Muhammad Azfar Faizan

Multiobjective Portfolio Optimization Including Sentiment Analysis

Master's thesis of mathematical information technology

May 28, 2019

University of Jyväskylä

Faculty of Information Technology

Author: Muhammad Azfar Faizan

Contact information: azfar-faizan@hotmail.com

Supervisors: Prof. Kaisa Miettinen and Dr. Markus Hartikainen

Title: Multiobjective Portfolio Optimization Including Sentiment Analysis

Project: Master's thesis

Study line: Operations Research

Page count: 56

Keywords: Multiobjective optimization, big data analysis, sentiment analysis, portfolio optimization, time series.

Table of Contents

List of Figures

List of Tables

Abstract

Volatility (or risk) in stock market is a crucial factor that has always been of great interest to investors to facilitate the decision making about their investments. The two core objectives of investors are optimization of volatility and generation of returns at the same time. One can also assume that news can be a factor which can determine volatility when combined with daily returns. In this study we used multiobjective optimization and sentiment analysis of news data together to create two models. In the first multiobjective optimization model, we optimize risk and returns using the conventional formulation and daily returns data. In the second multiobjective optimization model, we again optimize risk and returns but calculate returns differently using daily returns as well as sentiment analysis using news data to see if the model including news behaves differently as compared to the conventional model. The results of both the models have been analyzed in this study. It has been found that while keeping several factors constant, we found no difference in the risk and return of both the models.

Chapter 1: Introduction

The volatility of stock markets (Tan, L., Chen, Zheng, & Ouyang, 2016) has been a challenge for investors to invest money and make profits out of it. To counter these challenges in mind and various options available for investment, investors investing in various domains are always in search of a feasible portfolio where the objectives include are at minimum risk and maximum returns (Bata & Richardson, 2018; Kralewski, Gifford, & Porter, 1988; Peters & McKay, 2014; Popa, Holvoet, Van Montfort, Groeneveld, & Simoens, 2018). There are various techniques used by investors to consider these two objectives but the initial concept was proposed by Markowitz (1952) termed "mean-variance portfolio" which forms the basis for modern portfolio theory as a mathematical optimization model (Atta Mills, Yan, Yu, & Wei, 2016; Fan, Zhang, & Yu, 2012; Zhang, C., 2014).

Optimizations have always posed challenges due to the multiple aspects of a decision which may affect investors' decision making process. Previous attempts to solve the optimization related challenges focused mainly on single objective optimizations (Chang, 2015) which has its limitations in the current era to that poses multiple objective challenges. The implementation of modern portfolio theory (Liao et al., 2018) utilizes expected risk and returns of different assets and takes those results as an input to a mean-variance optimization model, which is a part of modern portfolio theory. The outcome of optimization through this model (Draviam, 2008) provides a curve of "feasible" combinations of risk and return which can be interpreted as maximum average returns that are expected to be obtained for a particular level of risk. If an investor expects higher returns on the investment, then the associated level of risk to that investment will be high.

For decades, the mean-variance optimization model has remained an incontrovertible quantitative model for investors. Although the theory was criticized during the financial crisis in 2007-08 (MARAKBI, 2016), Markowitz defended their theory (Markowitz, H. M., 2010) by denying the statement that the theory state predictability of the markets. In fact, according to them, markets are often very volatile for short time intervals, therefore, the theory is still valid and widely used by researchers and financial analysts (Markowitz, H. M., 2010).

With the increasing attention of researchers towards multi-objective optimization in the last few decades (Boada, Reynoso-Meza, Pico, & Vignoni, 2016; Dong, Zeng, & Chen, 2012; Dujardin & Chades,

2018; Ganesan, Elamvazuthi, Shaari, & Vasant, 2013; Liao et al., 2018; Song, Wang, Dai, & Vasile, 2015), there exists a rising need of additional objectives in the mean-variance optimization model (Woo Chang Kim, 2015). The reason for this is that investors today consider other objectives as well while selecting a feasible portfolio. According (Ralph E. Steuer, 2005) a traditional investor focuses on maximum returns without recognizing the importance of additional objectives, whereas a non-traditional investor considers additional objectives in the set of objectives to be considered in order to get better off. The claim of a traditional investor of not being concerned about objectives other than returns is because of efficient market hypothesis (EMH) (Farmer & Lo, 1999; Zhang, H., Wei, & Huang, 2014) which exhibits that other factors are captured in prices of securities and therefore no other objectives are required to be incorporate.

The efficient market hypothesis has been questioned in financial markets among researchers (Biondo, Pluchino, Rapisarda, & Helbing, 2013; Patzelt & Pawelzik, 2013; Yim, Oh, & Kim, 2016) and provides a clue that markets must not be expected to remain efficient all the time. (Winfried Hallerbach, 2002) also highlighted the benefits of incorporating multiple objectives in financial decision making (Spronk, 2003). In order to counter the traditional investor's claim, (Sun, 2005) gave another view for incorporating multiple objectives in portfolio selection. This view incorporates the accuracy of the investors investing in a security just for the sake of maximum returns. It has been argued that the investor does not fully trust the data which represents portfolio returns because of randomness of stock markets. Therefore, it has been concluded to be reasonable to consider multiple objectives in the portfolio selection.

Apart from stock market, investors are also interested in investing in gold, foreign exchange and other long term bonds depending on the level of risk and returns they target. For example, in foreign exchange investment (Sanchez-Granero, Trinidad-Segovia, Clara-Rahola, Puertas, & De Las Nieves, 2017) investors must be well informed since these investments are more volatile when compared to stock market (Piskorec et al., 2014). Gold prices have a comparatively low level of risk when compared to stocks and foreign exchange (Chu, Nadarajah, & Chan, 2015) and long term bonds are completely risk free (McKay & Peters, 2017). However, returns are significantly lower in bond investments and they are not liquid either (Arthur, Williams, & Delfabbro, 2016; Deluccia, 1989).

While there is a focus of researchers in incorporating multiple objectives in portfolio selection, an

investor also seeks market information from specific company news or other socio political news which may affect the market in a positive or negative way (Cohen-Charash, Scherbaum, Kammeyer-Mueller, & Staw, 2013). There is a possibility that an investment which performs well over some time is connected with a bad news or related industry is affected (Yang, W. et al., 2018) and this might result in the loss of confidence of investors and the prices will fall (Heiberger, 2015; Ranco, Aleksovski, Caldarelli, Grcar, & Mozetic, 2015). A very simple and well known example are the housing prices in USA which were continuously showing positive trends for 5 decades and suddenly started to fall during the financial crisis of 2007-08 (Tan, J. & Cheong, 2016). Investments were taken out as investors were losing their confidence in housing investments (Luc Eyraud, 2007)

Investors remain continuously busy in gathering and analyzing information for a variety of different assets (Yang, Z. H., Liu, Yu, & Han, 2017). They monitor the global economy and also read and analyze the news well before reaching any decision of making an investment (Gonzalez Mde, Jareno, & Skinner, 2016; Yang, Z. H. et al., 2017). It is believed that the more information an investor has, the more chances there are to turn an investment into profit (Gonzalez Mde et al., 2016). Investors are usually good in number crunching and quantitative models but recently a positive shift has been seen in analyzing news, blogs and domain expert articles at the same time (Harlow & Oswald, 2016; Preis, Moat, & Stanley, 2013) in order to build quantitative models including textual data (human perception data) which surround the market and then come up with a better decision for investing (Brown & Ridderinkhof, 2009; Wang, Vieito, & Ma, 2015).

There are large numbers of securities which investors usually handle (Copeland & Jacobs, 1981; Williams, Brown, & Healy, 2018) that influence investors' decision making. News and other sources of information which provide real time news about the socio politics and latest news related to a particular company can be one of the strong determinants of investor's decision making. Since it is humanly impossible for an investor to incorporate all the news, articles and blogs for a separate analysis, therefore, relating it to the market situation for making a buying or selling decision for an investment is challenging and comes with many risks (Lan, Xiong, He, & Ma, 2018; Wang, Jin, Vieito, & Ma, 2017). In order to overcome this problem, they can use sentiment analysis (Hopper & Uriyo, 2015; Ranco et al., 2015; Steinert & Herff, 2018) as a tool to determine whether a piece of news, a blog or an article reflects a positive or negative outcome of an investment.

Sentiment analysis and its use in portfolio selection have recently been gaining popularity (Carrillo-de-Albornoz, Rodriguez Vidal, & Plaza, 2018; Lv, Huang, Li, & Xiang, 2019; Ranco et al., 2016). Similarly, Shaun and Marco (Roache & Rossi, 2009) showed that gold prices react on scheduled announcements from the authorities but otherwise gold is a safe haven for investors due to low volatility. Investors are now also using sentiment analysis on news and social media data (Freedman, Viswanath, Vaz-Luis, & Keating, 2016; Gabarron, Dorronzoro, Rivera-Romero, & Wynn, 2019; Novak, Amicis, & Mozetic, 2018; Reyes-Menendez, Saura, & Alvarez-Alonso, 2018; Vickey & Breslin, 2017; Zheludev, Smith, & Aste, 2014) in order to determine the optimal portfolio not only by using multiple objectives but also incorporating social media sentiments and news sentiments in their analysis. Capturing market sentiments (Dhar, 2014) (discussed in chapter 4 in detail) has successfully allowed researchers to predict the values of their portfolio (Kim & Sayama, 2017) and therefore it has now been used on a regular basis (Pai, Hong, & Lin, 2018; Xu, Liu, Zhao, & Su, 2017) as an integral step towards the decision of making an investment.

The current study combines the two significant tools in optimal portfolio selection in order to facilitate the investor in decision making. The first tool is the multiobjective optimization approach to counter more than one objectives at the same time i.e. risk and return in our case and the second tool is performing sentiment analysis on the news data to determine whether to invest or not in a good looking investment by capturing and incorporating the surrounding news in our model.

The objective of the research is to determine that whether incorporating the news data in our optimization model can make a significant difference in selecting feasible proportions for each investment including stock, gold, foreign exchange (FOREX), real estate, bonds and bitcoins. The research has been performed in four steps. In the first step, data has been collected from various sources and cleaned such that it looks comparable. In order to process the textual news data and calculate sentiment score from it we have used IBM Watson. In the second step we normalize the data such that we get all the values between -1 and 1. In the third step, textual and quantitative data have been combined together using linear regression model to calculate expected returns. In the fourth step, multiobjective optimization methods (ε-constraint and reference point) have been applied to determine any significant differences in the model which incorporate news data and which does not incorporate news data.

The data were collected from various sources published during January 2013 till October 2017 contributing to more than 53 months of data. The reason for selecting the mentioned time period is that it allows us to have maximum availability of the data regarding the variables of interest. We have used several economic indicators including stock market returns, real estate, bitcoin, gold, foreign exchange, bonds with a specific focus on the news data. For this reason, the availability of all the data at the same time was a challenge. So by using this time period we make sure that we have maximum amount of data available for our optimization model.

The results of the multiobjective optimization 1) news data model and 2) without news data model we built in this research didn't show significant differences. There could be three possible reasons for the insignificant differences which are explained in section 6.5. So if we keep these factors constant we can reach to the conclusion that incorporating news data didn't make a significant difference in selecting feasible proportions for each investment.

Chapter 2: Basics of Multiobjective Optimization

2.1. Multiobjective Optimization

This chapter is based on Miettinen (Miettinen, 1999) to provide the basic theoretical concepts in multiobjective optimization (Miettinen, 1999). A general multiobjective optimization model can be formulated as follows:

$$\min\{f_1(x), f_2(x), \dots, f_k(x)\} \qquad (1)$$

subject to $x \in S$,

including $k \geq 2$ conflicting objective functions $f_i : S \to R$ that we want to minimize simultaneously for $x \in S$. The decision vectors $x = (x_1, x_2, \dots, x_n)^T$ belong to the region S $\subset R^n$ which is a non-empty feasible region. Objective vectors are the images of the decision vectors in the objective space having objective values $z = f(x) = \big(f_1(x), f_2(x), \dots, f_k(x)\big)^T$. Therefore, the image of the feasible region in the objective space can be defined as $Z = f(S)$.

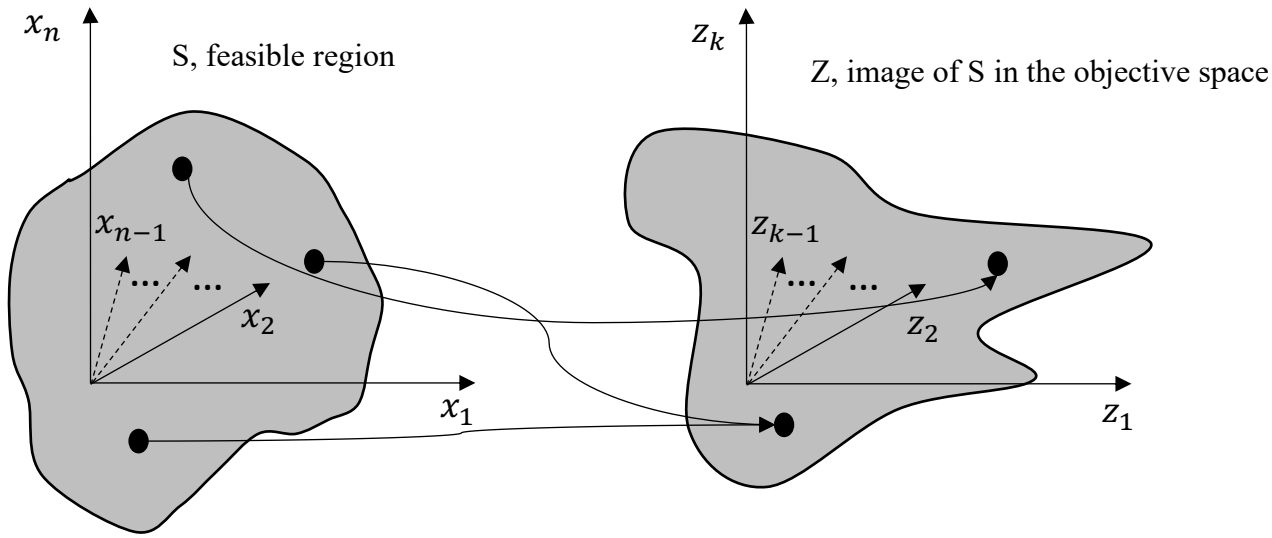

Figure 2.1: Decision space vs. objective space

A solution of a multiobjective optimization problem is considered to be optimal if no components z_i of the objective vector z can be improved without impairing at least one of the other components z_j $(j = 1,2, \dots, k \;\&\; j \neq i)$. In general, a decision vector $x^* \in S$ is called **Pareto optimal** if there does not exist another $x \in S$ such that $f_i(x) \leq f_i(x^*)$ for all $i = 1,2, \dots, k$ and $f_j(x) < f_j(x^*)$ for at least one index j. Similarly, an objective vector z is Pareto optimal if the decision vector x corresponding to it is Pareto optimal.

2.1.1. Some Basic Concepts

Lower and upper bounds of objective function values in the set of Pareto optimal solutions are called ideal and nadir objective vectors, respectively. Let an ideal objective vector be denoted by z^* and a nadir objective vector by z^{nad}. A vector strictly better than the ideal (z^*) is called a utopian objective vector denoted by z^{**}. These concepts are defined below:

- **Pareto optimal set** is a set of Pareto optimal solutions.
- **Payoff table** is calculated by minimizing each f_i for $i = 1,2, \dots k$ and using the same optimal values of decision variables $x \in S$ for all corresponding k objective functions. It is formed by minimizing each f_i which will form the rows and the optimized values after optimizing each f_i will become columns. Table 2.1 shows an example of a payoff table containing random values but meets the requirements to form a payoff table.

	z_1	z_2	z_3
z_1 value at min	10	1.5	150
z_2 value at min	40	0.2	120
z_3 value at min	30	0.75	100

Table 2.1: Payoff Table

- **Ideal objective vector or ideal vector (z^*)** provides the lower bounds of objective values in the Pareto optimal set and consists of the individual minima of each k objective functions. Thus, the ideal vector can be written as $z^* = (z_1^*, z_2^*, \dots, z_k^*)^T$, where z_i^* is the minimum of the i^{th}objective function subject to S, for all k objectives $i = 1,2, \dots, k$. In Table 2.1, components of the ideal vector lie on the diagonal which is $z^* = [10,\ 0.2,\ 100]$.
- **Nadir objective vector or nadir vector (z^{nad})** provides the upper bounds of objective values in the Pareto optimal set, that is constructed with worst value of objective functions in the complete Pareto-optimal set. In Table 2.1, nadir objective vector (z^{nad}) will be $z^{nad} = [40,\ 1.5,\ 150]$.
- **Utopian objective vector (z^{**})** is a vector which is strictly better than the ideal vector (z^*) and can be formulated as $z_i^{**} = z_i^* - \epsilon_i$ where $\epsilon_i > 0$ and $i \in 1,2, \dots, k$.
- **Weakly Pareto optimal** set is a set of decision vectors $x^* \in S$ for which there does not exist $x \in S$

such that $f_i(x) \leq f_i(x^*)$ for all $i = 1,2,\dots,k$

- $\boldsymbol{\varepsilon}$ **− Proper Pareto optimality** (from (Wierzbicki, A. P., 1980)) for a decision vector $x^* \in S$ and the corresponding objective vector $z^* \in Z$ is such that

$$(\boldsymbol{z^* - R_\varepsilon^k} \setminus \{\mathbf{0}\}) \cap \boldsymbol{Z} = \emptyset,$$

where $\boldsymbol{R_\varepsilon^k} = \{\boldsymbol{z} \in \boldsymbol{R^k} | \text{dist}(\boldsymbol{z}, \boldsymbol{R_+^k}) \leq \boldsymbol{\varepsilon}\|\boldsymbol{z}\|\}$ or $\boldsymbol{R_\varepsilon^k} = \{\boldsymbol{z} \in \boldsymbol{R^k} | \boldsymbol{max_{i=1,2\dots k} z_i} + \boldsymbol{\varepsilon} \sum_{i=1}^{k} \boldsymbol{z_i} \geq \mathbf{0}\}$ and $\boldsymbol{\varepsilon} > \mathbf{0}$ is a predetermined scalar.

- **Proper Pareto optimality** according to Geoffrion (Geoffrion, 1968) is such that a decision vector $x^* \in S$ is properly Pareto optimal if it is Pareto optimal and if there exists a real number M such that for each f_i and each $x \in S$ satisfying $f_i(x) < f_i(x^*)$ there exists at least one f_j such that $f_j(x) < f_j(x^*)$ and

$$\frac{f_i(x^*) < f_i(x)}{f_j(x) < f_j(x^*)} \leq M.$$

2.2. Types of multiobjective optimization methods

According to Miettinen (Miettinen, 1999) multiobjective optimization methods can be classified into four classes which include no-preference methods, a priori methods, a posteriori methods and interactive methods (Miettinen, 1999). The classification is based on the role of a **decision maker (DM)** in the solution process. A **DM** is a person who takes part in the solution process of a multiobjective optimization problem. The DM is not required to have prior knowledge of multiobjective optimization algorithms but he/she must be an experienced person in the specific problem domain. As a domain expert the DM can specify preference information or choose a solution from a set of Pareto optimal solutions provided by an **analyst** that satisfies him/her. An **analyst** is typically a person or an artificial agent that handles the technical side of the solution process. A very brief introduction to each of the classes of multiobjective optimization methods is given below.

- **No-preference methods**

The class of no-preference methods is used when there is no DM available to provide his/her preferences to the solution process and that is why it is called no-preference methods. The idea behind the class

of methods is to find a neutral compromise solution without any preference information from the DM. This means that instead of asking a DM about the preferences, some assumptions are made about what a "reasonable" solution could be like. Apart from this class of methods, the presence of DM in the solution process is required. An example of a no-preference method is the "method of global criterion" which is discussed in the later section.

- A priori methods

In a priori methods the DM will provide preference information in the beginning and then the solution process tries to find a Pareto optimal solution that is closest meeting the DM's preference. This approach of finding a Pareto optimal solution is straight-forward but in general, when the DM decides to solve a new kind of problem, the DM may not know all the possibilities and limitations of the problem. In this case, the DM's a priori preference information might be too optimistic or even pessimistic.

- **A posteriori methods**

A posteriori methods are used as an alternative to a priori methods in a way that a representative set of Pareto optimal solutions is first generated and then the DM selects the most preferred solution that is closest or meets his/her preferences. Although this method provides a good overview of available solutions, in the presence of $k > 2$ objectives, it might become difficult for the DM to analyze large amounts of information. This is because visualization is simple in a biobjective case but it will not remain simple for $k > 2$ objectives. In addition, producing a good representative set may be computationally expensive with a large number of objective functions.

- **Interactive methods**

The idea behind an interactive approach is to form an iterative solution process that is repeated more than once to satisfy the DM's preferences. The DM is supposed to specify preference information after each iteration. During the iterative process, the analyst aims to learn the pattern of DM's preferences in an interactive way. During the interactive solution process the DM can specify his/her preferences and learn about the interdependencies in the problem. As stated by (Kaisa Miettinen, 2008) interactive methods is the most extensive class of methods but we discuss only reference point method in this study because we shall be using a reference point based interactive approach to solve our problem. An example of interactive methods is

"reference point method" which is discussed in the later section.

2.3. General Solution Process of a Multiobjective Optimization Problem

In multiobjective optimization, an integral part of decision making is the presence of a DM and an analyst. The need of an analyst is usually to define the problem specification mathematically and apply possible methods to the problem at various stages of the optimization process. We start solving a multiobjective optimization problem by calculating ideal and nadir objective vectors so that the ranges of objective functions values can be obtained in the Pareto optimal set, if the objective functions are bounded over the feasible region. For the nadir objective vector, there is no constructive way of calculating it when $k > 2$ objectives, and therefore we use a payoff table to estimate it. This technique of finding a nadir objective vector works well for $k = 2$ but for $k > 2$ it might show over-or underestimation.

After calculating the ideal and nadir objective vectors, there are various classes of methods available as explained in Section 2.2 by which we can solve the multiobjective optimization problem. For example, there are widely used methods (that fall in the classes in Section 2.2) including weighting method, ε-constraint method, method of global criterion, and achievement scalarizing function and they will be discussed in the following section. The presence of an analyst discussed previously is important because he/she has to decide based on the knowledge of multiobjective optimization that which specific method should be used once the problem has been analyzed. The choice of a method depends naturally also on the availability of preference information and the type of preference information the DM can give.

The technique which is usually used for solving multiobjective optimization problems is scalarization. **Scalarization** means converting a multiobjective optimization problem and the preference information obtained a single objective optimization problem or a family of such problems. The converted single objective optimization problem has a real-valued objective function, therefore, it can be solved using traditional single objective optimizers. These real-valued functions are also called **scalarizing functions**. As stated in (Miettinen, 1999) it is justified to use such scalarizing functions that can generate Pareto optimal solutions and can find any Pareto optimal solution. The generation of locally or globally Pareto optimal solution depends on the nature of the problem and also whether a local or a global solver is used (Miettinen, 1999). Locally Pareto

optimal solutions are typically not of DM's interest, therefore, an appropriate solver must be used in order to get globally Pareto optimal solutions and that is why the presence of an analyst during the solution process is important. It should also be noted that the solutions obtained from numerical optimization methods might not be optimal in practice because it is possible that the method never converged or the global solver failed in finding the global optimum.

Sometimes it is assumed that the DM maximizes one's utility and makes decisions on the basis of an underlying function which represents the DM's preferences. The function is called a **value function** $v: R^k \rightarrow R$. It is assumed that a value function is non-increasing with the increase in objective values because from (1) we assume that all the objective functions are minimized but the value function (or utility function) has to be maximized for a DM. Alternatively it can assumed that the intention of a DM is to find a satisficing solution instead of the maximum of a value function. **Satisficing decision making** means that the DM does not intend to maximize any value function but the intention is to achieve aspiration levels that satisfy him/her (Sawaragi, 1985). **Aspiration levels** $\bar{z}_i$ $(i = 1,2, \dots, k)$ are the desired levels of objective functions in which the DM is particularly interested because of his/her past experience in the problem domain. The vector $\bar{z} \in R^k$ consisting of aspiration levels is called a **reference point**.

It is worth noting that according to the definition of Pareto optimality, the movement from one Pareto optimal solution to another in a Pareto optimal set results in a trade-off. It shows the ratio of how much a DM has to forgo in an objective in order to achieve a particular improvement in another objective. In the following section, we describe briefly some methods for solving multiobjective optimization problems. These are not all the methods that exist for solving multiobjective optimization problems but these are basic and widely used methods.

2.4. Some Multiobjective Optimization Methods

2.4.1. Weighting method

The general formulation of a weighting method from (Miettinen, 1999) is

$$\min \sum_{i=1}^{k} w_i f_i(x) \qquad (2)$$
$$\text{subject to } x \in S,$$

where w_i are the weights given to each of k objective functions such that $w_i \geq 0$ and $\sum_{i=1}^{k} w_i = 1$. The solution from the weighting method will be Pareto optimal only if $w_i > 0$ for all objectives $i = 1,2, \dots, k$, otherwise it will be weakly Pareto optimal. The weighting method can be used both as an a priori and a posteriori method. As an a posteriori method different weights are used to generate different Pareto optimal solutions and then the DM selects the most satisficing solution (Miettinen, 1999).

There is one misconception that if we evenly distribute the weights for the problem, it will produce evenly distributed Pareto optimal solutions but it is not true (I. Das, 1997). The implementation of this method is very easy as we can assign the weights for each objective function according to our preference. On the other hand, a serious disadvantage of this method is that it does not provide all Pareto optimal solutions for a non-convex problem.

2.4.2. ε-constraint method

The idea behind the ε-constraint method is to optimize one single objective $f_l(x)$ and the remaining $j = 1,2, \dots k$ objectives $(j \neq l)$ are put in the constraints. The general formulation of the ε-constraint method is

$$\begin{aligned} & \min f_l(x) \\ \text{subject to} \quad & f_j(x) \leq \varepsilon_j \quad \text{for all } j = 1,2, \dots k, \quad j \neq l, \\ & x \in S, \end{aligned} \tag{3}$$

where $l \in \{1,2, \dots k\}$ and ε_j are the upper bounds for each objective $(j \neq l)$.

The solution of (3) will always be weakly Pareto optimal but it can be Pareto optimal if and only if it can be solved for $x^* \in S$ for every $l = 1,2, \dots k$ such that $\varepsilon_j = f_j(x^*)$ for $j = 1,2, \dots k$, $(j \neq l)$. The method unlike weighting method works well for both convex and nonconvex problems but it may be difficult to specify the upper bounds for each $k - 1$ objectives such that the feasible region will be non-empty (Miettinen, 1999). The difficulty increases as the number of objectives increases. If the method is used as an a posteriori method, a set of upper bounds is needed. On the other hand, the method can be used as an a priori method where the DM provides upper bounds ε_j according to his/her preferences and experience in the problem domain.

2.4.3. Method of Global Criterion

The method of global criterion is intended to minimize the distance between the set of Pareto optimal solutions and a particular reference point. The reference point in this method is the ideal objective vector because we assume that there is no DM available during the solution process. The ideal objective vector z^* in most cases is the natural choice for an analyst with which he/she can minimize the distance. The reason for such a choice is that if no DM is available then any Pareto optimal solution closest to ideal would be preferred. The mathematical formulation is

$$\min \left(\sum_{i=1}^{k} |f_i(x) - z_i^*|^p \right)^{\frac{1}{p}} \qquad (4)$$
$$\text{subject to } x \in S,$$

where exponent $1/p$ can be dropped or

$$\min \left(\max_{1,2,\dots k} [|f_i(x) - z_i^*|] \right) \qquad (5)$$
$$\text{subject to } x \in S.$$

We solve (4) using the L_p-metric or (5) using the Chebyshev metric to measure the distance to the ideal objective vector z^* or the utopian objective vector z^{**}. The solution obtained from (4) can be proven to be Pareto optimal and the solution from (5) weakly Pareto optimal . Absolute value sign in (4) and (5) can be removed because $f_i(x) \geq z_i^*$ for all $i = 1,2, \dots k$. It was also demonstrated in (Miettinen, 1999) that different choices for p affect the solution of the problem differently. If objectives are of different magnitudes, then scaling the functions to a uniform and dimensionless scale will allow the method to work properly (Miettinen, 1999).

2.4.4. Achievement Scalarizing Functions

Achievement scalarizing functions (ASFs) are a special type of scalarizing functions which were introduced by Wierzbicki in 1982 (Wierzbicki, A., 1982). ASFs are based on an arbitrary reference point $\bar{z} \in R^k$ and the basic approach behind the method is to project the reference point on to the set of Pareto optimal solutions. Different reference points allow the DM to get different Pareto optimal solutions. The method has its basis in the method of global criterion but with some significant changes. An example of a mathematical

formulation of an ASF is

$$\text{minimize} \max_{i=1,2,3....k} [w_i(f(x_i) - \bar{z}_i)] + \rho \sum_{i=1}^{k} (f(x_i) - \bar{z}_i) \qquad (6)$$

subject to $x \in S$,

where (6) can be understood by dividing it into two components. The first component is the formulation before the plus sign which is somewhat derived from (5). The ideal objective vector (z^*) is replaced by the reference point $\bar{z} \in R^k$ which is provided by the DM. The reference point $\bar{z}$ can be based on the DM's past experience about the problem or it might be such that the DM wants to check different Pareto optimal solutions by changing the reference point $\bar{z}$ such that he/she reaches a solution that satisfies him/her. In (6), w_i is the scaling factor of choice for example, $w_i = \frac{1}{z_i^{nad} - z_i^{**}}$ is one type of scaling factor.

The second component is after the plus sign which contains ρ as the augmentation multiplier such that $\rho > 0$ is a small positive scalar. The reason for including the augmentation part in the function is to recall Pareto optimal solutions instead of weakly Pareto optimal solutions. The solution of (6) is ε-properly Pareto optimal but depending on the formulation ASFs can provide weakly or ε-properly Pareto optimal points. It is to be noted that the method contains the reference point that does not have to be fixed as the ideal or utopian vector as in method of global criterion. No matter how the reference point is selected in the objective space, it will generate a Pareto optimal solution.

2.4.5. Reference Point Method

The reference point method was presented by Wierzbicki (1982) (Wierzbicki, A., 1982) and is based on a reference point of aspiration levels. As mentioned in Section 2.3, a reference point in the objective space can be feasible or infeasible but it represents the desirable aspiration levels given by the DM. The reference point method solves an ASF and provides solutions that are weakly or ε-properly Pareto optimal. So, the method just needs a reference point to start without any specific assumptions on the problem to be solved. There are several methods which utilize the idea of reference points but the reference point method by Wierzbicki was the first among all.

The reference point algorithm is simple to understand and it works very much like other interactive

optimization methods. The DM is provided with initial information prior to the solution process. The desired information will be the ideal and (approximated) nadir objective vectors so that the DM can understand the ranges of the Pareto optimal set. The selection of an appropriate ASF from the analyst is also required. The reference point algorithm as stated in (Miettinen, 1999) is presented in Figure 2.2 in the form of a flow chart

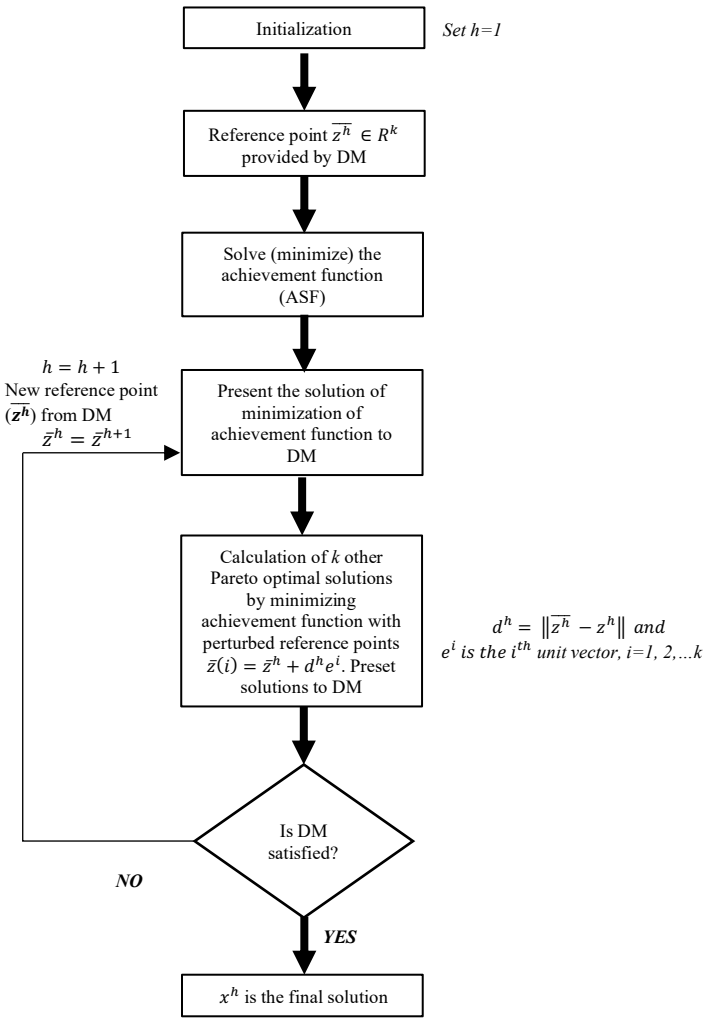

Figure 2.2: Reference Point Algorithm

2.5. Summary

Multiobjective optimization is a rapidly growing field among researchers due to the realization that real-life problems mostly require multiple objectives to be optimized at the same time. In this chapter an introduction to basic concepts in multiobjective optimization has been discussed to give an idea to the reader

of how multiobjective optimization works. Furthermore, few widely used methods were also discussed with their mathematical formulations. Another important set of methods are evolutionary algorithms which are used in multiobjective optimization a lot (Khin Lwin, 2014). The reason for not including evolutionary algorithms in our discussion is that their use comes into consideration when mathematical programming does not work. For example, in complex mathematical functions where traditional optimization methods usually give a local rather than a global optimum, evolutionary algorithms provide solutions near to global optimal solutions. Another example where we can use evolutionary algorithms are discontinuous functions. On the other hand, evolutionary algorithms are computationally expensive. In our study, evolutionary algorithms are not needed.

In our research problem we shall be using the reference point method. The reason for using the reference point method is that we are intending to solve the problem iteratively where DM provides the reference of a desired solution and the method returns a set of Pareto optimal solutions near that reference point. If the DM is not satisfied with the solution then he/she will provide another reference point interactively and will get another set of Pareto optimal solutions near the new reference point. The DM can interactively select one desired solution and then on the basis of those results, DM can change his/her reference point which will lead to different solution. We shall also be using concepts defined in this chapter regarding multiobjective optimization in the upcoming chapters.

Chapter 3: Time Series and Portfolio Optimization

3.1. Time Series & Autoregressive Models

Time series analysis is based on the collection of gathered observations over a period of time (Juang, Huang, Huang, Cheng, & Wann, 2017; Tsai, Cheng, Tsai, & Shiu, 2018; West et al., 2018). The main objective of a time series model is to collect the historical observations of an event and then develop a model which describes the inherent structure of the series. Time series analysis can also be explained as forecasting the future by understanding the past (Bertella et al., 2017; Frydman, Barberis, Camerer, Bossaerts, & Rangel, 2014). It is should be understood that an appropriate model fitting of a time series will only lead to predict the future accurately (Bertella, Pires, Feng, & Stanley, 2014; Feng, Li, Podobnik, Preis, & Stanley, 2012; Gao et al., 2014). A lot of researches have been done over the past years for developing efficient and accurate time series models. As a result, many time series models are now part of the time series literature. Figure 3.1 shows the examples of time series data.

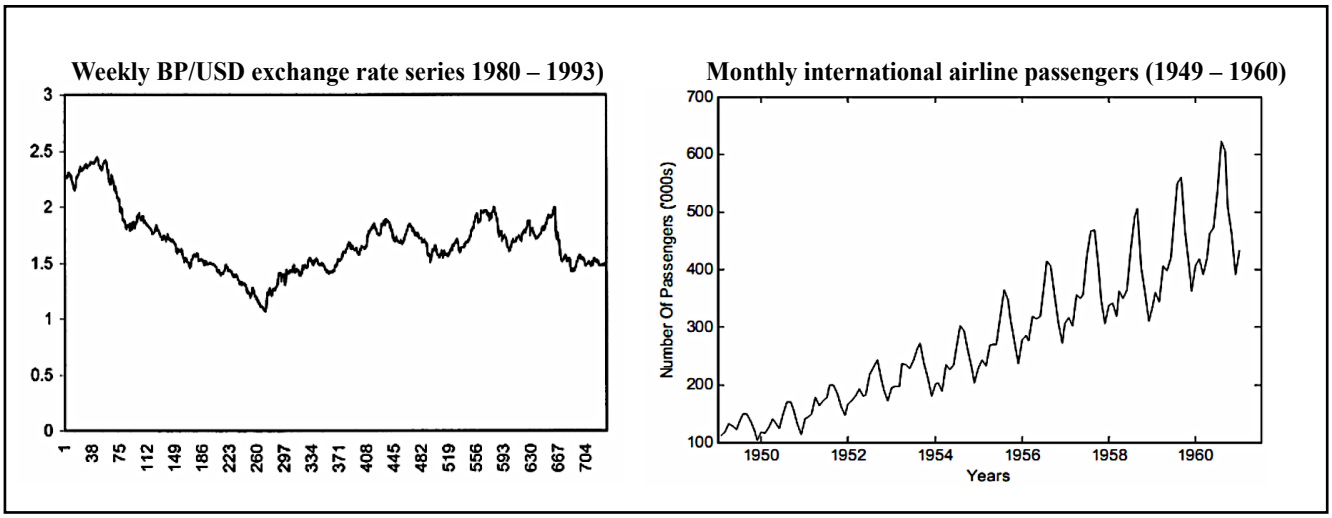

Figure 3.1: Examples of time series data. Adapted from (Hamzacebi, 2008; Zhang, G. P., 2003)

3.1.1. Concept of Stationarity

Stationarity can also be viewed as statistical equilibrium (K.W. Hipel, 1994). This includes statistical properties like mean and variance which do not depend upon time in stationarity process. One of the important applications of stationarity is its use in future forecasting as it is the crucial

factor for time series models commonly utilized in future forecasting. Also, mathematical complexity associated with the fitted model is reduced with this assumption. The process of stationarity is defined below.

A process for example $\{x(t), t = 0,1,2,\dots\}$ is categorized as strongly stationary or strictly stationary provided that the joint probability distribution function of $\{x_{t-s}, x_{t-s+1}, \dots, x_t, \dots x_{t+s-1}, x_{t+s}\}$ remians independent of t for all s. Therefore, for a strong stationary process the joint distribution of random variables of any possible set is time independents (Cochrane, 1997; K.W. Hipel, 1994). For some practical applications, the strong stationarity may not be required hence, we can consider weaker forms like stochastic process. A stochastic process is Weakly Stationary having order k provided that the statistical moments up to that order of the process exclusively depend on time differences rather than on the time of occurrences that are used to measure the moments (Cochrane, 1997; K.W. Hipel, 1994; Lee, J.). For instance a stochastic process $\{x(t), t = 0,1,2,\dots\}$ can be defined as a second order stationary (K.W. Hipel, 1994; Lee, J.) if it has time independent mean along with variance and the covariance values $Cov(x_t, x_{t-s})$that depend exclusively on s. It is important to consider that both strong or weak stationarity do not implement on others. Nevertheless, a weak stationary process that is followed by normal distribution can also be considered as strongly stationary (Cochrane, 1997).The general models used for time series data can be represented in many forms and variety of stochastic processes. Literature offers two universally used linear time series models that are, Autoregressive (AR) (K.W. Hipel, 1994; Lee, J.) models. We will emphasize on the AR model which is discussed below.

3.1.2. The Autoregressive (AR) models

The most basic and popular models in time series are autoregressive model or AR models (Zhang, X., Zhang, Young, & Li, 2014). In autoregressive model, the value of dependent variable (Y) at some instant of time is related directly with X or predictor variable (Di Lucca, Guglielmi, Muller,

& Quintana, 2013). The popularity of the AR models is mainly due to its flexibility to represent several varieties of time series with simplicity (Di Lucca et al., 2013; Moineddin, Upshur, Crighton, & Mamdani, 2003).

Schuurman, et al (Schuurman, Ferrer, de Boer-Sonnenschein, & Hamaker, 2016) described the AR process as the example of a stochastic process (Wada, Akaike, & Kato, 1986) that is the time series data appear to vary in a random manner. AR models are also recognized as the conditional models (Lee, D., 2011), transition model (de Vries, Fidler, Kuipers, & Hunink, 1998) or Markov models (Seifert, Abou-El-Ardat, Friedrich, Klink, & Deutsch, 2014). An AR (P) consists of particular lagged values of predictor variable as Y_t where lags obtained from one time period and have an impact on following period. The P value is entitled as order where AR1 is based on first order process of autoregressor and dependent variable for first order in same instant of time T can be directly related with time period that are apart from one period (the value at t-1). In the same manner, the second order can related with two periods apart and so on (Schuurman et al., 2016). The following equation defines the AR(P) model in explicit manner.

$$y_t = \delta + \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \cdots + \ + \varphi_p y_{t-p} + \varepsilon_t$$

where $Y_{t-1}, Y_{t-2}, \ldots, Y_{t-p}$ are lags or values of past series and ε_t is noise or randomness.

3.1.3. Autocorrelation

Defining a proper model for a given data of time series, a crucial step is to perform autocorrelation analysis. The statistical measures derived out of autocorrelation analysis show the relationship of observations in time series analysis. In order to perform modeling and forecasting, the most frequently implied useful plots are the autocorrelation function (ACF). These plots are generated against consecutive time lags. These plots are useful for determination of the order of AR and moving avergaes (MA). The mathematical definitions of AR and MA are described in the following:

For a time series $\{ x(t), t = 0,1,2,\ldots \}$ the Autocovariance (Woo Chang Kim, 2015; Zhang, H. et al., 2014) at lag k can be defined as:

$$\gamma_k = Cov(x_t, x_{t+k}) = E[(x_t - \mu)(x_{t+k} - \mu)]$$

The Autocorrelation Coeffient (Woo Chang Kim, 2015; Zhang, H. et al., 2014) at lag k can be defined as:

$$\rho_k = \frac{\gamma_k}{\gamma_0}$$

In the equation above, μ reflects mean of the time series, i.e. $\mu = [Ex_t]$. The autocovariance at lag zero i.e. γ_0 can be defined as the variance of the time series. The definition reflects that the autocorrelation coefficient ρ_k is dimensionless and therefore is independent of the scale of measurement. Also, clearly $-1 \leq \rho_k \leq 1$.

Principles of time series analysis including autoregressive models and autocorrelations have been used previously for portfolio optimization, however, the previous models do not incorporate news data for multiple objective portfolio optimization (described below). Therefore, we have utilized the principles of time series analysis for news data to perform portfolio analysis.

3.2. Portfolio Optimization

Investing in an asset or business is the process of allocating proportions (weights) of the money to be invested to a variety of different assets available for investment. Portfolio optimization is the process of selecting the best possible investment portfolio in such a way that there is no other better way of selecting weights for different investment which allows an investor to achieve his/her objectives. The objectives vary from investor to investor that is, some of them are interested in maximizing the profits and some of them are interested in slightly less profits but also with the less risk and some of them are interested in optimizing both the objectives at the same time. The optimal achievement of both the objectives simultaneously is practically impossible so there is always a trade-off, that is, one has to increase the desired risk levels in order to expect higher profit and vice versa. If we relate portfolio optimization to the previous section we observe that it is an example of multiobjective optimization where we select a portfolio based on investor's preference from a set of Pareto optimal portfolios. In portfolio optimization expected returns cannot be increased without increasing the related risk for each combination of risk and return.

3.3. Modern Portfolio Theory

The foundation of modern portfolio theory (MPT) was laid by Markowitz (Markowitz, H., 1952) through a mean–variance optimization model which can be formulated as

$$\begin{aligned} &\max\{\mu = E[R] = \textstyle\sum_{i=1}^{n} \mu_i x_i\} \\ &\min\{\sigma^2 = Var[R] = \textstyle\sum_{i=1}^{n}\sum_{j=1}^{n} \sigma_{i,j} x_i x_j\} \\ \text{subject to} \quad &\textstyle\sum_{i=1}^{n} x_i = 1 \\ &x_i \geq 0, \quad i = 1,2,3,\dots,n, \end{aligned} \tag{8}$$

where x_i is the weight of assets and n are the number of investment options. The assumption behind MPT is that investors are rational, that is, an investor will either choose a combination of risk (σ^2) and return (μ) for which he is well-off as compared to other combinations or at least he/she is indifferent. In other words if he/she has to choose between two portfolios with the same expected return, he/she will go for the one with minimum risk (Markowitz, H., 1952). As a multiobjective optimization problem (8) can be written as

$$\begin{aligned} &\boldsymbol{min}\ \{x^T V x,\ -x^T \mu'\} \\ \text{subject to} \quad &0 \leq x_i \leq 1 \\ &\textstyle\sum_{i=1}^{m} x_i = 1,\ i = 1 \dots n. \end{aligned} \tag{8a}$$

The set of Pareto optimal solutions which shows all feasible combinations of risk and return will be the same for all investors but the preferences will vary from investor to investor depending on the returns he/she wants to achieve with the related risk and vice versa. The set of Pareto optimal solutions is also called an **efficient frontier** in finance but mathematically it represents a set of Pareto optimal solutions. It consists of all possible solutions where risk cannot be reduced further without decreasing the expected returns. Figure 3.2 shows an example of the efficient frontier.

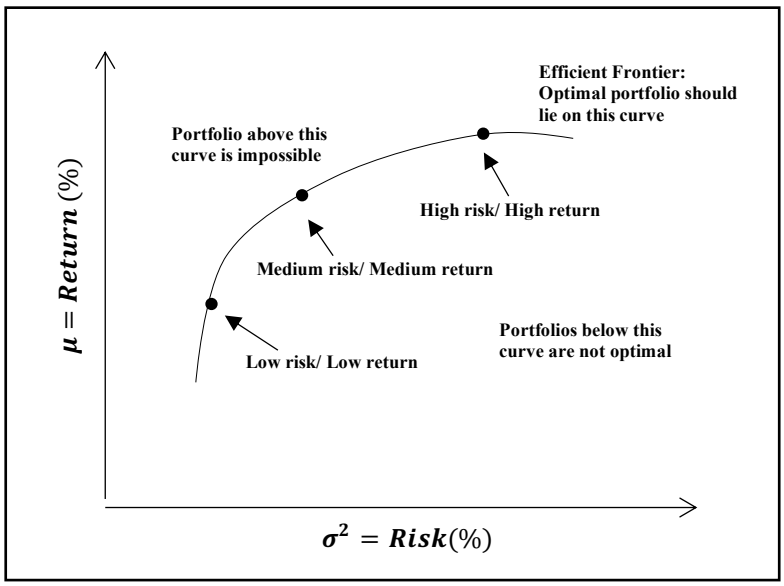

Figure 3.2: The efficient frontier of portfolio optimization

3.4. Multiobjective Portfolio Optimization

Multiobjective portfolio optimization or MPT in finance is the process of solving a multiobjective optimization problem where risk (σ^2) and return (μ) are optimized at the same time on the basis of investor's preferences. The concept of multiobjective portfolio optimization does not only hold for the optimization of σ^2 and μ but researchers have introduced various objectives other than risk and return. Steuer (Ralph E. Steuer, 2005) presented 12 conflicting objectives that can be optimized for a feasible portfolio selection. Additional objectives include liquidity, social responsibility, dividends, amount of short selling, amount invested in R&D, deviations from asset allocation percentages, number of securities in portfolio, turnover, maximum investment proportion weight and number of securities sold short. All of these objectives are not necessarily to be used at the same time but the selection depends on investor's interest and the availability of the data. The formulation of these objectives requires data which is not available all the time or it is hard to collect because companies usually do not share confidential data publicly.

3.5. Literature review

There are various assumptions behind MPT related to perfect market conditions which include no short selling, positive fractions in securities trading, infinite number of assets in the portfolio, no individual bias from the investors and that markets are not affected by rumors (O'Brien & Sculpher, 2000). The key issue in multiobjective financial decision making is to set multiple objectives according to the investor's needs and preferences which allows the investor to earn differently as compared to other investors, using the optimization problem (Winfried Hallerbach, 2002).

The mean-variance theory has been discussed since decades and there exists an enormous amount of research in modifying the basic model. These modifications over the decades have been done in three directions: (i) the model type has been simplified and the amount of data increased over the time therefore, less computationally expensive algorithms have been introduced; (ii) risk has been measured using alternative methods; and (iii) additional objectives other than risk and return or new constraints have been added. Anagnostopoulos and Mamanis (K.P. Anagnostopoulos, 2010) added the number of securities in their optimization model to have some diversification in their model for a particular level of risk and return (K.P.

Anagnostopoulos, 2010).

(Matthias Ehrgott, 2004) also extended Markowitz mean–variance model to five objectives. The authors extended risk further into volatility and his own created variables (S&P Star Ranking) which is defined in (K.P. Anagnostopoulos, 2010) further dividing risk into two objectives. S&P (Standard and Poors) index is a stock market index in the US based on market capitalizations of different types of companies whose common stocks are listed in the New York stock exchange (NYSE) and NADAQ. Similarly, for returns, (Matthias Ehrgott, 2004) split the returns into 12-months, 3-year performance and annual revenue. Hallerbach and Spronk (Winfried Hallerbach, 2002) discussed three areas of finance including corporate finance, financial investment and risk management and argued that many decision problems in these areas involve multiple objectives (Matthias Ehrgott, 2004)

3.6. Summary

There are various ways through which investors can decide to invest in different assets or businesses. Once we determine the feasible options for investments we go for multiobjective portfolio optimization to select our desired portfolio from the set of all feasible portfolio options available on the efficient frontier that is, the most preferred Pareto optimal solution. The choice of an optimal portfolio will vary from investor to investor depending on the level of risk and return he/she is expecting but the efficient frontier will remain the same for a specific data set for all investors.

Chapter 4: Portfolio Selection and Sentiment Analysis

4.1 Basics of Sentiment Analysis

Sentiment analysis (or text analysis) is the process of determining the essence of a text if that is positive, negative or sometimes neutral. The positive sentiment means that the text generally contains positive attributes and vice versa for negative sentiments. It is also possible that some part of the text contains positive sentiments and some part contains negative, but we can calculate total sentiments by adding both the negative and positive sentiments. Generally, sentiment is characterized into binary classifications that is, good/bad, like/dislike etc. We use sentiment analysis to quantify the sentiments (attitudes, emotions and opinions) hidden in a piece of text (usually a specific topic) using machine learning algorithms. It has many applications across industries including opinion mining, product review analysis, social media analysis and news analysis (Steve Y. Yang, 2014). Capturing of market sentiments using news, blogs and articles is a standard operating practice for many investors which keeps them a step ahead as compared to those who do not use sentiment analysis because they are not taking a significant factor into consideration. An overview of the steps using which the sentiments from a dataset are quantified by (Steven Bird, 2009) is described in Figure 4.1.

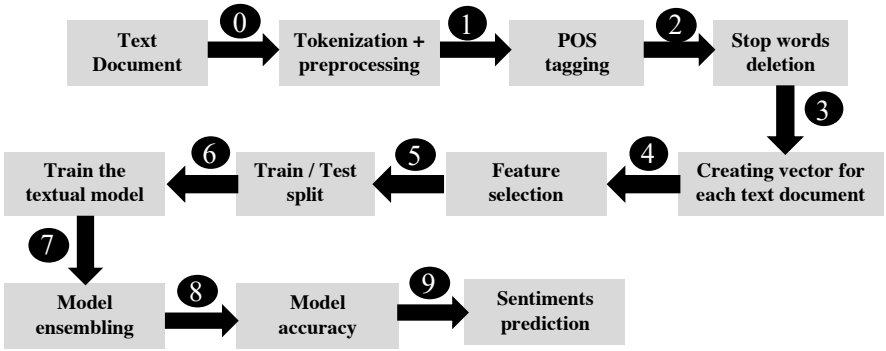

Figure 4.1: Process of calculating sentiments from the text documents

The process in Figure 4.1 shows how we can use machine learning to calculate the sentiments for data set (a collection of many text documents). In our explanation we shall be using a piece of text as an example document to give an intuitive idea about the process. Let our document be ***Document="Apples are very good"***. It is one of the simplest examples of a text document and in real applications the text length in a document can reach up to thousands of words. Using Figure 4.1 step 1 is to do tokenization and preprocessing as text data always contains a lot of mess including punctuations, special characters etc. Then we convert the document into tokens. The basic steps in preprocessing include lowercase all the words to make the document consistent,

removing punctuations etc. Tokenization is done after preprocessing where each word is split into an array or vector such that each word in the document represents *d* dimensional vector where *d* is the number of words in the document.

In step 2 each word is tagged with its relative parts of speech (POS). POS is a category to which a word is assigned in accordance with its syntactic functions. The reason for doing this is to get an idea about the word properties. For example, we know that words like *is, are, am, the, or* does not have specific meaning in a document i.e. they give the meaning when they are used with other words but does not make sense as separate words. They are just to describe the sentence in an understandable way. So in step 2 we apply POS tagging and remove the stop words from the document. In natural language processing, useless words, are referred to as stop words. The document after the second step will look like [('apples', 'NN'), ('are', 'SW'), ('very', 'SW'), ('good', 'ADJ')]. Here NN, SW and ADJ are POS tags which stands for nouns, stop words and adjectives respectively.

In step 3 all the stop words are removed from the document and the document is converted in to a vector. After the third step our *document* will be [apples, good] which has two elements in the vector and thus shows two dimensions for this document. As we apply the similar process to multiple documents, we get a matrix similar to Table 4.1. The **bag of words** is simply a long list of all the words that are available in the data set. In the table we see that Document 1 is the same as our example document and it contains an entry (1) only at two places showing it has only two dimensions and the other dimensions related to this document are 0. Similarly, entries 1 are marked on places where a specific word has occurred in a specific document and if there is no word in the bag of words in a specific document there will be an entry of 0.

Documents	***Bag of words***					
	apples	run	dance	good	….	laugh
Document 1	1	0	0	1	….	0
Document 2	0	1	1	0	….	0
….	….	….	….	….	….	….
Document N	1	1	0	0	….	1

Table 4.1: Bag of words

In the fourth step, feature selection is done which determines the best features that represent the data set. The data set is then split into training and testing. This split is usually based on the nature and volume of

the data. If the data is large enough, then according to (Eck, 2010) we can train our sentiment analysis model even with 60% of the data and test with the remaining 40% . If the data is small then the model training can be done with 90% data as well.

In step 6 and 7, the model is trained using either a single algorithm or multiple algorithms. In machine learning, combining the results of multiple models is called model ensembling. In model ensembling, the performance of one model was compared to ensemble predictions averaged over two, five, and seven different models. The reason model ensembling is to test the results with multiple algorithms and get more confidence from the results. In step 8 we calculate the accuracy by combining the results of ensemble model. Finally, in step 9 our sentiment analysis model is ready for predicting the sentiments when a text document is given as an input.

The process shown in Figure 4.1 is the basic method of calculating sentiments but there are various softwares and web services which use advanced machine learning models to calculate sentiments. It should also be noted that the results from advanced machine learning models will certainly beat the results of this basic calculation process. The sentiments can be calculated not only from the textual documents but also our discussions in social media for example Twitter, can also be used to give sentiments among users regarding a certain discussion. The major contribution in the field of natural language processing has been made by IBM Watson (Venkata Sasank Pagolu, 2016) and it is so far one of the best available cognitive machines which can perform sentiment analysis at scale (with huge amounts of data) and also provide significantly better analysis on the text as compared to other conventional sentiment analysis systems.

4.2. Significance of Sentiment Analysis in Portfolio Selection

The theory of efficient market hypothesis (EMH) (Yim et al., 2016) has been under consideration for decades in the financial markets. Financial economists believe that investor's sentiments play a key role in financial market returns (Matthias Ehrgott, 2004). EMH states that stock prices fully reflect all available information in the market and it is impossible to beat the market since stock prices adjust themselves on any new information available. On the other hand, a lot of research has been done which shows that stock market does act randomly and therefore it is very difficult to predict the market. Also, news may not represent stock market completely but the initial indicators extracted from social media, blogs and articles might provide

reasonable information about the future stock prices (Ranco et al., 2015).

For this reason, investors remain continuously busy in gathering and analyzing information for a variety of different assets. They monitor the global economy and also read and analyze the news well before making a buying decision for an investment. It is believed that the more information an investor has, the more chances there are to turn an investment into profit. Investors are usually good in quantitative models but recently a shift has been seen in analyzing news, blogs and articles at the same time in order to relate the quantitative models with the financial news which surround the market and then come up with a better decision for trading and investing (Wong, 2010).

Financial markets have relied heavily on estimations and quantitative models in order to capture movements in stocks and in other investments. However, investors are interested in accumulating qualitative information as well for example, ongoing news regarding a particular stock or industry and reviews of financial analysts about a particular investment certainly affects the prices of that particular stock. Financial organizations have realized the significance of public sentiments and therefore are encouraging their analysts to integrate massive datasets of news, blogs, social media feeds and online forums with the traditional quantitative models and come up with better investment options (Gajendra Shirsat, 2016).

Sentiment analysis has various applications in financial markets. Table 4.2 from (Gajendra Shirsat, 2016) shows different institutions in finance and how sentiment analysis helps these institutions in consulting their clients better.

Financial Exchanges	It helps in efficient market surveillance and in depth investigation of rumors and trending news. It allows indices creation which helps in adding or eliminating securities from the portfolio based on market sentiments.
Brokers and Traders	Market outlook can be tracked before and after policy statements from central banks or other regulatory authorities. It helps in analyzing the impact of investment recommendations on market positions.
Issuers	Public sentiments can be analyzed prior to initial offers which help in setting up better initial offers. Market behavior can be tracked for prior and post corporate actions.
Investors	Sentiment based portfolio selection is done in order to wipe out stocks with bad sentiments.

	Portfolio performance can be better analyzed.
Asset and Wealth Managers	It allows managers to identify new product and service ideas from investors as they have been sharing their views for better product design all the time. Acceptability of new products along with satisfaction levels can be measured through customer sentiments.

Table 4.2: Use of sentiment analysis in various financial institutions (Venkata Sasank Pagolu, 2016)

4.3. Literature review

Twitter is one of the largest social media platforms and its impact on stock market has been studied over the years successfully and according to Yang et al. adding the public sentiment into the analysis has helped in enhancing the stock market prediction (Anshul Mittal 2013). According to (Gajendra Shirsat, 2016), majority of the financial market professionals use social media for professional purposes. Monica and Viorel (Negru, 2015) also showed that integration of public sentiments in market analysis reduces the risk associated with the portfolio and if an investor knows beforehand that some positive action is going to occur then the investor will probably make an investment.

Anshul and Arpit (Anshul Mittal 2013) showed that emotions of individuals affect their decision making. This effect can be seen as a direct correlation between public sentiments (emotion and attitude of a rational man) and market sentiments (attitude of stock market based on investor's attitude). Anshul and Arpit also used twitter data to extract public sentiments and used them along with lagged day Dow Jones Industrial Average (DJIA) market standings to predict the future movements and then used those movements in selecting their portfolio. It was also claimed in the study that public sentiments can successfully capture socio-cultural events and those events are correlated with DJIA movements (Anshul Mittal 2013).

Hochreiter used sentiment analysis for a group of individuals and not a single individual because crowds combined together represent better sentiments (Hochreiter, 2015). The movements in financial market are a representation of the behavior of investors towards earning higher profits in the future which in general are the public sentiments. Twitter has been used in this regard and financial community was extracted which has its interest in financial markets and the community was also active in financial investments (Steve Y. Yang, 2014). On the basis of the sentiments extracted from that financial community, a significant correlation of public sentiments has been found with Dow Jones Industrial Index price and volatility (Steve Y. Yang, 2014).

4.4. Summary

Sentiment analysis is used in financial markets on a regular basis to gauge public sentiments by finance professionals. Many studies have shown that it has successfully helped in determining the market movements if used correctly. The correct use of sentiments means that the data used to analyze sentiments must contain a targeted audience. The targeted audience means that if we are aiming to research on financial markets, Twitter accounts which do discussion on such topics should be part of the sample rather than considering random accounts. It is because all Twitter accounts and their sentiments cannot determine market movements but if some particular groups are targeted then studies have shown positive results (Anshul Mittal 2013). Furthermore, capturing market sentiments is not an activity which has been done for decades. It is a recent advancement but its addition to the finance industry has certainly helped the researchers in gaining more stock market knowledge. On the other hand it is challenging as well because it requires significant amount of data to be able to get good results.

Chapter 5: Research Design and Process

5.1. Data collection and sources

The thesis is intended to focus on the US financial markets, so all the price data collected for various investment options (variables) is from the US. The various investment options include stock market or particularly NYSE, bond market, FOREX, gold, real estate market and bitcoin which is a very recent entry in terms of investment across the globe. The data has been gathered using Quandl API (See Table 5.1). Quandl provides access to open source economic and finance data using its API and it can be accessed directly or using any programming language. News data has been gathered from New York Times API (see Table 5.1) which has a huge collection of historical data along with the abstract of each news categorically. The category we selected for gathering our data is business and finance because we are assuming that other news will not have any significant impact on the financial sector.

The data has been collected from January 2013 till October 2017 which makes it more than 4.5 years of daily data. There are a couple of reasons for choosing the mentioned time period for data collection. The first reason is that bitcoin is a recent investment option for investors. It is not even a decade old so it was quite unstable and non-famous in the beginning but its prices started to increase and vary from 2013 onwards. Secondly, although the news data is available from the last 2 decades, there was some missing news for few days in the data. This was another reason to consider data from 2013 onwards because the missing news ratio was not high. The data for different investments has been gathered from different sources as shown in Table 5.1.

Investment Options	No. of years Data	Source
Real Estate, Bonds, FOREX, Gold	4	https://www.quandl.com/
NYSE	4	http://finance.yahoo.com/
Bitcoin	4	https:// www.coindesk.com/price
News	4	https://developer.nytimes.com/

Table 5.1: Data sources

5.2. Data preprocessing

The data has been gathered using different APIs so the structure of the data was not consistent and it was preprocessed to make it consistent. We collected the data by considering the time period in such a way that we have maximum availability of the data and also we tried to minimize the possibility of missingvalues but

still we found some and fixed them. There are many ways to handle the missing values but we used the 5-day rolling averages for each investment option since rolling average is a widely used method in handling the missing values in Finance (Yim et al., 2016).

All the data we are working with in this thesis is of time series type and therefore every API has a different format representing dates which was corrected and converted into a consistent date format. We are working with the returns on each investment but we collected the data for original prices. So for a particular day *t*, we calculated the returns using the percentage change formula $\frac{price_t}{price_{t-1}} - 1$. In other words, we normalized all the data between -1 and 1. The data for real estate was available on the monthly basis. We converted the data into daily data by keeping the prices constant for the whole month and change it in the following month. There are several ways to manage this kind of data but we considered the easiest way.

The processing of news data to calculate the sentiment score of daily news was performed using IBM Watson (AlchemyLanguage API). It provides various types of results including sentiment class, sentiment score etc but we are only interested in classification of the news such that if particular news is positive we assign a value 1 and if the news is negative we assign the sentiment score to be -1 and for neutral news we assign the value to be 0. The news data was initially gathered in the format described in Fig 5.1

A	B
Date	**News**
1/1/2013	Frequent Flier column features M Sanjayan, lead scientist for Nature Conservancy
1/1/2013	Some companies are adopting policies aimed at weaning employees from their sm
1/1/2013	Joe Sharkey On the Road column observes American Airlines is experimenting witl
1/2/2013	Vivian Marino 30-Minute Interview with Li Chung Pei, founding partner of Pei Part
1/2/2013	Marketing teams behind New York office buildings are starting to employ the type
1/2/2013	Russia's malls are luring shoppers and investors, even as such shopping centers ap
1/2/2013	Energy drinks, fastest-growing part of beverage industry, claim to offer mental anc

Figure 5.1: Raw news data

From Figure 5.1, it can be seen that there are multiple news for each day. So for each day we calculate the sentiment score for each news using IBM Watson API and sum the score in a way that if there are 3 news for a particular day, out of which two news are positive and one is negative then the total sentiment score for that day will be 1 + 1 + (– 1) = 1. Then "Total Score" values are normalized using the following formula so that the news data and investment option data are on the same scale.

$$Normalized = \frac{Total\ Score_i - \min(Total\ Score)}{\max(Total\ Score) - \min(Total\ Score)}.$$

The final news table with the sentiment score looks like the table in Figure 5.2. So the column "Normalized" is our news variable which we use in the autoregressive regression described in (12)

A	B	C	D	E	F	G	H
Date	News 1	News 2	News 3	...	News N	Total Score	Normalized
1/1/2013	0	1	1		0	2	0.46154
1/2/2013	-1	1	0		1	1	0.38462
1/3/2013	0	0	-1		-1	-2	0.15385
1/4/2013	1	-1	-1		0	-1	0.23077
1/5/2013	-1	1	-1		1	0	0.30769

Figure 5.2: Final results of sentiment analysis

Finally after cleaning all the data we merge the data together into a table so it is easily understandable and makes more sense for processing. The data we gathered from different sources was combined on the basis of dates as we are working with time series data. The news data we gathered was generic business and finance news data but not related to any specific investment option since this kind of data was not available for free and it will be considered as one of the limitations of the thesis.

5.3. Mathematical Concepts

In this thesis we shall be creating a model to optimize the risk and return on investments of the DM and these two factors will act as multiple objectives in our problem. Section 5.3.1 and 5.3.2 shows the mathematical formula and explanation of the two objectives.

5.3.1. Expected Returns

In finance returns are simply calculated as $\frac{price_t}{price_{t-1}} - 1$ where $price_t$ are the prices of a particular investment on a specific day and $price_{t-1}$ are the prices of the same investment on the previous day. Concisely, expected return is the weighted sum of the average returns of possible investment options. For example, if we have four different investment options including bitcoin, gold, stock and foreign exchange, the returns will be calculated by multiplying the weight vector (proportions applied to each investment that sums up to 1) and the average return vector (contains average returns for different investment options). The formula for calculating expected return is

$$E(R) = w^T \mu, \quad (9)$$

where μ is the average return vector such that μ = [average(daily returns on bitcoin), average(daily returns on gold), average(daily returns on stock), average(daily returns on foreign exchange)]. The average return vector μ in this example contains four elements where each element represents average return for each of the four investments over a given period of time. Similarly, w^T is the transposed weight vector indicating the proportion of money invested in each investment option that must sum up to 1 and each $w_i \geq 0$.

In this thesis we are calculating the vector μ differently than by just calculating simple averages. The reason for calculating μ differently is to analyze any significant differences in the returns calculated from simple averages and with the proposed technique. The vector μ in (9) can be represented as μ' = $[\alpha_1, \alpha_2 \ldots, \alpha_n]$ such that $i = 1,2, \ldots, n$ where each α_i is the average return on an investment option for example bonds, bitcoins etc and each α_i is calculated by using the following linear autoregressive regression model

$$\alpha_i = x_1 \alpha_{i(t-1)} + x_2 \alpha_{i(t-2)} + \epsilon \quad (10)$$

where x_1 and x_2 are the coefficients of autoregressive terms $\alpha_{i(t-1)}$ and $\alpha_{i(t-2)}$ respectively and ϵ is the residual. Using (10) we calculate expected returns for all elements in μ' and then use (9) to calculate overall expected returns by replacing μ with μ'. The final model for expected return without news data will be

$$E(R) = w^T \mu'. \quad (11)$$

The model (10) describes the returns that are calculated only on the basis of historical data. So in order to incorporate the news data in (10) we use the following linear autoregressive regression model. The vector μ in (9) can also be represented as μ'' = $[\beta_1, \beta_2 \ldots, \beta_n]$

$$\beta_i = x_1 \beta_{i(t-1)} + x_2 \beta_{i(t-2)} + x_3 N_{i(t-1)} + x_4 N_{i(t-2)} + \epsilon \quad (12)$$

where x_1 and x_2 are the coefficients of $\beta_{i(t-1)}$ and $\beta_{i(t-2)}$ and x_3 and x_4 are coefficients of the news variable described in Figure 5.2. Using (12) we calculate expected returns for all elements in μ'' and then use (9) to calculate overall expected returns by replacing μ with μ''. The final model for expected return with the news data will be

$$E(R) = w^T \mu''. \quad (13)$$

The reason for choosing models (10) and (12) which are AR(2) model is that we analyzed models including $t-1,\ t-2,\ t-3,\ t-4$ and $t-5$ using the similar autoregressive regression models described in Section 3.1.1 and $t-2$ shows the best results in terms of model fitting.

5.3.2 Risk

In finance we often refer to standard deviation as volatility or risk (Dong et al., 2012). If we have four different investment options like in Section 5.4.1, then we can calculate the risk by first calculating variance-covariance matrix based on the returns of different investment options. The variance-covariance matrix is a squared matrix that contains covariances and variances of all four investment options. The matrix will then be multiplied with the weight vector in the following way

$$E(\sigma^2) = w^T V w \quad (14)$$

where V is the variance-covariance matrix and w is weight vector.

5.4. Mathematical Model

At this point we combine the two objectives describes in Section 5.3 to formulate our multiobjective optimization problem. So the problem formulation we are focusing in this thesis for the model without news data using (11) is

$$\begin{aligned} & \boldsymbol{min}\ \{w^T V x\ ,\ -w^T \mu'\} \\ \text{subject to} \quad & 0\ \leq w_i \leq 1 \\ & \textstyle\sum_{i=1}^{m} w_i = 1,\ i = 1 \ldots n\ decision\ variables. \end{aligned} \quad (15)$$

Similarly, the problem formulation for news data model using (13) is

$$\begin{aligned} & \boldsymbol{min}\ \{w^T V w\ ,\ -w^T \mu''\} \\ \text{subject to} \quad & 0\ \leq w_i \leq 1 \\ & \textstyle\sum_{i=1}^{m} w_i = 1,\ i = 1 \ldots n\ decision\ variables. \end{aligned} \quad (16)$$

Here *n* is the number of decision variables which in our case is *n*=6. These decision variables are New York stock exchange (NYSE), bond market, foreign exchange market (FOREX), gold, real estate market and bitcoin. The models (15) and (16) from here onwards will be named as **model 1** and **model 2** respectively.

5.5. Research Problem and Design

Volatility in financial markets and specifically in the US is a problem that has been studied for decades in order to allow investors to make investments with confidence. There is a study in (Patzelt & Pawelzik, 2013) where Clifford showed that the volatility in financial markets is not random over time but it depends on various factors. These factors do not only include quantitative factors such as asset's own prices over the time but also qualitative factors as well such as market news and surrounding events (Dong et al., 2012). There are many researches where risk and return are optimized by only using asset's own prices for example in (Chang, 2015) but in this research we combine the two factors together i.e. asset's prices and market news and aim at building a new mathematical model for estimating the return on different investments. We then compare the results of the two multiobjective optimization models described in Section 5.4.

The problem is designed such that the DM is expected to optimize risk and return on the basis of feasible set of solutions called the Pareto front. An investor in our research will act as a DM where he/she is supposed to have prior knowledge of his/her investment domain. Therefore, the DM can choose the optimal portfolio from the Pareto front in an iterative way on the basis of experience and also according to his/her preferences.

Concisely, the objective of the research is to apply multiobjective optimization technique on our models in Section 5.4. First, a conventional model used to optimize portfolio risk described in (14) in Section 5.3.2 and autoregressive linear model for calculating expected returns using assets own prices described in (10) in Section 5.3.1. Secondly, our proposed autoregressive model where we calculate the risk in the same way but for expected returns we include asset's prices and also market news described in (12). The reason for only calculating the return in our proposed way but not the risk was due to the complexity of the risk formulation and it can be considered as one of the limitations of thesis. Finally we aim at analyzing differences between both the models to see if adding news data in our model makes difference or not.

5.6. Research Objectives

DM's are always in search of a tool that can help them in analyzing the risk and return on their investments. With the evolution of state-of-the-art technologies, these tools are becoming more and more complex but on the other hand, provide better insights as well. Also, the amount of data we have today is not only increasing at a fast pace but also becoming more complex and can be used to model the problems in a better way. For example, few years back financial research was only based on quantitative data but due to the availability of latest tools we can also process high volumes of qualitative data such as news and blogs etc. This processing of quantitative and qualitative data together provide better insights (Liao et al., 2018). Therefore, our objective in this thesis is to create a mathematical model not only based on the historical prices but also on the news data that can help DMs in optimizing risk and returns on their investments and compare it with a similar multiobjective portfolio optimization model that does not incorporate news data.

5.7. Process overview diagram

Figure 5.3 shows the flow chart in which the process of collecting and analyzing data is described. In the first stage we collected the data using four different sources mentioned in Table 5.1. After the preprocessing and cleaning explained in Section 5.2, we built a model for expected returns and risk described in Section 5.3.1 and 5.3.2. Then finally the two multiobjective optimization models described in Section 5.5 were used to generate sets of Pareto optimal solutions using ϵ-constraint method in Section 2.4.2 and compare the results of the two models. The reason for using ϵ-constraint method is to make sure that we generate the complete set of Pareto optimal solutions also for non-convex problems.

Furthermore, reference point method described in Section 2.4.5 was used to analyze any significant differences in the decision variables of the two models at different reference points. The reason for using reference point method is that we want analyze different combinations of risk and return of both the models 1 and 2. The different combinations include nadir returns vs ideal risk, ideal returns vs nadir risk, ideal returns vs ideal risk, midrange returns and risk (50%/50%) and high

returns and low risk (75%/25%).

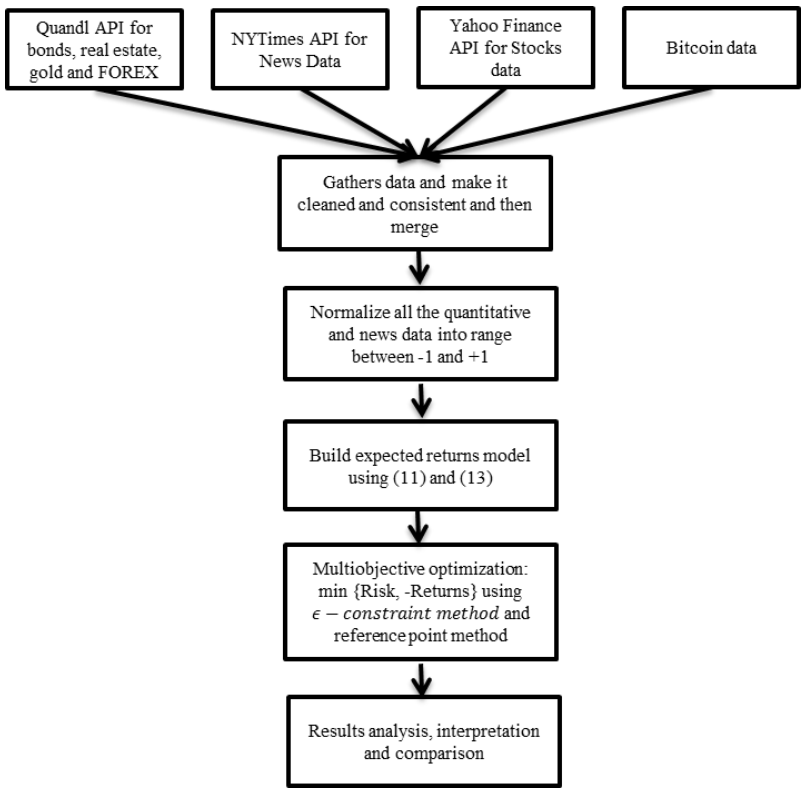

Figure 5.3: Process overview using a flowchart

Chapter 6: Analysis of Results

6.1. Analysis of descriptive statistics

The data we collected for more than 4.5 years consist of 1203 observations (i.e. approximately 261 working days/year) for all six variables including bonds, FOREX, bitcoin, gold, NYSE and real estate. All the calculations in the below tables are calculated using Python. Table 6.1 shows that the average returns for bonds, bitcoin and real estate over the time are positive and NYSE and gold shows negative returns over the time. Although FOREX has also showed negative returns its magnitude is significantly small as compared to other investment means and thus show approximately 0% returns. Bonds and bitcoin shows higher variation as compared to other investments but they also show higher returns as compared to other investments.

Bitcoin and bonds had the biggest losses (returns) over the period of time as compared to other investments which are -29.58% and -23.07% respectively. Similarly, biggest gains are also observed for bitcoin and bonds which are 64.81% and 30.0% respectively. On the other hand, real estate had the least losses of -1.13% and also least gains of 0.91% making real estate comparatively safe investment because the range of its historical returns is very small. Similarly biggest gains for gold and NYSE were 4.71% and 4.49% respectively and that for FOREX were 3.07%. Therefore bonds and FOREX showed the least gains as compared to other investments

	bonds	forex	bitcoin	gold	nyse	restate
count	1203	1203	1203	1203	1203	1203
mean	0.003474	-7.8E-05	0.006719	-0.00018	-0.00029	0.000279
std	0.057008	0.005493	0.058038	0.010268	0.007751	0.000873
min	-0.230769	-0.02388	-0.29581	-0.09354	-0.02879	-0.01135
max	0.3	0.030711	0.648166	0.047102	0.044941	0.009173

Table 6.1: Descriptive Statistics

Table 6.2 shows the cross correlation matrix between investments. Bonds and FOREX have a correlation of -0.167 which shows negative correlation between these two investments. Similarly, negative correlation has been observed between bonds and gold which is -0.172. FOREX and gold shows positive correlation of 0.267 and it is the highest positive correlation between other

investments. Positive correlation is also observed between bitcoin and gold which is 0.243. The correlation coefficients between other investments are significantly low and lie between -0.05 to 0.05. These correlation coefficients are considered insignificant since the correlation values are very close to zero and therefore can be ignored to make statistical conclusions.

	bonds	forex	bitcoin	gold	nyse	restate
bonds	1					
forex	-0.167731	1				
bitcoin	0.005608	-0.01513	1			
gold	-0.172253	0.267393	0.024392	1		
nyse	0.014255	0.020348	-0.04145	0.012318	1	
restate	0.027578	-0.02166	-0.0123	0.040425	0.015669	1

Table 6.2: Correlation Matrix

Table 6.3 shows the autocorrelation between the investments and their lags. As discussed earlier in Section 5.4.1 we tested different models and (10) and (12) performed better as compared to others. Therefore, autocorrelation for each investment with lags $t-1$ and $t-2$ has been calculated and the results show that there is significant autocorrelation between the lags. Investments including bonds, FOREX, bitcoin, NYSE and real estate show more than 99.5% correlation with $t-1$ and $t-2$ both. Gold returns shows slightly lower correlation with $t-1$ and $t-2$ which is 99.3% and 98.7% respectively. The overall results from Table 6.3 shows that significant autocorrelation exist in all investments with $t-1$ and $t-2$ which shows the evidence of stationarity and therefore time series models can be fit on all of the investments to predict the expected returns at time t.

	Corr (t-1)	Corr (t-2)
bonds	0.99899	0.998006
forex	0.998418	0.996984
bitcoin	0.997544	0.995191
gold	0.99356	0.987598
nyse	0.995181	0.99047
restate	0.999956	0.999912

Table 6.3: Autocorrelation

After analyzing the cross correlation and autocorrelation among the investments, the next

step is to analyze the variations between investments as compared to one another. Figure 6.1 shows the daily returns plot for all our investments. Bonds and bitcoin shows high variations over the time while other investments showed significantly low variations. From the plot it can be observed that the variations in bitcoin are significantly higher as compared to bonds but the results from Table 6.1 show the variations in bonds and bitcoin have no significant difference. Bitcoin shows higher variations on few days of the year which can be seen in Figure 6.1 as green spikes whereas variation in bonds remains consistent over the time. This is the reason for similar variations of these two investments although they look slightly less similar in the plot.

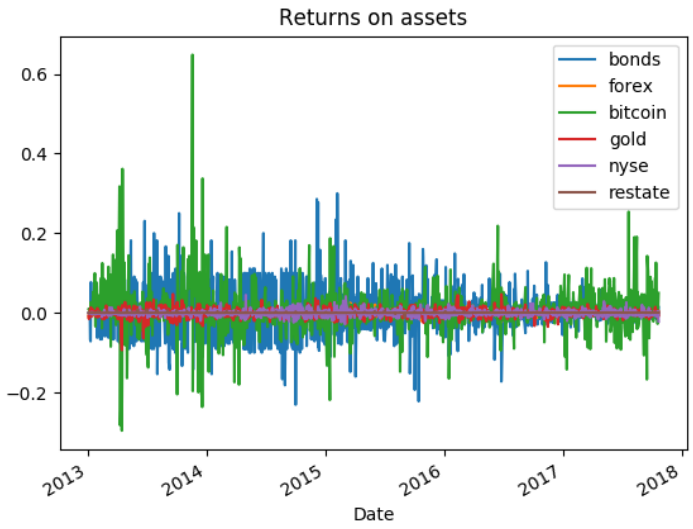

Fig 6.1: Normalized returns of all the assets

The analysis of descriptive statistics shows that bitcoin and bonds are the two investments which show higher risk and higher returns over the period of time. All the investments are independent of each other and have significant autocorrelation with $t-1$ and $t-2$ lags which allow us to create autoregressive time series models for each investment to calculate their expected returns.

6.2. Analysis of ideal and nadir vectors

In this section we shall analyze the results of ideal and nadir vectors obtained from model 1 and model 2. Table 6.4 shows the resulting ideal and nadir vectors in the form of a table. Ideal and nadir vectors were calculated for both the models by using concepts in Section 2.1.1. We used SLSQP optimizer (Scipy, 2019) to calculate components of the payoff table.

The return component of the ideal vector in model 1 shows approximately 0.75% returns and the value is in the Table 6.4 is -0.0075781. The negative values in return component in Table 6.4 implies from the fact that returns are maximized in our optimization model. Similarly, the return component of ideal vector in model 2 is approximately 1.2% or -0.0119053 is the exact value in the Table 6.4. The return components are approximately similar for both the models. The ideal risk component in both models 1 and 2 is close to 0% or 3.65×10^{-7} are the exacts value in the Table 6.4. The reason for similar risk components in both the models is that they were calculated with the same formula.

On the other hand, the return component of the nadir vector both in models 1 and 2 is approximately 0% or -5.92×10^{-5} and -5.81×10^{-5} are the exact values respectively in Table 6.4. Similarly, the risk component of the nadir vector both in model 1 and 2 is approximately 5.6% or 5.58×10^{-2} and 9.61×10^{-2} are the exact values in Table 6.4. As discussed in Chapter 2 ideal and nadir vectors are bounds of the objective function values in the Pareto optimal set and because we only have two objectives, payoff table works well for calculating the nadir objective vector. The values of all Pareto optimal solutions will lie in the range of ideal and nadir for each model respectively.

		Return	**Risk**
Model 1	**Ideal Vector**	-0. 0075781	6.65E-07
	Nadir Vector	-5.92E-05	5.58E-02
Model 2	**Ideal Vector**	-0. 0119053	6.65E-07
	Nadir Vector	-5.81E-05	5.61E-02

Table 6.4: Ideal and Nadir Vectors

6.3 Analysis of Pareto Front

After analyzing ideal and nadir vectors, in this section we shall analyze the Pareto optimal sets for models 1 and 2 using Figures 6.2 and 6.3 respectively. In model 1 we have solved the multiobjective optimization problem (15) using (10), (11) and (14). The expected return on our set of investments in model 1 is calculated using an autoregressive model without news data. The shape of the Pareto optimal set in Figure 6.2 is approximately linear which shows that both objectives have approximately a linear tradeoff.

The range of returns in the Pareto optimal set in model 1 lies between 0% and 0.75%. The range of returns is very small but we can validate it with Table 6.1 where average returns are also quite small. Similarly, the range of risk in the Pareto optimal set in model 1 lies approximately between 0% and 5.5%. The small range of the risk can also be validated by looking at the standard deviation values of all investments in Table 6.1

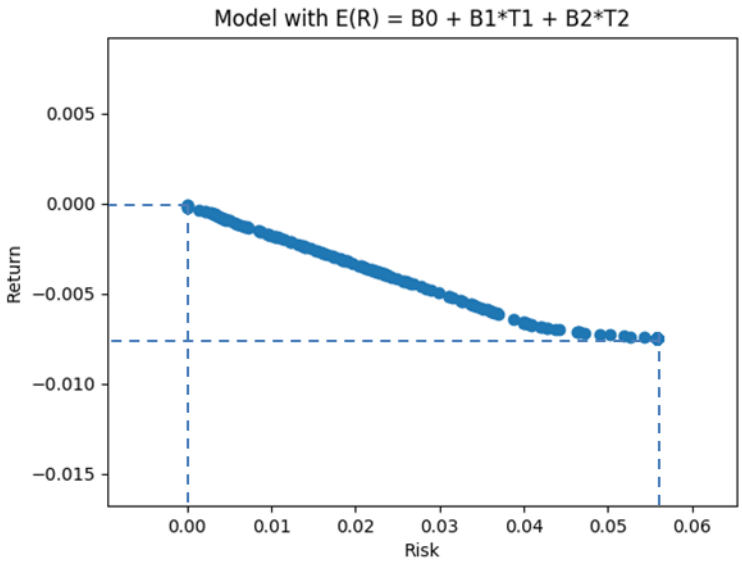

Figure 6.2: Pareto optimal solutions of model 1

In model 2 we solved the multiobjective optimization problem (16) using (12), (13) and (14). The expected return on our set of investments in model 2 is calculated using an autoregressive

model with news data. The shape of the Pareto optimal set in Figure 6.3 is also approximately linear.

The range of returns in the Pareto optimal set in model 2 lies between 0% and 1.2%. Similarly, the range of risk in the Pareto optimal set in model 1 lies approximately between 0% and 5.6%.

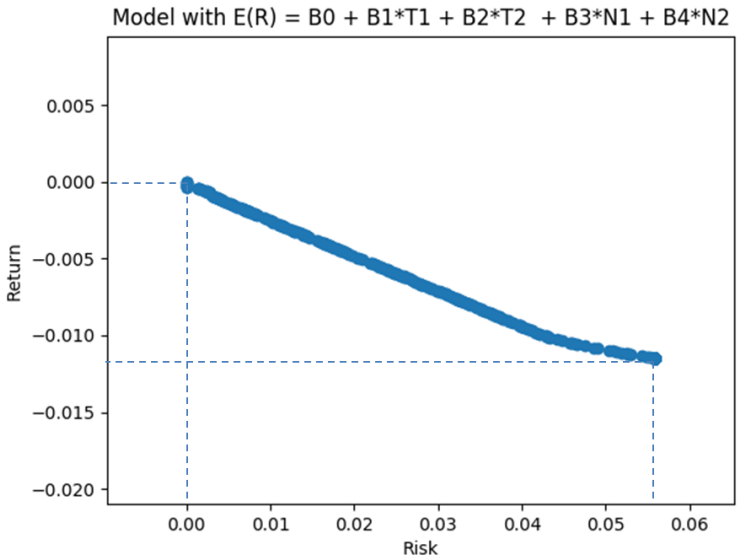

Figure 6.3: Pareto optimal solutions for model 2

The analysis of Pareto optimal sets obtained from models 1 and 2 shows the similar behavior which is approximately linear. There is slight difference in expected returns ranges which is approximately 0.75% in model 1 and 1.2 % in model 2 but the risk remains similar in both the models. The analysis of Section 6.2 and 6.3 shows that the results are approximately similar in both the models.

6.4. Analysis of reference points solutions

In Section 6.3 we analyse the results of models 1 and 2 by graphically analysing the set of Pareto optimal solutions and see that the results of both the models are approximately similar. Therefore, in this section we dive deep further into the analysis and use reference point method to

determine how decision variables behave with respect to different reference points. It is a possibility that the overall objectives are approximately similar but the distribution of investments in both the models are different. That is why this analysis will help in understanding the behavior of the variables in both the models.

The reference points are usually given from the DM and it is an iterative process but in order to analyse the behavior of decision variables we used reference points in a slightly different way. Our reference points are described in Table 6.5. The logic behind considering these reference points is to analyze the whole Pareto optimal set for each model 1 and 2 including extreme and middle values and determine how our decision variables behave at different reference points.

	Choices	**Return**	**Risk**
Model 1	Nadir returns vs ideal risk	-0.000059	0.000000665
	Ideal returns vs Nadir risk	-0.0075	0.056
	Ideal returns vs ideal risk	-0.0075	0.0000006.65
	Midrange returns and risk (50%/50%)	-0.00415	0.0246
	High returns and low risk (75%/25%)	-0.00631	0.00392
Model 2	Nadir returns vs ideal risk	-0.000058	0.0000006.65
	Ideal returns vs Nadir risk	-0.0119	0.055
	Ideal returns vs ideal risk	-0.0119	0.0000006.65
	Midrange returns and risk (50%/50%)	-0.00685	0.0251
	High returns and low risk (75%/25%)	-0.00902	0.0395

Table 6.5: Reference points and corresponding solution of returns and risk

The first three choices of reference point selection are simply the extreme values that a DM would want to analyze because these values cover the extreme point of the Pareto optimal set. The last two reference points have been used to analyze the behavior of decision variables in the middle range of Pareto optimal set. The choice "Midrange returns and risk (50%/50%)" in Table 6.5 shows the middle values of between ideal and nadir. So, the reference point is calculated using the formula $[\frac{(ideal-nadir)}{2}, \frac{(nadir-ideal)}{2}]$. Similarly, the reference point choice "High returns and low risk (75%/25%)" was calculated using $[(ideal - nadir) * 0.75, (nadir - ideal) * 0.25]$.

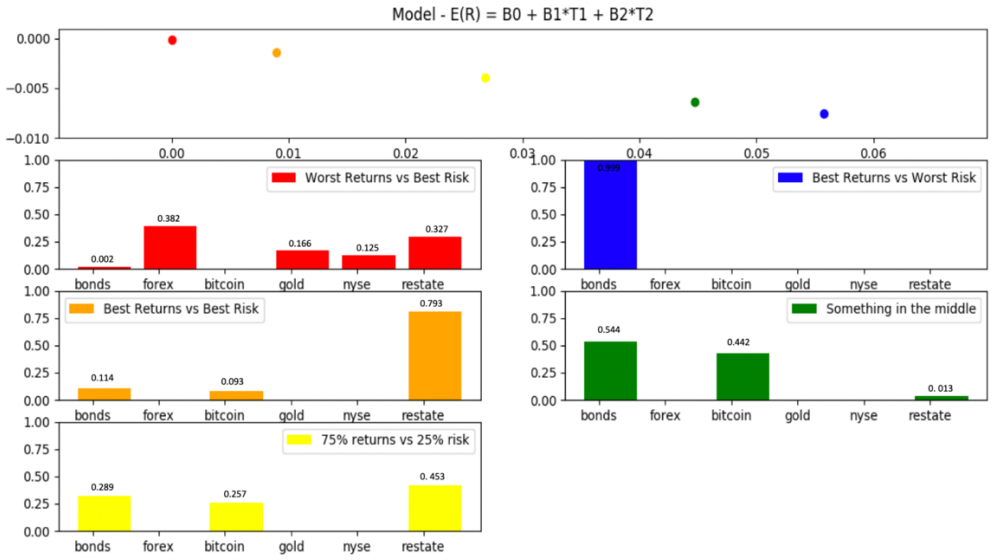

Figure 6.4: Scatter plot of function values based on reference points in Table 6.5 for model 1. Bar charts beneath the scatterplot shows the optimal weights of investments for each reference point

In model 1, the optimal investment corresponding to the first reference point uses bonds, FOREX, gold, NYSE and real estate. Bitcoin has not come up as an optimal investment option because it has the highest risk as compared to other investments but the given reference point has the risk value is close to zero. In the second reference point, bonds come up as the ultimate investment option because we have already seen that bonds are one of the investments with high returns. The third reference point returns real estate as a major investment options with almost 85% weight which is intuitive because we saw real estate as the safest investment in Section 6.1. The optimal investment from the fourth reference point majorly includes bonds and bitcoin and a very small proportion of real estate. This is also intuitive as we have seen bonds and bitcoins having higher returns than others. The optimal investment from the fifth reference point includes bonds, bitcoins and real estate. In this case real estate proportion is slightly higher as compared to bonds and bitcoins because the desired level of risk is significantly lower in the reference point. The reason for bond and bitcoin to be optimal investment options is because the expected return level was significantly higher in the reference point.

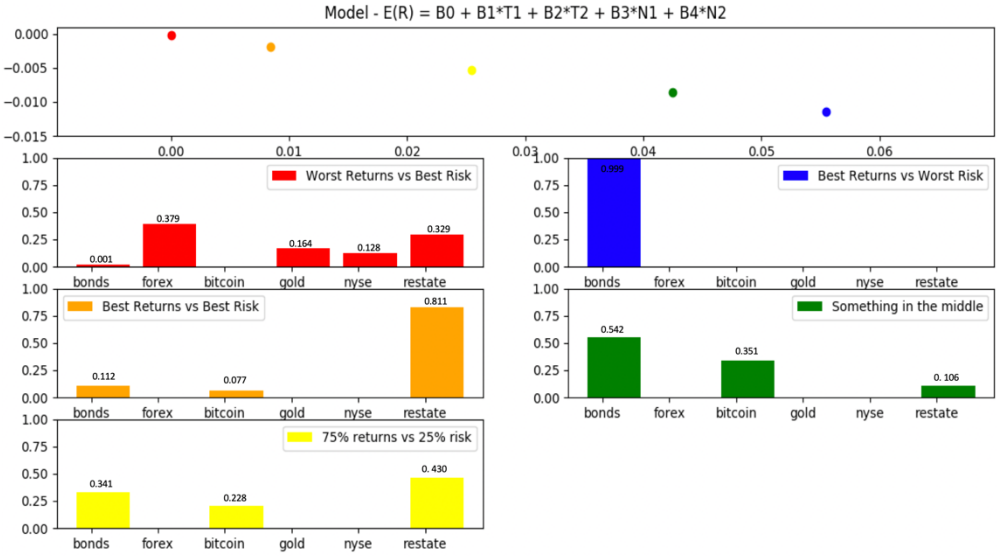

Figure 6.5: Scatter plot of function values based on reference points in Table 6.5 for model 2. Bar charts beneath the scatterplot shows the optimal weights of investments for each reference point

In model 2, the optimal weights for investment show similar behavior to that in model 1 but the return levels are increased from 0.75% to 1.12%. There were slight differences in the weights of the decision variables, but the overall behavior look similar in both the models. This shows that the news data didn't have an impact on the decision variables. This is because of the way the news data was incorporated. If the new data would be in such a way that we could have news segmented by investment options, then the results might be different.

6.5. Limitations and Future Research

Predicting the money market is a huge industry around the globe with trillions of dollars invested in it. There are extremely complex models on which organizations work when they predict the money market. This thesis has been done by assuming a lot of factors constant and therefore it contains several limitations and opens several doors for future. Some of the limitations and future research opening include:

1. The expected returns calculated in this thesis are based on very simple linear autoregressive model. The model formulation can be more complex including nonlinear models such that it represents the data in a better way.
2. The news data was available in a format where we could not separate news related to a particular investment option. So if we had the data structured in a way that for each investment option we have separate news sentiments, the added value of the sentiment analysis could be properly studied.
3. The third limitation is that there might be factors apart from historical data and news data which are impacting the returns, but we have not considered those factors in our model.

Chapter 7: Conclusions

Stock market in the US has trillions of dollar value and has millions of investors and naturally there are investors also in other countries. These investors are always concerned about the volatility of the market and want to choose best investment options with a low risk of losses. Therefore, research in this area is very old and a lot of powerful researches have been already conducted to address this problem. According to (Seifert et al., 2014) Markowitz addressed this problem for the first time and since then thousands of researchers have tried to analyse the behavior of the stock market. Quantitative data is not the only source for analyzing the trends but more sophisticated tools take other important factors into account for example, news data. The complexity of the problem can be understood in this way that after decades of research the problem is not yet completely solved but it is improving with the passage of time. With the invent of big data analysis a lot of new doors have been opened in this research. Financial researchers and consultants spend a lot of time in collecting big data i.e. huge amounts of historical data, news data etc. to make complex models. Therefore they can get better insights from the data and therefore can predict the market better to earn more money.

In this research we used the multiobjective optimization methods to solve the portfolio optimization problem by using time series data and news data. Our objective in the thesis was to minimize the risks and maximize the returns. We created two models: 1) model without news data and 2) model with news data or model with sentiment analysis. The difference in the two models was that the computation of expected return was done differently. In model 1 expected return was calculated using an autoregressive linear regression model based only on the historical data of different investment options. On the other hand, in model 2 we calculated the expected return with an autoregressive linear regression not only based on historical data of the stock but also included news data. The risk was calculated using the same formula in both the models.

The results obtained from both the multiobjective optimization models shows that both the

models gave similar results for risk and returns. The optimal weights of the decision variables including bonds, FOREX, bitcoin, gold, NYSE and real estate showed similar behavior. Although there was a slight difference in the expected return in both the models but the difference was marginal.

There are two major limitations of this study. First, is the unavailability of the news data in such a way that we can segment the news on the basis of each investment option. The news data was available in a simple raw text format without tagging of investment options (i.e. which news text relates to which investment option). If the news data we gathered was tagged with the related investment options, it might have made our modelling more reasonable and our analysis more conclusive. Second, is the dependency diversification of the risk and returns for an investment such that there could be many factors apart from the time series data and news data itself. The use of those hidden factors might provide better answers to our research question.

References

Anshul Mittal , A. G. (2013). Stock Prediction Using Twitter Sentiment Analysis.

Arthur, J. N., Williams, R. J., & Delfabbro, P. H. (2016). The conceptual and empirical relationship between gambling, investing, and speculation. *J Behav Addict, 5*(4), 580-591. doi:10.1556/2006.5.2016.084

Atta Mills, E. F., Yan, D., Yu, B., & Wei, X. (2016). Research on regularized mean-variance portfolio selection strategy with modified Roy safety-first principle. *Springerplus, 5*(1), 919. doi:10.1186/s40064-016-2621-7

Bata, S. A., & Richardson, T. (2018). Value of Investment as a Key Driver for Prioritization and Implementation of Healthcare Software. *Perspect Health Inf Manag, 15*(Winter), 1g.

Bertella, M. A., Pires, F. R., Feng, L., & Stanley, H. E. (2014). Confidence and the stock market: an agent-based approach. *PLoS One, 9*(1), e83488. doi:10.1371/journal.pone.0083488

Bertella, M. A., Pires, F. R., Rego, H. H., Silva, J. N., Vodenska, I., & Stanley, H. E. (2017). Confidence and self-attribution bias in an artificial stock market. *PLoS One, 12*(2), e0172258. doi:10.1371/journal.pone.0172258

Biondo, A. E., Pluchino, A., Rapisarda, A., & Helbing, D. (2013). Are random trading strategies more successful than technical ones? *PLoS One, 8*(7), e68344. doi:10.1371/journal.pone.0068344

Boada, Y., Reynoso-Meza, G., Pico, J., & Vignoni, A. (2016). Multi-objective optimization framework to obtain model-based guidelines for tuning biological synthetic devices: an adaptive network case. *BMC Syst Biol, 10*, 27. doi:10.1186/s12918-016-0269-0

Brown, S. B., & Ridderinkhof, K. R. (2009). Aging and the neuroeconomics of decision making: A review. *Cogn Affect Behav Neurosci, 9*(4), 365-379. doi:10.3758/CABN.9.4.365

Carrillo-de-Albornoz, J., Rodriguez Vidal, J., & Plaza, L. (2018). Feature engineering for sentiment analysis in e-health forums. *PLoS One, 13*(11), e0207996. doi:10.1371/journal.pone.0207996

Chang, K.-H. (2015). Design Optimization. *Design Theory and Methods Using CAD/CAE*.

Chu, J., Nadarajah, S., & Chan, S. (2015). Statistical Analysis of the Exchange Rate of Bitcoin. *PLoS One, 10*(7), e0133678. doi:10.1371/journal.pone.0133678

Cochrane, J. H. (1997). Time Series for Macroeconomics and Finance. *Graduate School of Business, University of Chicago*.

Cohen-Charash, Y., Scherbaum, C. A., Kammeyer-Mueller, J. D., & Staw, B. M. (2013). Mood and the market: can press reports of investors' mood predict stock prices? *PLoS One, 8*(8), e72031. doi:10.1371/journal.pone.0072031

Copeland, R., & Jacobs, P. (1981). Cost of capital, target rate of return, and investment decision making. *Health Serv Res, 16*(3), 335-341.

de Vries, S. O., Fidler, V., Kuipers, W. D., & Hunink, M. G. (1998). Fitting multistate transition models with autoregressive logistic regression: supervised exercise in intermittent claudication. *Med Decis Making, 18*(1), 52-60. doi:10.1177/0272989X9801800112

Deluccia, D. J. (1989). How to maximize return on bond proceeds. *Healthc Financ Manage, 43*(9), 87.

Dhar, V. (2014). Can Big Data Machines Analyze Stock Market Sentiment? *Big Data, 2*(4), 177-181. doi:10.1089/big.2014.1528

Di Lucca, M. A., Guglielmi, A., Muller, P., & Quintana, F. A. (2013). A Simple Class of Bayesian Nonparametric Autoregression Models. *Bayesian Anal, 8*(1), 63-88. doi:10.1214/13-BA803

Dong, X., Zeng, S., & Chen, J. (2012). A spatial multi-objective optimization model for sustainable urban wastewater system layout planning. *Water Sci Technol, 66*(2), 267-274. doi:10.2166/wst.2012.113

Draviam, T. C. a. T. (2008). Markowitz principles for multi-period portfolio selection problems

with moments of any order. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*.
Dujardin, Y., & Chades, I. (2018). Solving multi-objective optimization problems in conservation with the reference point method. *PLoS One, 13*(1), e0190748. doi:10.1371/journal.pone.0190748
Eck, P. H. a. D. (2010). *LEARNING FEATURES FROM MUSIC AUDIO WITH DEEP BELIEF NETWORKS*. Paper presented at the 11th International Society for Music Information Retrieval Conference (ISMIR 2010), Montreal.
Fan, J., Zhang, J., & Yu, K. (2012). Vast Portfolio Selection with Gross-exposure Constraints(). *J Am Stat Assoc, 107*(498), 592-606. doi:10.1080/01621459.2012.682825
Farmer, J. D., & Lo, A. W. (1999). Frontiers of finance: evolution and efficient markets. *Proc Natl Acad Sci U S A, 96*(18), 9991-9992.
Feng, L., Li, B., Podobnik, B., Preis, T., & Stanley, H. E. (2012). Linking agent-based models and stochastic models of financial markets. *Proc Natl Acad Sci U S A, 109*(22), 8388-8393. doi:10.1073/pnas.1205013109
Freedman, R. A., Viswanath, K., Vaz-Luis, I., & Keating, N. L. (2016). Learning from social media: utilizing advanced data extraction techniques to understand barriers to breast cancer treatment. *Breast Cancer Res Treat, 158*(2), 395-405. doi:10.1007/s10549-016-3872-2
Frydman, C., Barberis, N., Camerer, C., Bossaerts, P., & Rangel, A. (2014). Using Neural Data to Test A Theory of Investor Behavior: An Application to Realization Utility. *J Finance, 69*(2), 907-946. doi:10.1111/jofi.12126
Gabarron, E., Dorronzoro, E., Rivera-Romero, O., & Wynn, R. (2019). Diabetes on Twitter: A Sentiment Analysis. *J Diabetes Sci Technol, 13*(3), 439-444. doi:10.1177/1932296818811679
Gajendra Shirsat, I. C. (2016). Tuning in to the Emotions of the Capital Markets with Sentiment Analysis.
Ganesan, T., Elamvazuthi, I., Shaari, K. Z., & Vasant, P. (2013). An algorithmic framework for multiobjective optimization. *ScientificWorldJournal, 2013*, 859701. doi:10.1155/2013/859701
Gao, X., An, H., Fang, W., Huang, X., Li, H., & Zhong, W. (2014). Characteristics of the transmission of autoregressive sub-patterns in financial time series. *Sci Rep, 4*, 6290. doi:10.1038/srep06290
Geoffrion, A. M. (1968). Proper efficiency and the theory of vector maximization. *Journal of Mathematical Analysis and Applications, 22*(3), 618-630.
Gonzalez Mde, L., Jareno, F., & Skinner, F. S. (2016). Interest and Inflation Risk: Investor Behavior. *Front Psychol, 7*, 390. doi:10.3389/fpsyg.2016.00390
Hamzacebi, C. (2008). Improving artificial neural networks' performance in seasonal time series forecasting. *Information Sciences, 178*, 4550-4559.
Harlow, L. L., & Oswald, F. L. (2016). Big data in psychology: Introduction to the special issue. *Psychol Methods, 21*(4), 447-457. doi:10.1037/met0000120
Heiberger, R. H. (2015). Collective attention and stock prices: evidence from Google Trends data on Standard and Poor's 100. *PLoS One, 10*(8), e0135311. doi:10.1371/journal.pone.0135311
Hochreiter, R. (2015). Computing trading strategies based on financial sentiment data using evolutionary optimization. *Neural and Evolutionary Computing*. doi:10.1007/978-3-319-19824-8_15
Hopper, A. M., & Uriyo, M. (2015). Using sentiment analysis to review patient satisfaction data located on the internet. *J Health Organ Manag, 29*(2), 221-233. doi:10.1108/JHOM-12-2011-0129
I. Das, J. E. D. (1997). A closer look at drawbacks of minimizing weighted sums of objectives for Pareto set generation in multicriteria optimization problems. *Structural optimization, 14*(1),

63–69.
Juang, W. C., Huang, S. J., Huang, F. D., Cheng, P. W., & Wann, S. R. (2017). Application of time series analysis in modelling and forecasting emergency department visits in a medical centre in Southern Taiwan. *BMJ Open, 7*(11), e018628. doi:10.1136/bmjopen-2017-018628
K.P. Anagnostopoulos, G. M. (2010). A portfolio optimization model with three objectives and discrete variables. *Computers & Operations Research, 37*(7), 1285-1297.
K.W. Hipel, A. I. M. (1994). Time Series Modelling of Water Resources and Environmental Systems. *Elsevier Amsterdam*.
Kaisa Miettinen, F. R., Andrzej P. . (2008). *Introduction to Multiobjective Optimization: Interactive ApproachesIntroduction to Multiobjective Optimization: Interactive Approaches*. Paper presented at the Multiobjective Optimization, Interactive and Evolutionary Approaches.
Khin Lwin, R., Qu, Graham Kendall. (2014). A learning-guided multi-objective evolutionary algorithm for constrained portfolio optimization. *Applied Soft Computing*
24, 757-772.
Kim, M., & Sayama, H. (2017). Predicting stock market movements using network science: an information theoretic approach. *Appl Netw Sci, 2*(1), 35. doi:10.1007/s41109-017-0055-y
Kralewski, J., Gifford, G., & Porter, J. (1988). Profit vs. public welfare goals in investor-owned and not-for-profit hospitals. *Hosp Health Serv Adm, 33*(3), 311-329.
Lan, Q., Xiong, Q., He, L., & Ma, C. (2018). Individual investment decision behaviors based on demographic characteristics: Case from China. *PLoS One, 13*(8), e0201916. doi:10.1371/journal.pone.0201916
Lee, D. (2011). A comparison of conditional autoregressive models used in Bayesian disease mapping. *Spat Spatiotemporal Epidemiol, 2*(2), 79-89. doi:10.1016/j.sste.2011.03.001
Lee, J.). [Univariate time series modeling and forecasting (Box-Jenkins Method)].
Liao, Q., Sheng, Z., Shi, H., Zhang, L., Zhou, L., Ge, W., & Long, Z. (2018). A Comparative Study on Evolutionary Multi-objective Optimization Algorithms Estimating Surface Duct. *Sensors (Basel), 18*(12). doi:10.3390/s18124428
Luc Eyraud, A. F., Sophie RIVAUD. (2007). The impact of the housing slowdown on US consumption. *TRÉSOR-ECONOMICS*.
Lv, D., Huang, Z., Li, M., & Xiang, Y. (2019). Selection of the optimal trading model for stock investment in different industries. *PLoS One, 14*(2), e0212137. doi:10.1371/journal.pone.0212137
MARAKBI, Z. (2016). Mean-Variance Portfolio Optimization: Challenging the role of traditional covariance estimation.
Markowitz, H. (1952). PORTFOLIO SELECTION. *The Journal of Finance, 7*(1).
Markowitz, H. M. (2010). Portfolio Theory: As I Still See It. *Annual Review of Financial Economics, 2*, 1-23.
Matthias Ehrgott, K. K., Christian Schwehm. (2004). An MCDM approach to portfolio optimization. *European Journal of Operational Research, 155*(3), 752-770.
McKay, D. R., & Peters, D. A. (2017). The Midas Touch: Gold and Its Role in the Global Economy. *Plast Surg (Oakv), 25*(1), 61-63. doi:10.1177/2292550317696049
Miettinen, K. M. (1999). *Nonlinear Multiobjective Optimization*
Moineddin, R., Upshur, R. E., Crighton, E., & Mamdani, M. (2003). Autoregression as a means of assessing the strength of seasonality in a time series. *Popul Health Metr, 1*(1), 10. doi:10.1186/1478-7954-1-10
Negru, M. T. V. (2015). *Stock Market Trading Strategies Applying Risk and Decision Analysis Models for Detecting Financial Turbulence*. Paper presented at the 2015 17th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC).
Novak, P. K., Amicis, L., & Mozetic, I. (2018). Impact investing market on Twitter: influential users and communities. *Appl Netw Sci, 3*(1), 40. doi:10.1007/s41109-018-0097-9

O'Brien, B. J., & Sculpher, M. J. (2000). Building uncertainty into cost-effectiveness rankings: portfolio risk-return tradeoffs and implications for decision rules. *Med Care, 38*(5), 460-468.
Pai, P. F., Hong, L. C., & Lin, K. P. (2018). Using Internet Search Trends and Historical Trading Data for Predicting Stock Markets by the Least Squares Support Vector Regression Model. *Comput Intell Neurosci, 2018*, 6305246. doi:10.1155/2018/6305246
Patzelt, F., & Pawelzik, K. (2013). An inherent instability of efficient markets. *Sci Rep, 3*, 2784. doi:10.1038/srep02784
Peters, D. A., & McKay, D. A. (2014). Life insurance: Ownership and investment considerations. *Plast Surg (Oakv), 22*(1), 54-55.
Piskorec, M., Antulov-Fantulin, N., Novak, P. K., Mozetic, I., Grcar, M., Vodenska, I., & Smuc, T. (2014). Cohesiveness in financial news and its relation to market volatility. *Sci Rep, 4*, 5038. doi:10.1038/srep05038
Popa, C., Holvoet, K., Van Montfort, T., Groeneveld, F., & Simoens, S. (2018). Risk-Return Analysis of the Biopharmaceutical Industry as Compared to Other Industries. *Front Pharmacol, 9*, 1108. doi:10.3389/fphar.2018.01108
Preis, T., Moat, H. S., & Stanley, H. E. (2013). Quantifying trading behavior in financial markets using Google Trends. *Sci Rep, 3*, 1684. doi:10.1038/srep01684
Ralph E. Steuer, Y. Q., Markus Hirschberger. (2005). Multiple Objectives in Portfolio Selection *JOURNAL OF FINANCIAL DECISION MAKING, 1*(1).
Ranco, G., Aleksovski, D., Caldarelli, G., Grcar, M., & Mozetic, I. (2015). The Effects of Twitter Sentiment on Stock Price Returns. *PLoS One, 10*(9), e0138441. doi:10.1371/journal.pone.0138441
Ranco, G., Bordino, I., Bormetti, G., Caldarelli, G., Lillo, F., & Treccani, M. (2016). Coupling News Sentiment with Web Browsing Data Improves Prediction of Intra-Day Price Dynamics. *PLoS One, 11*(1), e0146576. doi:10.1371/journal.pone.0146576
Reyes-Menendez, A., Saura, J. R., & Alvarez-Alonso, C. (2018). Understanding #WorldEnvironmentDay User Opinions in Twitter: A Topic-Based Sentiment Analysis Approach. *Int J Environ Res Public Health, 15*(11). doi:10.3390/ijerph15112537
Roache, S. K., & Rossi, M. (2009). The Effects of Economic News on Commodity Prices: Is Gold Just Another Commodity? *IMF Working Paper No. 09/140*.
Sanchez-Granero, M. A., Trinidad-Segovia, J. E., Clara-Rahola, J., Puertas, A. M., & De Las Nieves, F. J. (2017). A model for foreign exchange markets based on glassy Brownian systems. *PLoS One, 12*(12), e0188814. doi:10.1371/journal.pone.0188814
Sawaragi, Y., H. Nakayama and T. Tanino. (1985). Theory of multiobjective optimization. *Academic Press, New York*.
Schuurman, N. K., Ferrer, E., de Boer-Sonnenschein, M., & Hamaker, E. L. (2016). How to compare cross-lagged associations in a multilevel autoregressive model. *Psychol Methods, 21*(2), 206-221. doi:10.1037/met0000062
Scipy. (Ed.) (2019) Scipy.
Seifert, M., Abou-El-Ardat, K., Friedrich, B., Klink, B., & Deutsch, A. (2014). Autoregressive higher-order hidden Markov models: exploiting local chromosomal dependencies in the analysis of tumor expression profiles. *PLoS One, 9*(6), e100295. doi:10.1371/journal.pone.0100295
Song, Z., Wang, M., Dai, G., & Vasile, M. (2015). A novel multiobjective evolutionary algorithm based on regression analysis. *ScientificWorldJournal, 2015*, 439307. doi:10.1155/2015/439307
Spronk, W. H. J. (2003). A multidimensional framework for financial–economic decisions. *Journal of Multi-Criteria Decision Analysis, 11*(3).
Steinert, L., & Herff, C. (2018). Predicting altcoin returns using social media. *PLoS One, 13*(12), e0208119. doi:10.1371/journal.pone.0208119

Steve Y. Yang, S. Y. K. M., Xiaodi Zhu. (2014). *An Empirical Study of the Financial Community Network on Twitter*. Paper presented at the 2014 IEEE Conference on Computational Intelligence for Financial Engineering & Economics (CIFEr).
Steven Bird, E. K., Edward Loper (2009). *Extracting Information from Text*. 1005 Gravenstein Highway North, Sebastopol, CA 95472: O'Reilly Media.
Sun, C. S. M. (2005). New Multiobjective Metaheuristic Solution Procedures for Capital Investment Planning. *Journal of Heuristics, 11*(3), 183–199.
Tan, J., & Cheong, S. A. (2016). The Regime Shift Associated with the 2004-2008 US Housing Market Bubble. *PLoS One, 11*(9), e0162140. doi:10.1371/journal.pone.0162140
Tan, L., Chen, J. J., Zheng, B., & Ouyang, F. Y. (2016). Exploring Market State and Stock Interactions on the Minute Timescale. *PLoS One, 11*(2), e0149648. doi:10.1371/journal.pone.0149648
Tsai, M. C., Cheng, C. H., Tsai, M. I., & Shiu, H. Y. (2018). Forecasting leading industry stock prices based on a hybrid time-series forecast model. *PLoS One, 13*(12), e0209922. doi:10.1371/journal.pone.0209922
Venkata Sasank Pagolu, K. N. R., Ganapati Panda, Babita Majhi. (2016). *Sentiment analysis of Twitter data for predicting stock market movements.* Paper presented at the 2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES).
Vickey, T., & Breslin, J. G. (2017). Online Influence and Sentiment of Fitness Tweets: Analysis of Two Million Fitness Tweets. *JMIR Public Health Surveill, 3*(4), e82. doi:10.2196/publichealth.8507
Wada, T., Akaike, H., & Kato, E. (1986). Autoregressive models provide stochastic descriptions of homeostatic processes in the body. *Nihon Jinzo Gakkai Shi, 28*(3), 263-268.
Wang, C., Jin, J., Vieito, J. P., & Ma, Q. (2017). Antiherding in Financial Decision Increases Valuation of Return on Investment: An Event-Related Potential Study. *Comput Intell Neurosci, 2017*, 4760930. doi:10.1155/2017/4760930
Wang, C., Vieito, J. P., & Ma, Q. (2015). A Neuroeconomics Analysis of Investment Process with Money Flow Information: The Error-Related Negativity. *Comput Intell Neurosci, 2015*, 701237. doi:10.1155/2015/701237
West, R., Coyle, K., Owen, L., Coyle, D., Pokhrel, S., & Group, E. S. (2018). Estimates of effectiveness and reach for 'return on investment' modelling of smoking cessation interventions using data from England. *Addiction, 113 Suppl 1*, 19-31. doi:10.1111/add.14006
Wierzbicki, A. (1982). A mathematical basis for satisficing decision making. *Mathematical Modelling, 3*, 391-405.
Wierzbicki, A. P. (1980). The Use of Reference Objectives in Multiobjective Optimization. *Multiple Criteria Decision Making Theory and Application*, 468-486.
Williams, I., Brown, H., & Healy, P. (2018). Contextual Factors Influencing Cost and Quality Decisions in Health and Care: A Structured Evidence Review and Narrative Synthesis. *Int J Health Policy Manag, 7*(8), 683-695. doi:10.15171/ijhpm.2018.09
Winfried Hallerbach, J. S. (2002). The relevance of MCDM for financial decisions. *Journal of Multi-Criteria Decision Analysis*, 23.
Wong, J. (2010). Market Predictions using sentiment analysis and state-space models. Retrieved from http://cs229.stanford.edu/proj2010/Wong-SentimentAnalysisForMarketMovements.pdf
Woo Chang Kim, J. H. K., Frank J. Fabozzi. (2015). *Shortcomings of Mean-Variance Analysis*.
Xu, Y., Liu, Z., Zhao, J., & Su, C. (2017). Weibo sentiments and stock return: A time-frequency view. *PLoS One, 12*(7), e0180723. doi:10.1371/journal.pone.0180723
Yang, W., Ma, J., Chen, H., Maglione, A. G., Modica, E., Rossi, D., . . . Babiloni, F. (2018). Good

News or Bad News, Which Do You Want First? The Importance of the Sequence and Organization of Information for Financial Decision-Making: A Neuro-Electrical Imaging Study. *Front Hum Neurosci, 12*, 294. doi:10.3389/fnhum.2018.00294

Yang, Z. H., Liu, J. G., Yu, C. R., & Han, J. T. (2017). Quantifying the effect of investors' attention on stock market. *PLoS One, 12*(5), e0176836. doi:10.1371/journal.pone.0176836

Yim, K., Oh, G., & Kim, S. (2016). Understanding Financial Market States Using an Artificial Double Auction Market. *PLoS One, 11*(3), e0152608. doi:10.1371/journal.pone.0152608

Zhang, C. (2014). Mean-variance portfolio selection for defined-contribution pension funds with stochastic salary. *ScientificWorldJournal, 2014*, 826125. doi:10.1155/2014/826125

Zhang, G. P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing, 50*, 159–175.

Zhang, H., Wei, J., & Huang, J. (2014). Scaling and predictability in stock markets: a comparative study. *PLoS One, 9*(3), e91707. doi:10.1371/journal.pone.0091707

Zhang, X., Zhang, T., Young, A. A., & Li, X. (2014). Applications and comparisons of four time series models in epidemiological surveillance data. *PLoS One, 9*(2), e88075. doi:10.1371/journal.pone.0088075

Zheludev, I., Smith, R., & Aste, T. (2014). When can social media lead financial markets? *Sci Rep, 4*, 4213. doi:10.1038/srep04213