Optimal portfolio allocation using Hierarchical Risk Parity and PCA

Tanaka Makuvaza (tanaka@aims.ac.za)
African Institute for Mathematical Sciences (AIMS)

Supervised by: Prof. Phillip Mashele
North West University, South Africa

14 May 2020

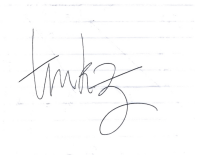
Submitted in partial fulfillment of a structured masters degree at AIMS South Africa

Abstract

Although Markowitz model is appealing from a mathematical and computational point of view, it has three major problems which are instability, concentration and under-performance. We are going to use Hierarchical Risk Parity (HRP) and Principal Component Analysis (PCA) for portfolio allocation and compare the results with the traditional Markowitz model. One major advantage of HRP is that it does not require the invertibility of the covariance matrix, thus can compute allocations on ill-generate or a singular covariance matrix which is an impossible task for quadratic optimizers. PCA makes algorithms faster with its dimensionality reduction.

Declaration

I, the undersigned, hereby declare that the work contained in this research project is my original work, and that any work done by others or by myself previously has been acknowledged and referenced accordingly.



Tanaka Makuvaza, 14 May 2020

Contents

1. Introduction

1.1 Background

Portfolio Optimisation is arguably the most recurrent problem in modern financial mathematics. It dates back to the 19th century when Harry Markowitz published a paper on how risk-averse investors can construct a portfolio or a set of portfolios that optimize return given a certain level of risk (Markowitz, 1952). Risk is the possibility that the actual return differs from the expected returns. It is the possibility of losing some or all of an original investment. Markowitz emphasised that risk is an essential part of a higher return. Portfolio optimisation is a giant problem in modern financial mathematics because it is impossible to make a higher return without exposure to more risk provided that the assets or portfolios are efficient. According to Markowitz (1952), "a portfolio is efficient if the investor cannot find a better one in the sense that it has either a higher expected return and the same (lower) variance, or a lower variance and the same (or higher) expected return". This rules out the idea of allocating all the capital to the assets with the highest returns because it will carry the highest risk. Instead, a diversified portfolio should be built taking into account the correlations of the assets in the portfolio. With his mean-variance optimisation (MVO), Markowitz (1952) showed that there is a set of optimal portfolios called the efficient frontier.

However the Markowitz (1952) model has faced a lot of criticism from investment practitioners and academics because of its underlying assumptions and some computational difficulties that might arise. Quadratic optimizers requires the inversion of the covariance matrix so cannot compute allocations for singular matrices (Meucci, 2007). It has been shown empirically that tiny changes in the forecasted returns causes MVO to produce very different portfolios thus making it unstable.

Despite its many pitfalls, MVO has remained the holy grail of investment management because there are a few portfolio selection methods that are not too complex and can be computed without great difficulty (Meucci, 2007). The parameters used in some of the alternative models are difficult to interpret. Some models use multiple objective functions which are more difficult to calculate than the MVO which uses the quadratic utility function.

The purpose of this essay is to use Hierarchical Risk Parity (HRP) and Principal Component Analysis (PCA) to solve the portfolio allocation problem. HRP uses modern mathematics (graph theory and machine learning techniques) to build a diversified portfolio using the covariance matrix (Prado, 2018). It uses is hierarchical clustering which groups similar features into clusters. In this essay, the features will be stocks of different companies. Risk parity portfolios have gained popularity amongst investment practitioners because they tend to perform better than the classical Markowitz model (Roncalli, 2014).

PCA is a commonly used dimension reduction method used to minimize the complexity in the data and minimizing information loss (Vidal and Ma, 2010). It has proven to be a powerful tool in finance and other industries because reducing the dimensions of data without too much information loss is a huge feat in data analysis. From a computational perspective, this makes algorithms faster (Kassambara, 2017). PCA has been widely used to gain an insight into financial markets. Some applications are analysing the shape of the yield curve, implementation of interest rate models such as the famous Heath Jarrow Morton (HJM) model and in forecasting to mention a few.

1.2 Objectives

In this essay we are going to use Hierarchical Risk Parity to compute an optimal portfolio. After finding the optimal portfolio, we then use Principal Component Analysis (PCA) and the classical Markowitz model to select optimal portfolios. We then compare results of the three approaches. We want a portfolio that is stable thus less risky. It is desirable to hold a portfolio with many assets so that we reduce idiosyncratic risk as much as possible. This is called diversification. We then choose the best approach among these three taking into consideration factors like stability, concentration and whether or not the portfolio is well diversified.

1.3 Outline

The rest of the essay is organised as follows: Chapter 2 introduces the Markowitz model, risk parity portfolios and literature review on Principal Component Analysis (PCA). Chapter 3 discusses Hierarchical Risk Parity (HRP), Principal Component Analysis (PCA) and experiments using these two models and the Markowitz model. Chapter 4 concludes the essay and gives an insight into possible future extensions to the work.

2. Literature Review

2.1 Markowitz Model

The Markowitz model also known as the mean-variance portfolio theory specifies a way to construct a portfolio that gives the maximum return for a given risk, or the minimum risk for a specified return (Markowitz, 1952).

2.1.1 Assumptions of the Markowitz Model.

- All expected returns, variances and covariances of pair of assets are known.
- Investment decisions are made purely on the basis of return and risk.
- Investors are non-satiated. This means that they prefer more to less.
- Investors are risk-averse. This means that investors dislike risk.
- There is a fixed single step time period.
- There are no taxes and transaction costs.
- Assets may be held in any amounts, i.e short selling is possible, we can have infinite divisible holdings, and there are no maximum investment limits.

2.1.2 The Model.

According to Roncalli (2014), we start by formulating the investor's financial problem :

1. maximising the portfolio expected return given a constraint in risk ($\sigma-$problem):

   $max\ \mu(x) \quad u.c \quad \sigma(x) \leq \sigma^*$ where $\mu(x)$ and $\sigma(x)$ are the portfolio mean and variance respectively (2.1.1)

2. minimising the portfolio risk under the return constraint (μ-problem):

$$min\ \sigma(x) \quad u.c \quad \mu(x) \geq \mu^* \tag{2.1.2}$$

The portfolio expected return is given by:

$$\mu_p = E(R_p) = \sum_{i=1}^{N} x_i \mu_i \tag{2.1.3}$$

Where μ_i is the expected return on security S_i.

The portfolio risk is given by:

$$\sigma_p^2 = Var(R_p) = \sum_{i=1}^{N} \sum_{j=1}^{N} x_i x_j C_{ij} \tag{2.1.4}$$

where C_{ij} is the covariance of the returns on security i and j, x_i and x_j are the weights of security i and j, and $C_{ii} = V_i$ where V_i is the variance of security i.

Alternatively, it can be written in matrix form as:

$$\sigma_p^2 = w^T \sum w \tag{2.1.5}$$

where w is a vector of allocations, $\sum$ is the covariance matrix and w^T is the transpose of w.

The risk minimisation problem thus becomes:

$$w_p = \underset{w \in \mathbb{R}^p}{\text{argmin}}\ w^T \sum w \tag{2.1.6}$$

$$s.t \quad w^T 1_p = 1 \tag{2.1.7}$$

$$(w^T \mu = E) \tag{2.1.8}$$

The return maximisation problem:

$$w_p = \underset{w \in \mathbb{R}^p}{\text{argmax}}\ w^T \mu \tag{2.1.9}$$

$$s.t \quad w^T 1_p = 1 \tag{2.1.10}$$

$$w^T \sum w = \sigma \tag{2.1.11}$$

The solution to this optimisation problem can then be found using numerical methods or analytical techniques like Langrange multipliers.

2.1.3 Problems of the Markowitz Model.

- Investment decisions are not made purely on the basis of risk and return. Some investors care about the maximum loss that they will incur on an investment, liquidity of the assets and skewness of the returns
- Not all investors are risk averse. The risk appetite depends on other factors like the level of wealth.
- In reality, there are taxes and transaction costs which have a large impact on investment decisions.
- The model has a lot of parameters that need to be computed and if the estimation of parameters is inaccurate, this may lead to wrong allocations.

2.1.4 Why Diversification is Important.

According to Markowitz (1952), diversification is important in a portfolio because it eliminates idiosyncratic or specific risk. Investors are not compensated for specific risk but only for systematic risk.

Using the Markowitz model, the portfolio risk is given as (Elton et al., 2014):

$$\sigma_p^2 = \sum_{i=1}^{N} x_i^2 \sigma_i^2 + \sum_{\substack{i=1 \\ i \neq j}}^{N} \sum_{j=1}^{N} x_i x_j C_{ij} \tag{2.1.12}$$

Where σ_i is the variance of security i, x_i and x_j are the weights of security i and j, C_{ij} is the covariance between securities i and j and the other terms are defined as before.

Let the allocations be equal for all the assets,

$$x_i = \frac{1}{N} \quad \forall i \tag{2.1.13}$$

$$\sigma_p^2 = \sum_{i=1}^{N}(\frac{1}{N})^2\sigma_i^2 + \sum_{\substack{i=1 \\ i\neq j}}^{N}\sum_{j=1}^{N}(\frac{1}{N})(\frac{1}{N})C_{ij} \tag{2.1.14}$$

$$\sigma_p^2 = \frac{1}{N}\sum_{i=1}^{N}\frac{\sigma_i^2}{N} + \frac{N-1}{N}\sum_{\substack{i=1 \\ i\neq j}}^{N}\sum_{j=1}^{N}\frac{C_{ij}}{N(N-1)} \tag{2.1.15}$$

$$\sigma_p^2 = \frac{\bar{\sigma}^2}{N} + \frac{N-1}{N}\bar{C} \tag{2.1.16}$$

Where $\bar{\sigma}^2$ is the average variance and $\bar{C}$ is the average covariance.

Letting the number of assets in our portfolio go to infinity gives:

$$\lim_{N\to\infty}\sigma_p^2 = \lim_{N\to\infty}\frac{\bar{\sigma}^2}{N} + \frac{N-1}{N}\bar{C} = \bar{C} \tag{2.1.17}$$

$\bar{C}$ is systematic risk. This is the risk that is inherent to the entire market Rachev et al. (2010). It is also called undiversifiable risk or market risk. According to Elton et al. (2014), this is the only risk investors are compensated for because the other one can be eliminated by holding a well diversified portfolio.

2.2 From Markowitz Model to Risk Parity Portfolios

Risk parity is an example of a heuristic method. Heuristic methods refer to experience-based techniques and trial-and-error methods to find an acceptable solution, which does not correspond to the optimal solution of the optimisation problem (Roncalli, 2014). The key idea of risk parity is to construct a balanced portfolio with the same risk contribution for different assets. This will make it an equally weighted portfolio in terms of risk not in terms of allocations. It is very difficult to trace the origins of the risk parity portfolio. The term risk parity was first used 2005 but there is empirical evidence to support that these portfolios were used before 2005 by some investment management companies (Roncalli, 2014). For example, it was the main investment technique of All Weather Fund managed by Bridgewater for a long time. Before the theoretical work of Maillard et al. (2008), this portfolio was only sensible from a practical perspective but he managed to show that this portfolio exists, is unique and it lies between the minimum variance portfolio and the equally weighed portfolio.

Over the years, there has been an increase in the application of risk parity portfolios to investment management (Roncalli, 2014). In modern investment management, risk parity represents different allocation techniques. For instance, some investment practitioners use the term risk parity when asset weights and asset volatility are inversely proportional. Others consider risk parity portfolios as portfolios with equally weighted risk contributions (ERC) (Roncalli, 2014). In some cases, risk parity is the same as risk budgeting (RB) portfolios (Grinold and Kahn, 1999).

We will start by defining the properties of a risk measure $R(x)$ that are acceptable in terms of a risk allocation principle (Roncalli, 2014).

2.2.1 Properties of a risk measure.

R is a coherent risk measure if it satisfies the following properties (Artzner, 1999)

1. Sub-additivity

$$R(x_1 + x_2) \leq R(x_1) + R(x_2) \tag{2.2.1}$$

   The risk of one portfolio should be less than the risk of two separate portfolios.

2. Homogeneity

$$R(\mu x) = \mu R(x) \quad \text{if} \quad \mu \geq 0 \tag{2.2.2}$$

   Leveraging or deleveraging of the portfolio increases or decreases the risk measure in the same magnitude

3. Monotonicity

$$\text{if} \quad x_1 \prec x_2, \quad \text{then} \quad R(x_1) \geq R(x_2) \tag{2.2.3}$$

   If portfolio x_2 has a better return than portfolio x_1 under all scenarios, the risk measure $R(x_1)$ should be higher than the risk measure $R(x_2)$.

4. Translation invariance

$$\text{if } m \in \mathbb{R}, \quad \text{then} \quad R(x + m) = R(x) - m \tag{2.2.4}$$

   Adding a cash position of amount m to the portfolio reduces the risk by m.

The homogeneity and sub-additivity conditions were replaced by a weaker condition called the convexity property:

$$R(\mu x_1 + (1 - \mu)x_2) \leq \mu R(x_1) + (1 - \mu)R(x_2) \tag{2.2.5}$$

This condition means that diversification should not increase the risk.

2.2.2 Examples of Risk Measures.

By definition, the loss on the portfolio is $L(x) = -\mu_p(x)$ where $\mu_p(x)$ is the return on the portfolio. Let's consider different risk measures (Rachev et al., 2010):

- Volatility of the loss

$$R(x) = \sigma L(x) = \sigma(x) \tag{2.2.6}$$

- Standard deviation-based risk measure

$$R(x) = SD_c(x) = \mathbb{E}[L(x)] + c.\sigma L(x) = -\mu(x) + c.\sigma(x) \tag{2.2.7}$$

  This risk measure is found by scaling the volatility by a constant c and subtract the expected portfolio return

- Value-at-Risk

$$R(x) = \text{VaR}_\alpha(x) = \inf\{l : Pr\{L(x) \leq l\} \geq \alpha\} \tag{2.2.8}$$

The value at risk is the α quantile of the loss distribution X and we denote it by $X^{-1}(\alpha)$

- Expected Shortfall

$$R(x) = \text{ES}_\alpha(x) = \frac{1}{1-\alpha}\int_\alpha^1 \text{VaR}_s(x)ds \tag{2.2.9}$$

This can also be denoted as the expected loss beyond the value-at-risk:

$$\text{ES}_\alpha(x) = \mathbb{E}[L(x)|L(x) \geq VaR_\alpha(x)] \tag{2.2.10}$$

It can be shown that the expected shortfall and the standard deviation risk measures satisfy the coherency and convexity conditions.

2.2.3 Euler's Allocation Principle.

The first step of measuring portfolio risk management is risk measurement. One important risk management question is how much risk each asset is contributing to the portfolio risk? This step is called risk allocation (Roncalli, 2014). Risk allocation is a fundamental concept in risk management because it allows us to understand the risk profile of each asset in the portfolio and their contributions to portfolio risk. The Euler's risk allocation principle is the most prominent allocation principle among investment managers and the most accepted one.

Let Π be the profit and loss (P&L) of the portfolio. We split it into a sum of the n-security:

$$\pi = \sum_{j=1}^{n} \Pi_i \tag{2.2.11}$$

We denote the risk measure associated with the (P&L) by $R(\Pi)$. Let us consider the risk-adjusted performance measure (RAPM) defined by:

$$\text{RAPM}(\Pi) = \frac{\mathbb{E}[\Pi_i]}{R(\Pi)} \tag{2.2.12}$$

The two properties of risk contribution that are desirable from an economic perspective (Tasche, 2008):

1. Risk contributions $R(\Pi_i|\Pi)$ to portfolio-wide risk $R(\Pi)$ satisfy the full allocation property if:

$$\sum_{i=1}^{n} R(\Pi_i|\Pi) = R(\Pi) \tag{2.2.13}$$

2. Risk contributions $R(\Pi_i|\Pi)$ are RAPM compatible if there are some $\epsilon_i > 0$ such that:

$$\text{RAPM}(\Pi_i|\Pi) > \text{RAPM}(\Pi) \implies \text{RAPM}(\Pi + h\Pi_i) > \text{RAPM}(\Pi) \quad \forall\ 0 < h < \epsilon_i \tag{2.2.14}$$

If there are no risk contributions that are RAPM compatible in the sense of two previous properties (2.2.13) and (2.2.14), then $R(\Pi_i|\Pi)$ is uniquely determined as (Tasche, 2008):

$$R(\Pi_i|\Pi) = \frac{\mathrm{d}}{\mathrm{d}h} R(\Pi + h\Pi_i)\big|_{h=0} \tag{2.2.15}$$

and the risk measure is homogeneous of degree 1. In the case of a sub-additive risk measure, one can also show that:

$$R(\Pi_i|\Pi) \leq R(\Pi_i) \tag{2.2.16}$$

This means that the risk contribution of asset i is always smaller than its stand-alone risk measure. This difference is due to diversification.

Measuring $R(x)$ in terms of weights. The risk contribution of security i is defined as:

$$RC_i = x_i \frac{\partial R(x)}{\partial x_i} \tag{2.2.17}$$

and the risk measure satisfies the Euler decomposition:

$$R(x) = \sum_{i=1}^{n} x_i \frac{\partial R(x)}{\partial x_i} = \sum_{i=1}^{n} RC_i \tag{2.2.18}$$

This is the Euler allocation principle. It is at the heart of risk parity portfolios and a fundamental concept used by investment practitioners.

2.3 Principal Component Analysis : Mathematical Aspects

The pivotal idea behind Principal Component Analysis (PCA) is the reduction of the dimensionality of a dataset without loosing too much information from the data (Kassambara, 2017). We start with an NxN matrix.

The first step is to calculate the mean for every dimension in the dataset. We use the sample mean formula to calculate the mean of each dimension (Kassambara, 2017):

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{2.3.1}$$

where n is the number of observations, x_i is an observation.

The second step is to calculate the covariance matrix for the whole dataset. The covariance of a pair of dimensions is given by the formula:

$$\text{cov}(X, Y) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \tag{2.3.2}$$

The $n-1$ is to make the estimator of the covariance unbiased.

The next step is to compute the eigenvalues and corresponding eigenvectors of the covariance matrix.

Let A be the covariance matrix, $\mathbf{v}$ be a vector and λ be a scalar that satisfies $A\mathbf{v} = \lambda\mathbf{v}$, then λ is the eigenvalue associated with the eigenvector $\mathbf{v}$ of A. The eigenvalues of A are the values of λ that satisfy the following characteristic equation (Vidal and Ma, 2010):

$$\det(A - \lambda I) = 0 \tag{2.3.3}$$

We then sort the eigenvectors starting with the one with the largest eigenvalue and choose m eigenvectors with the largest eigenvalues to form a Nxm dimensional matrix M. Roughly speaking, the eigenvectors

with the smallest eigenvalues bear less information about the distribution of the data and those ones can be dropped.

The last step is to transform the samples onto the new subspace. We use the matrix M to transform our samples onto the new subspace via the equation:

$$y = M' \mathrm{x} x \tag{2.3.4}$$

where M' is the transpose of M.

3. Methodology

3.1 Hierarchical Clustering

Hierarchical clustering (Hierarchical cluster analysis) is the algorithm that groups similar features into groups called clusters, for example, when analysing stock market data, hierarchical clustering involves grouping stocks with similar characteristics like risk, return and liquidity, to mention a few, into clusters (Aggarwal and Reddy, 2014). The output is a set of clusters where each cluster is disjoint from the other, and the items in each cluster are similar to each other. We can perform hierarchical clustering on raw data or on a distance matrix.

The hierarchical clustering algorithm starts by treating each feature as a separate cluster (Aggarwal and Reddy, 2014). It first identifies the clusters that are closest to each other and then combine the most similar clusters. It iterates this process until all the clusters are joined together. The distance between two clusters is usually taken as the straight line from one cluster to another. This is called the Euclidean distance.

Determining from where the distance is computed from can be done in several ways (Bouveyron et al., 2019). One way is to compute between two most similar parts of a cluster (single-linkage) or at the centre of the cluster (mean or average linkage) or some other criterion like median, centroid to mention a few.

The output from hierarchical clustering can be shown in the form of a dendrogram. A dendrogram is used to show the relationship between features. In this essay, we are going to use a dendrogram to show the relationship between stocks.

3.2 Hierarchical Risk Parity

In this chapter, we introduce the Hierarchical Risk Parity (HRP) approach. HRP portfolios addresses three major concerns of the quadratic optimisers and Markowitz's critical line algorithm (CLA) in particular: instability, concentration, and under-performance (Prado, 2018). HRP applies graph theory and machine learning techniques to build a diversified portfolio based on the covariance matrix. The main benefit of the HRP approach is that it does not require us to invert the covariance matrix. This means that even if we have a singular covariance matrix, we can still calculate the weights of our portfolio which is not the case for the CLA.

The instability issues of the Markowitz's CLA have led to the rise of risk parity portfolios and other alternative models. Most alternative models attempt to increase robustness by adding more constraints (Clarke et al., 2002), or using Bayesian priors, or improving the numerical stability of the covariance matrix's inverse. These methods above, although published quite recently come from very classical areas of mathematics like calculus, linear algebra and geometry. A correlation matrix is derived from linear algebra and it measures the cosine between two vectors in a vector space of returns (Prado, 2018). We can use some concepts from graph theory to explain why the quadratic optimisers are unstable. One explanation for this is that its vector space is modelled with a complete graph where the nodes have a possibility of substituting the other. Finding the inverse of a matrix in algorithmic terms is finding the partial correlations across a complete graph. If we consider a 100x100 covariance matrix, it can be written as a graph with 100 nodes and 4950 edges. This sophisticated set up enlarges estimation errors, leading to wrong solutions. Thus, it is important to drop unnecessary edges.

Lets consider the practical meaning of this complex structure. Consider a diversified portfolio with equity, bonds and property to mention a few. Using economic theory, these investments can be grouped into substitutes and compliments. Stocks can be classified by liquidity, size and industry to mention a few (Shahidi, 2015). In deciding the allocation to a large traded equity like Apple, we might need to consider adding or deducting allocations of another similar stock like Google rather than a small insurance company in South Africa or a pharmaceutical company in India. If we consider the covariance matrix, all assets have the possibility of substituting the others. In other words, the correlation matrix lack the notion of hierarchy. This lack of hierarchy causes weights to vary freely which is the major source of instability in the CLA.

In the next section we are going to introduce a portfolio construction technique that uses graph theory and clustering. The Hierarchical Risk Parity method uses the information from the covariance matrix without requiring inverting the matrix or it to be positive-definitive. The model operates in three stages: tree clustering, quasi-diagonalisation, and recursive bisection.

3.2.1 HRP algorithm.

We start with a TxN matrix of daily stock returns. We want to blend N-column returns vectors into a hierarchical structure of allocations. The main reason for this is to make allocations flow downstream through a tree graph.

We start by computing an NxN correlation matrix from the daily stock returns with the following entries $\rho = \{\rho_{i,j}\}_{i,j=1,2,...,N}$ where $\rho_{i,j} = \rho[X_i, X_j]$.

We define the distance measure:

$$d : (X_i, X_j) \subset C \to \mathbb{R} \in [0, 1] \tag{3.2.1}$$

$$d_{i,j} = d[X_i, X_j] = \sqrt{\frac{1}{2}(1 - \rho_{i,j})} \tag{3.2.2}$$

C is a cartesian product of items in $\{1, 2, ..., i, ..., N\}$. This now allows us to calculate the distance matrix D. We can show that D is a proper matrix space.

Let's consider an example:

$$\{\rho_{i,j}\} = \begin{bmatrix} 1 & 0.9 & 0.2 & -0.7 \\ 0.9 & 1 & -0.7 & 0.1 \\ 0.2 & -0.7 & 1 & 0.8 \\ -0.7 & 0.1 & 0.8 & 1 \end{bmatrix} \to \tag{3.2.3}$$

$$\{d_{i,j}\} = \begin{bmatrix} 0 & 0.2236 & 0.6325 & 0.9220 \\ 0.2236 & 0 & 0.9220 & 0.6708 \\ 0.6325 & 0.9220 & 0 & 0.3162 \\ 0.9220 & 0.6708 & 0.3162 & 0 \end{bmatrix} \tag{3.2.4}$$

We calculate the Euclidian distance of two column vectors in the distance matrix D as follows:

$$\bar{d} : (D_i, D_j) \subset C \to \mathbb{R} \in [0, \sqrt{N}] \tag{3.2.5}$$

$$\bar{d}_{ij} = \bar{d}[D_i, D_j] = \sqrt{\sum_{n=1}^{N} (d_{n,i} - d_{n,j})^2} \tag{3.2.6}$$

$$\tag{3.2.7}$$

This gives:

$$\{\bar{d}_{i,j}\} = \begin{bmatrix} 0 & 0.8282 & 1.2864 & 1.4142 \\ 0.8282 & 0 & 1.4118 & 1.3247 \\ 1.2846 & 1.4118 & 0 & 0.5890 \\ 1.4142 & 1.3247 & 0.5890 & 0 \end{bmatrix} \rightarrow u[1] = (1,2) \tag{3.2.8}$$

We thus define a distance between the a newly created cluster $u[1]$ and the single items, so that $\{\bar{d}_{ij}\}$ is updated. This is technique is called linkage criterion in hierarchical clustering. This procedure is shown below:

$$u[1] = (1,2) \rightarrow \{d_{i,u[1]}\} = \begin{bmatrix} \min[0, 0.8282] \\ \min[0.8282, 0] \\ \min[1.2846, 1.4118] \\ \min[1.4142, 1.3247] \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 1.2846 \\ 1.3247 \end{bmatrix} \tag{3.2.9}$$

We update the matrix by appending $d_{i,u[1]}$ and dropping the clustered columns and rows:

$$\{\bar{d}_{i,j}\}_{i,j=1,2,3,4,5} = \begin{bmatrix} 0 & 0.8282 & 1.2864 & 1.4142 & 0 \\ 0.8282 & 0 & 1.4118 & 1.3247 & 0 \\ 1.2846 & 1.4118 & 0 & 0.5890 & 1.2846 \\ 1.4142 & 1.3247 & 0.5890 & 0 & 1.3247 \\ 0 & 0 & 1.2846 & 1.3247 & 0 \end{bmatrix} \tag{3.2.10}$$

$$\{\bar{d}_{i,j}\}_{i,j=3,4,5} = \begin{bmatrix} 0 & 0.5890 & 1.2846 \\ 0.5890 & 0 & 1.3247 \\ 1.2846 & 1.3247 & 0 \end{bmatrix} \tag{3.2.11}$$

We then perform the same procedure to columns 3 and 4 and after removing the clustered columns and rows, we get a 2x2 matrix and we perform the procedure again for the last time and then we stop.

3.2.2 Quasi-Diagonalisation.

The second step of the algorithm is quasi-diagonalisation or matrix seriation. Seriation is an explanatory data analysis technique to reorder objects into a sequence along a one-dimensional continuum so that it best reviews regularity and patterning among the whole series (Liiv, 2010). Using Hierarchical clusters from the preceding steps, we arrange the rows and columns in such a way that the stocks with larger covariances are placed along the diagonal and smaller ones around the diagonal. This is called a quasi-diagonal covariance matrix because the off-diagonal elements are not entirely zeros. The quasi-diagonalisation does not require us to change the basis of the covariance matrix. It is important because it groups similar investments together, and the dissimilar ones are placed far apart. Algorithmically, each row in the linkage matrix combines two branches into one. The clusters are then substituted with their constituents until all of them are finished. These substitutions keeps the order of the clustering.

3.2.3 Recursive Bisection.

The last part of the algorithm is recursive bisection (Prado, 2018). The inverse-variance allocation is optimal for a diagonal covariance matrix. We use these ideas in two different ways:

- bottom-up, to define the variance of a contiguous subset as the variance of an inverse-variance allocation.
- top-down, to split allocations between adjacent subsets in inverse proportion to their aggregated variances.

The algorithm below implements these ideas:

1. We initialise the algorithm by:

   (a) creating a list of items: $l = \{l_0\}$, where $l_0 = \{n\}_{n=1,2,...,N}$

   (b) we assign a weight of one to all the items in the list: $\alpha_n = 1 \quad \forall n = 1, 2, ..., N$

2. If $|l_i| = 1, \quad \forall l_i \in l$, then stop

3. For each $l_i \in l$ such that $|l_i| > 1$:

   (a) bisect l_i into two subsets, $l_i^{(1)} \cup l_i^{(2)} = l_i$, where $|l_i^{(1)}| = \text{int}[\frac{1}{2}|l_i|]$, and the order is preserved.

   (b) define the variance of $L_i^{(k)}, \quad k = 1, 2$ as a quadratic form $V_i^k = \bar{\alpha}_i^{(k)'} V_i^{(k)} \bar{\alpha}_i^k$, where V_i^k is a covariance matrix between the constituents of the $l_i^{(k)}$ bisection, and $\bar{\alpha}_i^j = \text{diag}[V_i^{(k)}]^{-1} \frac{1}{\text{tr}[\text{diag}[V_i^{(k)}]^{-1}}$ where diag[.] and tr[.] are diagonal and trace operators.

   (c) calculate the split factor: $\beta_i = \frac{\bar{V}_i^{(1)}}{\bar{V}_i^{(1)} + \bar{V}_i^{(2)}}$ such that $0 \leq \beta_i \leq 1$

   (d) re-scale allocations α_n by a factor of $\beta_i, \quad \forall n \in l_i^{(1)}$

   (e) re-scale allocations α_n by a factor of $(1 - \beta_i), \forall n \in l_i^{(2)}$

4. Loop to step 2

3.3 Data Analysis

We first download the data from yahoo finance and our data consists of 19 stocks listed on the New York Stock Exchange (NYSE). The data consists of daily stock returns from 2009-2019. We use adjusted closing price of stock for our data analysis. Adjusted closing price is a bit more complex than closing price because it takes into account factors like dividends, stock splits and new stock issues to determine its value (Reilly and Brown, 2012). The companies that we selected were from different industries so as to diversify our portfolio. We have seen from the earlier chapters that diversification reduces the portfolio risk and we have showed using the Markowitz (1952) model that if we take sufficiently large stocks in our portfolio, we can eliminate all the specific risk and be left only with systematic risk which cannot be diversified away. We considered tech companies like Apple and Google, huge companies in the finance sector like JP Morgan and retail companies like Amazon and Walmatt to mention a few.

We start by cleaning the data before the computations. After that we calculate the average daily return for each stock. For that we use the following formula:

$$r_t = \frac{S_t}{S_{t-1}} \tag{3.3.1}$$

Where S_t is the stock price at time t, S_{t-1} is the stock price at time $t-1$ and r_t is the return at time t. Our daily returns matrix is a TxN matrix. We then calculate the correlation matrix. The correlation

of two stocks is given by the following formula:

$$\rho = \frac{cov(S_i, S_j)}{S_i S_j} = \frac{\mathbb{E}[(S_i - \mu_i)(S_j - \mu_j)]}{\sigma_i \sigma_j} \tag{3.3.2}$$

$$= \frac{\mathbb{E}[(S_i - \mu_i)(S_j - \mu_j)]}{s_i s_j} \tag{3.3.3}$$

Where $\sigma_x = s_x =$ standard deviation of x and $\sigma_y = s_y =$ standard deviation of y

We use this above formula to compute the covariance and correlation matrix.

The correlation matrix is shown as a heatmap below:

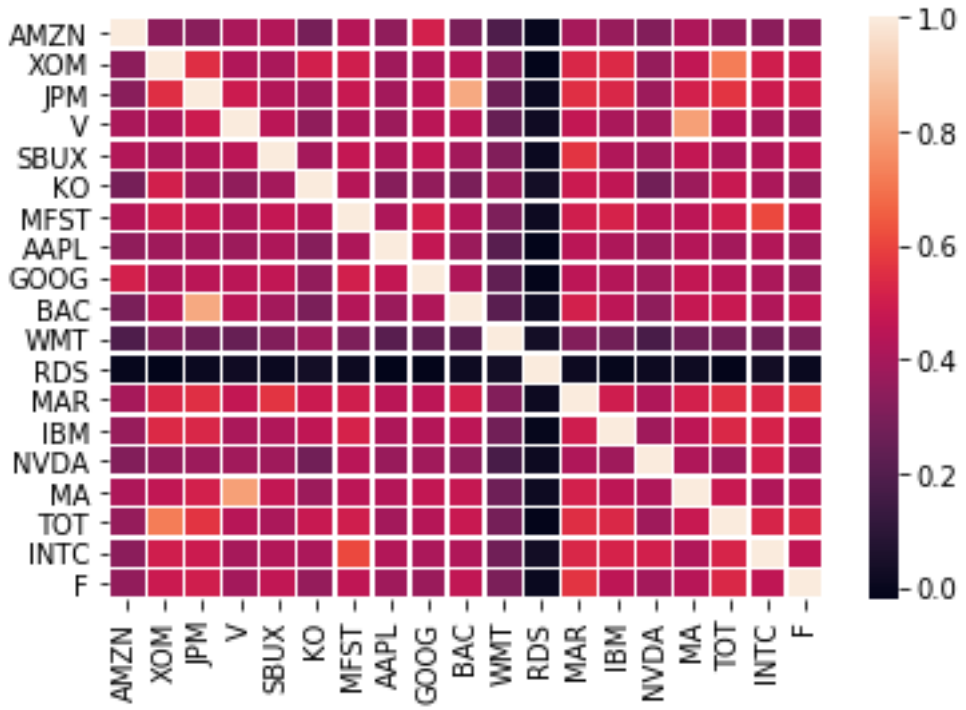


Figure 3.1: Heatmap of the correlation matrix

The other important part is the risk measure for each individual asset. We use the standard deviation as our risk measure for the following reasons (Fabozzi et al., 2007):

- It is mathematically tractable
- It leads to elegant solutions that are easy to intepret.
- Emperical evidence shows that it provides a good approximation to different methodologies.

However, it has the following drawbacks:

- investors are usually concerned with the downside risk. This is the risk that returns will fall below the mean. This is a problem because it leads to overstating of the risk that the investors are actually exposed to.

Standard deviation is given by the following formula:

$$\hat{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(S_i - \mathbb{E}(S_i))^2} \tag{3.3.4}$$

We use $n-1$ instead of n so as to make our formula an unbiased estimator for variance.

We use the correlation matrix for hierarchical clustering. We start by computing the distance matrix given by:

$$d = \sqrt{\frac{1}{2}(1 - c_{ij})} \tag{3.3.5}$$

where c_{ij} is an entry from the correlation matrix.

We can plot a tree structure from our example in the form of a dendrogram.

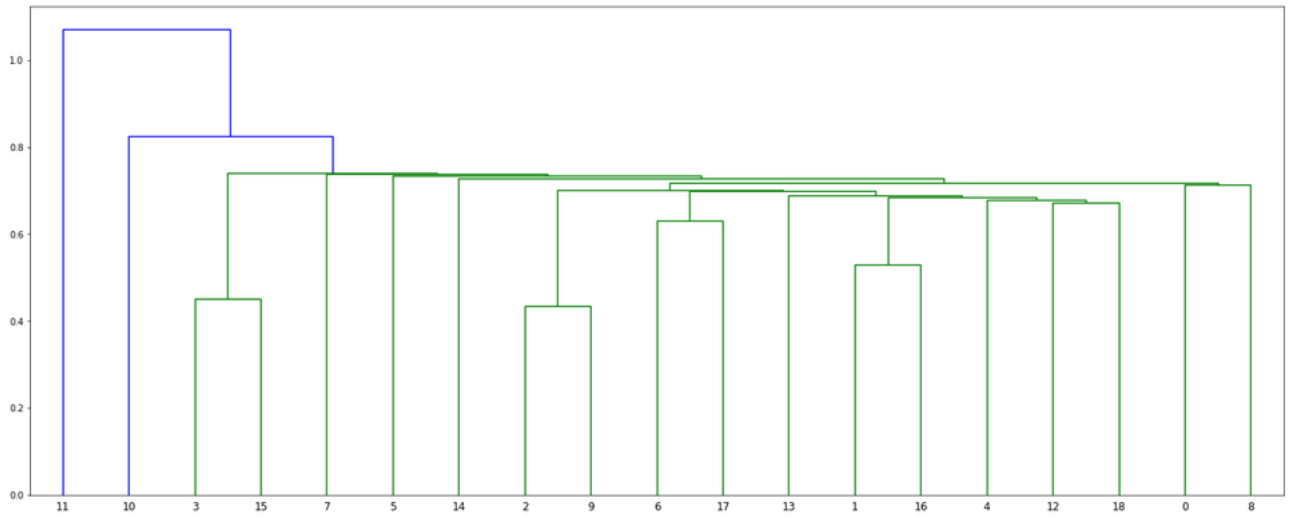


Figure 3.2: Dendrogram of cluster information

The $y - axis$ measures the distance between two clusters, in this case, it is measuring the distance between two stocks.

We then calculate the quasi-diagonal correlation matrix. The main advantage of quasi-diagonalisation is that it does not require us to change the basis which is not the case for principal component analysis and other similar procedures. HRP solves the allocation problem robustly, while working with original investments (Prado, 2018).

The last part is recursive bisection which takes advantage of the fact that the inverse-variance allocation is optimal for a diagonal covariance matrix. This is part that calculates the allocations using the algorithm defined in the previous section.

3.4 Principal Component Analysis (PCA)

PCA is at the heart of data analysis because of its ability to reduce the dimensionality of a data set without the loss of too much information Vidal and Ma (2010). It is also powerful from a computational perspective because it makes algorithms run faster.

The first step is to standardise the data. We standardise the data because PCA is so sensitive to the scaling of data. We standardise the data using the formula:

$$Z_i = \frac{X_i - \mathbb{E}(X)}{std(X)} \tag{3.4.1}$$

Where Z_i is the standardised observation, X_i is the initial observation, $\mathbb{E}(X)$ is the mean of the observation and $std(X)$ is the sample standard deviation from the mean. Again, we use:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{to estimate } \mathbb{E}(X) \tag{3.4.2}$$

and

$$std(X) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \tag{3.4.3}$$

Alternatively, we can use the sklearn built in function for standard scaling.

We calculate the covariance matrix of dimensions. We use the scaled data to calculate the covariance matrix. The covariance matrix is a 19x19 matrix because we have 19 stocks in our portfolio. We use the numpy command **np.cov(data)** to find the covariance matrix of the data.

We calculate the eigenvalues and eigenvectors of the covariance matrix. The eigenvectors determine the direction of the new feature space and eigenvalues determine the size. Alternatively, the eigenvalue explains the data along the new features and the corresponding eigenvector tells us how much variance is included in the new transformed feature. To calculate the eigenvalues and corresponding eigenvectors, we use the numpy command **np.linalg.eig(cov matrix)**.

We then verify if the sum of the square of each value in an eigenvector is 1. We do this to check if the eigenvectors were calculated correctly.

After checking we then select the principal components. The goal is to reduce the dimensionality of the matrix by by projecting it into smaller subspaces where the eigenvectors will form the axes. To decide which eigenvectors to drop without loosing too much information for the construction of the lower dimension subspace, we need to inspect the corresponding eigenvalues. The eigenvectors with the lowest eigenvalues bear less information about the distribution of the data and those ones can be dropped. To do so, we can rank the eigenvalues from highest to lowest and choose the top m eigenvalues. There is no rule of thumb to finding this m but we want to choose the principal components that contribute more to the total variance.

A useful measure used to choose the principal components is the "explained variance". This can be calculated from the eigenvalues. This tells use the amount of variance or information each principal component is contributing.

After doing this, we then pick the principal components that are contributing a lot to the total variance. For example, if there are 6 principal components with the following variances $40\%, 30\%, 20\%, 5\%, 3\%$ and 2%

respectively, I would pick the first 3 because they are contributing 90% to the total variance. From our experiments, we see that 4 principal components are explaining 95% of the data. This is show in the diagram below:

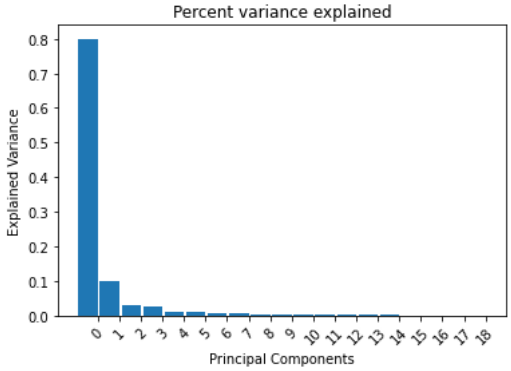


Figure 3.3: Percentage variance explained by principal components

However, all these steps can be done by a simple command called **PCA()** but the underlying idea is the same.

The next step is to calculate the Sharpe Ratio of single eigen-portfolios. A Sharpe ratio was developed by William Sharpe and is used by investors to compare risk and return. This ratio is the average return earned in excess of the risk-free rate per unit of volatility or total risk (Fabozzi et al., 2007). We calculate the annualised returns with the following formula:

$$r_a = [\prod_{i=1}^{n}(1 + r_i)]^{1/t} \tag{3.4.4}$$

Where t is the number of years, r_i are the returns on the eigen-portfolio.

The annualised volatility is:

$$std(r_a) = std(\text{returns})\sqrt{250} \tag{3.4.5}$$

where $std()$ is defined as before, 250 is the number of trading days in a year. The annualised Sharpe ratio is:

$$S_x = \frac{r_a}{std(r_a)} \tag{3.4.6}$$

We compute principal component's weights for each eigen-portfolio using an iterable loop. The principal components weights are given by:

$$w_i = \frac{\text{pcs(i)}}{\text{sum(pcs())}} \tag{3.4.7}$$

where w is the weight of principal component i, pcs(i) is the principal component i and sum(pcs) is the sum of all principal components. We then calculate the portfolio that maximises the Sharpe ratio. We use the Sharpe Algorithm to find the portfolio with the highest Sharpe ratio.

3.5 Results

Using the HRP model, we get the following allocations:

Asset	Allocation
RDS	0.0000
WMT	0.1981
V	0.0493
MA	0.0418
AAPL	0.0594
KO	0.1703
NVDA	0.0224
JPM	0.0205
BAC	0.0101
MSFT	0.0549
INTC	0.0499
IBM	0.0746
XOM	0.0474
TOT	0.0239
SBUX	0.0451
MAR	0.0385
F	0.0374
AMZN	0.0208
GOOG	0.0356

Table 3.1: Allocations using HRP

HRP allocates the most capital to Walmatt. 19% is allocated to this stock. The least capital is allocated Royal Dutch Shell. The proportion allocated is nearly 0%. Overall, we can observe that HRP diversifies across all investments and also across the cluster. Allocating capital across all investments makes the portfolio more resilient to idiosyncratic shocks. Allocating capital across the cluster takes into consideration the correlation structure will make the portfolio more resistant to systemic shocks.

Using the classical Markowitz (1952) model, we get the following allocations:

Asset	Allocation
RDS	0.1938
WMT	0.1981
V	0
MA	0
AAPL	0
KO	0
NVDA	0
JPM	0
BAC	0
MSFT	0
INTC	0
IBM	0
XOM	0
TOT	0
SBUX	0
MAR	0
F	0
AMZN	0.8062
GOOG	0

Table 3.2: Allocations using the Markowitz model

The efficient frontier is given by:

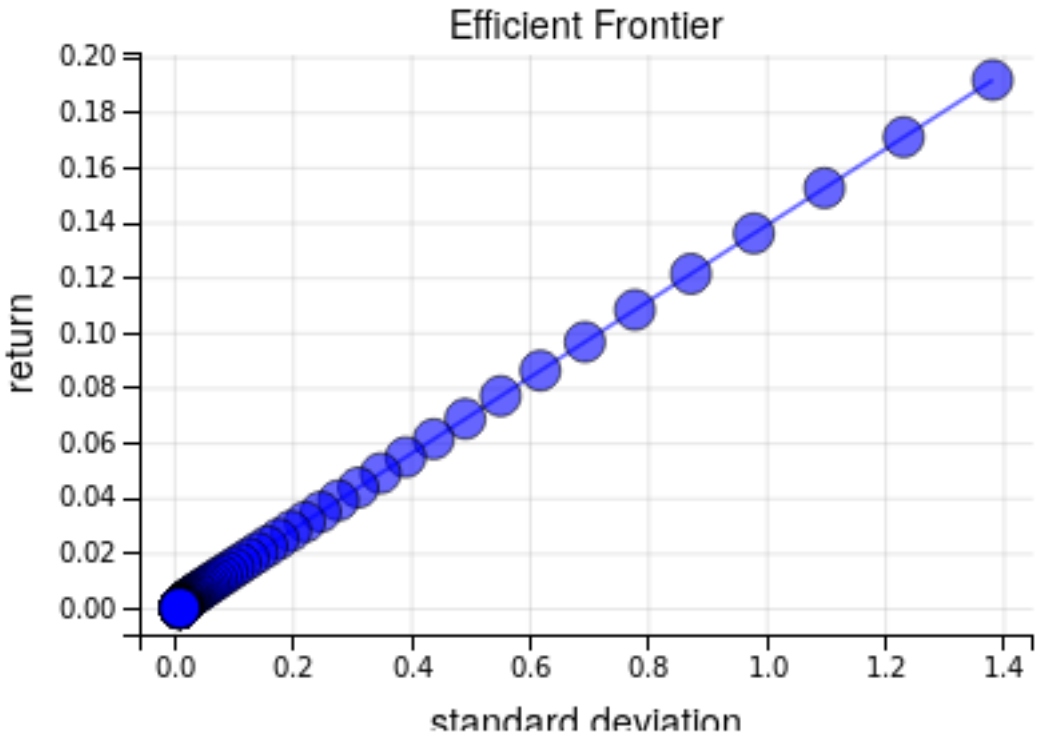


Figure 3.4: Efficient Frontier

From the Markowitz model, we only allocate capital to two assets which are Amazon and Royal Dutch Shell. This is a common problem with the Markowitz model since it allocates capital to fewer investments. This increases the exposure to idiosyncratic shocks. These are shocks that are associated with a particular stock of a particular company. It is a form of specific risk. This is why HRP is a better model because it diversifies across all asset classes and by doing so it makes the model stable model.

The allocations from PCA are:

Asset	Allocation %
RDS	3.49
WMT	1.87
V	1.65
MA	20.71
AAPL	2.72
KO	2.70
NVDA	16.05
JPM	1.87
BAC	0.74
MSFT	3.44
INTC	6.32
IBM	10.25
XOM	2.00
TOT	0.29
SBUX	0.29
MAR	9.95
SP500	10.42
AMZN	4.33
GOOG	1.19

Table 3.3: Allocations using PCA

The PCA is allocating capital to all assets and the allocations on most assets are between $0 - 5\%$. We are allocating more than 10% of the capital to four assets. As far as diversification is concerned, this model provides a well diversified portfolio because it's allocating capital to all the assets and no asset is completely dominating the others. We can get more insight into the allocations by sketching the diagram of allocations.

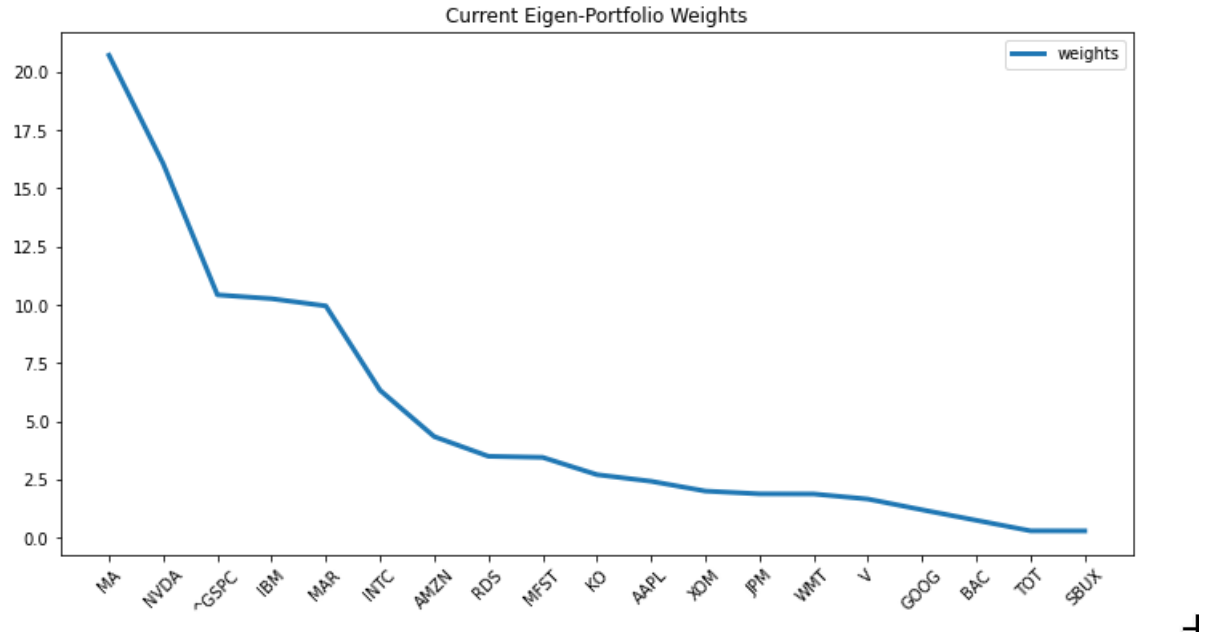


Figure 3.5: Allocations using PCA

We see that the graph is very steep from the first asset to the third indicating that the the first three allocations differ significantly from each other. From the fourth asset to the last we see a steady fall in the allocations with implies that the difference is less significant.

4. Conclusion

We have allocated capital using three methods which are the classical Markowitz model, Hierarchical Risk Parity (HRP) and Principal Component Analysis (PCA). In this Section, we are going to choose the best by taking factors like diversification, resilience to systematic and idiosyncratic shocks, flexibility and stability of the model into consideration. The most important property is how diversified a portfolio is. We already showed in Section 2 that diversification reduces the specific risk and can be achieved by increasing the number of assets in the portfolio. Specific risk is market related thus cannot be diversified away but can be controlled by taking into account the correlation structure of the portfolio. We can control it by spreading out the proportions of assets. Taking these factors into consideration, we can compare the three approaches.

4.1 Comparison of the Three Approaches

We observe that HRP and PCA are allocating capital to many assets and in different proportions. The weights from HRP and PCA are quite close to each other as compared to those from the classical Markowitz model. The Markowitz model is concentrating its weights to a few assets which is the reason for its instability. This is also increase the variance of the portfolio because we are not minimising the specific risk as much as possible. Thus, allocating capital using the Markowitz model is quite risky. However, this is not the case for HRP and PCA algorithms we have considered. We observe that they are quite stable because they are diversifying the risk across all assets leading to more stable portfolios.

HRP and PCA are great methods because they are resilient to both idiosyncratic and systematic shocks which is arguably the most important thing in investment management. Both produce a well diversified portfolio which is very pivotal in modern investment management because these diversified portfolios are less risky as compared to allocating all the capital to a few assets. We can conclude that these two methods lead to well diversified portfolios. Nevertheless, a sensitivity analysis on the PCA might be important to check for the stability of the weights.

However, between the two methods, HRP is better than PCA because it is very difficult to interpret results from a portfolio that has independent assets using PCA. This means that the covariance matrix will have a lot of values close to zero. Principal components are linear combinations of original features. Principal components are not as readable and interpretable as original features. PCA is also losing some information. Although the loss of information is not too much, this can lead sensitive portfolios which are unstable if not used carefully. We can also change the HRP algorithm to make it suitable for our needs. For example, we can choose an alternative distance function and allocation constraints to mention a few. This is not however the case for PCA because if we make some changes to it we end up with results that are difficult to interpret.

4.2 Future work

We can change the distance function of HRP, consider alternative clustering techniques like bi-clustering or introduce different allocation constraints. We can also test for the stability for the PCA method.

Acknowledgements

I want to thank God for giving me strength to do this project. Secondly, I want to thank my family and friends for their unconditional support. Lastly, I want to acknowledge AIMS and it's funders for supporting this work, as well as my supervisor, Prof Phillip Mashele from North West University.

References

Aggarwal, C. C. and Reddy, C. K. *DATA CLUSTERING Algorithms and Applications*. CRC Press, 2014.

Artzner, A. Coherent measures of risk. *Mathematical Finance*, 1999.

Ban, G.-Y., Karoui, N. E., and Lim, A. E. Machine learning and portfolio optimization. *Management Science Articles in Advance*, 2016.

Bouveyron, C., Celeux, G., Murphy, T. B., and Raftery, A. R. *Model Based Clustering and Classification for Data Science*. Cambridge University Press, 2019.

Clarke, Silva, D., and Thornley. Portfolio constraints and the fundamental law of active management. *Financial Analysts Journal*, 58:48–66, 2002.

Elton, E., Gruber, M., Brown, S., and Goetzmann, W. *Modern Portfolio Theory and Investment Analysis*. Wiley, 2014.

Emerson, S., Kennedy, R., O'Shea, L., and O'Brien, J. Trends and applications of machine learning in quantitative finance. Available from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3397005(Last accessed on 6 April 2020), 2019.

Fabozzi, F., Kolm, P., Pachamanova, D., and Focardi, S. *Robust Portfolio Optimization and Management*. Wiley, 2007.

Grinold, R. and Kahn, R. *Active portfolio management*. McGraw-Hill, 1999.

Kassambara, A. *Practical Guide to Principal Component Methods in R*. STHDA, 2017.

Liiv, I. *Serial and matrix reordering methods: A historical view*. Wiley, 2010.

Maillard, S., Roncalli, T., and Teiletche, J. On the properties of equally weighed risk contribution portfolios. *the journal of PORTFOLIO management*, 2008.

Markowitz, H. Portfolio selection. *The Journal of Finance*, 7:77–91, 1952.

Meucci, A. *Risk and Asset Allocation*. Springer, 2007.

Pankaj Gupta, G. M., Mukesh Kumar Mehlawat. Asset portfolio optimization using support vector machines. *J Glob Optim*, 53:297–315, 1948.

Perrin, S. and Roncalli, T. Machine learning optimisation algorithms and portfolio allocation. Available from https://arxiv.org/abs/1909.10233(Last accessed on 10 April 2020), 2019.

Prado, M. L. D. *Advances in Financial Machine Learning*. Wiley, 2018.

Rachev, S., Stoyanov, S., and Fabozzi, F. *Advanced Stochastic Models, Risk Assessment, and Portfolio Optimisation*. Wiley, 2010.

Reilly, F. and Brown, K. *Investment Analysis and Portfolio Management*. Cengage Learning, 2012.

Roncalli, T. *Introduction to Risk Parity and Budgeting*. Chapman and Hall, 2014.

Shahidi, A. *Balanced Asset Allocation*. Wiley, 2015.

Tasche, D. *Capital Allocation to all Business Units and Sub Portfolios*. Risk Books, 2008.

Vidal, R. and Ma, Y. *Generalised Principal Component Analysis*. Springer, 2010.