Clustering to Create Diversified Portfolios

Abstract

Table of Contents

1. Introduction

The purpose of this report is to delineate diversification’s role in the construction of asset portfolios and managing wealth. Diversification can be qualitatively defined as a method of capital allocation such that the portfolio is not heavily impacted by any individual asset’s risk. A definition such as this would not face much criticism, however any quantitative definition of diversification would. Not only does risk have a different meaning to different investors, “heavily impacted”, is also subjective. This report presents 5 methods of portfolio construction, their links to diversification, and how such construction has affected portfolio performance over the period from January 3 2007 to October 24 2022, where the assets in consideration are the stocks in the S&P 500. The 5 methods in question are naive diversification, mean-variance optimisation, hierarchical clustering, k-means clustering and financial clustering. Further, we analyse selected diversification and performance metrics to compare the aforementioned methods side-by-side.

2. Portfolio Construction

2.1 Portfolio Assumptions

Rf rate is 4.17% US 10 year treasury yield rf = 1.0417^(1/252)-1
For naive - no recomposition 1/N at beginning and leave it untouched for whole period
Short selling is allowed

a. Naive
b. MPT/MVO

2.2 MPT/MVO

Methods

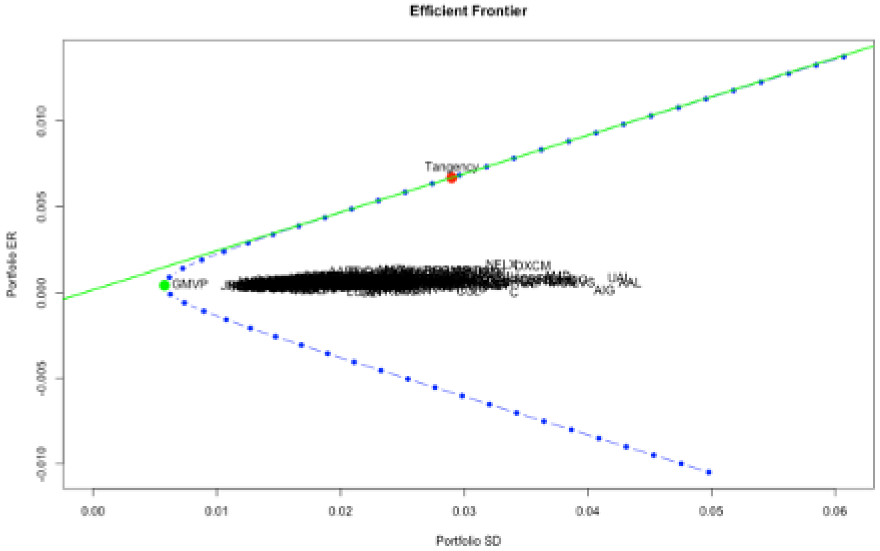
When constructing a portfolio in accordance with Modern portfolio theory (MPT), we aim to maximise expected return while minimising variance/standard deviation, which is the measure of risk under this framework. Firstly, the efficient frontier of risky stocks is defined as the set of portfolios such that no other portfolio of risky securities exists with a higher expected return but with the same standard deviation of return. To map the efficient frontier, we find the function of expected return that computes the minimum variance possible for such expected return (which turns out to be hyperbolic), then we take the top half of the hyperbolic function. To see how MPT utilises the idea of diversification, we just have to plot the individual risky assets, which all generally fit inside the efficient frontier; the efficient portfolios are usually comprised of many assets as this almost always reduces risk of the overall portfolio, without compromising expected return. When considering that investors also have the option of the risk free rate, we have the one fund theorem, which states that investors will only allocate funds in proportion to one risky portfolio and the rest in the risk free asset. This means that the efficient frontier is now the tangent line between the risk free asset and the efficient frontier of risk stocks.

Further, the assumption of the MPT that investors should maximise expected return while minimising variance must be addressed. This relies on assumptions of risk averseness and either quadratic utility or normality of returns. This may not be true of many investors.

Testing

We construct feasible portfolios in accordance with mean-variance optimisation on the past 4000 daily historical returns of S&P 500 stocks. We also use the current 10 year US treasury bond yield.

Results



2.3 Hierarchical Clustering

Portfolio diversification and asset allocation based on hierarchical clustering was performed using the method devised in Raffinot (2018). Hierarchical clustering is recursive and results in a tree-based representation of the clusters as a dendrogram. Ward's method is used as a dissimilarity measure between clusters, being one of the more popular measures. The Gap index is used to identify the optimal number of clusters. After clusters are formed, a portfolio is created by distributing capital equally to each cluster hierarchy, and giving an equal weight to each individual stock within a cluster. In principle, this should result in a more diversified portfolio than naïve allocation as correlated risks have a reduced allocation.

c. K-Means

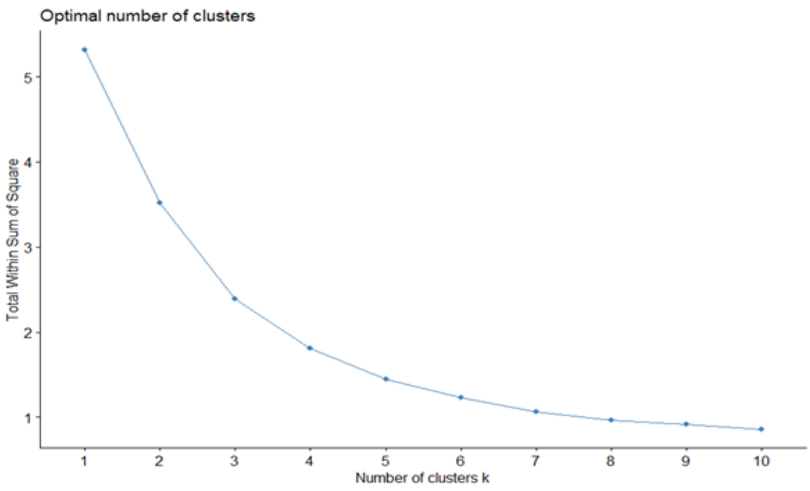
2.4 K-means Clustering Portfolio

Definition:

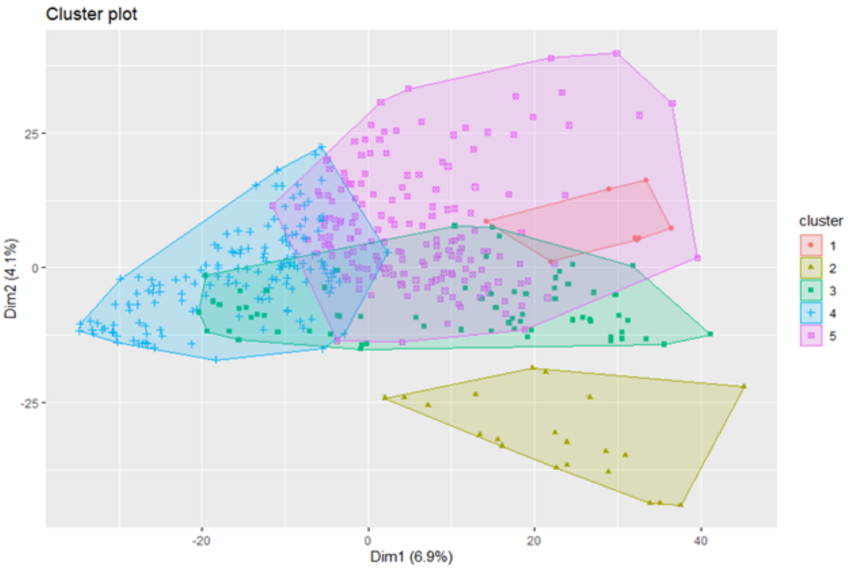
K-means clustering aims to sort “n” observations into k clusters with unique characteristics, with each observation belonging to the cluster whose centroids (mean) is the closest to its own. K-means clustering reduces inter-cluster variation. In this task, we aim to split all 424 stocks from S&P500 into clusters with similar characteristics.

Procedure:

To start, we should find the optimal number of clusters to sort our assets into, without making it excessive. This is done by the “Elbow method”.



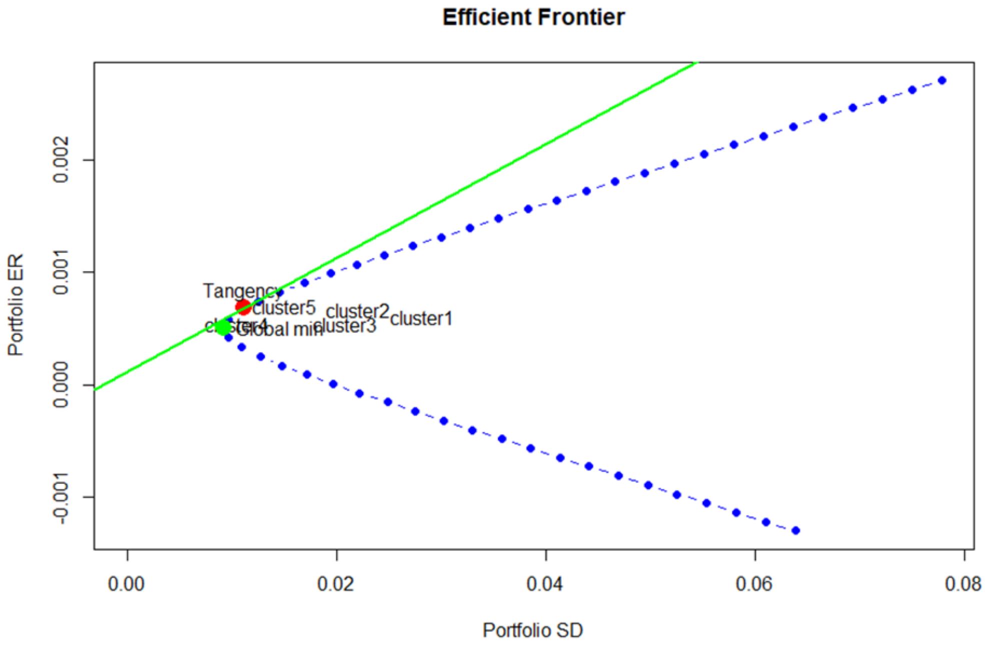
Elbow method essentially tells us the optimal number of clusters we should form to minimise the sum of square yet does not require very heavy computational power. The sum of square reduces exponentially as more clusters are formed. The graph generated from Elbow method will relatively resembles an elbow (hence the name). The process of choosing is quite intuitive, as we only need to choose the point where the number of clusters reduces the sum of squared substantially and adding more clusters won’t have too much effect on the reduction. From the graph generated by R, 5 clusters seem to be an adequate & reasonable choice.



The appropriate data will then be input into the function “fviz_nbclust” from the package “factoextra” which will then generate our 5 clusters, with each cluster containing stocks of similar characteristics. Below is a graph of the 5 clusters generated by R.

Designing a portfolio

From here, after creating our 5 clusters containing all 424 stocks, we treat each cluster as one individual “asset” and design a Tangency portfolio (red dot in the diagram) with our 5 new “assets”. (Inspired from Zhi Wei, who also treated each cluster as individual asset). This is supposedly a portfolio with Mean Variance Optimization, where returns are maximised for a certain level of risk.



2.5 Financial Clustering

For the 424 stocks in our data set, financial data was collected, using the ‘BatchGetStocks’ package, including Price to Book Ratio, Earnings Per Share, Last Price, Market Capitalization, and Price to Earnings Ratio. This is inspired by Marvin (2015), who utilises asset turnover ratio and return on assets ratios. For each financial metric, negative data values were omitted, as for certain ratios, negative data values may impact the accuracy of the analysis. For example Earnings Per Share does not make sense when earnings are negative.

Price to Book Ratio was selected as it shows which stocks are overvalued or undervalued relative to book value. Earnings Per Share and Last Price are control metrics as they inherently should not indicate financial performance, as market capitalization dictates the value of a company not the arbitrary stock price itself. Market capitalization is chosen to separate smaller companies from larger companies. Price to Earnings Ratio calculates the share price relative to the amount the company earns, showing which companies may be overvalued or undervalued.

For each financial ratio, the stocks (which have valid ratios, i.e. positive EPS) were split into 8 clusters. This is done by taking the stocks in each of the 8 clusters, and downloading price data using ‘tq_get’ from the tidyquant package. For each of the clusters the stocks are divided by the initial share price and then divided by the number of stocks in the cluster. They are then summed into one stock, representing a naive portfolio of stocks in the cluster. Returns were then calculated using ‘Return.calculate’ from the PerformanceAnalytics package. Thus 8 clusters were each naively diversified. These 8 clusters are then used to create a portfolio using the tangency portfolios.

3. Data Set

The S&P 500 stocks were chosen as the dataset because they have consistent publicly available price data over a sustained period of time. This accounts for changes in the economic environment over time, so our diversification model is most suited to respond to economic shocks. To keep consistency, we only used the constituent stocks that had pricing data within the entire time period.

Financial data of the S&P 500 stocks was sourced from Yahoo! Finance, from the time period January 3 2007 to October 24 2022. The ‘quantmod’ package was used to retrieve stock price history from a list of stock codes of S&P 500 stocks. After filtering stocks which had incomplete data within the time period, data from the remaining 424 stocks was obtained and daily returns were calculated using the ‘PerformanceAnalytics’ package.

4. Metrics

4.1 Sharpe Ratio

Measures risk adjusted relative returns

4.2 Adjusted Sharpe Ratio

Assuming a turnover cost of 0.5% we can use the formula to calculate the expected net returns. This incorporates turnover costs which is necessary for rebalancing portfolios into the Sharpe Ratio.

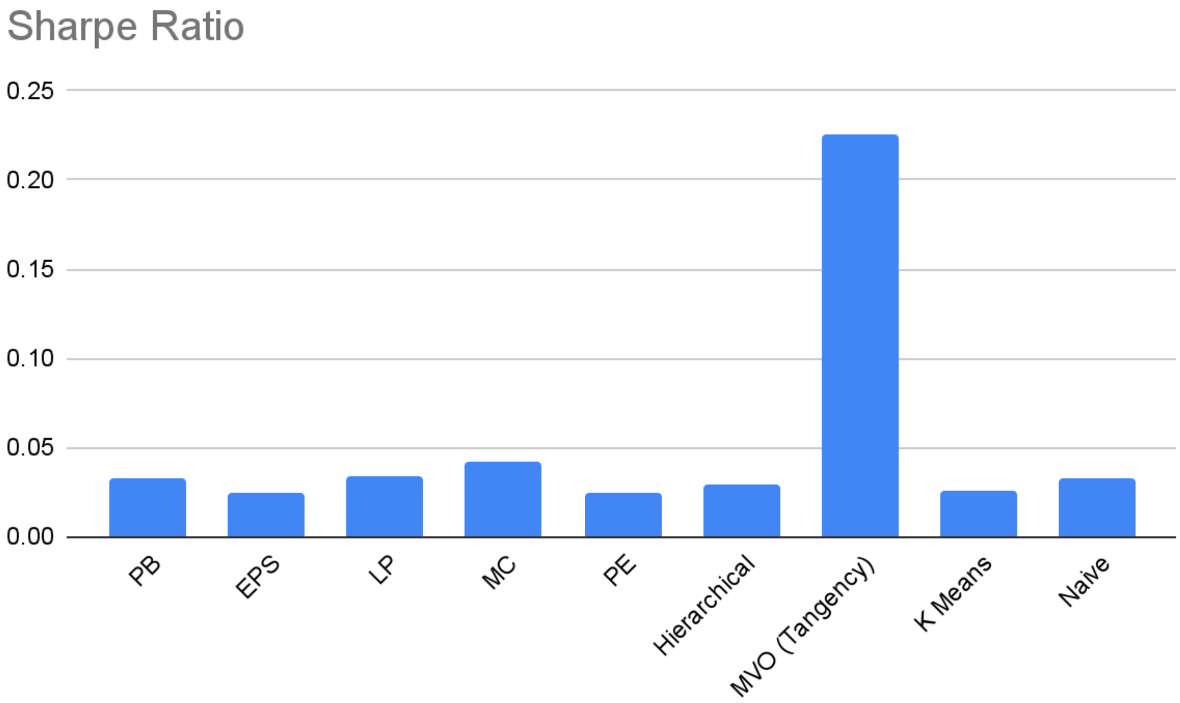
4.3 Diversification Ratio

The ratio betweens the weighted sum of volatility of individual assets over the overall portfolio volatility

4.4 Portfolio Volatility

Measures the extent of inconsistencies within returns's fluctuations

5. Results



6. Analysis

Preliminary results show that the mean variance optimisation portfolio has by far the highest rate of return and Sharpe ratio, while maintaining one of the higher diversification ratios. However, these results seem to be an extreme outlier, as it has a 5 times higher Sharpe Ratio than any other portfolio, which also results in a large negative Adjusted Sharpe Ratio, as that formula does not account for the abnormally large Sharpe Ratio. Further checking on methodology used to attain these results will have to be conducted.

7. Conclusion

No conclusion currently, unverified results.

8. References

https://www.cs.princeton.edu/sites/default/files/uploads/karina_marvin.pdf

https://arxiv.org/pdf/2005.03204.pdf

9. Appendices

Portfolio	Risk-Free Rate: 4.17% p/a		Returns Metrics				
			Sharpe Ratio	Annual Sharpe	Adjusted Sharpe Ratio	Sortino	Mean
	Financials	PB	0.03256495	1.373922	0.9683791	0.1144041	0.002162317
		EPS	0.02420593	0.7733817	0.7064031	0.08496385	0.002139666
		LP	0.03424295	1.501891	0.9980185	0.1204537	0.002748506
		MC	0.04209938	1.030127	0.7191023	0.0938652	0.001541393
		PE	0.02470218	0.924359	0.8034045	0.0830442	0.001075554
	Hierarchical		0.0289671	0.3648671	0.3355281	0.03504029	0.0005577
	MVO (Tangency)		0.2250398	7.91444	-16.49043	0.3805198	0.006686783
	K Means		0.02597789	0.763842	0.6171729	0.06788515	0.0006803588
	Naive		0.03257636	0.4253633	0.3823037	0.03961847	0.0006142174

Portfolio	Risk-Free Rate: 4.17% p/a	Risk/Diversification Metrics						
		Skewness	Standard Deviation	Kurtosis	Diversification Ratio	Downside Deviation (under mean)	Sum of Squared Portfolio Weights	Semi-Deviation

	Financials	**PB**	-0.1957542	0.02448729	3.182942	0.3627567	0.01756432	28.40742	0.01756432
		EPS	-0.07854093	0.03418451	3.06887	0.5847583	0.0243359	73.05172	0.0243359
		LP	-0.2376796	0.0298946	2.936562	0.3083256	0.02157957	39.91482	0.02157957
		MC	0.2400467	0.0208158	7.760739	0.438365	0.01476108	24.63515	0.01476108
		PE	-0.1200167	0.01517665	3.156106	0.5271472	0.01095221	4.911055	0.01095221
	Hierarchical		-0.1614675	0.01365517	12.72598	1.521951	0.009853796	0.004749616	0.009853796
	MVO (Tangency)		0.1331198	0.02898405	4.248761	0.8158495	0.02023473	11.37927	0.02023473
	K Means		0.05275863	0.01120523	8.174681	0.6149734	0.007972719	1.738307	0.008049028
	Naive		-0.2403732	0.01387771	11.16728	1.564676	0.01008512	0.002358491	0.01008512

Metric	Formula	Explanation
Sharpe Ratio	$S_a = \frac{E[R_a - R_b]}{\sigma_a}$	Measures risk adjusted relative returns.
Sharpe Ratio Adjusted for Turnover Cost	$\mathbf{r}_{t+1} = \left(1 - \kappa \sum_{i=1}^{N} \lvert w_{i,t} - w_{i,(t-1)*} \rvert \right) (\mathbf{w}_t)' \mathbf{r}_{\mathbf{t+1}},$	Assuming a turnover cost of 0.5% we can use the formula to calculate the expected net returns. This incorporates turnover costs which is necessary for rebalancing portfolios into the Sharpe Ratio.
Diversification Ratio	$DR = \frac{\sum w_i \sigma_i}{\sigma_p}$	Diversification Ratio: The ratio betweens the weighted sum of volatility of individual assets over the overall portfolio volatility
Portfolio Volatility	$\sigma_p = \sqrt{\sum_{i=1}^{N} w_i^2 \sigma_i^2 + \sum_{i=1}^{N} \sum_{j=1}^{N} w_i w_j \sigma_i \sigma_j \rho_{ij}}$	Measures the extent of inconsistencies within returns's fluctuations
Adjusted Sharpe Ratio (Pezier and White [2008])	$\mathcal{ASR} = \mathcal{SR}[1 + (\frac{\mu_3}{6})\mathcal{SR} - \frac{(\mu_4 - 3)}{24})\mathcal{SR}^2]$	Based on the Sharpe Ratio but adjusts for skewness and kurtosis by including a penalty factor

Sum of Squared Portfolio Weights	$\mathcal{SSPW} = \frac{1}{F}\sum_{t=2}^{F}\sum_{i=1}^{N} w_{i,t}^{2}$	Measures underlying level of diversification in a portfolio from 0 to 1
Diversification Ratio	$DR = \frac{\sum w_i \sigma_i}{\sigma_p}$	Diversification Ratio: The ratio betweens the weighted sum of volatility of individual assets over the overall portfolio volatility

Add sortino, downside deviation, semi deviation