École Polytechnique Fédérale de Lausanne

Cluster and RTM analysis for portfolio optimization

Financial Big Data

Benedek Harsányi

January 4, 2023

1 Introduction

Portfolio optimization is a technique used by investors to maximize the return on their investment portfolio while minimizing the risk. It involves selecting a mix of investments that are expected to provide the best combination of risk and return. This is typically achieved by diversifying the portfolio across a range of assets, such as stocks, bonds, and commodities, and by carefully considering the trade-off between risk and return for each individual investment. By using portfolio optimization, investors can create a balanced portfolio that is tailored to their specific investment goals and risk tolerance. In this report we investigate different techniques on how to solve the optimization problem, putting emphasis on the statistical uncertainty of the correlation matrix and its cleaning procedures. The report is organized as the following. In section 2 we introduce the mathematical formulation of the portfolio optimization problem, along with the summary statistics, we need in order to evaluate our results. In section 3 we explain different covariance cleaning methods. Section 4 and 5 present the numerical results and the research questions we were trying to answer. The conclusion is drawn in section 6.

2 Problem Formulation

2.1 Portfolio optimization

In this section, we introduce the Markowitz portfolio optimization problem [1].Given N risky assets, with mean returns $\mathbb{E}[r_i] = \mu_i$ we would like to find the optimal portfolio weights w_i, such that, the portfolio volatility $\sigma^2 = \sum_{i=1}^{N}\sum_{j=1}^{N} w_i w_j \sigma_{ij}$ is minimized. Here we denote the covariance of asset i and j with σ_{ij}. The problem can be written with matrix notation as

$$\begin{aligned} \min_{w} \quad & w^T \Sigma w, \\ \text{s.t.} \quad & w^T \mathbb{1} = 1. \end{aligned}$$

where Σ is the covariance matrix, and the last constraint ensures that the portfolio weights are summed up to one, meaning it represents the fraction of our wealth to be invested in each asset. Using Lagrangian-multipliers, the minimum variance is achieved by setting

$$w^* = \frac{\Sigma^{-1}\mathbb{1}}{\mathbb{1}^T \Sigma^{-1} \mathbb{1}}.$$

2.2 Metrics

Given a dataset consisting of stock returns, we can pick a time window $2T$, which we can devide into the in-sample period consisting of T, days, and the out-sample period. The in-sample (or train set) is used to calibrate the model, and the out-sample (or test set) is used to check the performance. Using the in-sample data, we can calculate the sample mean and the sample covariance between any two assets

$$\mu_i = \bar{r}_i = \frac{1}{T}\sum_{t=1}^{T} r_i,$$

$$\hat{\Sigma}_{ij} = \hat{\sigma}_{ij} = \frac{1}{T}\sum_{t=i}^{T} (r_{i,t} - \bar{r}_i)(r_{j,t} - \bar{r}_j).$$

We will regard the sample covariance matrix as the ground truth. For a given portfolio weights, we can calculate the realized volatility of the portfolio by $\hat{\sigma} = w^T \hat{\Sigma} w$. We will investigate different prediction methods other than

the sample covariance for the covariance matrix computation. These estimates will be denoted by Σ, the estimated volatility for a portfolio weight is $\sigma = w^T \Sigma w$. For different predictions we can calculate the reliability, by the formula

$$R = \frac{|\hat{\sigma} - \sigma|}{\sigma}.$$

The smaller R is, the more reliable the prediction on some out-sample test interval. In every prediction scenario, we will use Markovitz's minimal variance portfolio weights w^*. Another question we will investigate is the composition of the portfolio. The metric, proposed in [2], captures the number of effective stocks in the portfolio $N_{eff} = \frac{1}{\sum_{i=1}^{N} w_i^2}$. Indeed if we only invest in one asset, the effective value is one, on the other hand in case of the uniform portfolio, this quantity will be equal to n. Using the out-sample data, we can calculate the out-sample-risk, which is $\sigma_{out} = w^T \hat{\Sigma}_{out} w$, where $\hat{\Sigma}_{out}$ is the empirical covariance matrix of the out-sample data.

3 Methods

3.1 Random Matrix Theory

A cleaning procedure, based on Random Matrix Theory (RTM) [3] has been proposed to clean the empirical covariance matrix. We assume N independent assets with Gaussian returns, with zero mean, σ^2 variance, through a time window of T days, and assume that the ratio $q = T/N \geq 1$ (although, we will investigate scenarios, when this assumption does not hold). The covariance matrix is symmetric, thus the eigenvalues are real and the eigenvectors can be chosen to form an orthonormal basis, $\Sigma = Q \Lambda Q^T$, where $\Lambda = \mathrm{diag}(\lambda_i)$ contains the eigenvalues. The spectral density of the covariance matrix can be given by

$$\rho(\lambda) = \frac{Q}{2\pi\sigma^2\lambda} \sqrt{(\lambda_{\max} - \lambda)(\lambda - \lambda_{\min})}$$

where the Marcenko-Pasture edge is given by $\lambda_{\min}^{\max} = (1 \pm \sqrt{Q})^2$. The spectral density is nonzero on this interval. This gives us a cleaning procedure, which involves clipping the eigenvectors below the Marcenko-Pastur edge, while keeping all eigenvalues above the threshold. clipped eigenvalues should be considered as corrupted with noise, they cannot be distinguished from the spectrum of the correlation of a random matrix.

3.2 Hierarchical clustering

Correlation-based clustering approach has been proposed to clean the correlation matrix [2] [4]. This method reduces the degree of freedom of the correlation matrix. A correlation between two time series can be considered a similarity measure. One can define the distance matrix between stock i and stock j as $d_{ij} = 1 - \rho_{ij}$, where ρ_{ij} is the Pearson correlation coefficient. Based on the distance matrix, we can perform an agglomerative clustering using the average linkage between two clusters $\mathcal{C}_p$ and $\mathcal{C}_q$, defined as

$$\alpha_{pq} = \frac{\sum_{i \in \mathcal{C}_p} \sum_{j \in \mathcal{C}_q} d_{ij}}{|\mathcal{C}_p||\mathcal{C}_q|}.$$

In an iteration of the algorithm, we calculate the average linkage between all clusters, and we merge the two clusters with the smallest linkage. The algorithm terminates, after $N-1$ merge operations and it outputs a stream $(\mathcal{C}_p^n, \mathcal{C}_q^n, \alpha_{pq}^n)_{n=1}^{N-1}$, called a dendrogram, where in the nth step, $\mathcal{C}_p^n$ and $\mathcal{C}_q^n$ clusters are merged. We obtain the filtered matrix by averaging out for all (i,j) pairs $i \in \mathcal{C}_p^n$ and $j \in \mathcal{C}_p^n$, the value of the original matrix

$$c_{ij}^{<} = 1 - \alpha_{pq}^n.$$

Finally, the diagonal elements are filled with ones. This procedure is called Hierarchical Average Linkage Clustering (HALC), the resulting matrix will have $N-1$ distinct element instead of the original $N(N-1)/2$.

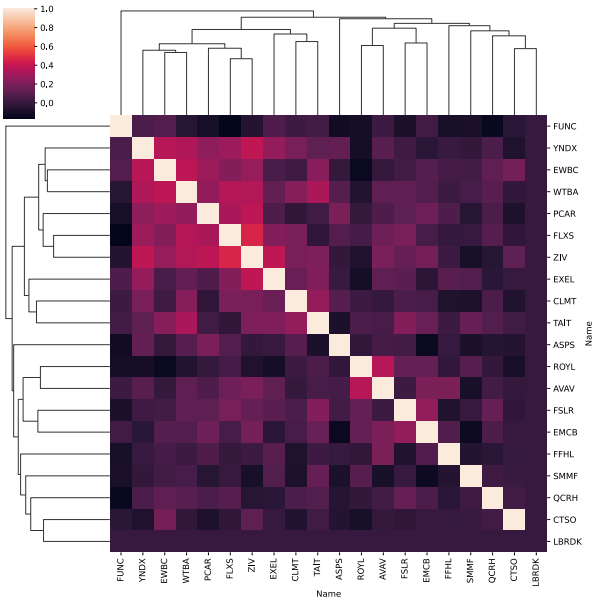

Figure 1: Clustering of 20 stocks based on HALC

4 Data set

For testing the methods, we will use a dataset, consisting of the daily adjusted closing price of 1729 stocks listed on NASDAQ, between 2010 January 1 and 2020 December 31.

5 Numerical experiments

Experiments were run on a computer with an Intel(R) Core(TM) i7-6500U CPU @ 2.50GHz processor, using 8GB RAM, on Ubuntu 20.04.3 LTS operating system. Python 3.10.4 general-purpose programming language was used along with various packages, including pandas, numpy and matplotlib. The hierarchical clustering implementation is based on the package [5], and RTM-related implementations are based on [6]. The source code is available on Github, further instructions on how to reproduce the results can be found in the README.md file. We investigated the following three questions.

5.1 How does the out-of-sample risk and N_{eff} evolve for different covariance estimates?

A rolling time window was picked $2T = 252$, meaning we used T for in-sample and out-sample calculations. For each window, we calculated the optimal Markowitz portfolio weights and report the effective number of stocks in the portfolio, along with the out-of-sample risk and the reliability of the risk.

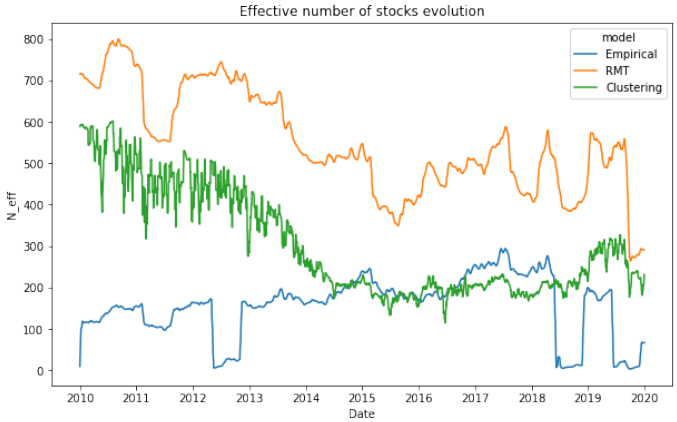

Figure 2: N_{eff} smoothed between 2010 and 2020

Each time series consist of 2520 entries. On average the RTM calculation takes 17.12s (total runtime was around 13 hours, because of the IOs and additional calculations performed), the clustering approach takes 0.99s (total runtime was about 4 hours), and the empirical calculation takes 1.16 s (total runtime was about 3 hours). In order to make the plot more readable, we applied a moving average filter with a window length of 10 on each time series. Fig 2. shows the results of the effective number of stocks evolution. It can be seen, that the RTM method, made use of the large pool of possible stocks over the 10 year period. The empirical calculation yields the lowest amount of the effective number of stocks, however between 2015 and 2018, it is on par with the clustering method.

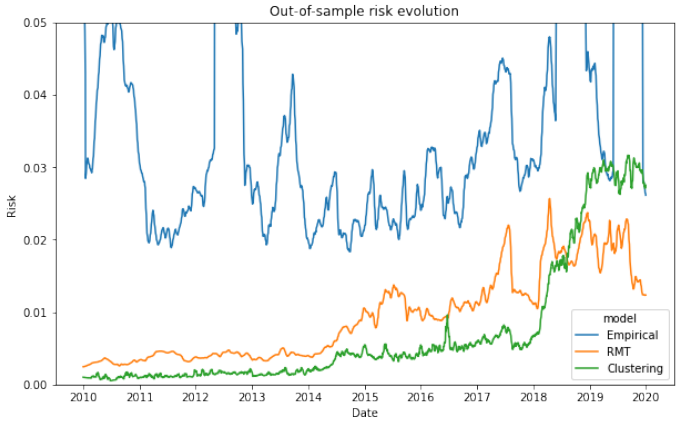

Figure 3: σ_{out} smoothed between 2010 and 2020

Fig 3, shows the results of the out-of-sample evolution. Again, a moving average smoothing was applied to each time series. Both RTM and the clustering method clearly outperform the empirical calculation. Between 2010 and 2018, the clustering method seems to be superior, however between 2018 and 2020, RTM produces better results.

5.2 How does the reliability change in RTM for different cut-off values?

In the RTM method, the cut-off value can be considered as a hyperparameter, we can vary what percentage of eigenvalues we keep. For 9 different values in $[0, 1]$ along with the Marcenko-Pasture edge. A rolling window of $T = 200$ was chosen along with $N = 300$ sampled stocks. In each time window, we estimate the covariance matrix, calculate the Markovitz weights and report the reliability of the estimator. We report the distribution of the different methods in a boxplot, along with a line plot comparing the MP method with the best-performing 0.3 cutoff value.

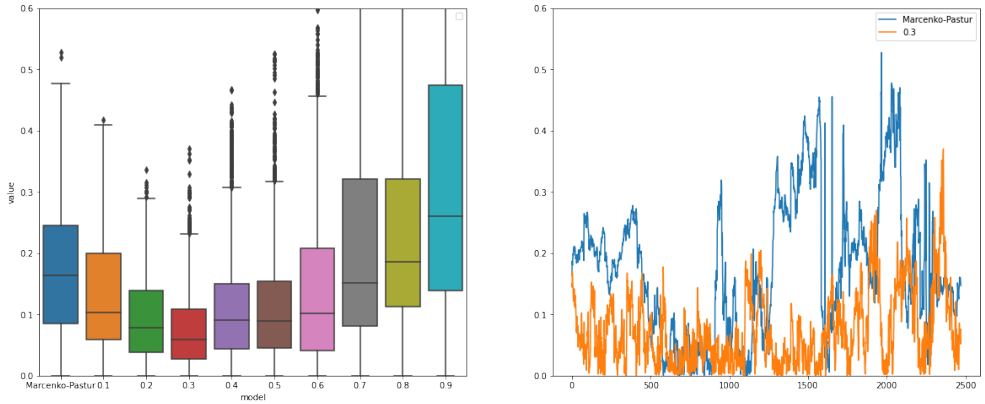

Figure 4: Boxplot of the distributions of the reliabilities and time evolution comparison of the Marcenko-Pasture and 0.3 cutoff model

5.3 How does the reliability of the models changes in the function of T and N?

We perform a similar analysis, presented in [2]. The reliability R, of a model measures, how close the risk prediction is to the empirical risk, smaller R means a better estimate. For a fix time window, and a number of assets, we can

calculate the optimal weights, the predicted and the true risk of the portfolio and compare them. In each scenario, we also estimate the covariance matrix using RTM and hierarchical clustering. We vary $T \in [0, 1000]$ and $N \in [0, 500]$ for 10 evenly spaced numbers over these intervals. For each configuration, we perform 50 different portfolio optimization (bootstrap experiments) for different t_0 starting time, and count how many times the clustering approach outperformed the RTM approach, meaning $R_{RTM} > R_{cluster}$, we report the percentage of successes of the clustering method over the RTM method.

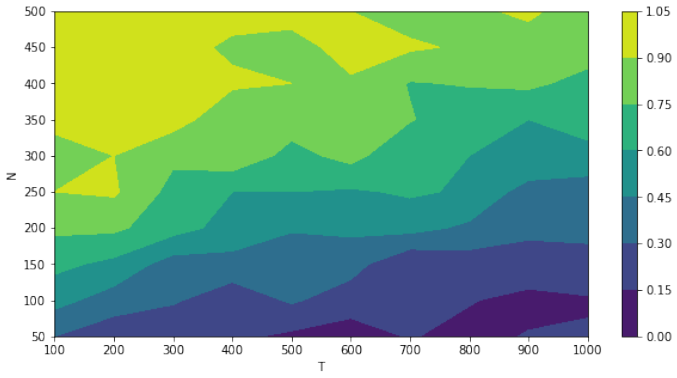

Figure 5: Contour plot of the percentage of success of HALC over RTM for different (T, N) pairs

We can observe from Fig 5, that the RTM method works better if the T/N ratio is higher. On the other hand, if we include more stocks on the optimization and choose a smaller time horizon, the clustering method is more reliable. This result coincides with the results of Section 5.1, where we observed that for 1729 stocks with a time window of 126, the clustering approach produces lower out-of-sample risk for most of the time.

6 Discussion

In this report we investigated different filtering procedures for the covariance matrix in a minimum variance portfolio optimization setting. We performed three experiments. In a rolling window fashion, we reported the out-of-sample risk and the effective number of stocks evolutions for the different methods, we investigated the different cut off values for the RTM method, and finally we tried to answer the question, what is the role of the parameters T and N in the optimization procedure. From the above analysis, we can conclude that the best cleaning procedure, depends on the choice of other hyperparameters in the optimization. Higher N (number of stocks), will make the covariance matrix bigger, therefore introducing more noise to the problem, resulting in poor performance on the RTM method. On the other hand, the larger the lookback window, the more reliable the RTM results are. Furthermore, in case of RTM, the Marcenko-Pasture edge is not necessarily the best cutoff choice, as far as we compare the portfolios based on their reliabilities.

References

[1] Harry M Markowitz. Foundations of portfolio theory. *The journal of finance*, 46(2):469–477, 1991.

[2] Vincenzo Tola, Fabrizio Lillo, Mauro Gallegati, and Rosario N Mantegna. Cluster analysis for portfolio optimization. *Journal of Economic Dynamics and Control*, 32(1):235–258, 2008.

[3] Jean-Philippe Bouchaud and Marc Potters. Financial applications of random matrix theory: a short review. *arXiv preprint arXiv:0910.1205*, 2009.

[4] Christian Bongiorno and Damien Challet. Covariance matrix filtering with bootstrapped hierarchies. *PloS one*, 16(1):e0245092, 2021.

[5] Christian Bongiorno. bootstrapped average linkage clustering. `https://pypi.org/project/bahc/`.

[6] G. Giecold and L. Ouaknin. pyrmt. `https://github.com/GGiecold/pyRMT`, 2017.