Portfolio Optimization using various correlation estimation techniques

Ashutosh Mulchandani
200100037

Desai Utsav Manojkumar
200100054

Dhruvkumar Jagan Patil
200100056

Pragyesh Gupta
200020094

Abstract

This Project aims to achieve Portfolio Optimization using various Correlation Matrices. The Problem is to Maximize the Expected Returns for a given risk tolerance limit. Our base model is the Markowitz Model and we have tried to implement various correlation matrices including Distance Correlation, Spearman Rank Correlation, Kendall Rank Correlation and Pearson's Correlation. We have attempted to solve a relaxed Mixed Integer Nonlinear Programming Problem to calculate weights distribution for the stocks of companies included in SENSEX.

Introduction

- Portfolio optimization is the process of selecting the best portfolio (asset distribution), out of the set of all portfolios being considered, according to some objective. The objective typically maximizes factors such as expected return, and utility of the investor and minimizes costs like financial risk.
- Factors being considered may range from tangible (such as assets, liabilities, earnings or other fundamentals) to intangible (such as selective divestment).
- This particular project entails maximizing the returns while considering the risk tolerance specified. The returns and risks associated are derived from the famous Markowitz Model, which estimates the return of an asset as the time-averaged asset price and the risk of an asset as the standard deviation associated with the asset over the considered time interval.
- The portfolio return is simply the sum of all optimized weights times the return of an asset while the portfolio risk involves correlation between assets.

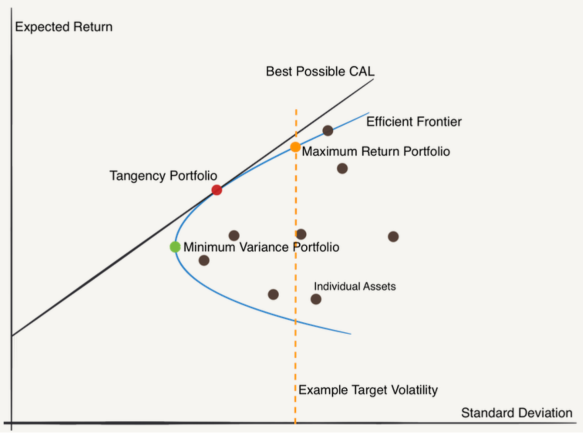


Figure 1 : Efficient Frontier: (Reference: Markowitz Model-(wallstreetmojo.com))
The Green Point corresponds to Minimum variance Portfolio while the yellow point represents maximum return portfolio for the given risk target. The blue line is the efficient frontier and the various assets are marked in brown.

- The **efficient frontier** is a concept in modern portfolio theory (MPT) that refers to a set of optimal portfolios that offer the highest possible expected return for a given level of risk or the lowest possible risk for a given level of expected return.
- It is the set of portfolios that provide the highest expected return for a given level of risk, or the lowest risk for a given level of expected return. The efficient frontier is determined by plotting the expected return and standard deviation (or volatility) of all possible portfolios on a graph. The efficient frontier represents the upper boundary of the graph, and any portfolio lying on this boundary is considered to be an efficient portfolio.
- The correlation between two assets is described as $Corr(X, Y) = Cov(X, Y)/(\sigma_X * \sigma_Y)$, where σ_X is the Standard deviation or the risk associated with the asset X and $Cor(X, Y),\ and\ Cor(X, Y)$ are Correlation and covariances between assets, X and Y, respectively.

Problem Definition

The problem is to maximize the investment return R(w), by diversifying the amount of risk between *k* assets, given the historical data of assets for ***N*** days.

Decision Variable:

The distribution of investments across different asset classes (e.g. stocks, bonds, etc.). Weights will be assigned to assets where

$$\mathbf{w_i} = \frac{Capital\ invested\ in\ asset\ `i'}{Total\ Capital} \quad , 0 < w_i <= 1$$

Objective Function: Maximize Portfolio returns

$$\mathbf{R(w)} = argmax\ (w^T.E) \quad , \quad w \in R^k$$

Constraints:

- **Portfolio Risk** to be less than **Investor's Risk tolerance:** $w^T.Cov.w <= \gamma^2$
- **Capacity Constraint:** $\Sigma w = 1\ \&\ w_i >= 0 \quad \forall\ i = 1,2,3 \ldots.k$

Parameters:

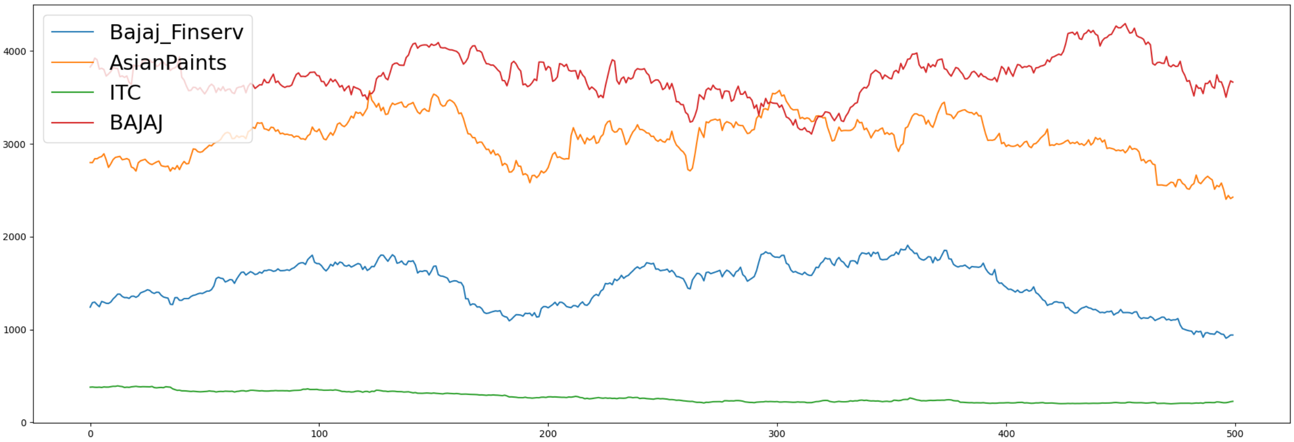
- **Daily return (ith day):**

  $R_i = P_{i+1}/P_i - 1 \quad , R_i \in \mathbb{R}^k$

  P is the Price matrix where jth column is the prices of jth asset and P_i is the row matrix of all the prices of the assets on day i
- **Expected Returns:** $E = \sum_{i=1}^{N} R_i \quad ,E \in \mathbb{R}^k$
- **Target Risk:** Investor's risk tolerance = γ

Data Collection

We had chosen the same 30 companies which are included in SENSEX index calculation for our project, since their data was readily available and the results comparison would also be easier. We downloaded historical data of one year for these companies from Yahoo Finance. Then the different fragments of data(like 120 or 60 days somewhere) was used for different time series analysis..



Example Data plot for four stocks

As visible in the graph below, the order of prices of stocks was also very different, ranging from 10's to 1000's. Although, this will not affect the optimization results since eventually we are considering the ratio of relative differences to closing prices as the returns to calculate the covariance matrix.

Below is the list of the companies considered for the project:

HDFC	AXISBANK	HCLTECH	TCS	ULTRACEMCO	HEROMOTOCO
HDFCBANK	LT	SBIN	HINDUNILVR	TITAN	INDUSINDBK
RELIANCE	MARUTI	NESTLTIND	ITC	TECHM	NTPC
INFY	BAJAJFINSV	SUNPHARMA	KOTAKBNK	BAJAJ-AUTO	ONGC
ICICIBANK	ASIANPAINT	M&M	BHARTIARTL	POWERGRID	TATASTEEL

Literature Review

We referred to some of the papers based on portfolio optimisations and their recent research items and also about different correlation matrices and their uses. Some of them as follows::

1) **"A Simplified Perspective of the Markowitz Portfolio Theory"** by **Myles E Margram** highlights the information about the famous Markowitz model and its intricacies. As presented in the paper, Modern Portfolio Theory is based on Markowitz' Portfolio Selection theory and William Sharpe's

contributions to the theory of financial asset price formation, which came to be known as the Capital Asset Pricing Model. Essentially, MPT is an investment framework for the selection and construction of investment portfolios based on the maximization of expected returns of the portfolio and the simultaneous minimization of investment risk.

a) It talks about different types of risk, namely systematic (common for many assets) and unsystematic risk (specific for one asset). Also, the concept of 'Risk and Return trade-off' relates to Markowitz' basic principle that the riskier the investment, the greater the required potential return. This paper also gives us insight about the various ways to determine the volatility (risk) of a particular security's return. The two most common measures are variance and standard deviation.
b) Answers how MPT attempts to analyze the interrelationship between different investments. It utilizes statistical measures such as correlation to quantify the diversification effect on portfolio performance. If the correlation between the securities is positive, then the variables are positively correlated and if it is negative, then they are negatively correlated and if the correlation is zero, then the variables are determined to be uncorrelated. In real life, no two assets are uncorrelated.
c) It discusses the problems in the model like not including transaction costs, assumption of efficient markets etc. But it gives immense appreciation to the model for always being the basis for so many other developments in the field.

2) The paper **"On relationships between the Pearson and the distance correlation coefficients"** shows us the relationship between the Pearson correlation coefficient and the distance correlation coefficient.\

a) The authors first provide an overview of the Pearson correlation coefficient and the distance correlation coefficient and their respective properties. They then derive a formula that expresses the Pearson correlation coefficient in terms of the distance correlation coefficient and vice versa.
b) The paper also discusses the strengths and limitations of both coefficients and provides examples to illustrate their differences in practice. Additionally, the authors explore the relationship between the coefficients in high-dimensional settings, demonstrating that the distance correlation coefficient is better suited for measuring dependence between variables in such contexts.

3) The paper **"Portfolio Optimization using Rank Correlation"** by **Wenjun Zhou** and **Chanaka Edirisinghe** presents a novel approach to portfolio optimization based on rank correlation.

a) This paper shows using Spearman's rank correlation coefficient, a non-parametric measure of dependence between two variables, to select stocks with high correlation and low risk.
b) The paper begins by introducing the concept of portfolio optimization and the role of correlation in constructing optimal portfolios. Then they explain the limitations of traditional correlation measures such as Pearson's correlation coefficient and propose the use of rank correlation as a more robust alternative than normally calculating correlations.
c) The paper then describes the methodology for portfolio optimization using rank correlation, including the selection of a suitable ranking system and the calculation of a rank correlation matrix. They depict the effectiveness of their approach through empirical examples, showing that portfolios constructed using rank correlation outperform those constructed using traditional methods.

Mathematical Background

Return:

We have calculated the expected returns of the assets in portfolio in three ways :-

1. 60 Days returns average - Taking the average value of the asset returns over last 60 days (i.e., most recent 60 days)
2. 120 Days returns average - Taking the average value of the asset returns over last 120 days (i.e., most recent 120 days)
3. Exponentially Weighted Moving Average - Taking average as follows:

$$\text{EMA}_{\text{Today}} = \text{Price Today} \times \left(\frac{\text{Smoothing}}{1 + \text{Days}}\right) + \text{EMA}_{\text{Yesterday}}\left(1 - \left(\frac{\text{Smoothing}}{1 + \text{Days}}\right)\right)$$

The most common choice for Smoothing factor = 2

That gives the most recent observation more weight. If the smoothing factor is increased, more recent observations have more influence on the EMA.

Risk: Risk can be estimated by some measure of dispersion. In the markowitz model, standard deviation, σ, is used to represent risk. Portfolio Risk is estimated by the following expression:

$$\mathbf{w^T.Cov.w}, \qquad \text{where Cov is the covariance of k assets}$$

In this text, we have discussed various methods for estimating correlation matrices, which lead to different covariance.

Correlation Matrices:

1. **Pearson's Correlation:** The Pearson correlation method is the most common method that is used for finding correlation in the portfolio optimisations and is part of the most basic Markowitz Portfolio model. It assigns a value between −1 and 1, where 0 is no correlation, 1 is perfect positive correlation and −1 is perfect negative correlation. This draws a line through the data of two variables to show their relationship as it assumes that there is a linear relationship between them.

$$r = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x})^2 \Sigma(y_i - \bar{y})^2}}$$

However, in the volatile field of stock prices, the relationship between two assets cannot be mapped accurately as having a linear nature. Thus a better estimate for the correlation between stocks is needed. So the below mentioned correlation matrices opens up the potential to better capture nonlinear dependencies between variables.

2. **Distance Correlation:** It goes beyond Pearson's correlation because it can spot more than linear associations and it can work multi-dimensionally. Distance correlation ranges from 0 to 1, where 0 implies independence between the two variables and 1 implies that their are highly correlated.

$$dCor(X,Y) = \frac{dCov(X,Y)}{\sqrt{dVar(X)\,dVar(Y)}}$$

$$Cor(X, Y) \ = \ Correlation\ between\ variables\ X\ and\ Y$$
$$Cov(X, Y) \ = \ Covariance\ of\ variables\ X\ and\ Y$$
$$Var(X) \ = \ Variance\ of\ variable\ X$$

It calculates the coefficient of correlation by the distance covariance, which is given by

$$dCov_n^2(X,Y) = \frac{1}{n^2}\sum_{j=1}^{n}\sum_{k=1}^{n} A_{j,k}B_{j,k}$$

Where $A_{j,k} = ||X_j - X_k||$, $B_{j,k} = ||Y_j - Y_k||$

and distance variance, which is simply the distance covariance of two identical variables.

3. **Spearman's Rank Correlation:** The Spearman correlation measures the strength of a monotonic relationship between two variables with the same scaling as the Pearson correlation. It is calculated by determining the ranks of the values of those two variables. As in real scenarios, the data is not linearly related and hence non parametric rank correlation methods are used like spearman and kendall.

$$r_s \ = \ 1 - \frac{6\sum D^2}{n(n^2 - 1)}$$

4. **Kendall Rank Correlation:** It is often referred to by Kendall's tau test. It is a statistic used to measure the ordinal association between two measured quantities or variables. The intuition for the test is that it calculates a normalized score for the number of matching or concordant rankings between the two samples and hence is also called Kendall's concordance test. It means that it measures how similar are the ranks of the values of variables and not values exactly.

$$\tau \ = \ \frac{n_c - n_d}{\frac{1}{2}n(n-1)}$$

$$n \ = \ Sample\ size$$
$$n_c \ = \ Number\ of\ concordant\ pairs$$
$$n_d \ = \ Number\ of\ discordant\ pairs$$

Various Approaches

- Moving forward with the project we tried various software for modelling our problem. We started with the AMPL software, which is an algebraic modeling language used for solving high complexity problems. But moving forward, we found some issues with CSV files handling in AMPL. Thus we had to shift to python language, where file handling is easier.

- In python we tried various libraries throughout the course of the project, which include, pyomo, baron, scipy, gurobipy, etc. Due to some limitations of various libraries and some issues while modelling the problem, we finalised to gurobipy library, which is a mathematical optimization library for solving mixed-integer linear and quadratic optimization problems.

In gurobipy, we implemented the optimization for the weights in four different ways, each being better than the previous, which are as follows :

1. **Continuous:** Here, we assumed the weights variation to be continuous from 0 to 1. The issue here was that we were getting some weights to much less as compared to others i.e. order of 10^{-10}, which is practically not possible to invest such a low amount.

2. **Relaxation:** To solve this problem we were guided by our professor to try out for Mixed Integer Programming and that has been the reason for selecting gurobipy as our model framework for solving the problem. Initially, there were some implementation problems for solving Mixed Integer Non Linear Programming Problem, then we attempted to solve a relaxation for the same problem. We attempted to solve with weights being continuous variables and then enforced all the weights which are less than 10^{-3} to be zero, i.e, we added the constraint to the model for all weights less than 10^{-3} to be 0 and again solved the problem.

3. **Integer Weights:** This is a somewhat improved version of previous assumptions. We assumed that the sum of all the weights is 1000, and the weights can only be positive integers. This ensured that no fraction of weights is less than 10^{-3} and the all optimized weights turned out to be greater than 0.001.

4. **Integer Stocks:** This is the most practical way of implementing the problem. Here, we assumed that the stocks can be only in integer amounts(one can't buy a fraction of stocks). So, here our final output was how much amounts of a particular stock should we be considering in our portfolio.

Results & Discussion

After assuming weights to be continuous, the problem was attempted to be solved with Gurobi solver and the results are shown below :

1) Continuous

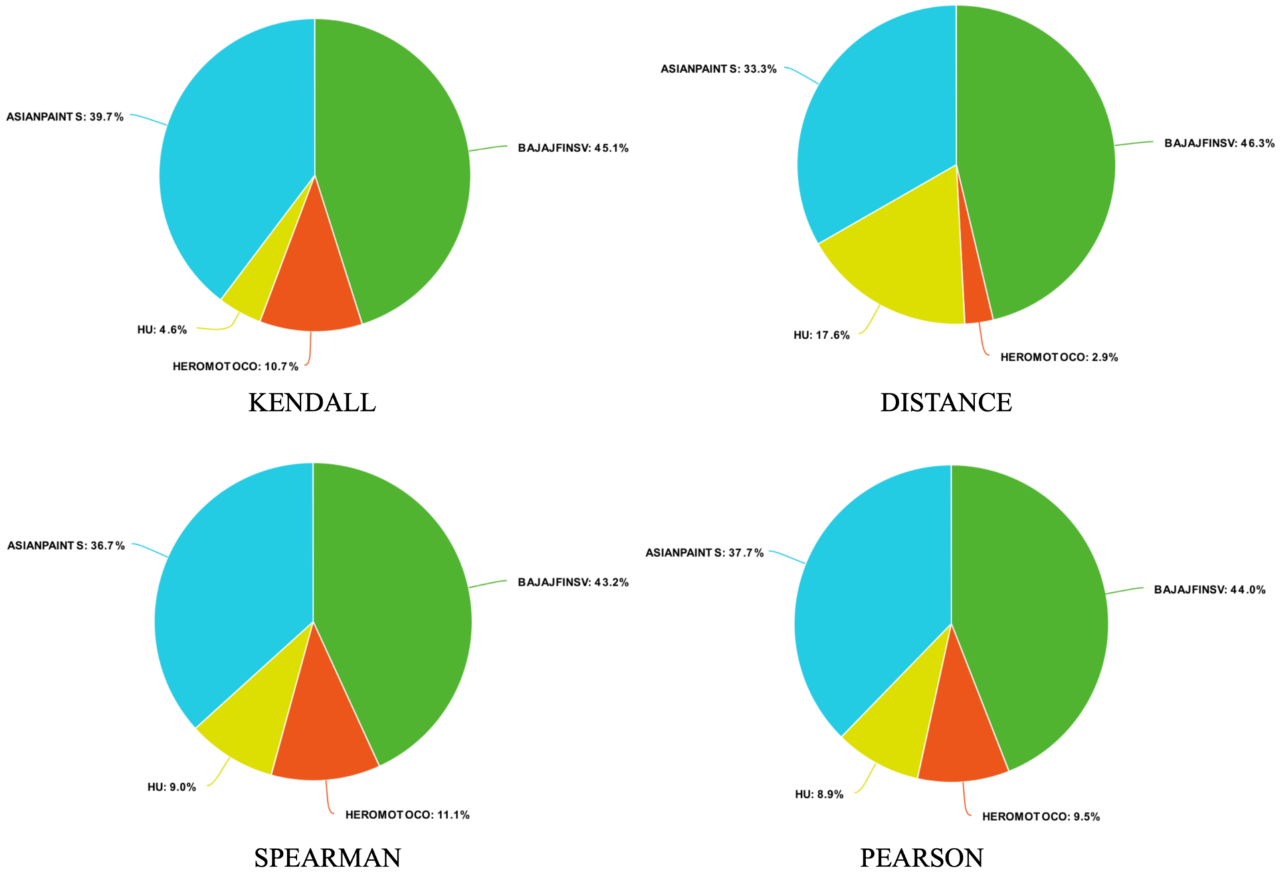
INDUSINDBK	0.000962	BHARTI ARTL	1.53E-06	BAJAJ AUTO	5.60E-10
RELIANCE	1.13E-08	HDFCBANK	7.75E-10	HUL	0.047949
BAJAJ FINSERV	0.451626	NESTLE	3.24E-09	INFOSYS	3.56E-09
AXISBANK	5.30E-10	TCS	1.07E-09	ULTRACEMCO	4.03E-10
KOTAKBANK	7.59E-08	ITC	4.77E-10	SBI	2.39E-09
TITAN	1.73E-09	HEROMOTOCO	0.104851	SUNPHARMA	2.54E-09
TECHM	8.85E-10	ICICI	2.33E-09	NTPC	7.05E-10

LT	4.18E-10	ONGC	5.69E-10	M&M	1.13E-05
MARUTI	1.12E-08	HCL	6.03E-10	ASIANPAINTS	0.394599
TATASTEEL	1.97E-09	POWERGRID	1.09E-09	HDFC	6.74E-10

- The noticeable thing in these weights is that the weights associated with some assets turned out to be very less, some even in the order of 10^{-10}, which isn't realistic for the real world scenarios.
- The weights assigned to each asset in the portfolio are based on its expected return and its covariance with the other assets in the portfolio.
- As the model is giving weights of many of the assets as very small and only a few as big, it implies that the model has identified a subset of assets that have a relatively high expected return or low covariance with the other assets in the portfolio.
- This subset of assets is likely to have a relatively large weight in the portfolio, while the other assets that have a lower expected return or higher covariance with the other assets will have smaller weights or even be excluded from the portfolio altogether.
- The assumption of continuous weights means that the model can allocate fractional amounts of capital to each asset, rather than being restricted to integer values. This allows for more fine-grained allocation of capital and can lead to more optimal portfolios.

2) Relaxation

Thus, we incorporated the relaxation of the problem and the results of the optimization are as follows



KENDALL

DISTANCE

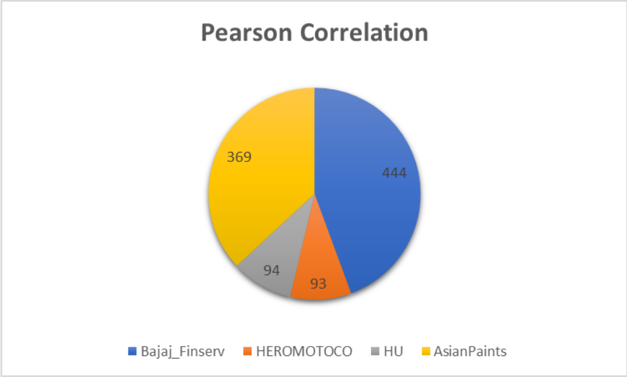
SPEARMAN

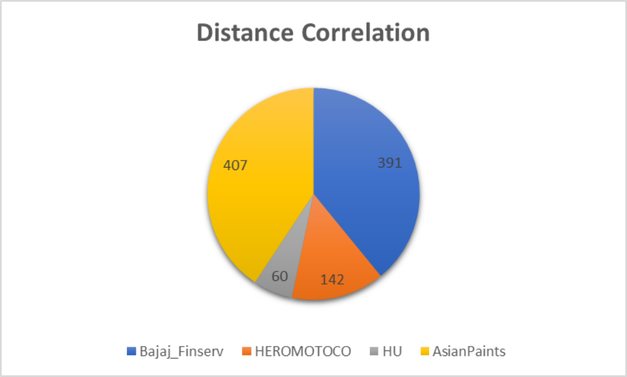
PEARSON

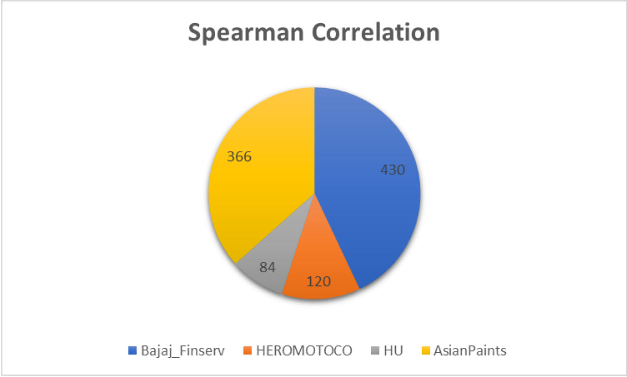
Optimized weights distribution for various correlation matrices*

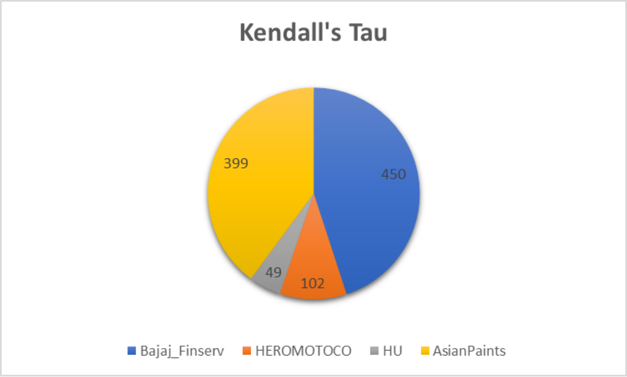
*Weights for only four stocks are visible since the weights for all the other companies are 0 due to the constraints employed in our model

3) Integer Weights



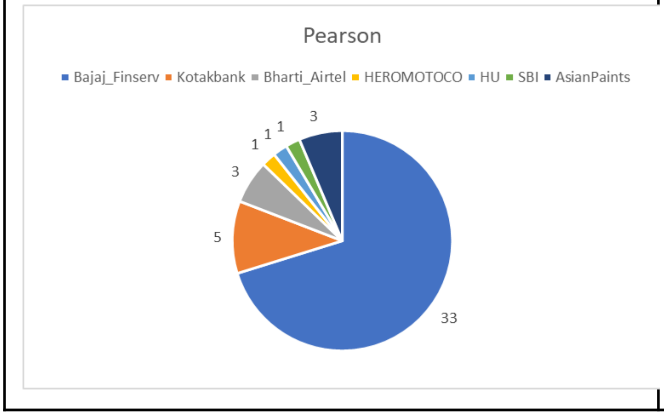


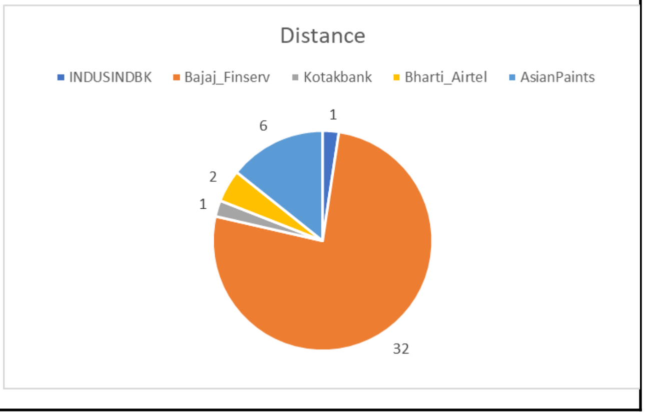


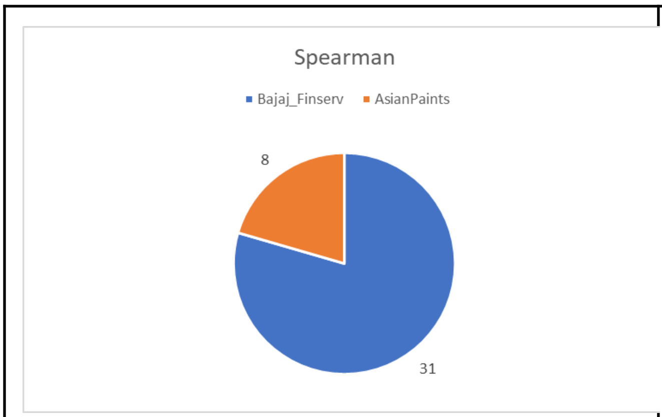


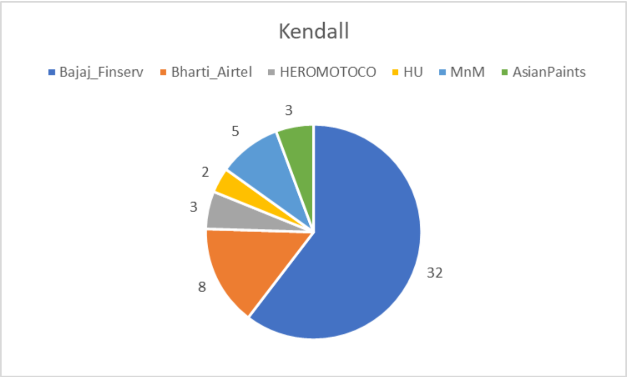
4) Integer Stocks

The best way to model the problem is when we can choose some integer quantities of stocks. It is incorporated as taking the latest prices (the last trading day in our data) for the asset prices and correspondingly number of different stock assets are shown below.



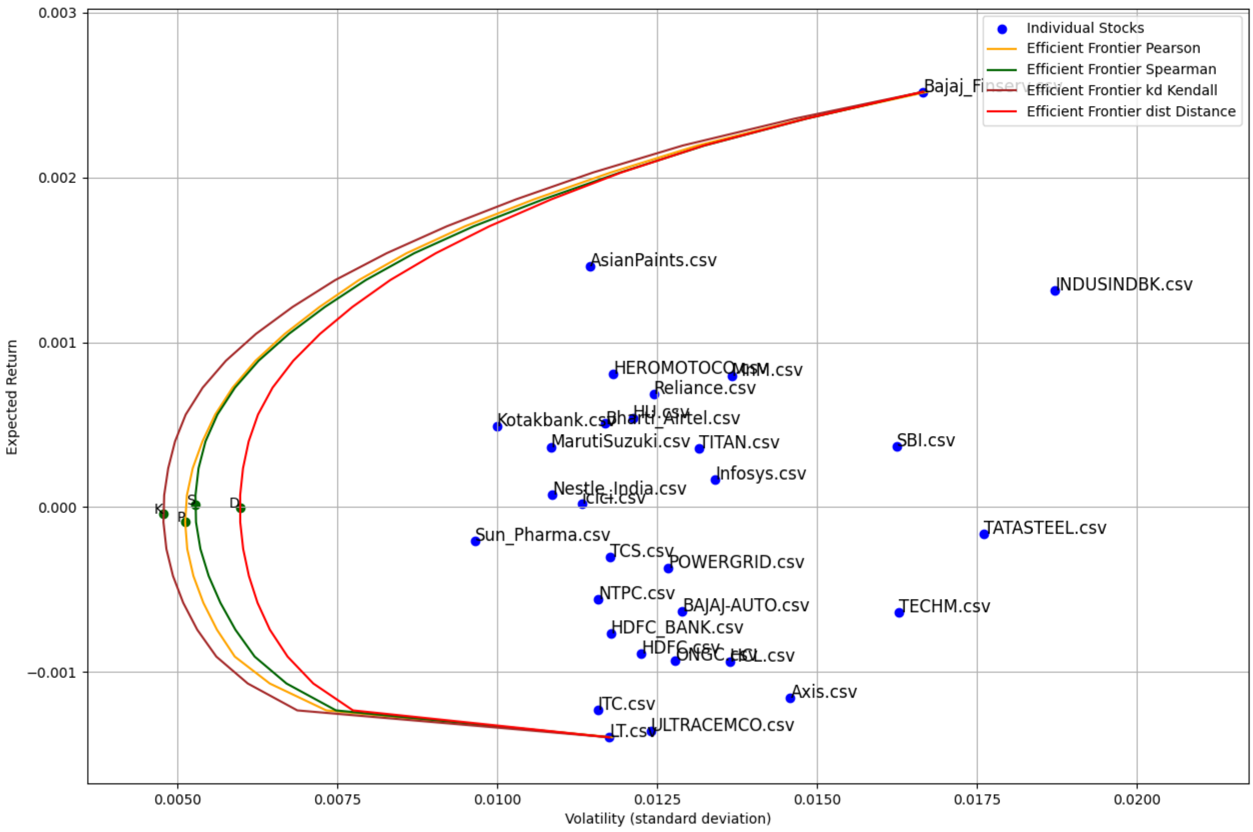






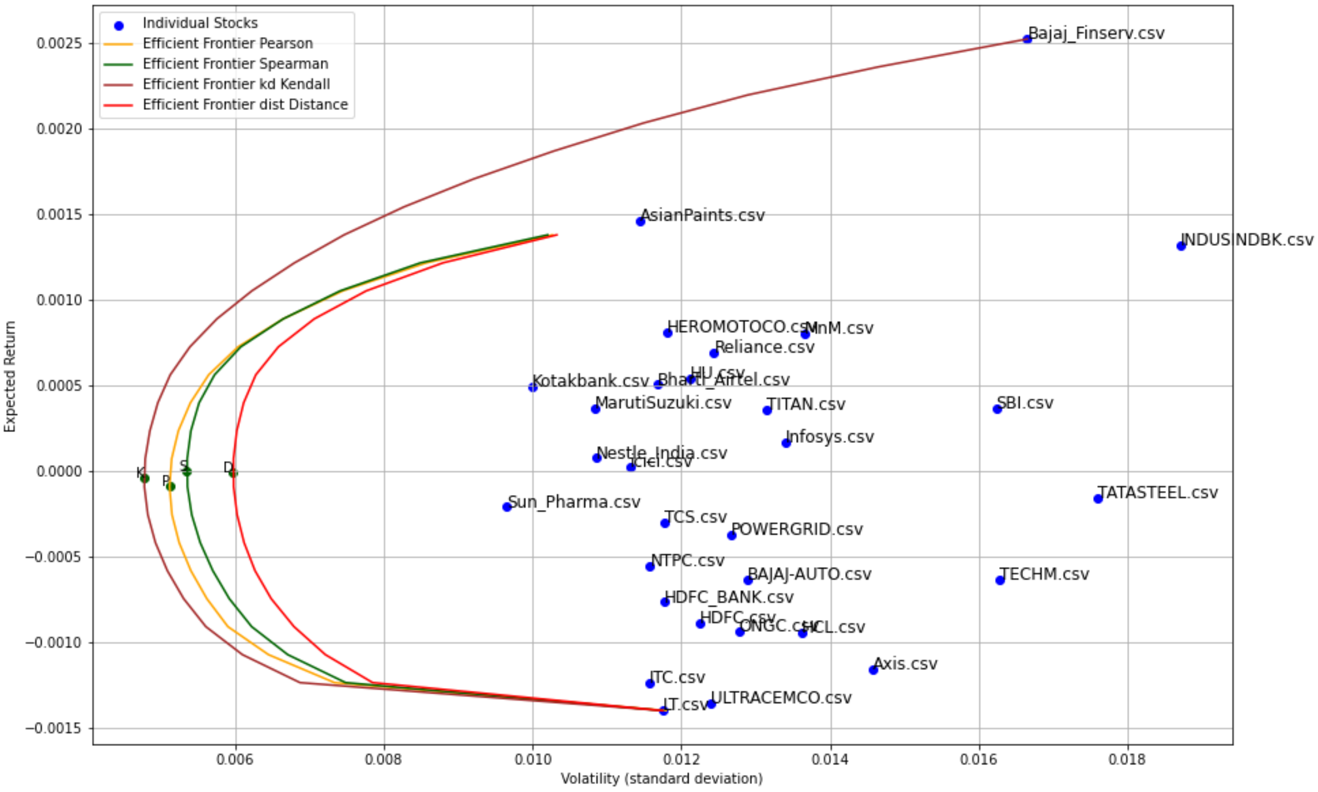
- ## Efficient Frontier

The Efficient Frontier for the four different ways of calculating correlation is as follows



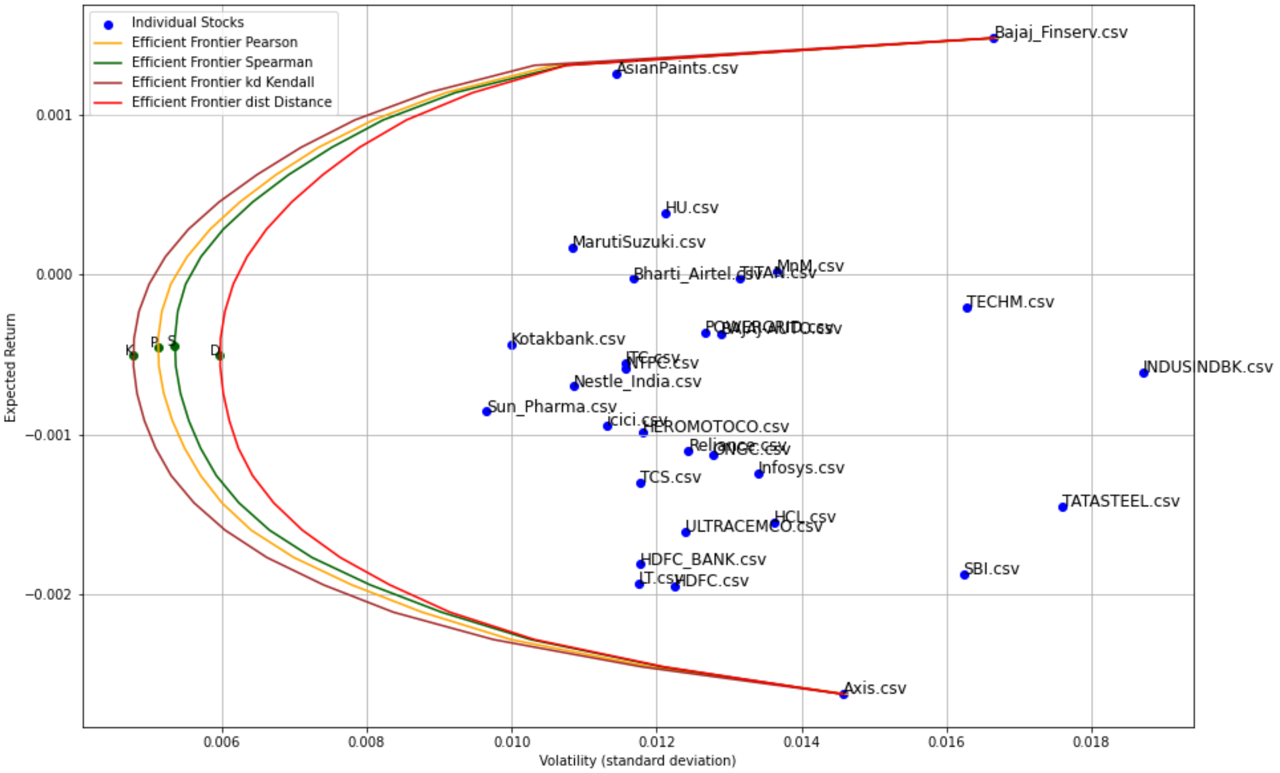
Efficient frontier figure -1 :

Mean-Variance (weights as continuous variables) with Expected return as mean value of 120 days



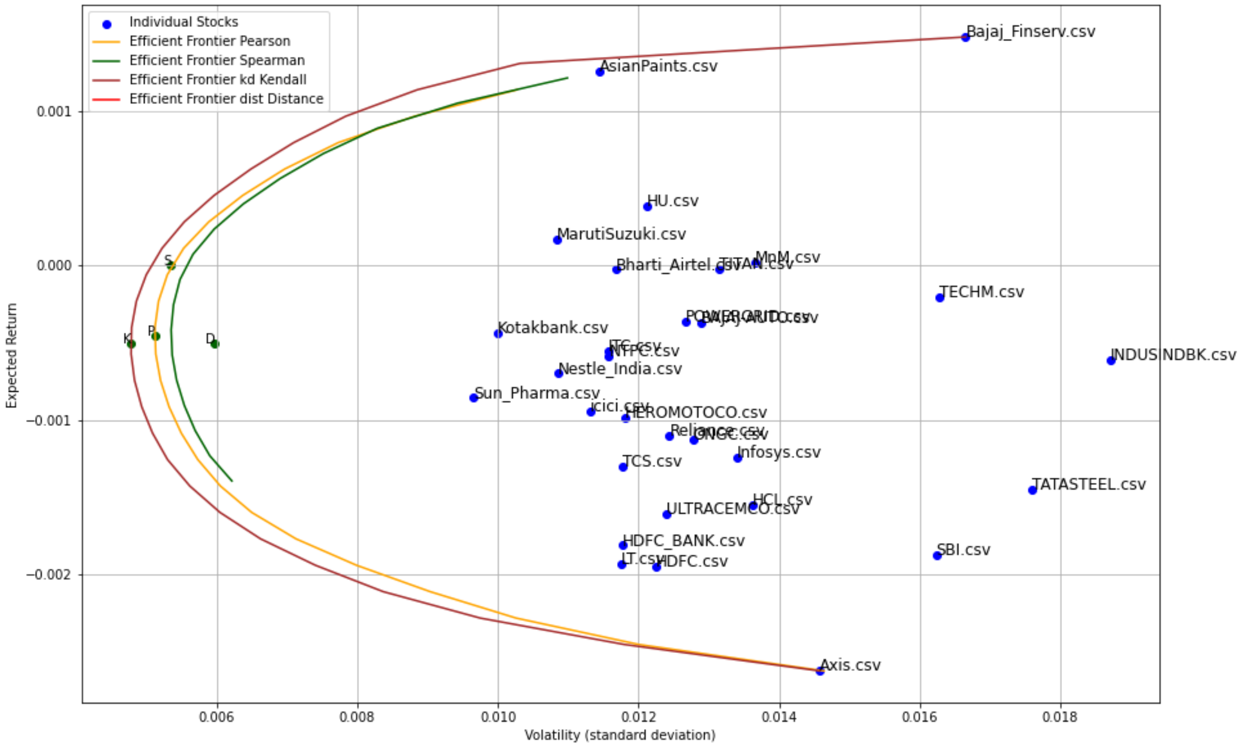
Efficient frontier figure -2 :

Mean-Variance (weights as continuous variables) with Expected return as exponential moving average (exponential smoothing)



Efficient frontier figure -3 :

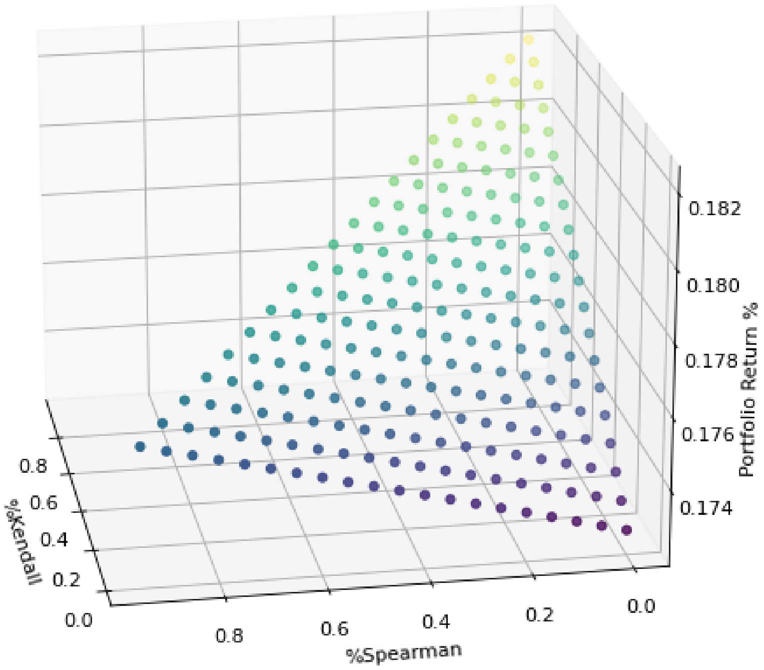
Relaxation (MIP) Model with Expected return as mean value of 120 days



Efficient frontier figure -4 :

Relaxation (MIP) Model with Expected return as exponential moving average (exponential smoothing)

Convex Combination of Covariance :-



Correlation = a_1 . Kendall + a_2. Spearman + a_3. Distance

s.t $a_1 + a_2 + a_3 = 1$

The Peak corresponds to pure combination of Kendall matrix, i.e. $a_1 = 1$, $a_2 = 0$, $a_3 = 0$ (**Highest returns of all**)

Conclusion

As we can see from the above plots, the Kendall correlation coefficient is giving higher returns than other correlation matrices for a specific level of risk in all types of cases which tells us that Kendall may be a more appropriate measure of correlation for those specific assets in our portfolio than others.

As discussed above, the Kendall correlation coefficient is a measure of the similarity of the rankings of two variables, regardless of the actual values of those variables. It measures the extent to which the relative orderings of the variables are similar. This is in contrast to other correlation measures like Pearson or Spearman, which measure the degree of linear and monotonic association between the variables, respectively.
So, from this we conclude that if the assets in our portfolio have non-linear or non-monotonic relationships, then using Kendall as the measure of correlation may be more beneficial. It is also possible that there may be some unique features of our particular set of assets that can make Kendall a better choice for our specific case.

So, by this we get to the fact that Kendall is consistently giving higher returns for a given level of risk across all types of cases making it a very good method of finding correlation. However, from the literature review that we studied, we also learned that there isn't any better correlation matrix and that may be dependent upon the type of dataset and its features.

Feedback

Our initial problem was to perform portfolio optimization and include diversification and other factors to solve it, however, our Professor advised us to take on a challenging problem of estimating the correlation matrix between different assets instead of doing what everyone does in the field of finance, that is using the sample correlation. Thereafter, we met regularly with our TA guide and the Professor to help define the problem clearly and map out our path and remove any roadblocks to using libraries and also received constructive feedback to put in more effort. We were also advised by our Professor to use Gurobi Py instead of other solvers and consider our problem as a Mixed Integer Programming Problem.

While showcasing our poster, we received very positive feedback from the TA's whom we presented our project to. We were able to sufficiently answer the questions asked regarding the work done. Some of them also appreciated the problem chosen by correlation matrices.

Contributions

All of us collaborated together, organised weekly meetings among ourselves and with TA Guide, along with biweekly meetings with the professor. Eventually we all are having equal contributions in all the work done throughout the semester.

References

- https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2147880
- https://www.sciencedirect.com/science/article/pii/S0167715220302637
- https://www.researchgate.net/publication/277666521_Portfolio_Optimization_using_Rank_Correlation
- https://www.freecodecamp.org/news/how-machines-make-predictions-finding-correlations-in-complex-data-dfd9f0d87889/
- https://finance.yahoo.com/
- https://www.gurobi.com/
- https://realpython.com/numpy-scipy-pandas-correlation-python/#pearson-correlation-numpy-and-scipy-implementation
- Markowitz Model - What Is It, Assumptions, Diagram, Formula (wallstreetmojo.com)