École Polytechnique Fédérale de Lausanne

Cluster and RMT analysis for portfolio optimization

Financial Big Data

Benedek Harsányi

January 23, 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. Financial big data often contains a large amount of noise and standard estimators tend to overfit the noisy data, resulting in unreliable portfolio weights. On the other hand, covariance filtering enables us to build robust portfolios, outperforming the standard methods on out-of-sample data. 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 divide 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. By doing so, the returns are stored in the data matrix $R \in \mathbb{R}^{N \times T}$. For numerical stability we will normalize the data such that, the covariance matrix coincides with the Pearson correlation matrix, for each asset k, we divide the returns by its standard deviation, making the diagonal entries in the sample covariance matrix $\sigma_{kk} = 1$. 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 weight, 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 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 the 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

A major concern in practice is that the covariance between two stock returns is rarely computable precisely since the true distribution of the vector is not known, we only observe a finite sample. One possibility is to use the empirical covariance matrix since in the case of a large enough population, it converges (almost surely) to the true covariance. However, in practice, only a finite amount of data is available, so a better choice is to use different filtering or cleaning procedures to estimate the true covariance. In this section, we introduce different filtering procedures and compare their performances on different metrics. The procedures can be broadly categorized into two groups. The shrinkage methods start from the empirical covariance matrix, and apply some transformations, to prevent the matrix from overfitting to noise. Filtering procedures rely on the so-called Random Matrix Theory (RMT). These methods use the properties of the underlying spectrum of the matrix.

3.1 Linear shrinkage

The oldest cleaning procedure is the so-called linear shrinkage covariance matrix estimator, which is a linear combination of the sample covariance and the identity matrix:

$$\Sigma_{\text{LS}} = \alpha\hat{\Sigma} + (1 - \alpha)I,$$

this method can be seen as a heuristic way to control the diversification of the portfolio. $\alpha = 1$, corresponds with the no shrinkage case, while $\alpha = 0$, does not take into account the data itself and treats the stocks uncorrelated, resulting in extreme shrinkage.

3.2 Random Matrix Theory

A cleaning procedure, based on Random Matrix Theory (RMT) [3] has been proposed to clean the empirical covariance matrix. The theory suggests, that only the largest eigenvalues of the covariance matrix give information about the cross-correlations. We assume N independent assets with Gaussian returns, with zero-mean, unit 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 $\Lambda = \text{diag}(\lambda_i)$ contains the eigenvalues, and Q is the square matrix whose ith column is the eigenvector u_i. The spectral density of the covariance matrix can be given by

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

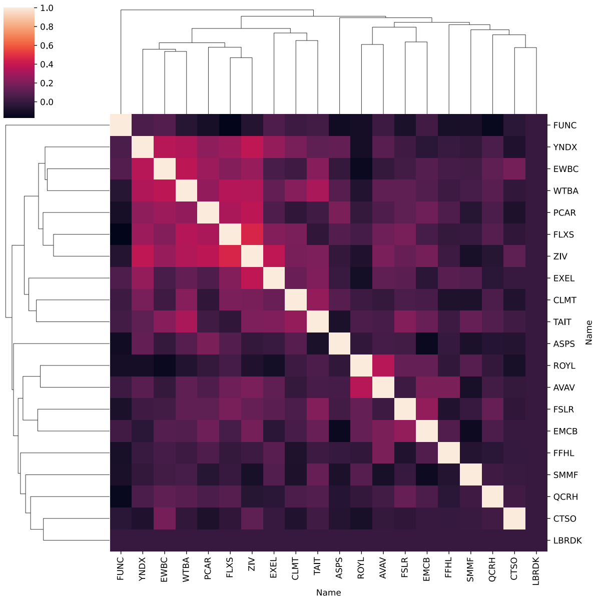

Figure 1: Clustering of 20 stocks based on HALC

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. So we can recover the clipped correlation matrix by

$$\Sigma_{\text{RMT}} = \sum_{k=1}^{N} \lambda_k^{\text{clip}} u_k u_k^T, \qquad \lambda_k^{\text{clip}} = \lambda_k \mathbb{1}_{\{k \leq C\}},$$

where u_k are the orthogonal eigenvectors from the eigendecomposition, and C is the clipping constant, usually chosen as the Marcenko-Pastur edge.

3.3 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 the 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 using the average linkage for all (i, j) pairs $i \in \mathcal{C}_p^n$ and $j \in \mathcal{C}_p^n$, so the filtered matrix becomes

$$(\Sigma_{\text{HALC}})_{ij} = 1 - \alpha_{pq}^n.$$

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 $N(N-1)/2$.

Fig 1 visualizes the cleaning procedure, showing the dendogram, along with the cleaned covariance matrix of 20 randomly picked stock.

3.4 Bootstrap Average Linkage

BAHC, which stands for Bootstrap-averaged hierarchical clustering [5], tries to improve on the simple HALC. By resampling, a more persistent structure is captured and the eigenvectors are more stable. First, a higher-order estimate is calculated by recursively filtering the residuals. The k order residue matrix is $E_k = \hat{\Sigma} - \Sigma_{HALC,k}$. The recursion starts with $k = 0$, $\Sigma_{HALC,0} = 0$ and $E_0 = \hat{\Sigma}$. We perform the hierarchical clustering on the E_k, introduced in 3.3, and after each step we increment the estimator by

$$\Sigma_{HALC,k+1} = \Sigma_{HALC,k} + E_k.$$

By choosing $k = 1$, we get back the HALC cleaning procedure. We can improve upon this estimate by bootstrapping. m copies of the data matrix $R \in \mathbb{R}^{N \times T}$ are maintained and each copy, and we perform a sampling by replacement along the time dimension for each copy with length T. For each copy, we compute the empirical correlation matrix and perform the higher order hierarchical clustering procedure, we just described earlier, resulting in the estimates $\Sigma^{(i)}_{HALC,k}$, where $i \in \{1, \dots, m\}$. The BAHC correlation estimate is given by averaging out the m different estimates,

$$\Sigma_{BAHC} = \sum_{i=1}^{m} \frac{\Sigma^{(i)}_{HALC,k}}{m}.$$

This means, when performing a BAHC estimation, the two hyperparameters are k, controlling how many times we want to perform a higher order recursive clustering, and m, the number of bootstrap sampling with replacements.

3.5 Rotationally invariant, optimal shrinkage

The RIE estimator, proposed by [6], uses the fact, it is possible to compute the overlap between true and sample eigenvectors. Let us denote in this section the true correlation matrix with C, and the estimate with E. The estimator is similar to the RMT method given by eigendecomposition

$$\Sigma_{\mathrm{RIE}} = \sum_{k=1}^{N} \lambda_k^{\mathrm{RIE}} u_k u_k^T, \qquad \lambda_k^{\mathrm{RIE}} = \frac{\lambda_k}{|1 - q + q_k \lim_{\nu \to 0^+} g_E(z_k)|^2},$$

where $z_k = \lambda_k - i\nu$. and g_M is the Stieltjes transform of a matrix M, defined as

$$g_M(z) = N^{-1} \mathrm{tr}(zI - M)^{-1}$$

The Marcenko-Pastur equation, which establishes the connection between the true and the sample correlations

$$\begin{aligned} z g_E(z) &= Z(z) g_C(Z(z)), \\ Z(z) &= \frac{z}{1 - q + qz(g_E(z))}. \end{aligned}$$

The RIE estimator is a method for estimating the true eigenvalues of a matrix by inverting the Marcenko-Pastur equation. However, this approach can be unstable and does not take into account the overlap between the sample eigenvectors and the true eigenvectors. Ledoit and Peche [7] were the first to explain how to calculate these overlaps, and Bun, Bouchaud and Potters [8] later expanded on this by also correcting for a bias in small eigenvalues.

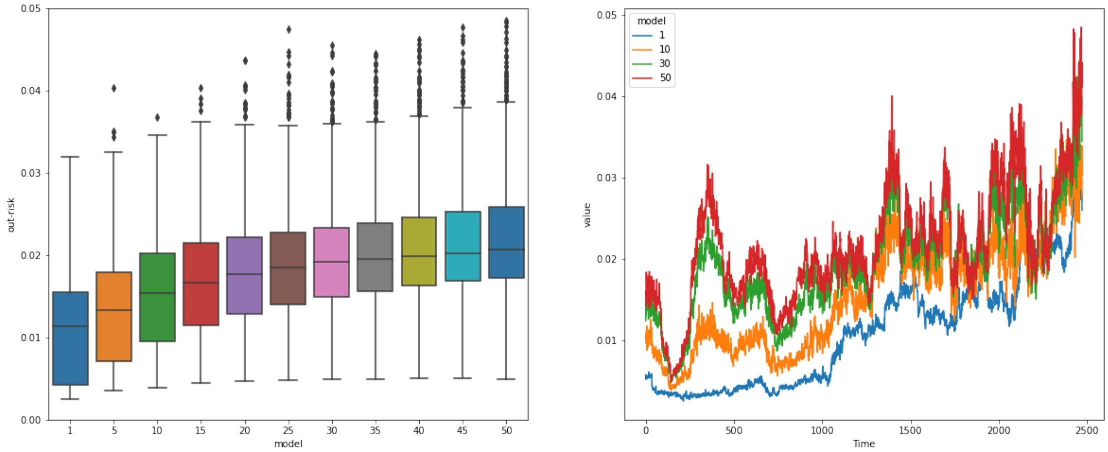

Figure 2: Boxplot of the distributions of the out-of-sample risks for different αs using the linear shrinkage method and the corresponding time series

4 Data set

For testing the methods, we will use a dataset, consisting of the daily adjusted closing price of 1720 stocks listed on NASDAQ, between 2010 January 1 and 2020 December 31 (2772 days). In order to gather the necessary data for our analysis, we employed the use of yfinance, an open-source tool that utilizes Yahoo's publicly available API to facilitate the mass download of market data. During the preprocessing stage, the stock returns were calculated based on the adjusted closed prices, undefined values were imputed as zeros, in periods, where the stock is not issued or no data is available.

5 Numerical experiments

The experiments in this study were carried out on a computer with a high-performance processor and plenty of memory to ensure efficient and accurate results. We used the Ubuntu 20.04.3 LTS operating system and the Python 3.10.4 programming language, along with various packages such as pandas, numpy, and matplotlib. In addition, we made use of specialized packages for hierarchical clustering and RMT-related implementations, as cited in the references [9] and [10]. The entire source code for the experiments is available on Github, and a detailed guide on how to reproduce the results can be found in the README.md file. Through these experiments, we aimed to address the following research questions.

5.1 How does the out-of-sample risk changes for different shrinkage coefficients?

We chose 11 evenly spaced values for α in the range $[0, 1]$, controlling how much we shrink the correlation matrix. A rolling time window was picked $2T = 300$, meaning we used T for in-sample and out-sample calculations. All calculations were performed with only randomly sampled $N = 200$ stocks. For each window, we calculate the Pearson correlation coefficients and shrink the matrix with all possible values of alphas. In each scenario we calculated the optimal Markowitz portfolio weights and using these weights, the out-of-sample risk of the portfolio was calculated. We report the results in Fig 2. $\alpha = 0.6$ achieved the best result, with an average out-of-sample risk of 0.015. The worst performance was achieved by $k = 1$, a.k.a without shrinkage, with an average out-of-sample risk of 0.1754. This experience shows, that linear shrinkage can improve upon using empirical correlations.

5.2 Does higher-order clustering and bootstrapping improve HALC?

In this question, we started by sampling $N = 200$ stocks. A rolling time window was picked $2T = 300$, meaning we used T for in-sample and out-sample calculations. We fixed the number of bootstrap calculations as $m = 50$, and varied the higher-order clustering coefficient $k \in \{1, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50\}$. In every case, we calculated and cleaned

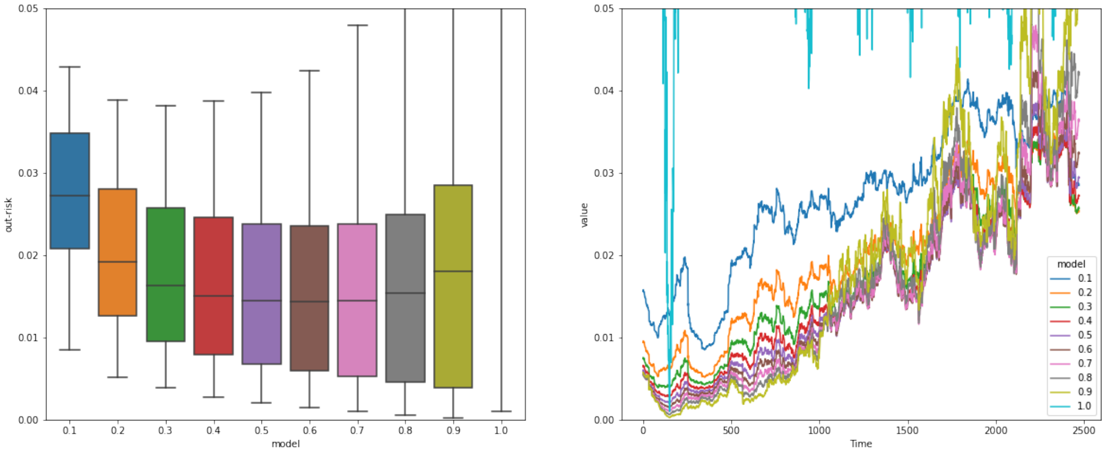

Figure 3: Boxplot of the distributions of the out-of-sample risks for different ks using the BAHC method with $m = 50$ bootsraps and the corresponding time series

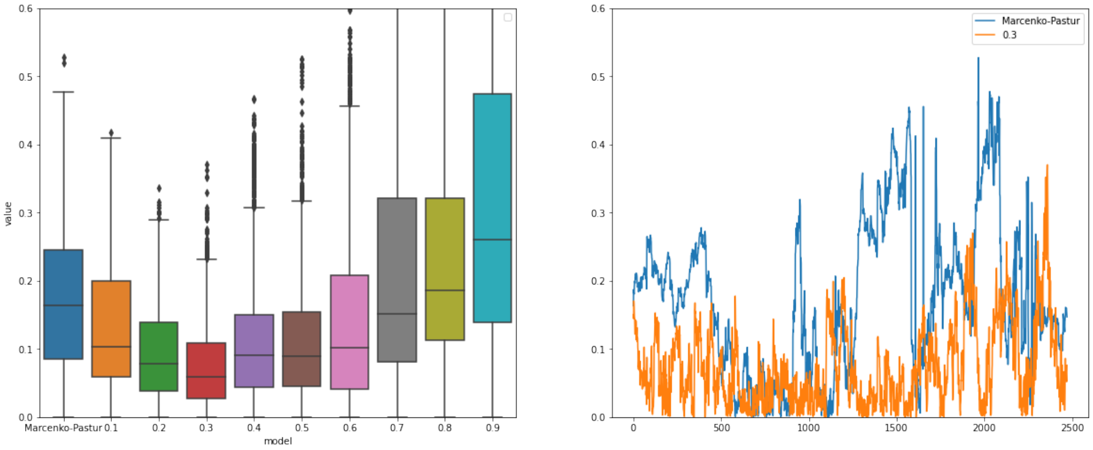

Figure 4: Boxplot of the distributions of the reliabilities and time evolution comparison of the Marcenko-Pasture and 0.3 cutoff model

the correlation matrix, applied for the higher-order hierarchical clustering cleaning, and report the out-of-sample risk using the Markovitz portfolio weights. The results are reported in Table 3. As we can see, with this particular pair of (T, N), the recursive filtering does not improve the results, the best-performing model is with $k = 1$, achieving an average out-of-sample risk of 0.011. The total runtime of this script was 5.083 hours.

5.3 How does the reliability change in RMT for different cut-off values?

In the RMT 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 correlation 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 in Fig 4 .

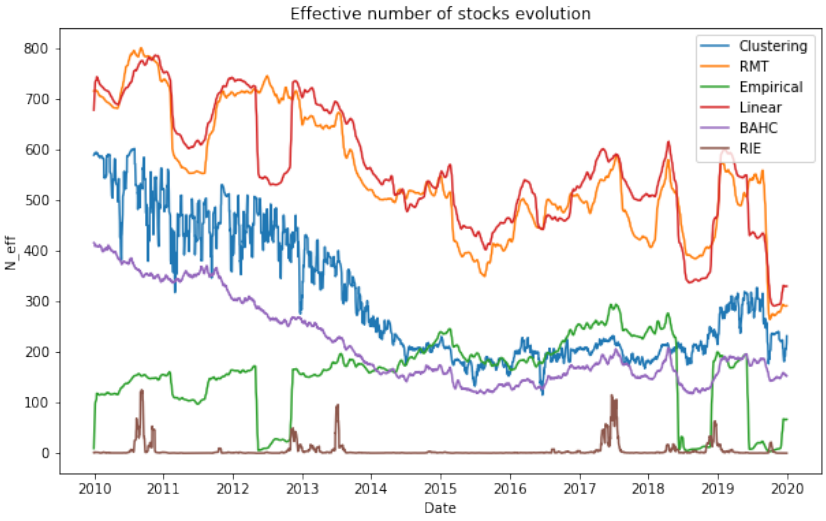

Figure 5: N_{eff} smoothed between 2010 and 2020

Model	σ_{out}	R	N_{eff}	T time
Empirical	0.042	0	**156.768**	2939.827
Linear shrinkage	0.010	0.952	561.334	**2249.348**
RMT	0.009	1.963	549.457	43156.202
HALC	0.007	4.661	304.289	2510.652
Bootsrapped-AHC	**0.005**	**0.494**	215.081	27648.644
RIE	7944857	4345665000	5.624	39857.749

Table 1: Average model performances with $T = 252/2$

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

In this question, we use the entire dataset. 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. We compare the empirical estimator, the linear shrinkage method with $\alpha = 0.6$, RTM with the MP cutoff value, the basic hierarchical clustering approach, the bootstrapped version with $k = 1$ and $m = 10$, and the rotationally invariant optimal estimator. We report all average summary statistics in Table 1. Each time series consist of 2520 entries. All methods, except for RIE are showing similar competitive results. The RIE estimator may fail to produce good results because the assumption $N < T$ clearly does not hold in our settings. In order to make the plots more readable, we applied a moving average filter with a window length of 10 on each time series. Fig 5. shows the results of the effective number of stocks evolution. It can be seen, that the RMT 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. The simple HALC and BAHC evolve similarly, and BAHC seems to be less noisy, thanks to bootstrapping. On the other hand, the linear shrinkage and RMT seem to comove as well both producing less concentrated portfolios. Fig 6, shows the results of the out-of-sample evolution. Again, a moving average smoothing was applied to each time series. We can observe similar behaviors as before, the Clustering approach and its bootstrapped version evolving similarly, BAHC always outperforming the standard method. The linear shrinkage and RMT produce similar results to one another, but until 2018, these estimators are significantly worse compared to the clustering approach. Overall, according to mean performance, BAHC achieves the lowest average out-of-sample risk, along with the lowest reliability. However, the tradeoff is its runtime, being ten

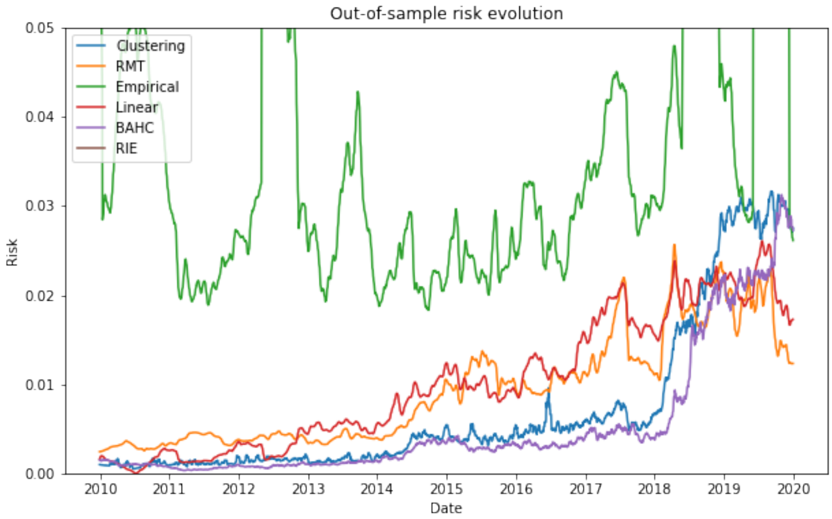

Figure 6: σ_{out} smoothed between 2010 and 2020

times larger compared to others, due to the $m = 10$ bootstrap operations.

5.5 How does the reliability of RMT and HALC change in the function of T and N?

We perform a similar analysis, presented in [2] to compare the two methods. 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 fixed 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 using reliability. In each scenario, we also estimate the correlation matrix using RMT 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 randomly sampled t_0 starting times, and count how many times the clustering approach outperformed the RMT approach, meaning $R_{RMT} > R_{cluster}$, we report the percentage of successes of the clustering method over the RMT method.

We can observe from Fig 7, that the RMT method works better if the T/N ratio is higher. On the other hand, if we include more stocks in the optimization and choose a smaller time horizon, the clustering method is more reliable. This result coincides with the results of Section 5.4, where we observed that for 1700 stocks with a time window of 126, the clustering approach produces lower out-of-sample risk most of the time.

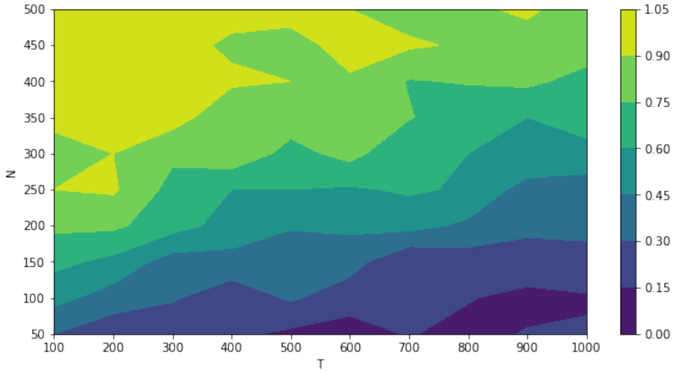

Figure 7: Contour plot of the percentage of success of HALC over RMT for different (T, N) pairs

6 Discussion

In this report, we investigated different filtering procedures for the covariance matrix in a minimum variance portfolio optimization setting. We performed five 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 RMT method, the different shrinkage coefficients for linear shrinkage method, different k parameters for the higher order clustering methods with bootstrapping, 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 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 RMT method. On the other hand, the larger the lookback window, the more reliable the RMT results are. When considering our entire stock universe, clustering clearly outperforms shrinkage, RIE, and RMT methods. Higher order clustering approximation does not improve the results, on the other hand, bootstrapping does, with $m = 10$, achieve the best out-of-sample risk, by making the runtimes significantly larger. Overall, according to our experiments in these specific settings, almost all methods (except for RIE) outperform the simple empirical estimator.

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

[6] Joël Bun, Romain Allez, Jean-Philippe Bouchaud, and Marc Potters. Rotational invariant estimator for general noisy matrices. *IEEE Transactions on Information Theory*, 62(12):7475–7490, 2016.

[7] Olivier Ledoit and Sandrine Péché. Eigenvectors of some large sample covariance matrix ensembles. *Probability Theory and Related Fields*, 151(1):233–264, 2011.

[8] Joël Bun, Jean-Philippe Bouchaud, and Marc Potters. Cleaning large correlation matrices: tools from random matrix theory. *Physics Reports*, 666:1–109, 2017.

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

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