster Analysis of Some Portfolio Optimization

MS Thesis

by

Muhammad Bilal

CIIT/SP22-RMT-009/LHR

COMSATS University Islamabad
Pakistan

Fall 2023

On Cluster Analysis of Some Portfolio Optimization

A thesis submitted to
COMSATS University Islamabad

In partial fulfillment
of the requirement for the degree of

Master of Science
in
Mathematics

by
Muhammad Bilal

CIIT/SP22-RMT-009/LHR

Department of Mathematics
Faculty of Science

COMSATS University Islamabad
Pakistan

Fall 2023

On Cluster Analysis of Some Portfolio Optimization

This thesis is submitted to the department of Mathematics in partial fulfillment of the requirement for the award of degree of Master of Science in Mathematics

Name	Registration Number
Muhammad Bilal	CIIT/SP22-RMT-009/LHR

Supervisor

Dr. Sarfraz Ahmad

Professor,
Department of Mathematics

COMSATS University Islamabad, Lahore Campus

January, 2024

Certificate of Approval

This thesis titled

On Cluster Analysis of Some Portfolio Optimization

By

Muhammad Bilal

CIIT/SP22-RMT-009/LHR

Has been approved

For COMSATS University Islamabad, Lahore Campus.

External Examiner:_____

Prof. Dr. Shabnam Malik

Department of Mathematics

Forman Christian College (A Chartered University) Lahore

Supervisor:_____

Prof. Dr. Sarfraz Ahmad

Department of Mathematics, (CUI) Lahore Campus

Head of Department:_____

Prof. Dr. Muhammad Hussain

Department of Mathematics, (CUI) Lahore Campus

Author's Declaration

I Muhammad Bilal, CIIT/SP22-RMT-009/LHR, hereby declare that I have produced the work presented in this thesis, during the scheduled period of study. I also declare that I have not taken any material from any source except referred to wherever due to that amount of plagiarism is within an acceptable range. If a violation of HEC rules on research has occurred in this thesis, I shall be liable to punishable action under the plagiarism rules of HEC.

Date:_____ _____

Muhammad Bilal

CIIT/SP22-RMT-009/LHR

Certificate

It is certified that Muhammad Bilal, CIIT/SP22-RMT-009/LHR has carried out all the work related to this thesis under my supervision at the Department of Mathematics, COMSATS University Islamabad, Lahore Campus and the work fulfills the requirement for award of MS degree.

Date:_____

Supervisor

_____

Dr. Sarfraz Ahmad
Professor,
Mathematics,
COMSATS University Islamabad
Lahore Campus

Dedication

To My Parents and All Family

Acknowledgements

Praise to be ALLAH, the Cherisher and Lord
of the World, Most gracious and Most Merciful

First and foremost, I would like to thank ALLAH Almighty (the most beneficent and most merciful) for giving me the strength, knowledge, ability and opportunity to undertake this research study and to preserve and complete it satisfactorily. Without countless blessing of ALLAH Almighty, this achievement would not have been possible. May His peace and blessings be upon His messenger Hazrat Muhammad (PBUH), upon his family, companions and whoever follows him. My insightful gratitude to Hazrat Muhammad (PBUH) Who is forever a track of guidance and knowledge for humanity as a whole. In my journey towards this degree, I have found a teacher, an inspiration, a role model and a pillar of support in my life, my kind.

Muhammad Bilal
CIIT/SP22-RMT-009/LHR

Abstract

On Cluster Analysis of Some Portfolio Optimization

By

Muhammad Bilal

The classic mean-variance portfolio optimization approach is criticized in large part for its propensity to overstate estimate error. An estimated inaccuracy of a few percent can skew the entire package. The Black-Litterman technique (Bayesian method) and the resampling method are two common approaches to solving this problem. A more recent approach to the issue's solution is the clustering method. By clustering, we initially combine the stocks that have a strong correlation and handle the group as a single stock. Following the grouping of the stocks, we will have a few stock clusters. For these clusters, we do the standard mean-variance portfolio optimization. By using the clustering approach, the influence of estimating error may be minimized and the portfolio's stability can be increased. In this project, we'll examine how it functions and run experiments to see if clustering techniques enhance the portfolio's performance and stabilities.

Table of Contents

List of Figures

List of Tables

Chapter 1
Introduction

Individuals and businesses use financial investments throughout the world. A financial investment business that manages investor cash is one example of an organization that makes investments for a variety of reasons. As a result, the players conduct their investments using varying requirements, objectives, and strategies. Both simple techniques and sophisticated algorithms can be used to generate investment strategies. All investors, nevertheless, share the same goal of obtaining a high return with little risk. It is more difficult to achieve such a portfolio.

Analyzing data is the initial stage in developing an investment plan. Financial data, which is frequently gathered from day-to-day activities, must typically be processed to make it easier to handle and analyze. In addition, the data must be analyzed using one or more methodologies, taking into account some factors. Risk is one thing that needs to be taken into account. While some performers can handle it well, others can't stand it. As previously said, the goal is to strike the ideal balance between risk and reward. There are several methods for determining and managing risk. Diversification is a well-known and often utilized strategy for managing portfolio optimization. Therefore, by building more diverse portfolios, the risk can be decreased. By incorporating riskier data from several sources into a portfolio, diversification ultimately aims to lower portfolio volatility.

Theoretically, a cluster analysis of financial assets, including bonds, futures, and stocks, can yield a diversified portfolio. Cluster analysis is applied in many fields, and prior studies suggest that it may also be suitable for financial objectives. Disorganized data is systematized into subsets of groups, or clusters, using a cluster analysis. The idea behind cluster analysis is that while the distance between clusters should be considerable, the difference, or more particularly, the gap between stocks inside a cluster, should be as minimal as feasible. Therefore, the goal of cluster analysis is to identify correlations within the data. Clusters of related data sets are formed by grouping the most comparable data. Investment plans can then be developed using these clusters. To identify a trustworthy technique,

some alternative cluster analysis techniques will be contrasted and examined. The investment strategy will be theoretically reviewed and appraised, and the cluster analysis will be assessed using data sets that Nordea has given.

One of the most serious problems in asset management is the portfolio optimization problem [12]. Numerous additional research studies have focused on various elements of portfolio optimization from both a theoretical and applied perspective since Markowitz's groundbreaking work [37]. Here we focus our attention on the role of the correlation coefficient matrix in portfolio optimization. Because the asset return time series has a finite duration, there is inevitably a statistical uncertainty associated with the correlation matrix calculation. Recently, quantifying the level of statistical uncertainty in a correlation matrix has been the subject of several contributions in the econophysics literature. A method to filter the information in the correlation coefficient matrix that is robust for the inevitable statistical uncertainty has recently been developed using the RMT quantification of the statistical uncertainty associated with the estimation of the correlation coefficient matrix of a finite multivariate time series (in the econophysics literature it has been used as noise dressing)[1]. The process of filtering has yielded correlation matrices, which have been employed in portfolio optimization. According to some research, the difference between the realized and anticipated optimum portfolios, assuming flawless forecasting of future returns and volatilities, is less for the filtered correlation matrix than for the original one, given a specific level of portfolio return.

Other correlation coefficient matrix filtering techniques, carried out using correlation-based clustering techniques, have also been presented in the econophysics literature in recent years. The correlation coefficient matrix, which is indicative of the complete matrix and frequently less impacted by statistical uncertainty and thus more stable than the entire matrix during the system's temporal development, is another set of information selected by these approaches [17, 29, 44, 30, 19, 11, 48, 43, 40, 47, 20].

In this thesis, we examine how various filtering techniques applied to the correlation coefficient matrix affect the portfolio optimization process. We specifically take into account correlation-based clustering processes and RMT-based filtering procedures. The structure of this thesis is outlined as follows.

In Chapter 2, we have given an overview of descriptive statistics, which are useful for summarising data. To start, we defined the many kinds of data that may be found, discussed why data collecting is necessary, and listed some popular sources. In Chapter 3, we provide a concise overview of the mean-variance optimization problem, laying out the notation used and summarizing the challenge related to estimating the correlation matrix. In Chapter 4, we describe the clustering algorithms used to perform the portfolio optimization. These algorithms are average linkage and single linkage. we describe two methods based on these clustering algorithms to build asset portfolios that are robust and reliable.

Chapter 2
Basic Definitions

In this chapter, We have given an overview of descriptive statistics, which are useful for summarising data. We started by outlining the reasons for data gathering, characterizing the many kinds of data that may be encountered, and offering a few standard resources for data. We defined distribution and described how to create cumulative, relative, and frequency distributions for data. We also showed how to graphically represent the data distribution using histograms. Then, we discussed location metrics for data distribution, such as mean, median, mode, and geometric mean, and measures of variability, such as range, variance, standard deviation, and coefficient of variation. Measures of the two variables' connection were discussed. The relationship between variables may be seen via a scatter plot. A single number can represent the linear relationship between variables: covariance and correlation coefficient.

2.1 Types of Data

We will discuss some categories of data. In statistics and data analysis data can be classified into many categories according to their nature, attributes, and methods of collection or representation. Among the main categories of data are:

2.1.1 Data

Definition 1. *Information, facts, or observations that are gathered, documented, or displayed for examination, reference, or deduction are referred to as data.*
It can take on several forms, including as text, numbers, pictures, sounds, or any other interpretable or processable format.
Data provides the basis for inference, insight generation, and well-informed decision-making in a variety of domains, including business, science, research, and daily life. Based on its attributes, It may be divided into a variety of categories, such as cross-sectional data, time series data, qualitative or quantitative data, population data, sample data, and more.

2.1.2 Population Data

Definition 2. *The term "population data" describes the entire group of people, things, or components that have particular qualities and are relevant to a particular investigation or analysis. It speaks for the whole group under investigation.*
For instance, all adult men in a nation would make up the population if you were researching the average height of all adult males in that nation.
Another instance may be the complete group of clients who have purchased a certain item from a business.
In statistics and research, an understanding of the population data is crucial since it serves as the foundation for inferences, forecasts, and generalizations on the features of the wider group under investigation.

2.1.3 Sample Data

Definition 3. *A subset of data chosen from a broader dataset or population is referred to as sample data. This smaller set was selected such that it accurately reflects the traits or qualities of the total population.*
Making assumptions or judgments about the larger population using sampling information allows for conclusion-making without requiring full dataset analysis.
For instance, Consider a situation where you need to find the average age of every worker at a firm that employs thousands of people. Rather than gathering age data from each employee, you may choose to sample, say, 200 people.
This sample was selected to serve as a miniature representation of the company's personnel, taking into account different departments, job titles, and levels of experience. You may estimate or conclude the average age of all corporate employees by examining the age data from this sample.
For any conclusions obtained from the sample to be appropriately applied to the wider group, the sample data should ideally be typical of the population. Selecting samples that most closely reflect the characteristics of the population involves using a variety of sampling approaches, such as stratified sampling or random sampling.

2.1.4 Quantitative Data

Definition 4. *Quantitative data refers to information that is expressed in numerical terms and can be measured or counted. This kind of data can be expressed as numerical values and deals with numbers or quantities.*

It is frequently used to carry out mathematical computations, comparisons, and analyses in statistical analysis and research.

Examples of quantitative data include:

1. ***Measurements***

(a). *Height: 165 centimeters*

(b). *Weight: 68 kilograms*

(c). *Temperature: 25 degrees Celsius*

2. ***Counts***

(a). *Number of books in a library: 500*

(b). *Daily sales revenue: $2,500*

(c). *Population of a city: 1,000,000*

Quantitative data can be further categorized into two subtypes

Discrete Data:

Discrete data is made up of countable whole numbers.

For instance, the number of automobiles in a parking lot or the number of pupils in a classroom.

Continuous Data:

Measurements that fall within a certain range might be considered continuous data.

For instance,

height, weight, temperature, and time.

Quantitative data allows for precise mathematical operations, statistical analyses, and graphical representations, aiding in the understanding and interpretation of numerical patterns, relationships, and trends within the data.

2.1.5 Categorical Data

Definition 5. *Qualitative or categorical data also referred to as characteristic or attribute data, depicts traits or attributes that fall into particular groupings or categories. Categorical data is made up of non-numeric information and is often descriptive, in contrast to quantitative data, which includes numerical values.*
It is used to group and classify observations according to characteristics or labels.
Examples of categorical data include:

1. ***Nominal Data:***

(a). *Colors: green, blue, and red.*

(b). *Relationship Status: Not married, married, divorced*

2. ***Ordinal Data:***

(a). *Educational Levels: Bachelor's degree, master's degree, and high school*

(b). *Survey Responses: Poor, fair, good, excellent*

Categorical data is often utilized in a variety of sectors, including the social sciences, market research, and demography.

2.1.6 Cross-Sectional Data

Definition 6. *Information gathered from various people, entities, or subjects at one moment or over a predetermined period is referred to as cross-sectional data.*
It takes a momentary snapshot of the observations and does not follow changes or trends for the same subjects over time.
For instance, Consider carrying out a family income survey in a specific city in January 2023. Cross-sectional data is what was gathered at this particular period from different homes. The income of each family is recorded at that specific instant, giving a snapshot of the income distribution among various families in that city at that specific moment.
Another example could be a study that collects data on various companies' earnings, costs, and workforce size within a given industry sector on a given date, without accounting for changes over time.

Cross-sectional data is useful for examining traits, distinctions, or connections between various people or things at a certain moment in time. On the other hand, it misses changes or advancements over time for the same individuals, This may be examined using time-series or longitudinal data that track changes over a certain amount of time.

2.1.7 Time Series Data

Definition 7. *A time series is a collection of measurements or observations that are made, recorded, or observed throughout time at regular intervals that are evenly spaced apart. This type of analysis uses chronologically ordered data points, often at regular intervals, to monitor changes, trends, or patterns across time in a particular event or variable.*
Examples of time series data include:
1. ***Stock Prices:*** *Daily closing prices of a company's stock recorded over several months or years.*
2. ***Temperature Readings:*** *Hourly temperature measurements recorded over a year at a weather station.*
3. ***Sales Figures:*** *Monthly sales data for a product recorded over several years.*
4. ***Population Growth:*** *Yearly census data tracking population changes in a city over decades.*
Time series data analysis involves examining, modeling, and forecasting the behavior of the variable being measured over time. Statistical techniques and methods such as trend analysis, and moving averages, are commonly used to analyze and make predictions based on time series data.

2.2 CREATING DISTRIBUTIONS FROM DATA

Distributions, which indicate the frequency with which particular values for a variable occur in a data collection, aid in summarising several features of a data set. For both quantitative and categorical data, distributions may be generated, and they help the analyst measure

variance.

2.2.1 Frequency Distributions for Categorical Data

For categorical data, frequency distributions show the number or percentage of observations that fit into various categories. The distribution of qualitative or categorical variables may be summarised and understood using this distribution, which displays the frequency or occurrence of each category within a dataset.

Procedure for producing a frequency distribution for classification data:

1. **List Categories:** Identify and list all distinct categories or groups present in the dataset.
2. **Count Frequencies:** Count the number of occurrences of each category in the dataset.
3. **Create a Table:** Create a table or list that displays the categories along with their corresponding frequencies or counts.

Soft Drink	**Frequency**
Coca-Cola	19
Diet Coke	8
Dr. Pepper	5
Pepsi	13
Sprite	5
Total	$\sum f = 50$

Frequency distributions help in summarizing categorical data, identifying the most common categories, and understanding the distribution or variation among different groups.

Visual representations such as bar charts or pie charts are often used to present frequency distributions graphically, providing a clear visualization of the distribution of categorical data.

2.2.2 Relative Frequency and Percent Frequency Distributions

Frequency distribution variants that are used to describe categorical data include percent and relative frequency distributions. Along with the counts or frequencies of the categories, they also provide the percentage or proportion of each category to the overall number of observations, which adds more information.

Relative Frequency Distribution: The percentage or share of observations in each category to the total number of observations in the dataset is shown by this distribution. By dividing the frequency of every category by the total number of observations, it is computed.

Percent Frequency Distribution: The percentage of observations in each category as a percentage of all observations is shown in this distribution. The percentage of observations in each category as a percentage of all observations is shown in this distribution.

Example 1. *consider the same dataset of survey responses on preferred modes of transportation:*

Frequency Distribution:

Mode of Transportation	Frequency
Car	5
Bus	2
Train	2
Bicycle	1
Total	$\sum f = 10$

Relative Frequency Distribution:

Mode of Transportation	Relative Frequency
Car	5/10 = 0.5
Bus	2/10 = 0.2
Train	2/10 = 0.2
Bicycle	1/10 = 0.1

Percent Frequency Distribution:

Mode of Transportation	Percent Frequency
Car	50%
Bus	20%
Train	20%
Bicycle	10%

To facilitate comparisons and interpretations, relative frequency and percent frequency distributions provide a more lucid picture of the percentage of each category in the dataset.

They are especially helpful for displaying the relative weight or frequency of various dataset types.

2.2.3 Frequency Distributions for Quantitative Data

Distributions of frequencies in quantitative data organize numerical values into intervals or ranges and display the count or frequency of observations falling within each interval. This distribution allows you to summarize and understand the distribution of numerical data by grouping values into classes or bins.

Steps to create a frequency distribution for quantitative data:

1. **Determine the Number of Intervals (Bins):** Decide on the number of intervals or ranges to divide the data into. Commonly used rules include the square root rule to determine the number of intervals.

2. **Calculate Interval Width:** Find out how wide each interval is. It is computed by dividing the number of intervals by the range of the data.

3. **Create Intervals:** Create non-overlapping intervals that cover the range of the data. Ensure that each value falls into exactly one interval.

4. **Count Frequencies:** Count the number of observations falling within each interval.

5. **Create a Table or Histogram:** Present the intervals along with their corresponding frequencies.

For example, consider a dataset of test scores:

Test Scores: 65, 72, 80, 85, 78, 90, 92, 68, 76, 82, 88, 72, 96, 85, 78, 82, 90, 94

Assuming we want to create a frequency distribution with 5 intervals:

(i). Identify the data's range: Range: = Highest value - Lowest value = 96 - 65 = 31

(ii). Calculate the interval width: Interval width = Range / Number of intervals = 31 / 5 = 6.2 (approx. 6)

Frequency Distribution:

Class Interval	Frequency
65-70	2
71-76	2
77-82	4
83-88	5
89-94	5

2.2.4 Histograms

Definition 8. *A graph of neighboring rectangles built on an XY plane is called a histogram. It's a frequency distribution graph. In practical use, histograms are used to depict both discrete and continuous frequency distributions.*

The vertical columns of a histogram are drawn without any gaps between them, which sets it apart from a bar graph.

Construction of a Histogram

The table below displays the estimated millimeter lengths (mm) of forty leaves sampled from various sections of a given species.

Length (mm)	Number of leaves
25-30	1
30-35	4
35-40	8
40-45	10
45-50	8
50-55	7
55-60	2

Represent the data in the form of a histogram.

The horizontal and vertical axes are drawn to produce a histogram. Put "Length of the leaves" as the graph's title. List the intervals across the horizontal axis on an appropriate scale and label the horizontal axis "Length (mm)". Write "Number of leaves" on the vertical axis. Mark the x-axis by fives. Draw a vertical column corresponding to the proper frequency value for each interval on the horizontal axis. Remember that there are never any gaps between vertical columns on a histogram.

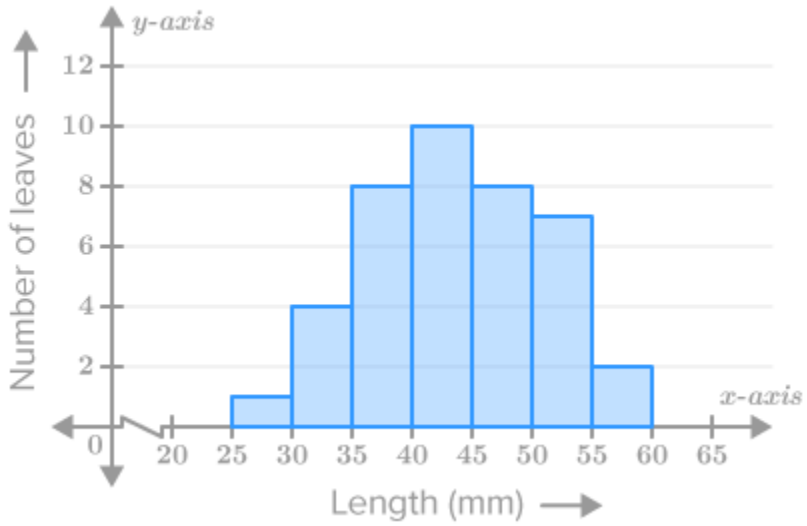


Figure 2.1: Histogram

2.2.5 Cumulative Distributions

Definition 9. *When an observation in a dataset falls below a given value, the cumulative count or proportion is displayed via cumulative distributions, also known as cumulative frequency distributions.*

They provide a running total of frequencies as values increase, showing the accumulation of observations up to a particular point in a dataset.

To create a cumulative distribution:

1. ***Calculate Cumulative Frequencies:*** *Start by organizing the data in ascending order.*
2. ***Count Frequencies:*** *For each value, determine the number of observations up to and including that value.*
3. ***Create a Table or Graph:*** *Present the values along with their corresponding cumulative frequencies.*

Example 2. *For the given data, create a* ***Cumulative frequency distribution****.*

Class Interval	Frequency
20-24	1
25-29	2
30-34	26
35-39	22
40-44	20
45-49	15
50-54	14

*The **cumulative frequency distribution** is constructed below*

Class Interval	Frequency	Relative Frequency
20-24	1	1
25-29	2	1 + 2 = 3
30-34	26	3 + 26 = 29
35-39	22	29 + 22 = 51
40-44	20	51 + 20= 71
45-49	15	71 + 15 = 86
50-54	14	86 + 14 = 100

In the table, each value represents the cumulative count of observations up to that point in the data.

2.3 MEASURES OF LOCATION

Definition 10. *Measures of location, also known as measures of central tendency, are statistical metrics used to describe the central or typical value in a data set.*

They provide a single numerical number that indicates the central tendency of the data. The following are the four main location metrics such as:

2.3.1 Mean (Arithmetic Mean)

Definition 11. *The Arithmetic Mean, sometimes referred to as the Mean, is a metric that is calculated by dividing the total of all the values (observations) of the variable by the number of observations.* $\overline{X}$ *is used to represent the arithmetic mean defining symbols as follows:*

$$A.M. = \overline{X} = \frac{\sum X}{n} = \frac{\textit{Sum of all values of observation}}{\textit{No. of observation}}$$

The formula for the arithmetic mean of n values $x_1 + x_2 + x_3 + ... + x_n$ *is:*

$$\textit{Arithmetic mean} = \overline{X} = \frac{x_1 + x_2 + x_3 + ... + x_n}{n}$$

Example 3. *Examine a dataset that shows a collection of people's ages:*

Ages: 25, 30, 35, 40, 45

To calculate the arithmetic mean:

1. *Add all the values together:*

$$25 + 30 + 35 + 40 + 45 = 175$$

2. *Count the number of values in the dataset, which is 5 in this case.*

3. *The sum divided by the total number of values:*

$$\textit{Arithmetic mean} = \bar{X} = \frac{175}{5} = 35$$

Therefore, the arithmetic mean (or average) age of this group of people is 35 years old.

The arithmetic mean is widely used due to its simplicity and interpretability. Its sensitivity to the dataset's extreme values, or outliers, can have a big impact on the mean. Because of this, it's crucial to take into account other central tendency measurements like the median or mode.

2.3.2 Median

Definition 12. *The middle observation in a collection of arranged data is called the median.*

A statistical measure of central tendency known as the median is used to reflect the middle value when values in a dataset are arranged in either ascending or descending order. With half of the values lying below and half above the median, the dataset is evenly divided in half.

Case.1 *The formula below can be used to get the median, or middlemost observation, of a collection of data ordered in order of magnitude when the number of observations is odd.*

$$Median = \left(\frac{n+1}{2}\right)^{th} observation$$

Case.2 *The median, which is the average of the two middle observations, is found when the number of observations in a set of data in order of magnitude is even. that is, median is average of* $\left(\frac{n}{2}\right)^{th}$ *and* $\left(\frac{n}{2}+1\right)^{th}$ *values.*

$$Median = size\ of\ \frac{1}{2}\left[\left(\frac{n}{2}\right)^{th} + \left(\frac{n}{2}+1\right)^{th}\right] observation$$

Example 4. *let's take a dataset representing the incomes of ten individuals in a community:*
Incomes: $25,000, $30,000, $35,000, $40,000, $45,000, $50,000, $55,000, $60,000, $65,000, $70,000

To find the median:

1. *Arrange the incomes in ascending order: $25,000, $30,000, $35,000, $40,000, $45,000, $50,000, $55,000, $60,000, $65,000, $70,000*

2. *Since there are ten values (an even number), the median will be the average of the two*

middle values.

3. *The two middle values are $45,000 and $50,000.*

4. *The median income would be the average of these two values:*

$$Median = \frac{45,000 + 50,000}{2} = \frac{95,000}{2} = 47,500$$

$$Median = 47,500$$

2.3.3 Mode

Definition 13. *In statistics, The mode of a dataset is the value that appears the most frequently. That's the observation or category that shows up most frequently.*

Example 5. *Let us consider a dataset that shows students' test scores:*

Scores: 78, 85, 92, 78, 75, 85, 78, 90, 78

The most common number in this collection is 78, which occurs four times.

Consequently, the mode of this data set is 78.

To determine the most frequent or prominent value of data, the mode is especially helpful. The mode explicitly indicates the value that happens most frequently, instead of the mean (average) or median (middle value), which considers all values equally. Depending on if two or more values occur with the same greatest frequency, a dataset may have different modes (bimodal, trimodal, etc.).

2.3.4 Geometric Mean

Definition 14. *The geometric mean of a dataset is the n^{th} positive root of the product of the $x_1, x_2, x_3, ..., x_n$ observations. In symbols, we write*

$$G.M. = (x_1, x_2, x_3, ..., x_n)^{1/n}$$

OR

$$G.M. = \sqrt[n]{x_1.x_2.x_3.....x_n}$$

Example 6. *Consider a set of values representing the growth rates of a company over five years:*

$5\%, 8\%, 6\%, 7\%, and 9\%$

To find the geometric mean:

1. *Multiply all the values together:* $5\% \times 8\% \times 6\% \times 7\% \times 9\%$

2. *There are five values, so take the fifth root of the product:* $\sqrt[5]{\% \times 8\% \times 6\% \times 7\% \times 9\%}$

Calculating the product gives $0.05 \times 0.08 \times 0.06 \times 0.07 \times 0.09 = 0.0001512$ *Then the fifth root of 0.0001512 is approximately 0.0651 or 6.51% when converted to percentage.*

So, the geometric mean of these growth rates over five years is approximately 6.51%

2.4 MEASURES OF VARIABILITY

Definition 15. *The spread or dispersion of data points within a dataset is described by measures of variability, which are also referred to as measures of dispersion.*

They offer information on the distribution and variability of the dataset and measure the degree to which individual data points vary or diverge from the central tendency (mean, median, or mode).

Common measures of variability include:

2.4.1 Range

Definition 16. *The simplest measure of variability is the range, which is the difference between a dataset's highest and lowest values. Although it might be vulnerable to outliers, it provides a broad idea of the dispersion.*

2.4.2 Variance

Definition 17. *The average squared departure of the data points from the mean is measured by variance. The average of the squared deviations between each data point and the mean is computed.*

Greater variability is indicated by a bigger variance, but it's not directly interpretable in the original units of the data.

2.4.3 Standard Deviation

Definition 18. *The average departure of data points from the mean is expressed as the standard deviation, which is the square root of the variance. Because it gives a clearer picture of dispersion and is in the same units as the original data, it is commonly utilized.*

2.4.4 Coefficient of Variation

Definition 19. *An indicator used to compare is the coefficient of variation (CV) the variability or dispersion of data relative to its mean, especially when dealing with different datasets with varying scales or units.*
It is computed as the percentage representation of the standard deviation to the mean.
The coefficient of variation formula is:

$$\textit{Coefficient of Variation (CV)} = \left(\frac{\textit{StandardDeviation}}{\textit{Mean}}\right) \times 100$$

To demonstrate how to calculate the coefficient of variation, let's look at an example:
Suppose we have two datasets representing the heights of two groups:
Group A:
(i). *Mean height: 160 cm*
(ii). *Standard deviation: 8 cm*
Group B:
(i). *Mean height: 175 cm*
(ii). *Standard deviation: 12 cm*
Calculating the coefficient of variation for each group:
For Group A:

$$\textit{CV for Group A} = \left(\frac{8}{160}\right) \times 100 = 5\%$$

For Group B:

$$\textit{CV for Group } B = \left(\frac{12}{175}\right) \times 100 = 6.86\%$$

In this example, Group A has a coefficient of variation of 5%, indicating that the standard deviation is 5% of the mean height. Group B has a higher coefficient of variation at 6.86%, suggesting relatively higher variability in heights compared to the mean when compared to Group A.

2.5 MEASURES OF ASSOCIATION BETWEEN TWO VARIABLES

Measures of association quantify the relationship or strength of association between two variables in a dataset. By revealing patterns, directions, and strengths of relationships between variables, they aid in the understanding of how changes in one are related to changes in another.

2.5.1 Scatter Charts

Definition 20. *Graphs called scatter plots show the connection between two variables in a dataset.*

It shows data points as a two-dimensional plane or as a Cartesian system. Plotting the independent variable, or attribute, on the X-axis corresponds to plotting the dependent variable on the Y-axis. These images are sometimes called scatter graphs or scatter diagrams.

Here are some instructions for making a scatter diagram.:

1. *Gather Bivariate Data.*

2. *Data should be displayed as a chart with the dependent variable always in the second column and the independent variable always in the first column.*

3. *Convert the table's rows into data points.*

4. *Label the independent variable on the x-axis and the dependent variable on the y-axis to begin producing a graph.*

5. *Based on the data gathered, choose the suitable scale for each axis (if the data spans*

from 0 to 10, consider a scale of 1; if the data extends from 0 to 500, consider a scale of 50). To ensure that the scale will appropriately display the data, experiment with it..

6. *Plot the points.*

Here is an example. For this example, examine the depth that a scuba diver goes versus water temperature.

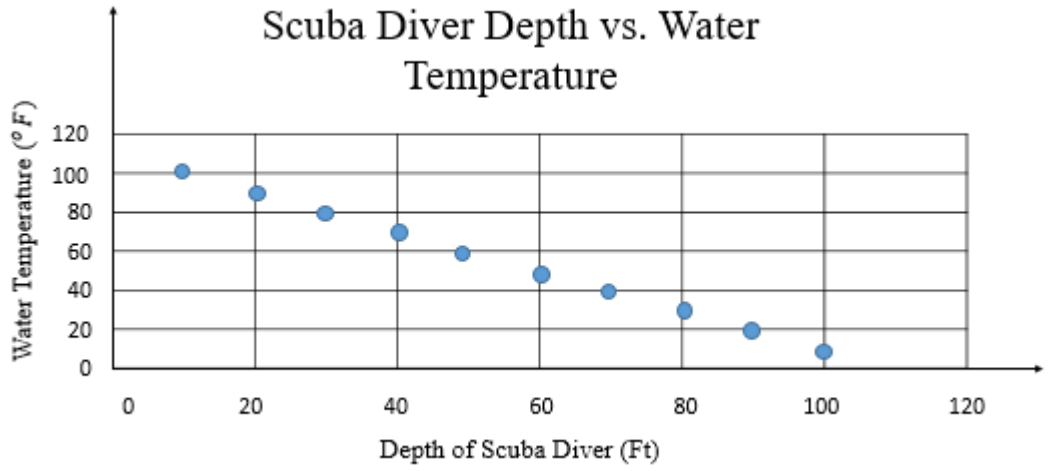
First, gather the data. The following data has been collected: at 10 ft down the water temperature is 80F, at 20 ft down it is 77°F, at 30 ft down it is 70°F, at 40 ft down it is at 68°F, and at 50 ft down it is 67°F.

Now organize that information into a table.

Depth of a Scuba Diver (ft.)	Water Temperature (°F)
10	80 °F
20	77 °F
30	70 °F
40	68 °F
50	67 °F

Now that a table has been created, organize this information into the following points: (10, 80), (20, 77), (30, 70), (40, 68), (50, 67).

Finally, draw the graph, label the axes, create an appropriate scale, and plot the points. Remember to always label the axes and give the scatter plot a title.



2.5.2 Covariance

Definition 21. *A statistical metric used to characterize the connection between two variables is covariance. It shows how much two random variables vary simultaneously.*
The formula for the sample covariance between two variables X and Y in a dataset with n observations is:

$$Cov(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$$

Where:
X_i *and* Y_i *are individual data points.*
$\bar{X}$ *and* $\bar{Y}$ *are the means of X and Y respectively.*
n is the number of observations.
Interpreting covariance:
(i). *Positive covariance indicates that the two variables tend to rise together as one increases. On the other hand, as one declines, the other also tends to decline.*
(ii). *If the covariance is negative, it means that one variable tends to decrease when the other grows, and vice versa.*
(iii). *There isn't a linear relationship between the variables if the covariance is 0.*

2.5.3 Correlation Coefficient

Definition 22. *The correlation coefficient is a statistical measure that may be used to determine the direction and strength of a linear relationship between two continuous variables. This assesses the degree to which the connection between the variables may be accurately represented by a straight line.*
Pearson's correlation coefficient, which is frequently represented by the symbol r, is the most often used. It has a range of -1 to 1:
(i). *A complete positive linear connection is shown by* $r = 1$, *meaning that if one variable grows, the other increases proportionately.*

(ii). *A complete negative linear connection is shown by* $r = -1$, *meaning that when one variable grows, the other drops proportionately.*

(iii). *There is no linear relationship between the variables when* $r = 0$.

The formula for Pearson's correlation coefficient is:

$$r = \frac{\sum(X_i - \bar{X})\sum(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2}\sqrt{\sum(Y_i - \bar{Y})^2}}$$

Where:

X_i *and* Y_i *are individual data points.*

$\bar{X}$ *and* $\bar{Y}$ *are the means of* X *and* Y *respectively.*

The covariance is effectively normalized by the square roots of the sum of squares in the denominator.

Chapter 3
Markowitz's Portfolio theory

In this chapter, we covered nonlinear optimization models. Permitting nonlinear terms significantly expands the pool of significant applications that may be modeled as optimization problems because so many business analytics applications incorporate nonlinear functions. Nonlinear models may be used to solve a wide range of issues in portfolio optimization, option pricing, marketing, economics, facility placement, forecasting, and scheduling.

3.1 Introduction

A little over seventy years ago, an economics doctoral student at the University of Chicago was looking for a topic for his dissertation when he stumbled upon a stockbroker, who suggested he research the stock market. Harry Markowitz's ideas led to the development of a theory that changed investor behavior and established the basis of financial economics. His achievements earned him a share of the 1990 Nobel Prize in Economics. A fundamental tenet of economics is that all economic choices include trade-offs since resources are scarce. The investor faces a trade-off between risk and projected return, which Markowitz recognized. Choosing which securities to hold is only one aspect of investing; another is deciding how to allocate the investor's money among the assets. According to the title of Markowitz's seminal study published in the Journal of Finance in March 1952, this is the "Portfolio Selection" problem. By extending the use of linear programming techniques, Markowitz develops the critical line approach in that study and later publications.

Using variance or standard deviation to measure risk, the critical line technique determines all potential portfolios that minimize risk for a given level of expected return and maximize anticipated return for a given level of risk. The standard deviation vs. anticipated return space displayed by these portfolios indicates the efficient frontier. An investor must weigh risk vs expected return while constructing his portfolio, and this trade-off is represented by the efficient frontier. Bound to be well-diversified are the majority of the efficient frontier. The rationale for this is that diversity is an effective way to lower risk.

Mean-variance analysis was created by Markowitz to choose a portfolio of common equities. The use of mean-variance analysis in asset allocation has grown within the past 20 years. The process of choosing an investing portfolio in which each element is an asset class rather than a single securities is known as asset allocation. Mean-variance analysis is more suited for asset allocation than stock portfolio selection. Mean-variance analysis requires knowledge of not only the expected return and standard deviation for each asset but also the correlation of returns for each pair of assets. While a stock portfolio selection issue may involve hundreds of stocks (and therefore thousands of correlations) (e.g., stocks, bonds, cash, real estate, and gold), an asset allocation problem typically involves a small number of asset classes. Furthermore, there is an opportunity to reduce portfolio risk overall due to the lack of correlation among assets. Since stocks frequently move in tandem, the benefits of variety within a stock portfolio are constrained. On the other hand, there is typically little to no connection across asset classes, and occasionally even a negative correlation. To find significant chances for risk reduction through diversification, mean-variance is a potent instrument in asset allocation.

3.2 Markowitz Mean-Variance Portfolio Theory

3.2.1 Portfolio Return Rates

Definition 23. *In the financial market, an asset is something that is frequently bought and sold.*

Let's say we buy an asset for y_0 dollars one day and sell it for y_1 dollars the next. We refer to the ratio as

$$R = \frac{y_1}{y_0}$$

the return on the asset.

The asset's rate of return is determined by

$$r = \frac{y_1 - y_0}{y_0} = R - 1$$

. Therefore,

$$y_1 = Ry_0$$

and

$$y_1 = (1 + r)\, y_0$$

Occasionally, we might be able to sell a non-owned asset. We call it short selling. In a way, it functions like this. Let's say you want to short sell or short a specific stock. To begin with, find out if every stock that a customer of your stock broker has is owned by the broker's firm. You can ask the brokerage to sell any amount of shares up to the amount they possess if they do hold (or manage) some stock. The amount of the debt from this transaction, which is the number of stocks they sell on your behalf, is credited to your account. In other words, rather than being expressed in dollars, The number of stocks you are shorting—that is, the number of stocks by which your account is short—expresses your debt. This short sale is shown as a negative figure associated with the shorted asset on your account asset sheet. Recall that the amount of stocks or other assets that are shorted, rather than money, is what determines this negative figure. You have got y_0 dollars as a result of the sale of stock. At some point, you will need to request that the brokerage purchase back the same quantity of stock that you first requested they sell and add it back to the asset pool they are holding for their clients. On the day you return to the brokerage, you ask your broker to purchase the stock back for y_1 dollars, the going rate at the time. if $y_1 < y_0$, then this deal has resulted in a profit for you; otherwise, a loss. The return and rate of return on this transaction is given by

$$R = \frac{-y_1}{-y_0} = \frac{y_1}{y_0}$$

and

$$r = \frac{(-y_1) - (-y_0)}{-y_0} = \frac{y_1 - y_0}{y_0}$$

respectively. Since short selling is so dangerous, many brokerage houses do not permit it. Still, it may be lucrative.

Now let's think about building a portfolio with n assets. We would want to provide an initial budget of y_0 dollars to these assets. The amount that we assign to asset i is $y_{0i} = w_0 y_0$ for $i = 1,2,3,...,n$, where w_i is a weighting factor for asset i. We allow the weights to have negative values; this indicates that there is a shortage of the asset in our portfolio. To maintain the financial constraints we require that the weights sum to 1.

$$\sum_{i=1}^{n} w_i = 1$$

.

$$\text{the sum of the investments} = \sum_{i=1}^{n} w_i y_0 = y_0 \sum_{i=1}^{n} w_i = y_0$$

Notice that When we short a stock, we gain its market value right away, and we may use that money to buy other assets or to reinvest it elsewhere. This allows us to free up more cash for the acquisition of other stocks. If R_i denotes the return on asset i, then the total receipts from our portfolio is

$$y_1 = \sum_{i=1}^{n} R_i w_i y_0 = y_0 \sum_{i=1}^{n} R_i w_i$$

, and so the portfolio's overall return is

$$R_i = \sum_{i=1}^{n} R_i w_i$$

. In addition, we have that the rate of return from asset i is $r_i = R_i - 1$, $i = 1,2,3,...,n$. Hence rate of return on the portfolio is

$$r = R - 1 = \left(\sum_{i=1}^{n} R_i w_i \right) - \left(\sum_{i=1}^{n} w_i \right) = \sum_{i=1}^{n} (R_i - 1) w_i = \sum_{i=1}^{n} r_i w_i$$

,

3.2.2 The Basics of Markowitz Mean-Variance Portfolio Theory

The Markowitz mean-variance portfolio theory models the rate of return on assets as a random variable. Selecting the portfolio weighting components as effectively as feasible is the next stage. Markowitz states that the optimal weight combination for a portfolio generates an adequate baseline expected rate of return with little volatility. In this case, the volatility of an item is substituted by the variation in its rate of return.

For each $i = 1, 2, ..., n$, let r_i be the random variable corresponding to the rate of return for asset i. Then, define the random vector

$$z = \begin{pmatrix} r_1 \\ r_2 \\ \vdots \\ r_n \end{pmatrix}$$

Set $\lambda_i = E(r_i)$, $m = (\lambda_1, \lambda_2, ..., \lambda_n)^T$, and $cov(z) = \sum$. If $w = (w_1, w_2, ..., w_n)^T$ is a collection of weights connected to a portfolio, then the rate of return of this portfolio $r = \sum_{i=1}^{n} r_i w_i$, is another arbitrary variable that has a mean $m^T w$ and variance $m^T \sum w$. If λ_b represents the appropriate baseline anticipated rate of return, then any portfolio that solves the following quadratic program qualifies as an optimum portfolio according to Markowitz theory:

$$M \text{ Minimize } \tfrac{1}{2} w^T \Sigma w$$

$$\text{Subject to } m^T w \geq \lambda_b, \text{ and } e^T w = 1,$$

In this case, every element of e is the number 1, meaning that e always represents the vector of ones. The quadratic program's KKT criteria are as follows:

$$0 = \Sigma w - um - ve$$
$$\lambda_b \leq m^T w, \quad e^T w = 1 \quad 0 \leq u,$$
$$u^T (m^T w - \lambda_b) = 0$$

For some $u, v \in \mathbb{R}$. We may determine that w is a solution to M if (w, u, v) is any triple that satisfies the KKT criteria since the covariance matrix is positive definite and symmetric.

It's rather simple to demonstrate that if M is possible, then there must always be a solution to M, meaning that a KKT triple may be constructed for M.

3.3 Assumptions

Like every model, mean-variance analysis has assumptions that must be understood to be used successfully. First of all, a single-period investment model serves as the foundation for mean-variance analysis. The investor divides his money into several asset classes at the start of the term, giving each item a nonnegative weight. Every investment produces a different rate of return at different times over the term, resulting in a weighted average of returns at the end that affects the portfolio's value. An investor must choose asset weights while taking into account some linear limitations, one of which is that the weights must add up to one. According to economic theory, humans maximize the projected value of a developing concave utility function of consumption while making decisions in the face of uncertainty. Von Neumann and Morgenstern's game theory research serves as the foundation for this theory. In a one-period model, consumption is end-of-period wealth. Generally speaking, selecting portfolio weights to maximize the predicted utility of ending period wealth is a challenging stochastic nonlinear programming issue.

In brief, assumptions:

1. Increasing the projected return on the total amount of money is the aim of investing.
2. It is believed that each investor has the same time horizon for investments.
3. Every investor has a low-risk tolerance, meaning they will only accept greater risk if the expected return is higher.
4. The projected return and risk are the main factors that investors consider when making investments.
5. Every market is completely efficient (i.e., there are no transaction costs or taxes).

It is assumed that the utility function is concave and rising. In terms of the approximation utility function, this translates into expected utility expanding in expected return (more is

better than less) and falling in variance (less risk is better). As a result, the investor should consider only those potential portfolios that minimize variance for a given level of expected return or maximize predicted return for a given level of variation. These are the portfolios that comprise the mean-variance efficient set.

3.4 Optimal Portfolio selection Model

Considering that N assets in the portfolio have returns R_i $i = 1, 2, 3, ..., N$

Let,

R_p = Portfolio return

R_i = Asset return i

w_i = The component asset's weight i (that is, the asset's share i in the portfolio)

σ_i = The asset's standard deviation i

Portfolio return:

$$R_p = \sum_i w_i \cdot E(R_i)$$

Portfolio return variance:

$$\sigma_p^2 = \sum_i \sum_j w_i w_j \sigma_i \sigma_j \rho_{ij}$$

between the returns on assets i and j, where ρ_{ij} is the correlation coefficient.

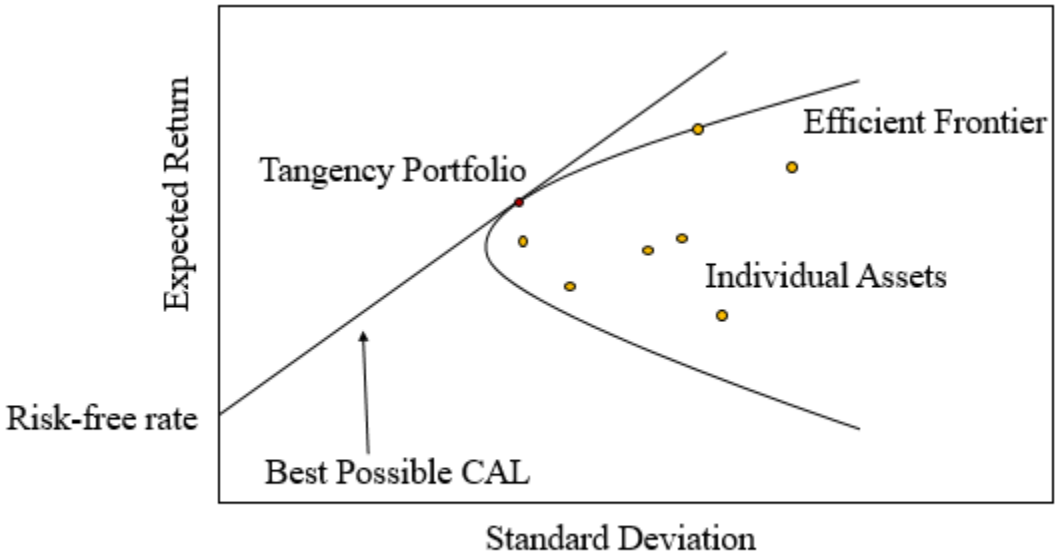


Fig 1.

Markowitz demonstrated that the correlation (covariance) between the asset returns and their weights, as well as the standard deviations of those returns, determine the risk of an asset portfolio.

The link between the expected return from a portfolio and its volatility, or riskiness, is described by the efficient frontier. On a risk graph, it can be shown as a curve versus the expected return of a portfolio. The efficient frontier indicates the lowest amount of risk required to get a particular projected rate of return or the best return that can be anticipated for a given level of risk. The idea of portfolio design and valuation heavily relies on the efficient frontier. One way to show the advantages of variety is through the idea of an efficient frontier. An undiversified portfolio can be moved closer to the efficient frontier by diversifying. Therefore, diversification can increase returns without increasing risk or decrease risk without decreasing expected returns.

The CAL illustrates how the risk of the portfolio can be further decreased in Fig. 1 if the investor has access to investments that don't carry risk. The (fictitious) asset that pays a risk-free rate is the risk-free asset. Short-term government securities, like US Treasury bills, are often utilized as risk-free assets due to their extremely low default risk and fixed rate of interest. Since the risk-free asset's variance is zero, it is by definition risk-free and uncorrelated with any other asset. It also has zero return variance.

In the portfolio theory application, the following two quantities must be computed using the appropriate units of measurement:

1. **Historical Values:** These are the absolute or relative source data points.

2. **Expected Returns:** The expected return on investment for the period under review, which has to be specified and will be reimbursed in units (absolute or relative) based on past performance.

The units used within the historical values must or will be consistent with the expected return values that are evaluated or supplied.

"We need to have processes in place for determining suitable μ_i and ρ_{ij} to apply the E-V rule to the selection securities. I think these processes ought to integrate the assessment of practical knowledge with statistical methodologies. Using observed data from a certain period in the past is one recommendation for the tentative μ_i and ρ_{ij}. I think there are better methods out there that take into account more information. Security analysis has to be rewritten in a "probabilistic" way, in my opinion.

3.5 Later Developments

An essential component of the present theory of asset allocation is Markowitz's selection model. Many models of portfolio selection have been created since Markowitz's model was first introduced, to improve and complete portfolio theory in multiple ways. Huang [21] and Markowitz [36] are two researchers who have constructed models to minimize semivariance in various instances. Other researchers, including Konno and Suzuki [4], Liu, Wang, and Qiu [33], and Pornchai, Krushnan, Shatid, and Arun [10], have included skewness in their analysis of portfolio selection.

The prevalent presumptions are that investors possess sufficient historical data and that historical data can accurately forecast the state of asset markets in the future. Since this isn't always the case, real-world issues come up. For instance, there is no historical data available for newly listed equities on the stock exchange. Investors found that using random, fuzzy, and random fuzzy optimization models helped them deal with uncertainty. Several

studies have demonstrated that mean-variance efficient portfolios that rely on estimations are extremely vulnerable to changes in these estimates. Jobson, Korkie, and Ratti [22] as well as Jobson and Korkie [23] go into length about these issues and recommend using shrinkage estimators.

Certain authors Carlsson, Fuller, and Majlender [9], Leon, Liern, and Vercher [32], and Vercher, Bermudez, and Segura [52] replace the unpredictable returns of the securities with fuzzy numbers. Possibilistic distributions were employed by Tanaka and Guo [49] and Tanaka, Guo, and Tìurksen [50] to model return uncertainty. Rigid targets for return rate, risk, and liquidity based on projected intervals were proposed by Arenas-Parra, Bilbao-Terol, and Rodriguez-Ura [46]. Feiring, Wong, Poon, and Chan [15], as well as Konno, include a measure of downside risk. An approximation to the lower semi-third instant is used by Shirakawa and Yamazaki [28] in their Mean-Absolute Deviation-Skewness portfolio model. As an alternative to the mean-variance (MV) model, Konno and Yamazaki developed the mean absolute deviation (MAD) model, arguing that it reduces computation time, preserves all the good aspects of the mean-variance model, and does not require the covariance matrix.

Frost and Savarino [16] address the estimate risk issue by demonstrating that limiting portfolio weights minimizes estimation error by limiting the action space during the optimization. A resampling technique for estimating error is proposed by Jorion [24]. To account for parameter uncertainty and maintain the choice simplicity of the efficient frontier, Michaud [38] proposes a sampling-based method for estimating a resampled efficient frontier. Polson and Tew [45] make the case that posterior predictive moments should be used in place of point estimates for the sample model's mean and variance.

Britten-Jones [7] suggests explicitly setting relevant prior densities on the portfolio weights using a Bayesian technique. As per the findings of Chopra and Ziemba's study [35], errors in means are almost ten times more significant than errors in variances, and errors in variances are nearly twice as important as errors in covariances. It was demonstrated by Best and Grauer [3] that the amount of projected returns has a significant impact on optimum portfolios. Jorion [25] uses a shrinkage method, but Treynor and Black [51] support combining investors' perspectives with past data. When a risk-averse Bayesian investor

allocates their portfolio between stocks and cash, Kandel and Stambaugh [26] look at the predictability of stock returns. It is emphasized by Zellner and Chetty [53], Klein and Bawa [27], and Brown [8] to use a predictive probability model. Píastor and Stambaugh [41] investigate the effects of different pricing models on ideal portfolios, reevaluating assumptions in light of example data. Asset pricing models are suggested for use in providing useful prior distributions for future returns by Pastor [42] and Black and Litterman [5].
Bollerslev et al. are concentrating their research on conditional covariances and correlations. [6] or Engle [13] as a way to describe time fluctuations in the conditional dependence of asset returns. Goetzmann, Li, and Rouwenhorst [18] discovered that correlations between equities returns fluctuate significantly over time and peak during times when financial markets are strongly interconnected. Their findings were based on data spanning the previous 150 years. The idea that there is a permanent link between worldwide stock markets was disproved by Longin and Solnik [34], who examined changes in the correlation structure of global equity markets.
Moreover, Ang and Chen [2] confirmed connections between stock returns and a US market aggregate market index. Some people associate the economic cycle stage with variations in stock return correlations. Time-varying correlations are contingent upon the state of the economy and tend to be stronger during recessions, as Ledoit, Santa-Clara, and Wolf [31] and Erb, Harvey, and Viskanta [14] have shown. Similarly, Moskowitz [39] connects NBER recessions to temporal variation of volatilities and covariances.

3.6 Konno Yamazaki, 1991

3.6.1 Introduction

Using mean absolute deviation (MAD) as a risk indicator, Konno and Yamazaki (1991) created a new model to overcome the limitations of Markowitz's mean-variance model. The computational complexity of solving a large-scale quadratic problem with a dense covariance matrix is one of the main causes of the problems. They claim that to determine a link between the rate of return and assets and market portfolio, equilibrium models need to

apply some unreasonable assumptions. However, according to data from the Tokyo Stock Exchange, this link is quite unstable, therefore using the CAPM data is best done as a first-order approximation.

Most of Markowitz's model's shortcomings were addressed by Konno and Yamazaki while maintaining its advantages over equilibrium models by using L1-mean absolute deviation as a risk indicator instead of variance. Here are a few problems that are rarely fixed in real-world situations:

Computational burden: Taking up complicated, large-scale quadratic problems can be difficult.

Perception of investors: Reliability of the standard deviation as a risk indicator was questioned by a sizable segment of investors.

Transaction fees and the cut-off effect: This suggests that because he will have to pay the transaction costs, the investor who makes little purchases in a range of equities would be upset. Additionally, Since stocks cannot be bought in fractions, the investor must round down to whole numbers.

3.6.2 Model

First, they presented the L1 risk function.

$$w(x) = E\left[\,|\sum_{j=1}^{n} R_j x_j - E\left[\sum_{j=1}^{n} R_j x_j\right]|\,\right]$$

Where,

$R_j =$ The rate of return on asset S_j represented by a random variable

$x_j =$ Amount invested in S_j

$M_o =$ Total amount of funds

$E[.] =$ the random variable's expected value in the bracket

They continue by stating and demonstrating the subsequent theorem.:

If $(R_1, \ldots\ldots, R_n)$ are multivariate normally distributed , then

$$w(x) = \sqrt{\frac{2}{\pi}}\sigma(x)$$

Where $\sigma(x)$ = The standard deviation They demonstrated that in the case of multivariate regularly distributed $(R_1–R_n)$, these two measures $(w(x) and R_i)$ are identical.

Thus, the Model assumes the following form:

$$\begin{aligned} Min \qquad & w(x)E\left[\left|\sum_{j=1}^{n} R_j x_j - E\left[\sum_{j=1}^{n} R_j x_j\right]\right|\right] \\ ST \qquad & \begin{cases} \sum_{j=1}^{n} E\left[R_j\right] x_j \geq \rho M_0, \\ \sum_{j=1}^{n} x_j = M_0, \\ 0 \leq x_j \leq u_j, j = 1,2,3,...,n. \end{cases} \end{aligned} \tag{3.6.1}$$

Konno and Yamazaki assumed that the average of the data might be used to estimate the anticipated value of the random variable.

Therefore,

$$r_j = E\left[R_j\right] = \sum_{t=1}^{T} r_{jt}/T$$

Now,

$$E\left[\left|\sum_{j=1}^{n} R_j x_j - E\left[\sum_{j=1}^{n} R_j x_j\right]\right|\right] = \frac{1}{T}\sum_{t=1}^{T}\left|\sum_{j=1}^{n}(r_{jt} - r_j)x_j\right|$$

Let

$$a_{jt} = r_{jt} - rj, \qquad j = 1,2,3,...,n; \qquad t = 1,2,3,...,T.$$

Model in (3.6.1) can be stated as,

$$\text{Min} \qquad \sum_{t=1}^{T}\left|\sum_{j=1}^{n} a_{jt}x_j\right|/T$$

$$ST \quad \begin{cases} \sum_{j=1}^{n} r_j x_j \geq \rho M_0, \\ \sum_{j=1}^{n} x_j = M_0, \\ 0 \leq x_j \leq u_j, j = 1,2,3,...,n. \end{cases}$$

This corresponds to the linear program that follows:

$$\text{Min} \qquad \sum_{t=1}^{T} y_t / T$$

$$ST \quad \begin{cases} y_t + \sum_{j=1}^{n} a_{jt} x_j \geq 0, & t = 1,2,3,...,T, \\ y_t - \sum_{j=1}^{n} a_{jt} x_j \geq 0, & t = 1,2,3,...,T, \\ \sum_{j=1}^{n} r_j x_j \geq \rho M_0, \\ \sum_{j=1}^{n} x_j = M_0, \\ 0 \leq x_j \leq u_j, j = 1,2,3,...,n \end{cases}$$

The following are some of Konno-Yamazaki's benefits over Markowitz's model:

1. There is no need to figure out the covariance matrix.

2. Compared to completing a quadratic program, tackling a linear program is considerably simpler.

3. A smaller solution size is ideal.

4. T can be used as a control variable to limit the portfolio's total number of assets.

3.7 Black and Litterman, 1992

3.7.1 Introduction

The Black-Litterman asset allocation model was created by Fischer Black and Robert Litterman. It is a method for creating portfolios that tackle the problems of highly concen-

trated portfolios, input sensitivity, and optimizing estimate error. Their methodology uses a Bayesian strategy to construct a new, mixed estimate of expected returns by combining the market equilibrium vector of anticipated returns (the prior distribution) with the investor's subjective beliefs about the expected returns of one or more assets.

The Black-Litterman asset allocation model was first introduced by Black and Litterman in 1990, and it was further refined by Black and Litterman in 1991 and 1992. The Black Litterman model combines the universal hedge ratio / Black's global CAPM (Black, 1989a, 1989b), mean-variance optimization (Markowitz, 1952), mixed estimation (Theil, 1971, 1978), and reverse optimization (Sharpe, 1974). The process integrates recent and historical data to provide the updated expected return distribution. If an investor has any subjective judgments, the weights on individual assets deviate from the weights in the market equilibrium; if not, the weights are established by the facts in the market equilibrium.

Investor opinions and market equilibrium returns are the main inputs to the Black and Litterman model. Investor views are included in this framework to assist investors in managing the size of tilts brought about by views.

3.7.2 Market Equilibrium Returns

The Black and Litterman model is based on the capital asset pricing model (CAPM) or market equilibrium weights. The CAPM is created by forming the efficient frontier of the market portfolios and the capital market line (CML). No other combination of risky and non-risky assets can produce greater returns at a given level of risk since the CML is tangent to the efficient frontier at the market portfolio.

CAMP;

$$E(r_i) = r_j + \frac{\sigma_i}{\sigma_m}(r_m - r_j)$$

$$E(r_i) = r_j + \beta_i(r_m - r_j)$$

Where,

$E(r_i) = Expectedreturnonasseti$

$r_j = Risk\,free\,asset\,return$

$r_m = Return\,on\,market\,portfolio$

$\sigma_i = Standard\,deviation\,of\,returns\,on\,asset\,i$

$\sigma_m = Standard\,deviation\,of\,returns\,on\,market\,portfolio$

$\beta_i = \frac{\sigma_i}{\sigma_m}$

CAPM is used in reverse by the model. It is based on the assumption that mean-variance investors hold the market portfolio and uses optimization to back out the optimal predicted returns. Market equilibrium returns are defined by them as:

$\pi = v \sum \omega$

Where,

N = Quantity of assets

π = Implied access return vector (N, 1)

$\sum$ = Returns' covariance matrix (N, N)

ω = Weights of the assets based on market capitalization vector (N, 1)

v = Coefficient of risk aversion

$v = \frac{r_m - r_j}{\sigma_m^2}$

3.7.3 Black-Litterman equation

The Black-Litterman formula integrates investor perspectives and equilibrium returns into a single formula to calculate predicted returns, which are then utilized to calculate the ideal portfolio weights.

$$E[R] = \left[(\tau \sum)^{-1} + P' \Omega^{-1} P \right]^{-1} \left[(\tau \sum)^{-1} \prod + P' \Omega^{-1} Q \right]^{-1}$$

Where,

$E[R]$ = Vector of aggregated outcomes (N, 1)

τ = Scalar representing the CAPM prior's uncertainty

$\sum$ = Equilibrium access returns' covariance matrix (N, N)

P = Investor viewpoint matrix (N, 1)

Ω = View error terms' diagonal covariance matrix (K, N)

$\prod$ = Equilibrium access return vector (N, 1)

Q = Vector of investor view (K, 1)

If the investor is unrestricted, we calculate the ideal portfolio weights using the Black-Litterman formula by,

$$w^* = \left(v\sum\right)^{-1}\lambda$$

Where λ is the vector derived from above equation

$$Max_w \qquad w'\lambda - vw'\sum w/2$$

Generally speaking, the Black-Litterman method lets users realize the benefits of the Markowitz paradigm by overcoming the mean-variance optimization's most commonly cited shortcomings, which include highly concentrated portfolios, input sensitivity, and estimate error maximization.

Chapter 4
Portfolio Optimization with Clustering Algorithms

4.1 Introduction

4.1.1 Background

Across the world, both people and organizations have embraced the concept of financial investing. These investments support a variety of goals, including financial institutions managing investor assets. Because of this, many participants in this market have unique needs, objectives, and strategies regarding their financial commitments. Investing strategies range from simple methods to complex algorithms. All investors, however, have the same objective in mind: to strike a balance between high returns and low risk. However, assembling such a portfolio is a difficult task.

Data analysis is the first step in creating an investing plan. Financial data must usually be processed and transformed into more digestible formats that are suitable for study. This data is frequently obtained from daily activities. After that, the information has to be examined using one or more approaches while taking into account different aspects. Risk is one of these; whilst some participants show a greater willingness to take risks, others adopt a more cautious approach. The goal is still to strike the ideal balance between risk and reward. There are several approaches to assessing and controlling risk. When it comes to portfolio optimization, diversity is a popular and well-respected strategy. By using data that is vulnerable to risk from a variety of sources, diversified portfolios help to reduce risk and decrease portfolio susceptibility. By incorporating risky data from several sources into the portfolio, diversification aims to lower portfolio volatility.

In theory, financial assets such as bonds, futures, and stocks can be subjected to a cluster analysis to create a diversified portfolio. Previous research indicates that cluster analysis, a method used in many other fields, has potential in financial situations. This method divides unorganized data into distinct components known as clusters. Essentially, cluster analysis aims to minimize the difference and maximize the distance between separate groups or dis-

tance, between stocks inside a cluster. Finding correlations in the data is its goal; the most related datasets are grouped into clusters. Investment plans can then be developed based on these clusters. We will compare and examine several cluster analysis techniques to find a reliable strategy. Using datasets from Nordea, the evaluation of cluster analysis will be carried out, evaluating the investment plan within a theoretical framework.

4.2 Previous Research

To identify patterns in data, cluster analysis is applied in many fields. However, few cluster studies have been conducted using stock indexes or financial data in general.

A study that has been written about the topic looks at the best way to measure the variation in stocks. The study uses a self-composed distance metric for cluster analysis to lessen the drawbacks associated with other well-known techniques. One such is the correlation approach, which sometimes varies during difficult times financially and may yield inaccurate findings. The study concludes that while diversity is beneficial, there isn't a single, well-defined method for creating and managing a diverse portfolio.

Another research presents cluster analysis as a novel method for resolving issues with estimation errors in stock-based portfolios. This study compares the estimation errors using cluster analysis with a resampling approach, which was used more frequently in the past. The results show that a portfolio's resilience and performance may be enhanced by using cluster analysis. Additionally, it highlights the paucity of research on the topic. Moreover, certain inquiries concerning the application and interpretation of the cluster analysis technique remain unresolved.

One paper discusses cluster analysis using a data set that is comparable to the one used in this thesis. According to the study's conclusion, investing methods developed from the cluster analysis of stock data might result in both profit and loss.

As is evident, research on cluster analysis with stock data sets has been conducted. Even though the prior research did not employ the same approach as this thesis, it does indicate that further investigation into the cluster analysis method is warranted.

4.3 Mathematical theory

Since mathematical models may forecast future financial market outcomes based on past performance, they are frequently the foundation of investment strategies. Decisions made by the mathematical model are based on some elements that are beyond the scope of human analysis. In addition, it is simple to make incremental modifications to a mathematical model as a result of unforeseen market developments. Since not all elements can be anticipated and accounted for in a mathematical model, it is not always possible to predict the direction of the market. Even with mathematical models, human intervention is still necessary to adjust the model for unforeseen events.

Many mathematical models are available to generate investing strategies. Each approach has unique benefits and drawbacks. For instance, certain techniques work well for forecasting an investment plan in a particular industry, while others perform better for handling anomalies.

This thesis focuses on using cluster analysis, a mathematical technique, to create an investing strategy. The benefits of diversifying portfolios make this approach popular. There are several approaches available for performing cluster analysis, each with its own set of benefits. Some of the criteria that affected the approaches selected for this thesis were the type of data, the purpose of the study, and the definition of the distances between the stock returns.

4.3.1 Cluster analysis

One method for organizing observations into clusters is cluster analysis. For the selected features, the resulting clusters are homogenous. Based on the selected criteria, every cluster will differ from the others. The characteristics chosen differ depending on the data and the cluster analysis's objective. Cluster analysis is a widely used technique in genetics, marketing, finance, and mapping of disease categories, such as breast cancer kinds.

In addition to the selection of features, a unique set of methodologies has developed since cluster analysis may be applied in a multitude of domains. Hierarchical and non-hierarchical clustering are the two basic techniques used in cluster analyses. Each of the two methods has advantages that vary based on the type of data and the goal of the grouping.

Hierarchical clustering

One or more data point smaller clusters can be combined into bigger clusters or divided into smaller clusters by using hierarchical clustering. Agglomerative clustering refers to the more typical practice of combining smaller clusters into bigger ones. A tree, or dendrogram, is typically used to illustrate the result of hierarchical clustering (see Figure 4.1). Data points 1 and 4 comprise the lowest connected nodes (shown in Figure 4.1), which are made up of the most related data points. As one navigates up the tree, one can see that the data points are connected at higher altitudes. Depending on the type of data being utilized, several linking techniques are available for carrying out hierarchical clustering. Because it doesn't require a set number of clusters, hierarchical clustering may be used for a variety of issues.

Non-hierarchical clustering

On the other hand, non-hierarchical clustering instantly divides the data into a few different groups. K-means is the most widely used algorithm in non-hierarchical clustering. The K-means clustering technique may be used to create K-distinct clusters elegantly and with simplicity. To do K-means clustering, the number of clusters, K, has to be chosen beforehand. The fact that figuring out how many clusters to use might occasionally be difficult is one of the primary issues with the K-means clustering technique. Another drawback of this approach is that, unlike the hierarchical technique, it is not feasible to show the outcome in a figure.

4.3.2 Choice of main method

Hierarchical clustering is the main technique of choice in this thesis. Since the focus of this thesis is stock index data, hierarchical clustering is a better fit than non-hierarchical since the relationship between the stock indices has never been studied before. Furthermore, it

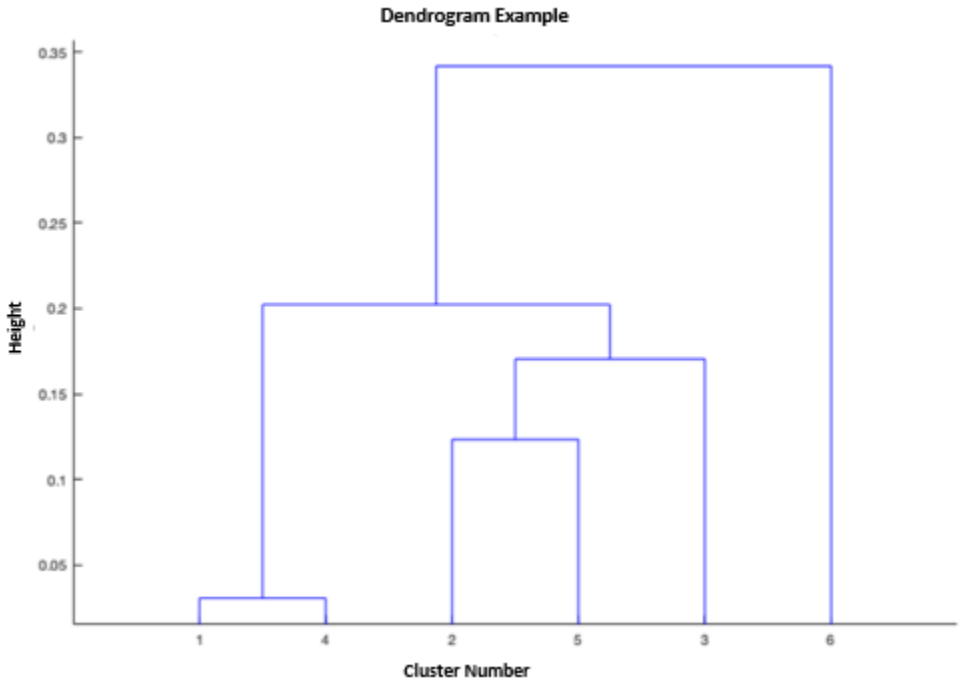


Figure 4.1: Example of a dendrogram

is challenging to estimate the number of clusters ahead of time, which is required for the non-hierarchical technique.

4.3.3 Methods for performing hierarchical clustering

Performing a hierarchical cluster analysis may be done using various connection techniques and distance metrics. To arrive at a trustworthy result, this thesis looks at two distance metrics and three distinct connection techniques. The centroid method, the average linkage method, and Ward's approach are the recommended linking techniques. The approaches have been selected based on their widespread usage and numerous benefits, which raise the likelihood of obtaining a dependable outcome. The correlation measure and the Euclidean distance measure are the distance measurements that are employed. Since the Euclidean distance metric is widely used and applicable to all connection techniques, It was chosen. In addition, the Correlation measure will be utilized to investigate the differences between the Euclidean and Correlation distances. The Correlation measure can only be matched

with the Average approach among the three connecting strategies that were chosen. Below, we'll go into more depth about the characteristics of the linking techniques and measurements.

Linkage methods

Ward

A hierarchical linking approach is Ward's method. By combining the sum of squares and determining the amount that the within-cluster sum of squares rises, It establishes the separation between two stocks, clusters, or observations.

$$d_{ward}(A,B) = \sqrt{\frac{2n_A n_B}{n_A + n_B}} ||\overline{m}_A - \overline{m}_B|| \tag{4.3.1}$$

Equation (4.3.1) is the equation for the sum of squares to rise, it stands for the expense of combining clusters A and B. The number of data points in the cluster is denoted by n_j, while the centroid of cluster j is represented by m_j. The $\|\|$ is a representation of the Euclidean distance, which is detailed below. Since every point is a separate cluster, the total of squares begins at zero and grows as the clusters combine. The increase is intended to be as little as feasible by employing Ward's technique. Ward's approach will combine the cluster with the fewest data points if the merging costs of the two clusters are the same.

The method's greed and limitations stem from prior grouping decisions. Once a data point is allocated to a cluster, it cannot move clusters.

Only the Euclidean distance, one of the techniques for determining the separation of stocks, may be used in conjunction with Ward's approach. This is due to the algorithm's requirement for the Euclidean computation during the initial setup, which occurs when each data point is a separate cluster.

Centroid

The distance between each cluster's centroids is measured using the centroid technique. The cluster's centroid, or average element, is its most representative location. The average value of the data points inside the cluster determines the value. The centroid, which is a cluster's center of mass, will serve as a benchmark for measuring the separation between clusters.

$$d_{centroid}(A,B) = \|\overline{m}_A - \overline{m}_B\|, \text{where } \overline{m}_j = \frac{1}{n_j}\sum_{j=1}^{n_i} m_{ji}, \;\; i = 1,...,n \tag{4.3.2}$$

m_j = centroid of cluster j. Further i portrays the cluster's total amount of data points.

One often utilized technique, particularly in genomics, is the centroid approach. But there is a drawback to it. When two clusters are gathered at a height less than the height of one cluster that is already visible in the dendrogram, an inversion takes place. The dendrogram may be more challenging to see and comprehend as a result of the inversion.

Average linkage

The Average Linkage Method calculates the average distance between each pair of clusters to get the exact distance between them. To calculate average linkage, the distances between each data point in one cluster and each data point in the other cluster are compared. The findings are then averaged.

$$d_{average}(A,B) = \frac{1}{n_A n_B}\sum_{i\in A, j\in B} d_{ji} \tag{4.3.3}$$

where n_i is the cluster's total amount of data points.

Equation (4.3.3) provides the average linkage calculation formula. The distance between clusters A and B is the average distance between each data point in a cluster.

Unlike the other approaches, the Average linkage method may make use of measures such as the Correlation measure in addition to the Euclidean distance metric.

Distance measures

Euclidean

One popular technique for determining the separation between data to find commonalities is the Euclidean distance metric. The method, which calculates the shortest path between observations using a straight line distance, is based on the Pythagorean theorem and has its roots in ancient geometry. By taking the square root of the total squared differences between two clusters, one may determine the Euclidean distance measure, p, and q, in the ith dimension.

$$\text{Euclidean distance} = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2} \tag{4.3.4}$$

Correlation For the Average linkage approach, the Correlation measure can be utilized as a distance metric. The variance between data points is measured by the correlation matrix. But in a technical sense, the Correlation matrix is not a metric. To do a cluster analysis using the correlations, all that needs to be done is to translate the coefficients into distance units. For cluster analysis, the correlation measure is a useful method when examining stock-based data. It's common for the stock prices of the other stocks in a cluster to drop in response to a decline in the price of one of the stocks. This is an effective method for building and refining a portfolio.

Matrix 4.1 illustrates the Correlation matrix, which contains comparisons between the equities. σ^2 represents the variance of the stocks, and X_m indicates the stock m. The diagonal components are all one. Additionally, each component offers the correlation between stock m and stock n.

$$\begin{bmatrix}
1 & \frac{\sigma^2(X_1,X_2)}{\sqrt{\sigma^2(X_1)\times\sigma^2(X_2)}} & \frac{\sigma^2(X_1,X_3)}{\sqrt{\sigma^2(X_1)\times\sigma^2(X_3)}} & \cdots & \frac{\sigma^2(X_1,X_n)}{\sqrt{\sigma^2(X_1)\times\sigma^2(X_n)}} \\
\frac{\sigma^2(X_2,X_1)}{\sqrt{\sigma^2(X_2)\times\sigma^2(X_1)}} & 1 & \frac{\sigma^2(X_2,X_3)}{\sqrt{\sigma^2(X_2)\times\sigma^2(X_3)}} & \cdots & \frac{\sigma^2(X_2,X_n)}{\sqrt{\sigma^2(X_2)\times\sigma^2(X_n)}} \\
\frac{\sigma^2(X_3,X_1)}{\sqrt{\sigma^2(X_3)\times\sigma^2(X_1)}} & \frac{\sigma^2(X_3,X_2)}{\sqrt{\sigma^2(X_3)\times\sigma^2(X_2)}} & 1 & \cdots & \frac{\sigma^2(X_3,X_n)}{\sqrt{\sigma^2(X_3)\times\sigma^2(X_n)}} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\frac{\sigma^2(X_m,X_1)}{\sqrt{\sigma^2(X_m)\times\sigma^2(X_1)}} & \frac{\sigma^2(X_m,X_2)}{\sqrt{\sigma^2(X_m)\times\sigma^2(X_2)}} & \frac{\sigma^2(X_m,X_3)}{\sqrt{\sigma^2(X_m)\times\sigma^2(X_3)}} & \cdots & 1
\end{bmatrix}$$

Matrix 4.1 Correlation matrix

4.4 Method and Model

In our research work, we find out the optimal portfolio for minimizing the risk and maximizing the return using Microsoft Excel, Excel Solver, and matrix multiplication to compute variance and covariance matrix using five stocks.

4.4.1 Data Collection

In this demonstration, we are going to use five years of monthly stock price data for five companies such as:

1. McDonald's Corporation (MCD)
2. The Coca-Cola Company (KO)
3. PepsiCo, Inc. (PEP)
4. Nestlé S.A. (NSRGY)
5. Yum! Brands, Inc. (YUM)

For sixty months from 01-01-2015 to 01-01-2020. To find out the optimal portfolio for minimizing the risk and maximizing the return using Microsoft Excel, Excel Solver, and matrix multiplication to compute variance and covariance matrix using five stocks.

Firstly we collect data from Financeyahoo and then use Microsoft Excel to set the close price of five stocks for sixty months.

Table 4.1 shows the monthly price of five stocks (MCD), (KO), (PEP), (NSRGY), and (YUM)

Firstly we calculate the McDonald's Corporation (MCD) return using the formula of return.

$$R = \frac{y_1}{y_0}$$

The asset's rate of return is determined by

$$r = \frac{y_1 - y_0}{y_0} = \left(\frac{y_1}{y_0} - 1\right) \times 100$$

. the same formula is used for the remaining companies such as The Coca-Cola Company (KO), PepsiCo, Inc. (PEP), Nestlé S.A. (NSRGY), and Yum! Brands, Inc. (YUM). Table 4.2 shows the return rate of the monthly price of five stocks.

Table 4.3 shows the calculations of Average monthly return, Monthly return, and Annual return.

Table 4.1: Monthly Prices of Five Stocks

Monthly Prices of five stocks					
Date	MCD	KO	PEP	NSRGY	YUM
1/1/2015	92.44	41.17	93.78	76.51	51.96262
2/1/2015	98.9	43.3	98.98	78.15	58.31057
3/1/2015	97.44	40.55	95.62	75.22	56.59238
4/1/2015	96.55	40.56	95.12	77.62	61.79727
5/1/2015	95.93	40.96	96.43	77.44	64.78073
6/1/2015	95.07	39.23	93.34	72.16	64.75916
7/1/2015	99.86	41.08	96.35	75.6	63.0913
8/1/2015	95.02	39.32	92.93	73.65	57.34723
9/1/2015	98.53	40.12	94.3	75.24	57.47664
10/1/2015	112.25	42.35	102.19	76.22	50.97772
11/1/2015	114.16	42.62	100.16	73.94	52.12796
12/1/2015	118.14	42.96	99.92	74.42	52.51617
1/1/2016	123.78	42.92	99.3	73.74	52.02732
2/1/2016	117.19	43.13	97.82	69.86	52.09921
3/1/2016	125.68	46.39	102.48	74.61	58.84256
4/1/2016	126.49	44.8	102.96	74.57	57.19626
5/1/2016	122.06	44.6	101.17	73.9	59.0151
6/1/2016	120.34	45.33	105.94	77.31	59.61179
7/1/2016	117.65	43.63	108.92	80.23	64.28468
8/1/2016	115.66	43.43	106.75	79.5	65.21207
9/1/2016	115.36	42.32	108.77	79.02	65.28397
10/1/2016	112.57	42.4	107.2	72.65	62.02732
11/1/2016	119.27	40.35	100.1	67.3	63.39
12/1/2016	121.72	41.46	104.63	71.74	63.33
1/1/2017	122.57	41.57	103.78	73.22	65.53
2/1/2017	127.65	41.96	110.38	73.69	65.32
3/1/2017	129.61	42.44	111.86	76.9	63.9
4/1/2017	139.93	43.15	113.28	77	65.75
5/1/2017	150.89	45.47	116.87	85.16	72.64
6/1/2017	153.16	44.85	115.49	87.2	73.76
7/1/2017	155.14	45.84	116.61	84.29	75.48
8/1/2017	159.97	45.55	115.73	84.76	76.82
9/1/2017	156.68	45.01	111.43	84.01	73.61

10/1/2017	166.91	45.98	110.23	84.23	74.45
11/1/2017	171.97	45.77	116.52	85.41	83.47
12/1/2017	172.12	45.88	119.92	85.97	81.61
1/1/2018	171.14	47.59	120.3	86.4	84.59
2/1/2018	157.74	43.22	109.73	79.53	81.38
3/1/2018	156.38	43.43	109.15	79.05	85.13
4/1/2018	167.44	43.21	100.94	77.36	87.1
5/1/2018	160.01	43	100.25	75.67	81.33
6/1/2018	156.69	43.86	108.87	77.43	78.22
7/1/2018	157.54	46.63	115	81.51	79.29
8/1/2018	162.23	44.57	112.01	83.75	86.89
9/1/2018	167.29	46.19	111.8	83.2	90.91
10/1/2018	176.9	47.88	112.38	84.28	90.41
11/1/2018	188.51	50.4	121.94	85.22	92.22
12/1/2018	177.57	47.35	110.48	80.96	91.92
1/1/2019	178.78	48.13	112.67	87.24	93.98
2/1/2019	183.84	45.34	115.64	90.33	94.5
3/1/2019	189.9	46.86	122.55	95.32	99.81
4/1/2019	197.57	49.06	128.05	96.53	104.39
5/1/2019	198.27	49.13	128	99.16	102.35
6/1/2019	207.66	50.92	131.13	103.4	110.67
7/1/2019	210.72	52.63	127.81	106.08	112.52
8/1/2019	217.97	55.04	136.73	112.39	116.78
9/1/2019	214.71	54.44	137.1	108.4	113.43
10/1/2019	196.7	54.43	137.17	107.16	101.71
11/1/2019	194.48	53.4	135.83	103.94	100.67
12/1/2019	197.61	55.35	136.67	108.26	100.73

Table 4.2: Rate of Return of Five Stocks

Returns				
MCD	KO	PEP	NSRGY	YUM
6.988317	5.173673	5.544897	2.143511	12.21639
-1.47624	-6.35104	-3.39463	-3.7492	-2.94662
-0.91338	0.024666	-0.5229	3.190643	9.197151
-0.64216	0.986188	1.377205	-0.2319	4.827823
-0.89649	-4.22363	-3.2044	-6.81818	-0.03329
5.038394	4.715784	3.224772	4.767175	-2.57548
-4.84679	-4.28433	-3.54956	-2.57936	-9.10437
3.693961	2.034585	1.474231	2.158854	0.225646
13.92469	5.558323	8.366913	1.302503	-11.3071
1.701563	0.637547	-1.98649	-2.99134	2.256376
3.48633	0.797748	-0.23962	0.64917	0.744725
4.773997	-0.09311	-0.62049	-0.91373	-0.93087
-5.32396	0.489289	-1.49044	-5.26173	0.138181
7.244644	7.558539	4.763855	6.799313	12.94329
0.644492	-3.42746	0.46838	-0.05361	-2.7978
-3.50225	-0.44643	-1.73854	-0.89848	3.179993
-1.40914	1.63678	4.71484	4.614338	1.011082
-2.23533	-3.75028	2.812909	3.777008	7.838874
-1.69146	-0.4584	-1.99229	-0.90989	1.442631
-0.25938	-2.55584	1.892269	-0.60378	0.110243
-2.41852	0.189041	-1.44341	-8.06124	-4.98844
5.95185	-4.83491	-6.62313	-7.36407	2.196906
2.054166	2.750932	4.525474	6.597318	-0.09465
0.698323	0.265318	-0.81238	2.06301	3.473862
4.144572	0.938174	6.359605	0.641902	-0.32046
1.535448	1.143947	1.340826	4.356086	-2.17391
7.962342	1.672957	1.269442	0.130036	2.895145
7.832492	5.376591	3.169142	10.59741	10.47909
1.504411	-1.36354	-1.1808	2.395483	1.541854
1.292762	2.207362	0.969784	-3.33715	2.331888
3.113318	-0.63264	-0.75465	0.5576	1.775301

-2.05664	-1.18551	-3.71555	-0.88485	-4.1786
6.529239	2.155081	-1.07691	0.261875	1.141144
3.031572	-0.45672	5.706245	1.400927	12.11552
0.087221	0.240334	2.917955	0.655657	-2.22835
-0.56937	3.727112	0.316882	0.500176	3.651507
-7.82984	-9.1826	-8.78637	-7.95139	-3.79477
-0.86218	0.485884	-0.52857	-0.60354	4.608012
7.072514	-0.50656	-7.52176	-2.13789	2.314109
-4.43741	-0.486	-0.68358	-2.1846	-6.62457
-2.07487	2.000002	8.598507	2.325891	-3.82393
0.542467	6.315549	5.630566	5.269278	1.367937
2.977024	-4.41776	-2.6	2.748127	9.585065
3.119027	3.63473	-0.18748	-0.65672	4.626545
5.744516	3.658805	0.518778	1.298079	-0.54999
6.563031	5.26316	8.506856	1.115332	2.001988
-5.8034	-6.05159	-9.39806	-4.99883	-0.32531
0.681417	1.647314	1.982255	7.756916	2.241085
2.830293	-5.7968	2.636018	3.541958	0.553306
3.296344	3.35245	5.975445	5.524187	5.619046
4.038975	4.694836	4.487964	1.269407	4.58872
0.354303	0.142682	-0.03905	2.724547	-1.95421
4.735966	3.643389	2.445316	4.275916	8.128969
1.473561	3.358215	-2.53184	2.591876	1.671636
3.440585	4.579137	6.979108	5.948338	3.785995
-1.49562	-1.09012	0.270614	-3.55014	-2.86864
-8.38806	-0.01837	0.051052	-1.14391	-10.3324
-1.12862	-1.89234	-0.97689	-3.00485	-1.02252
1.609423	3.651678	0.618417	4.156244	0.059606

Table 4.3: Statistical Measures of Five Stocks

Statistics					
	MCD	KO	PEP	NSRGY	YUM
Average monthly return	1.38055	0.562743	0.717233	0.664673	1.252719
Monthly variance	17.44747	12.0968	15.6372	15.18934	25.63664
Annual Return	16.5666	6.752912	8.606792	7.976075	15.03263
Annual variance	209.3697	145.1616	187.6464	182.2721	307.6397

Table 4.4: Excess Return of Five Stocks

Excess Returns				
MCD	KO	PEP	NSRGY	YUM
5.607767	4.61093	4.827664	1.478838	10.96367
-2.85679	-6.91378	-4.11186	-4.41387	-4.19934
-2.29393	-0.53808	-1.24014	2.52597	7.944432
-2.02271	0.423446	0.659972	-0.89657	3.575104
-2.27704	-4.78637	-3.92163	-7.48285	-1.28601
3.657844	4.153041	2.507539	4.102503	-3.8282
-6.22734	-4.84707	-4.26679	-3.24403	-10.3571
2.313411	1.471843	0.756998	1.494181	-1.02707
12.54414	4.99558	7.64968	0.63783	-12.5598
0.321013	0.074804	-2.70373	-3.65601	1.003657
2.105781	0.235005	-0.95686	-0.0155	-0.50799
3.393447	-0.65585	-1.33772	-1.57841	-2.18359
-6.70451	-0.07345	-2.20767	-5.9264	-1.11454
5.864094	6.995796	4.046622	6.13464	11.69057
-0.73606	-3.99021	-0.24885	-0.71829	-4.05052
-4.8828	-1.00917	-2.45577	-1.56315	1.927274
-2.78969	1.074038	3.997608	3.949665	-0.24164
-3.61588	-4.31302	2.095676	3.112335	6.586155
-3.07201	-1.02115	-2.70952	-1.57456	0.189912
-1.63993	-3.11858	1.175036	-1.26845	-1.14248
-3.79907	-0.3737	-2.16065	-8.72592	-6.24116
4.5713	-5.39766	-7.34037	-8.02875	0.944187
0.673616	2.188189	3.808241	5.932645	-1.34737
-0.68223	-0.29742	-1.52962	1.398337	2.221143
2.764022	0.375432	5.642372	-0.02277	-1.57318
0.154898	0.581204	0.623594	3.691413	-3.42663
6.581792	1.110215	0.552209	-0.53464	1.642426
6.451942	4.813848	2.451909	9.932735	9.226367
0.123861	-1.92629	-1.89804	1.73081	0.289135
-0.08779	1.64462	0.252551	-4.00182	1.079169
1.732768	-1.19538	-1.47188	-0.10707	0.522582
-3.43719	-1.74826	-4.43278	-1.54952	-5.43132

5.148689	1.592339	-1.79414	-0.4028	-0.11158
1.651022	-1.01946	4.989012	0.736254	10.8628
-1.29333	-0.32241	2.200722	-0.00902	-3.48106
-1.94992	3.164369	-0.40035	-0.1645	2.398788
-9.21039	-9.74534	-9.5036	-8.61607	-5.04749
-2.24273	-0.07686	-1.2458	-1.26821	3.355293
5.691964	-1.06931	-8.23899	-2.80256	1.06139
-5.81796	-1.04874	-1.40081	-2.84927	-7.87728
-3.45542	1.43726	7.881274	1.661219	-5.07665
-0.83808	5.752807	4.913333	4.604605	0.115218
1.596474	-4.9805	-3.31723	2.083454	8.332346
1.738477	3.071987	-0.90472	-1.32139	3.373826
4.363966	3.096062	-0.19845	0.633406	-1.80271
5.182481	4.700417	7.789624	0.450659	0.749268
-7.18395	-6.61434	-10.1153	-5.6635	-1.57803
-0.69913	1.084571	1.265022	7.092243	0.988366
1.449743	-6.35954	1.918785	2.877285	-0.69941
1.915794	2.789708	5.258212	4.859514	4.366326
2.658425	4.132093	3.770731	0.604734	3.336001
-1.02625	-0.42006	-0.75628	2.059874	-3.20693
3.355416	3.080646	1.728084	3.611243	6.87625
0.093011	2.795472	-3.24908	1.927203	0.418916
2.060035	4.016395	6.261876	5.283665	2.533276
-2.87617	-1.65286	-0.44662	-4.21481	-4.12136
-9.76861	-0.58111	-0.66618	-1.80858	-11.5851
-2.50917	-2.45508	-1.69412	-3.66953	-2.27524
0.228873	3.088935	-0.09882	3.491571	-1.19311
-1.38055	-0.56274	-0.71723	-0.66467	-1.25272

Table 4.5: Variance-Covariance Matrix

Variance-Covariance Matrix					
	MCD	KO	PEP	NSRGY	YUM
MCD	209.764	96.36237	86.09831	82.81609	87.51948
KO	96.36237	145.2271	111.9512	97.35758	57.56158
PEP	86.09831	111.9512	187.7528	119.2149	45.37838
NSRGY	82.81609	97.35758	119.2149	182.3635	98.58033
YUM	87.51948	57.56158	45.37838	98.58033	307.9644

Risk-Free Rate

To estimate investment returns and figure out the best way to distribute assets within a portfolio, the risk-free rate is a critical factor.

Risk-Free Rate = 3.88

Table 4.6: Annual Returns of Five Stocks

Annual Returns	
MCD	16.5666
KO	6.752912
PEP	8.606792
NSRGY	7.976075
YUM	15.03263

Equally-Weighted Portfolio	
MCD	0.2
KO	0.2
PEP	0.2
NSRGY	0.2
YUM	0.2
Total	1

Expected Return	10.99
Risk	10.58
Sharp Ratio	0.67

Optimally -Weighted Portfolio	
MCD	**0.704547**
KO	0
PEP	0
NSRGY	0
YUM	**0.295453**
Total	1
Expected Return	**16.11**
Risk	**12.94**
Sharp Ratio	**0.95**

In conclusion, if we agree with this result, to optimize our portfolio, we only invest 70% of our money in MCD and the remaining 30% in YUM. For the rest of the stocks, we do not invest.

Chapter 5
Conclusion

Since there are several approaches to doing the analysis and interpreting the findings, the cluster analysis method is unable to yield a single, accurate response. To determine if cluster analysis is a useful tool for developing investing strategies, more investigation is required.

Due to the ongoing changes in the financial market, this investment plan is already out of date and must be modified frequently. This investment plan should be approached cautiously as it could not be validated, as just one method produced an analytical result. Even more dubious was the outcome of the non-scientific method used to determine the number of clusters. The status of the economy has an impact on the investment strategy, which might be challenging to account for when doing the calculations. Reliability increases with the amount of data used. A cluster analysis-based investing approach is likely to be successful and yield a modest but good return if further research is done in this field.

Chapter 6
References

References

[1] Anderson, G. W., Guionnet, A., & Zeitouni, O. (2010). An introduction to random matrices (No. 118). Cambridge university press.

[2] Ang, A., & Chen, J. (2002). Asymmetric correlations of equity portfolios. Journal of financial Economics, 63(3), 443-494.

[3] Best, M. J., & Grauer, R. R. (1991). On the sensitivity of mean-variance-efficient portfolios to changes in asset means: some analytical and computational results. The review of financial studies, 4(2), 315-342.

[4] Bhattacharyya, R., & Kar, S. (2011). Possibilistic mean-variance-skewness portfolio selection models. International Journal of Operations Research, 8(3), 44-56.

[5] Black, F., & Litterman, R. (1992). Global portfolio optimization. Financial analysts journal, 48(5), 28-43.

[6] Bollerslev, T., Engle, R. F., & Wooldridge, J. M. (1988). A capital asset pricing model with time-varying covariances. Journal of political Economy, 96(1), 116-131.

[7] Britten-Jones, M. (2002). Portfolio optimization and Bayesian regression. Unpublished working paper, London Business School.

[8] Brown, S. J. (1978). The portfolio choice problem: Comparison of certainty equivalence and optimal Bayes portfolios. Communications in Statistics-Simulation and Computation, 7(4), 321-334.

[9] Carlsson, C., Fullér, R., & Majlender, P. (2002). A possibilistic approach to selecting portfolios with highest utility score. Fuzzy sets and systems, 131(1), 13-21.

[10] Chunhachinda, P., Dandapani, K., Hamid, S., & Prakash, A. J. (1997). Portfolio selection and skewness: Evidence from international stock markets. Journal of Banking & Finance, 21(2), 143-167.

[11] Drożdż, S., Kwapień, J., Grümmer, F., Ruf, F., & Speth, J. (2001). Quantifying the dynamics of financial correlations. Physica A: Statistical Mechanics and its Applications, 299(1-2), 144-153.

[12] Elton, E. J., & Gruber, M. J. (2003). Modern portfolio theory and investment analysis. Language, 14(705p), 28cm.

[13] Engle, R. (2002). Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models. Journal of Business & Economic Statistics, 20(3), 339-350.

[14] Erb, C. B., Harvey, C. R., & Viskanta, T. E. (1994). Forecasting international equity correlations. Financial analysts journal, 50(6), 32-45.

[15] Feiring, B. R., Wong, W., Poon, M., & Chan, Y. C. (1994). Portfolio selection in downside risk optimization approach: application to the Hong Kong stock market. International Journal of Systems Science, 25(11), 1921-1929.

[16] Frost, P. A., & Savarino, J. E. (1986). An empirical Bayes approach to efficient portfolio selection. Journal of Financial and Quantitative Analysis, 21(3), 293-305.

[17] Galluccio, S., Bouchaud, J. P., & Potters, M. (1998). Rational decisions, random matrices and spin glasses. Physica A: Statistical Mechanics and its Applications, 259(3-4), 449-456.

[18] Goetzmann, W. N., Li, L., & Rouwenhorst, K. G. (2001). Long-term global market correlations.

[19] Gopikrishnan, P., Rosenow, B., Plerou, V., & Stanley, H. E. (2001). Quantifying and interpreting collective behavior in financial markets. Physical Review E, 64(3), 035106.

[20] Guhr, T., & Kälber, B. (2003). A new method to estimate the noise in financial correlation matrices. Journal of Physics A: Mathematical and General, 36(12), 3009.

[21] Huang, X. (2008). Mean-semivariance models for fuzzy portfolio selection. Journal of computational and applied mathematics, 217(1), 1-8.

[22] Jobson, J. D. (1979). Improved estimation for Markowitz portfolios using James-Stein type estimators. In Proceedings of the American Statistical Association, Business and Economics Statistics Section (Vol. 71, pp. 279-284).

[23] Jobson, J. D., & Korkie, B. (1980). Estimation for Markowitz efficient portfolios. Journal of the American Statistical Association, 75(371), 544-554.

[24] Jorion, P. (1992). Portfolio optimization in practice. Financial analysts journal, 48(1), 68-74.

[25] Jorion, P. (1986). Bayes-Stein estimation for portfolio analysis. Journal of Financial and Quantitative analysis, 21(3), 279-292.

[26] Kandel, S., & Stambaugh, R. F. (1996). On the predictability of stock returns: an asset-allocation perspective. The Journal of Finance, 51(2), 385-424.

[27] Klein, R. W., & Bawa, V. S. (1976). The effect of estimation risk on optimal portfolio choice. Journal of financial economics, 3(3), 215-231.

[28] Konno, H., Shirakawa, H., & Yamazaki, H. (1993). A mean-absolute deviation-skewness portfolio optimization model. Annals of Operations Research, 45(1), 205-220.

[29] Laloux, L., & Cizeau, P. (1999). J-. P Bouchaud, M. Potters. Phys. Rev. Lett, 83, 1467.

[30] Laloux, L., Cizeau, P., Potters, M., & Bouchaud, J. P. (2000). Random matrix theory and financial correlations. International Journal of Theoretical and Applied Finance, 3(03), 391-397.

[31] Ledoit, O., Santa-Clara, P., & Wolf, M. (2003). Flexible multivariate GARCH modeling with an application to international stock markets. Review of Economics and Statistics, 85(3), 735-747.

[32] León, T., Liern, V., & Vercher, E. (2002). Viability of infeasible portfolio selection problems: A fuzzy approach. European Journal of Operational Research, 139(1), 178-189.

[33] Liu, S. Y. W. S., Wang, S. Y., & Qiu, W. 2. (2003). Mean-variance-skewness model for portfolio selection with transaction costs. International Journal of Systems Science, 34(4), 255-262.

[34] Longin, F., & Solnik, B. (1995). Is the correlation in international equity returns constant: 1960–1990?. Journal of international money and finance, 14(1), 3-26.

[35] MacLean, L. C., & Ziemba, W. T. (2013). Handbook of the fundamentals of financial decision making (Vol. 4). World Scientific.

[36] Markowitz, H., Todd, P., Xu, G., & Yamane, Y. (1993). Computation of mean-semivariance efficient sets by the critical line algorithm. Annals of operations research, 45, 307-317.

[37] Markovitz, H. M. (1959). Portfolio selection: Efficient diversification of investments. John Wiley.

[38] Michaud, R. O., & Michaud, R. (2007). Estimation error and portfolio optimization: a resampling solution. Available at SSRN 2658657.

[39] Moskowitz, T. J. (2003). An analysis of covariance risk and pricing anomalies. The Review of Financial Studies, 16(2), 417-457.

[40] Pafka, S., & Kondor, I. (2003). Noisy covariance matrices and portfolio optimization II. Physica A: Statistical Mechanics and its Applications, 319, 487-494.

[41] Pástor, Ľ., & Stambaugh, R. F. (2000). Comparing asset pricing models: an investment perspective. Journal of Financial Economics, 56(3), 335-381.

[42] Pástor, Ľ. (2000). Portfolio selection and asset pricing models. The Journal of Finance, 55(1), 179-223.

[43] Plerou, V., Gopikrishnan, P., Rosenow, B., Amaral, L. A. N., Guhr, T., & Stanley, H. E. (2002). Random matrix approach to cross correlations in financial data. Physical Review E, 65(6), 066126.

[44] Plerou, V., Gopikrishnan, P., Rosenow, B., Amaral, L. A. N., & Stanley, H. E. (1999). Universal and nonuniversal properties of cross correlations in financial time series. Physical review letters, 83(7), 1471.

[45] Polson, N. G., & Tew, B. V. (2000). Bayesian portfolio selection: An empirical analysis of the S&P 500 index 1970–1996. Journal of Business & Economic Statistics, 18(2), 164-173.

[46] Ravipati, A. (2012). Markowitz's portfolio selection model and related problems (Doctoral dissertation, Rutgers University-Graduate School-New Brunswick).

[47] Rosenow, B., Gopikrishnan, P., Plerou, V., & Stanley, H. E. (2003). Dynamics of cross-correlations in the stock market. Physica A: Statistical Mechanics and its Applications, 324(1-2), 241-246.

[48] Rosenow, B., Plerou, V., Gopikrishnan, P., & Stanley, H. E. (2002). Portfolio optimization and the random magnet problem. Europhysics Letters, 59(4), 500.

[49] Tanaka, H., & Guo, P. (1999). Portfolio selection based on upper and lower exponential possibility distributions. European Journal of operational research, 114(1), 115-126.

[50] Tanaka, H., Guo, P., & Türksen, I. B. (2000). Portfolio selection based on fuzzy probabilities and possibility distributions. Fuzzy sets and systems, 111(3), 387-397.

[51] Treynor, J. L., & Black, F. (1973). How to use security analysis to improve portfolio selection. The journal of business, 46(1), 66-86.

[52] Vercher, E., Bermúdez, J. D., & Segura, J. V. (2007). Fuzzy portfolio optimization under downside risk measures. Fuzzy sets and systems, 158(7), 769-782.

[53] Zellner, A., & Chetty, V. K. (1965). Prediction and decision problems in regression models from the Bayesian point of view. Journal of the American Statistical Association, 60(310), 608-616.