

Markowitz Portfolio Selection Using Various Estimators of Expected Returns and Filtering Techniques for Correlation Matrices

András London, Imre Gera, Balázs Bánhelyi

University of Szeged, Institute of Informatics, 6720 Szeged, Árpád tér 2, Hungary
london@inf.u-szeged.hu

Abstract: In this study we examine the performance of the Markowitz portfolio optimization model using stock time series data of various stock exchanges and investment period intervals. Several methods are used to estimate expected returns, then different “noise” filtering techniques are applied on the correlation matrix containing the pairwise correlations of the time series. The performance of the methods is compared using the estimated and realized returns and risks, respectively. The results show that the estimated risk is closer to the realized risk using filtering methods in general. Bootstrap analysis shows that ratio between the realized return and the estimated risk (Sharpe ratio) is also improved by filtering. In terms of the expected return estimation results show that the James-Stein estimator improves the reliability of the portfolio, which means that the realized risk is closer to the estimated risk in this case.

Keywords: Portfolio optimization; Markowitz model; Random matrix theory; Hierarchical clustering

1 Introduction

The portfolio optimization is one of the fundamental problems in asset management that aims to reduce the risk of an investment by diversifying it into assets expected to fluctuate independently [7]. In his seminal work [17], Markowitz formulated the problem as a quadratic programming task: given the expected return of the portfolio, the risk, a quadratic function that is measured via the covariances of the asset time series, has to be minimized. Recently, the investigation of the correlation coefficient matrix, that is a normalization of the covariance matrix appears in the objective function of the model, has received a big amount of attention, see, without being exhaustive, e.g. [4, 6, 12, 13, 22, 24]. The question of quantifying the degree of statistical uncertainty, called “noise” especially in the statistical physics community, present in the correlation matrix and filter the part of information which is robust against this uncertainty has been addressed and tested [4, 10, 12, 13, 14]. Filtered

correlation matrices have been successfully used in portfolio optimization for risk reduction [13, 22, 24]. However, most of these studies assumed that the investor has perfect knowledge on the future returns at the time of optimization.

In this study we investigate and test the portfolio optimization problem by using several filtering procedures applied to the correlation matrix given by the pairwise asset correlations. The performance of the procedures is simply measured by comparing the predicted and realized risk and return they provide, respectively. For more details on performance analysis of portfolio selection, see [26] for example. In this work we assume that future returns are not known at the time of the investment. Moreover, besides the maximum likelihood estimator (i.e. the average of daily returns) we try other methods to calculate the expected returns.

The structure of this paper is the following. In Section 2 we describe the Markowitz portfolio optimization problem together with some possible estimations of the expected returns (Sec. 2.1) and several filtering procedures that can be performed on the correlation matrices (Sec. 2.2). In Section 3 we present our results with the detailed description of data sets, the experimental setup and evaluation metrics we used (Sec. 3.1, Sec. 3.2 and Sec. 3.3). Finally we draw some conclusions and indicate potential future work.

2 Markowitz portfolio optimization model

Given n risky assets, a portfolio composition is determined by the weights p_i ($i = 1, \dots, n$), such that $\sum_i^n p_i = 1$, indicating the fraction of wealth invested in asset i . The expected return and the variance of the portfolio $\mathbf{p}$ are calculated as

$$r_p = \mathbf{p}\mathbf{r}^T = \sum_{i=1}^n p_i r_i \quad (1)$$

and

$$\sigma_p^2 = \mathbf{p}\Sigma\mathbf{p}^T = \sum_{i=1}^n \sum_{j=1}^n \sigma_{ij} p_i p_j, \quad (2)$$

respectively, where r_i is the expected return of asset i , σ_{ij} is the covariance between asset i and j and Σ is the covariance matrix. Vectors are considered as row vectors in this paper. We should point out that only the proportions $p_1, \dots, p_n$ are needed to determine the performance of the portfolio. It means that the values r_p and σ_p^2 are the same for any investment volume if the weights are the same.

In the classical Markowitz model [17] the risk is measured by the variance providing a quadratic optimization problem consists in finding vector $\mathbf{p}$, such that $\sum_{i=1}^n p_i = 1$ which minimizes σ_p^2 for a given “minimal expected return” value of r_p . Here, we assume that short selling is allowed and therefore p_i can be negative. The solution of this problem, found by Markowitz, is

$$\mathbf{p}^* = \lambda \Sigma^{-1} \mathbf{1}^T + \gamma \Sigma^{-1} \mathbf{r}^T, \quad (3)$$

where $\mathbf{1} = (1, \dots, 1)$ while the other parameters are

$$\lambda = (C - r_p B)/D \text{ and } \gamma = (r_p A - B)/D$$

using the notations

$$A = \mathbf{1}\Sigma^{-1}\mathbf{1}^T, B = \mathbf{1}\Sigma^{-1}\mathbf{r}^T, C = \mathbf{r}\Sigma^{-1}\mathbf{r}^T \text{ and } D = AC - B^2.$$

However, Eq. 3 is rarely used to solve the Markowitz portfolio optimization problem due to numerical stability problems with matrix inversion [5]. Instead, we used the Lagrange multiplier method for optimization (see Sec. 3.3). Next we will describe three possible methods to calculate the expected stock returns in a given period.

2.1 Estimators for the expected returns

Considering the price time series of n assets and denoting the closure price of asset i in time t ($t = 0, 1, \dots, T$) by $P_i(t)$, the daily logarithmic return of i is defined as

$$r_i(t+1) = \log \frac{P_i(t+1)}{P_i(t)} = \log P_i(t+1) - \log P_i(t). \quad (4)$$

In case of stationary independent normal returns (as random variables) the maximum likelihood estimator is the sample mean of the past observations of r_i as it was defined by

$$\hat{r}_i^{ML} = \frac{1}{T} \sum_{t=1}^T r_i(t). \quad (5)$$

Hence, for the portfolio we define

$$\hat{\mathbf{r}}_{ML} = (\hat{r}_1^{ML}, \dots, \hat{r}_n^{ML}), \quad (6)$$

The maximum likelihood return estimation can be highly inefficient since assets with high past returns are likely to contain more positive estimation errors than others. The positive part trimming could further reduce the risk and the James-Stein estimator [11] provides a constructive shrinkage estimator in order to do it. The James-Stein estimation for the expected return for asset i is

$$\hat{\mathbf{r}}_{JS} = (1 - w)\hat{\mathbf{r}}_{ML} + wr_0\mathbf{1}, \quad (7)$$

where

$$r_0 = \frac{\mathbf{1}\Sigma^{-1}\hat{\mathbf{r}}_{ML}^T}{\mathbf{1}\Sigma^{-1}\mathbf{1}^T}, w = \frac{\lambda}{\lambda + T} \text{ and } \lambda = \frac{(n+2)(T-1)}{(\hat{\mathbf{r}}_{ML} - r_0\mathbf{1})\Sigma^{-1}(\hat{\mathbf{r}}_{ML} - r_0\mathbf{1})^T}.$$

In this calculation, each sample mean is shrunk toward the average return of the minimum variance portfolio r_0 .

For small sample size, usually below 50, it was observed that there is no evidence that common asset expected returns are different. If all expected returns are assumed to be equal, the minimum-variance portfolio is efficient and

$$\hat{\mathbf{r}}_{MV} = r_0 \mathbf{1}. \quad (8)$$

Finally, the covariance between asset i and j is estimated by the formula

$$\hat{\sigma}_{i,j}^2 = \frac{1}{T-1} \sum_{t=1}^T (r_i(t) - \hat{r}_i)(r_j(t) - \hat{r}_j), \quad (9)$$

where $\hat{r}_i$ is denotes the estimated value of the with respect the estimator used.

2.2 Filtering the statistical uncertainty

Random matrix theory

The correlation coefficient between asset i and j is defined as $\rho_{ij} = \sigma_{ij} / \sqrt{\sigma_i \sigma_j}$, where $\sigma_i = \sigma_{ii}$ is the standard deviation (often called average volatility) of asset i . A simple random matrix is a matrix whose elements are random numbers from a given distribution [19]. In context of asset portfolios random matrix theory (RMT) can be useful to investigate the effect of statistical uncertainty in the estimation of the correlation matrix [24]. Given the time series of length T of the returns of n assets and assuming that the returns are independent Gaussian random variables with zero mean and unit variance ($\sigma^2=1$), in the limit $n \rightarrow \infty$, $T \rightarrow \infty$ such that $Q = T/n$ is fixed, the distribution $\mathcal{P}_{rm}(\lambda)$ of the eigenvalues of a random correlation matrix ($\mathbf{C}_{rm}$) is given by

$$\mathcal{P}_{rm}(\lambda) = \frac{Q}{2\pi\sigma^2} \frac{\sqrt{(\lambda - \lambda_{\min})(\lambda_{\max} - \lambda)}}{\lambda}, \quad (10)$$

where $\lambda_{\min}$ and $\lambda_{\max}$ are the minimum and maximum eigenvalues, respectively [23], given in the form

$$\lambda_{\max,\min} = \sigma^2 \left(1 + \frac{1}{Q} \pm 2\sqrt{\frac{1}{Q}} \right). \quad (11)$$

Previous studies have pointed out that the largest eigenvalue of correlation matrices from returns of financial assets is completely inconsistent with Eq. 10 and refers the common behavior of the stocks in the portfolio, i.e. the behavior of the market itself. [12, 20]. Since Eq. 10 is strictly valid only for $n \rightarrow \infty$, $T \rightarrow \infty$, we constructed random matrices for the certain n and T values of the data sets we used and we compared the largest eigenvalues and the spectrum $\mathbf{C}$ and $\mathbf{C}_{rm}$. Since $\text{Trace}(\mathbf{C}) = n$ the variance of the part not explained by the largest eigenvalue can be quantified as $\sigma^2 = 1 - \lambda_{\text{largest}}/n$. We can recalculate $\lambda_{\min}$ and $\lambda_{\max}$ in Eq. 11 and construct a filtered diagonal matrix get by setting to zero all eigenvalues of $\mathbf{C}$ smaller than $\lambda_{\max}$ and transform it to the basis of $\mathbf{C}$ with setting the diagonal elements to one. A possible RMT approach for portfolio optimization, following [22], is to use Σ_{rm} (that can be easily calculated from $\mathbf{C}_{rm}$) instead of Σ in the Markowitz model.

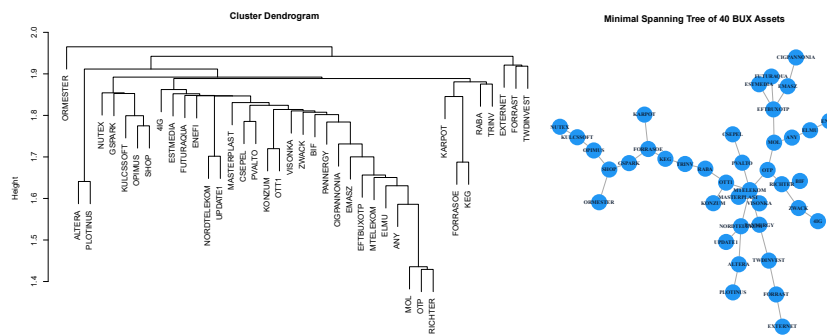


Figure 1

Indexed hierarchical tree - obtained by the single linkage clustering algorithm - and the associated MST of the correlation matrix of 40 assets of the Budapest Stock Exchange

Hierarchical clustering

Correlation based clustering can be considered as a filtering procedure transforming the correlation matrix such that a smaller number of distinct elements retains. The correlation matrix $\mathbf{C}$ has $n(n-1)/2 \sim n^2$ element therefore it contains a large amount of information even for a small number of assets considered in the portfolio. Mantegna and others showed that the single linkage hierarchical clustering algorithm (closely related to minimal spanning trees (MST) of graphs) provide meaningful economic information using only $n-1$ elements of the correlation matrix [15]. The effectiveness of clustering methods have been shown in many studies, e.g. in [2, 9, 18, 25]. To construct the MST, the correlation matrix $\mathbf{C}$ is converted into a distance matrix $\mathbf{D}$, for instance following [15, 16], using $d_{ij} = \sqrt{2(1-\rho_{ij})}$ ultrametric distance. Ultrametric distances are such distances that satisfy the inequality $d_{ij} \leq \max\{d_{ik}, d_{kj}\}$, which is a stronger assumption than the standard triangular inequality. The distance matrix $\mathbf{D}$ can be seen as representing a fully connected graph of the assets with edge weights d_{ij} representing a similarity of the time series of assets i and j . For this graph (i.e. a distance matrix) one can use the Kruskal algorithm in order to obtain the MST of $n-1$ elements and then construct the filtered correlation matrix $\mathbf{C}_{sl}$ using just the $n-1$ correlation coefficients converted back from the $n-1$ distances in the MST. Figure 1 shows an illustrative example for hierarchical clustering and the associated spanning tree obtained using the Budapest Stock Exchange data set. For portfolio optimization, we used Σ_{sl} instead of Σ in the Markowitz model.

In [24] the authors proposed a new portfolio optimization method using another widespread hierarchical clustering procedure, namely the average linkage algorithm. While the single linkage clustering procedure basically follows the greedy Kruskal MST method, the average linkage algorithm, in an iteration step, defines the distance between an element and a cluster as the average distance between the element and each element in the cluster. For detailed description, see e.g. [1]. For portfolio

lio optimization, we can use Σ_{al} constructed using average linkage clustering in the Markowitz model.

3 Results

3.1 Data description

To compare the performance of the methods we analyzed two different data sets. The first data set consists of $n = 40$ stocks traded in the Budapest Stock Exchange (BSE) in the period 1995-2016, using 5145 records of daily returns per stock. The second data set contains the stock time series of $n = 48$ companies of the Information Technology sector (Hardware + Software) that are available on Yahoo Finance (YF) (<https://finance.yahoo.com/>) in almost the same period as the BSE data with 5395 records of daily returns of each stock.

We consider $t = t_0$ as the time when the optimization is performed. Since the covariance matrix has $\sim n^2$ distinct elements while the number of records used in the estimation is nT , the length of the time series need to be $T \gg n$ to get small errors on the covariance. On the other hand, for large T the non-stationarity of the time series more likely appears. The problem is known as the “curse of dimensionality” [27]. To handle this, we computed the covariance matrix and expected returns using the $[-T, 0]$ interval with $T = 50 \approx n$, $T = 100 > n$ and $T = 500 \gg n$ days preceding $t = 0$. The calculation of the expected returns, the covariance matrix and filtered covariance matrices was performed using the time series data of this interval. Then, the realized returns and realized risk (for each method) were calculated using the data on the $[0, T]$ interval. To quantify and compare the different methods considered, we used the measures described in the next section.

We should also mention here, that solving the Markowitz portfolio selection method as a quadratic programming problem is particularly simple when Σ (in. Eq. 2) is positive semi-definite and the constraints are equalities (as in Eq. 1). It is not difficult to see that the positive semi-definiteness is true for the original covariance matrix and also for the filtered matrix obtained by the RMT method. In [1] it was proved that the filtered correlation matrix obtained by the single linkage clustering procedure is always positive definite if all the elements of the obtained filtered correlation matrix are positive. This is usually the case for correlations of stock time series and it has been observed for all the matrices we have used. Moreover, it was proved in the same paper that the filtered correlation matrix obtained by using the average linkage clustering method is also positive definite under the same conditions as in the case of the single linkage procedure.

3.2 Performance evaluation

To measure the performance of a portfolio selected by the different models, we use the following measures to investigate how the estimated and the realized quantities

relate to each other. For portfolio p , the Sharpe ratio measures the excess return (realized) per unit of risk (estimated):

$$S_p = \frac{r_p - r_f}{\hat{\sigma}_p^2} \quad (12)$$

The portfolio risk, due to the estimation of the correlation matrix is calculated as

$$R_p = \frac{|\hat{\sigma}_p^2 - \sigma_p^2|}{\hat{\sigma}_p^2} \quad (13)$$

where $\hat{\sigma}_p^2$ is the predicted risk, while σ_p^2 is the realized risk of the portfolio.

3.3 Simulation setup and results

We implemented our simulation environment in R [21]. We are given a data set of stock time series and the input parameters the `timeInterval` T , vector of `startingTimes` $\mathbf{t_0} = (t_0^1, \dots, t_0^k)$ and $\mathbf{r_p} = (r_p^1, \dots, r_p^\ell)$ vector of `expectedReturns` (equal steps between the average return and the maximal return over all asset by default). The simulation procedure is done via the following steps:

1. For each starting time t_0^j the `asset.solve.Complete.SelectTimes()` subroutine checks whether the portfolio optimization can be done for that starting time on interval $[-T, t_0^j]$
 - if yes, it calculates the optimal portfolio using `asset.solve.Complete.R()`
 - if not¹, it goes to the next starting time t_0^{j+1}
2. The subroutine stores portfolio weights and the data required for performance evaluation

The subroutine `asset.solve.Complete.R()` works as follows:

1. Determines the expected returns using maximum likelihood, James-Stein and minimum variance estimations
2. Determines the covariance matrix of stock time series
3. Calculates the filtered covariance matrices using the RMT, the single linkage and average linkage procedures
4. Portfolio optimization is performed for each return estimation
 - using the Lagrange multipliers method of the 'Rsolnp' package [8] calculates the optimal weights for each covariance matrix
 - calculates the portfolio risk according to the optimal weights

¹ Usually, data with lots of missing (NA) values results in a singular covariance matrix and optimization cannot be performed

- determines the realized risk and Sharpe-ratio

In order to improve the running times the ‘doParallel’ R package [3] was used (here we do not describe the details of parallelization).²

To check the robustness of the methods, a standard bootstrap experiment was performed. We considered 50 starting times randomly and solved the optimization problem using the time series on the intervals $[-T, t_0^j]$ ($T = 50, 100, 500$, $j = 1, \dots, 50$). For each portfolio, the predicted risk was calculated according to Eq. 2 for fixed expected returns from the average $\sum_{i=1}^n r_i/n$ to the maximum expected return $\max\{r_i : i = 1, \dots, n\}$ with equal spans. The Lagrange multiplier method, that is available in ‘Rsolnp’ R package, was used for the optimization. In each case, the portfolios with realized returns in the top and bottom 10% were dropped. The realized risk using the determined stock weights at t_0^j , the realized covariance matrix and realized returns were calculated on $[t_0^j, T]$.

Fig. 2 and Fig. 3 show the ratio of the realized risk σ_p^2 (continuous line) and the predicted risk $\hat{\sigma}_p^2$ (dashed line) as the function of the expected return r_p obtained by the different procedures for the BSE data set and Yahoo data set, respectively. For each T , the time of the investment t_0^j ($j = 1, \dots, 50$) and the set of stocks were the same.

For the BSE data set, the classic method and the RMT method provide similar realized returns that are always higher using hierarchical clustering (single and average linkage). On the other hand, the risk ratio R_p (i.e. the reliability of the portfolio) is also significantly decreased (see Fig. 2, and Tab. 1 “Risk Ratio” column), but the deviations of the realized returns were increased. The Sharpe ratio of the hierarchical clustering methods were smaller than using the other methods, since the estimated risk was often higher than the risk obtained when using the classic and the RMT methods. It can be observed that each method provided better expected returns and smaller risk ratio (i.e. better reliability) for the smaller values of T ($T = 50, 100$, see Tab. 1). The results show that the James-Stein return estimation, although it increases the deviation of the realized returns, provides smaller risk ratios and improvements on the Sharpe ratio. The Sharpe-ratio of the minimum variance portfolio (see Tab. 1 last four column) was the highest due to very small expected risk the method estimated, while its reliability is significantly smaller than using the other return estimators.

For the Yahoo data set similar is true for the realized returns as in the case of BSE data set. Here, the smallest risk ratio was obtained when $T = 100$ days (Fig. 3 middle left and right). It can also be observed, that usage the James-Stein return estimator provided better results (realized returns, Sharpe ratio), while the usage minimum variance estimator decreased the risk ratio in some cases.

² We used ‘doParallel’ and its dependencies to create a parallel back end for the loop construction provided by the ‘forEach’ package.

Conclusions

In this paper, we investigated the Markowitz portfolio selection problem using filtered correlation matrices obtained using different filtering procedures, namely a random matrix theory approach and hierarchical clustering approaches. Furthermore, we used several estimators to determine the expected return of a portfolio. A large set of experiments have shown that using filtered covariance matrices the classic Markowitz solution can be outperformed in terms of realized returns and reliability, meaning that the realized risk and the estimated risk are closer to each other in case of filtering. Our simulations show that the different filtering procedures provide different portfolio optimization results: the most useful method can be different depending on the risk level of the portfolio, the investment period size and reliability of the risk and return estimation. We think that other filtering procedures combined with different return estimators could also provide interesting or better results for different parameters (e.g. expected returns, portfolio size, investment period length) of the optimization problem.

References

- [1] ANDERBERG, M. R. Cluster analysis for applications. monographs and textbooks on probability and mathematical statistics, 1973.
- [2] BASALTO, N., BELLOTTI, R., DE CARLO, F., FACCHI, P., AND PASCAZIO, S. Clustering stock market companies via chaotic map synchronization. *Physica A: Statistical Mechanics and its Applications* 345, 1 (2005), 196–206.
- [3] CALAWAY, RICH, W. S., AND TENENBAUM, D. *Foreach Parallel Adaptor for the 'parallel' Package*, 2015. R package version 2.14.
- [4] CONLON, T., RUSKIN, H. J., AND CRANE, M. Random matrix theory and fund of funds portfolio optimisation. *Physica A: Statistical Mechanics and its applications* 382, 2 (2007), 565–576.
- [5] DU CROZ, J. J., AND HIGHAM, N. J. Stability of methods for matrix inversion. *IMA Journal of Numerical Analysis* 12, 1 (1992), 1–19.
- [6] EL ALAOUI, M. Random matrix theory and portfolio optimization in moroccan stock exchange. *Physica A: Statistical Mechanics and its Applications* 433 (2015), 92–99.
- [7] ELTON, E. J., GRUBER, M. J., BROWN, S. J., AND GOETZMANN, W. N. *Modern portfolio theory and investment analysis*. John Wiley & Sons, 2009.
- [8] GHALANOS, A., AND THEUSSL, S. *Rsolnp: General Non-linear Optimization Using Augmented Lagrange Multiplier Method*, 2015. R package version 1.16.
- [9] GIADA, L., AND MARSILI, M. Data clustering and noise undressing of correlation matrices. *Physical Review E* 63, 6 (2001), 061101.
- [10] GUHR, T., AND KÄLBER, B. A new method to estimate the noise in financial correlation matrices. *Journal of Physics A: Mathematical and General* 36, 12 (2003), 3009.
- [11] JAMES, W., AND STEIN, C. Estimation with quadratic loss. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability* (1961), vol. 1, pp. 361–379.
- [12] LALOUX, L., CIZEAU, P., BOUCHAUD, J.-P., AND POTTERS, M. Noise dressing of financial correlation matrices. *Physical review letters* 83, 7 (1999), 1467.

- [13] LALOUX, L., CIZEAU, P., POTTERS, M., AND BOUCHAUD, J.-P. Random matrix theory and financial correlations. *International Journal of Theoretical and Applied Finance* 3, 03 (2000), 391–397.
- [14] MALEVERGNE, Y., AND SORNETTE, D. Collective origin of the coexistence of apparent random matrix theory noise and of factors in large sample correlation matrices. *Physica A: Statistical Mechanics and its Applications* 331, 3 (2004), 660–668.
- [15] MANTEGNA, R. N. Hierarchical structure in financial markets. *The European Physical Journal B-Condensed Matter and Complex Systems* 11, 1 (1999), 193–197.
- [16] MANTEGNA, R. N., AND STANLEY, H. E. *Introduction to econophysics: correlations and complexity in finance*. Cambridge university press, 1999.
- [17] MARKOWITZ, H. Portfolio selection: Efficient diversification of investments. cowles foundation monograph no. 16, 1959.
- [18] MASLOV, S. Measures of globalization based on cross-correlations of world financial indices. *Physica A: Statistical Mechanics and its Applications* 301, 1 (2001), 397–406.
- [19] MEHTA, M. L. *Random matrices*, vol. 142. Academic press, 2004.
- [20] PLEROU, V., GOPIKRISHNAN, P., ROSENOW, B., AMARAL, L. A. N., AND STANLEY, H. E. Universal and nonuniversal properties of cross correlations in financial time series. *Physical Review Letters* 83, 7 (1999), 1471.
- [21] R CORE TEAM. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013.
- [22] ROSENOW, B., PLEROU, V., GOPIKRISHNAN, P., AND STANLEY, H. E. Portfolio optimization and the random magnet problem. *EPL (Europhysics Letters)* 59, 4 (2002), 500.
- [23] SENGUPTA, A. M., AND MITRA, P. P. Distributions of singular values for some random matrices. *Physical Review E* 60, 3 (1999), 3389.
- [24] TOLA, V., LILLO, F., GALLEGATI, M., AND MANTEGNA, R. N. Cluster analysis for portfolio optimization. *Journal of Economic Dynamics and Control* 32, 1 (2008), 235–258.
- [25] TUMMINELLO, M., ASTE, T., DI MATTEO, T., AND MANTEGNA, R. N. A tool for filtering information in complex systems. *Proceedings of the National Academy of Sciences of the United States of America* 102, 30 (2005), 10421–10426.
- [26] URB N, A., AND ORMOS, M. Performance analysis of equally weighted portfolios: Usa and hungary. *Acta Polytechnica Hungarica* 9, 2 (2012), 155–168.
- [27] ZIMEK, A., SCHUBERT, E., AND KRIEGEL, H.-P. A survey on unsupervised outlier detection in high-dimensional numerical data. *Statistical Analysis and Data Mining* 5, 5 (2012), 363–387.

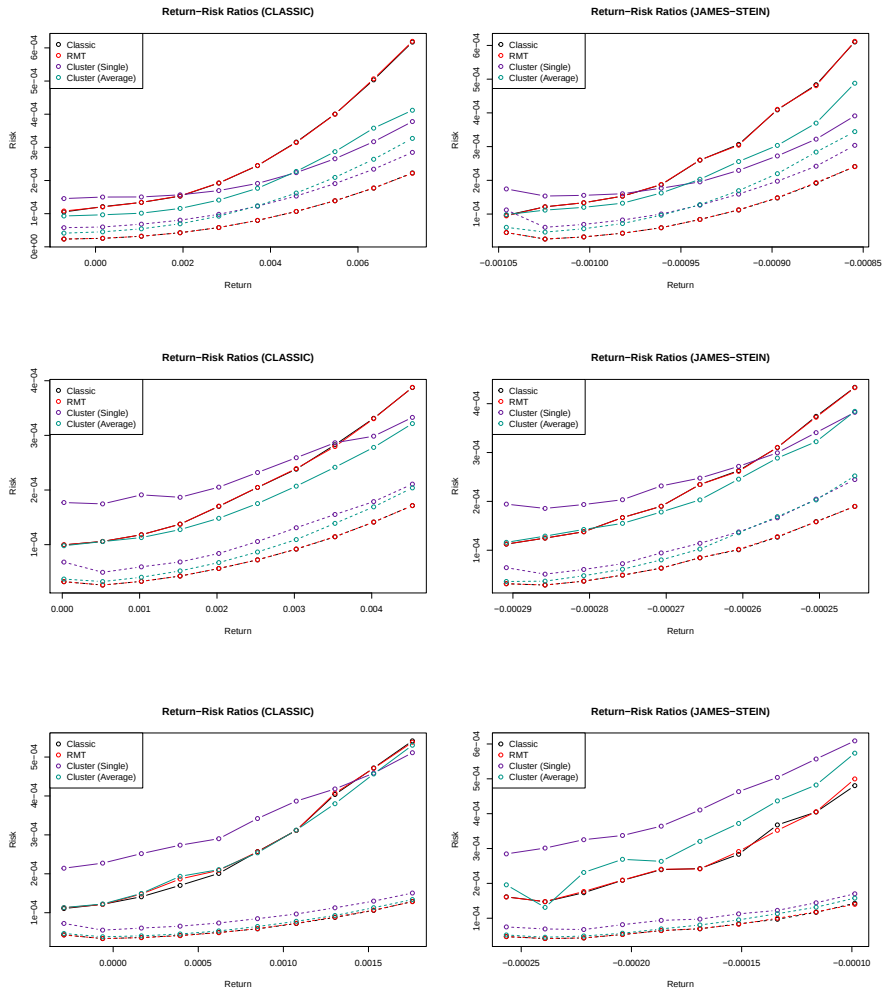


Figure 2

The ratio of the realized risk σ_p^2 and the predicted risk $\hat{\sigma}_p^2$ as the function of expected portfolio return (continuous line) and realized return (dashed line) for the different procedures as $T = 50, 100, 500$ (top-down) using the maximum likelihood estimator (left panels) and the James-Stein estimator (right panel). The data set contains 40 BSE stocks in the period 1995-2016.

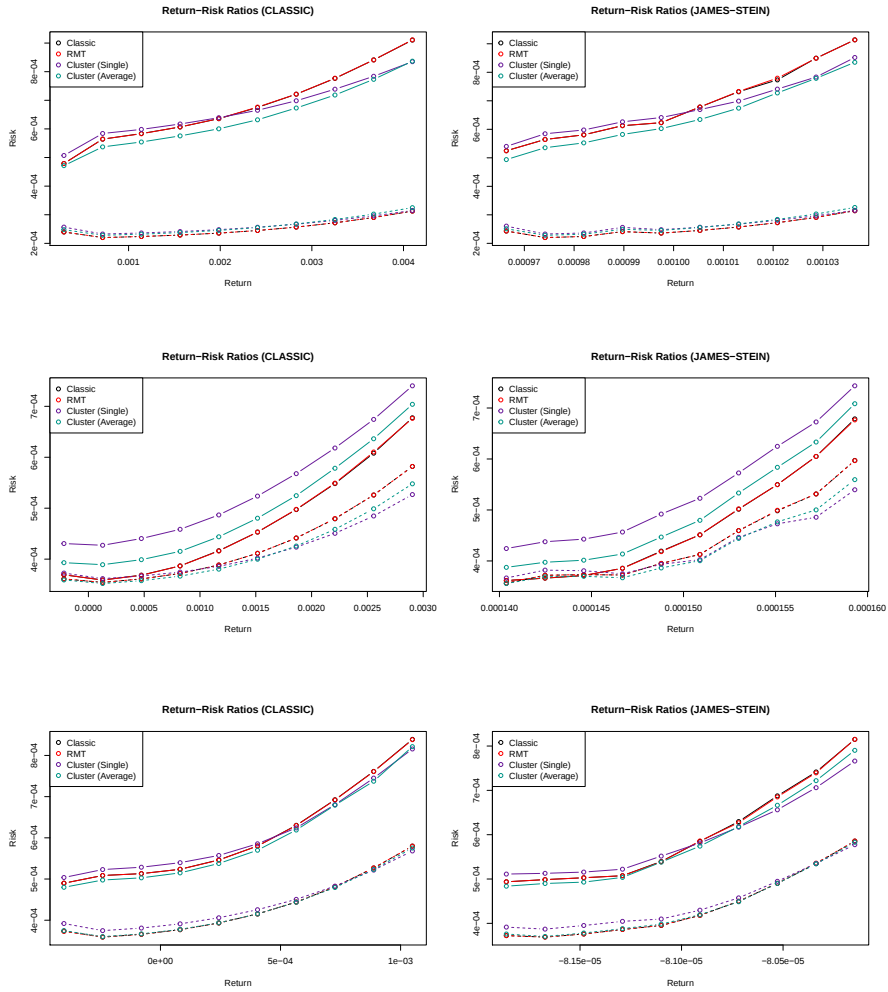


Figure 3

The ratio of the realized risk σ_p^2 and the predicted risk $\hat{\sigma}_p^2$ as the function of expected portfolio return (continuous line) and realized return (dashed line) for the different procedures as $T = 50, 100, 500$ (top-down) using the maximum likelihood estimator (left panels) and the James-Stein estimator (right panel). The data set contains 48 IT sector companies with available historical time series data in the Yahoo finance page in the period 1995-2016

BSE data set	Filtering	Average return estimator			James-Stein estimator			Min variance estimator					
		Realized Return	Realized Return (sd)	Sharpe ratio	Risk Ratio	Realized Return	Realized Return (sd)	Sharpe ratio	Risk Ratio	Realized Return	Realized Return (sd)	Sharpe ratio	Risk Ratio
T = 50	Classic	0.00123	0.00465	8.18922	2.30546	0.00117	0.00466	6.85066	2.14136	0.00084	0.00461	298776	3.32719
	RMT	0.00124	0.00465	7.92646	2.31340	0.00118	0.00466	7.14867	2.16424	0.00084	0.00461	214031	3.32716
	Single linkage	0.00121	0.00304	4.77064	0.73042	0.00192	0.00502	11.82008	0.76601	0.00107	0.00452	155939	1.17548
	Average linkage	0.00073	0.00189	0.51529	0.52383	0.00185	0.00475	10.52615	0.68557	0.00100	0.00447	266844	1.34028
T = 100	Classic	0.00013	0.00153	7.66392	2.01878	0.00101	0.00328	12.57921	1.99415	0.00070	0.00280	520868	3.52232
	RMT	0.00015	0.00154	7.87615	2.02137	0.00099	0.00325	12.34801	2.00979	0.00069	0.00280	508169	3.52259
	Single linkage	0.00119	0.00294	12.90469	1.44612	0.00164	0.00352	13.56478	1.45191	0.00080	0.00282	406972	2.65070
	Average linkage	0.00017	0.00114	8.95616	1.20554	0.00133	0.00339	13.76679	1.22242	0.00072	0.00280	422477	2.68602
T = 500	Classic	-0.00047	0.00121	-11.41798	3.06961	-0.00015	0.00115	-1.61974	2.70563	0.00026	0.00080	23937	4.75457
	RMT	-0.00052	0.00127	-14.16135	3.29680	-0.00016	0.00115	-1.86695	2.72909	0.00026	0.00080	24031	4.75513
	Single linkage	-0.00011	0.00121	-2.19925	2.94493	0.00013	0.00132	1.26614	2.88015	0.00034	0.00077	20362	3.17707
	Average linkage	-0.00053	0.00127	-13.73743	2.96084	-0.00009	0.00108	-0.25504	2.79321	0.00028	0.00075	22544	4.24691
Yahoo data set	Classic	-0.00058	0.00335	-3.76019	1.69646	-0.00055	0.00331	-3.46238	1.64100	-0.00075	0.00318	-9.20659	1.44308
	RMT	-0.00057	0.00334	-3.62685	1.69695	-0.00055	0.00331	-3.45527	1.64236	-0.00075	0.00317	-9.19923	1.44350
	Single linkage	-0.00048	0.00335	-0.55588	1.57894	-0.00049	0.00335	-0.55640	1.58649	-0.00074	0.00331	-4.67897	1.45398
	Average linkage	-0.00045	0.00328	0.37105	1.44763	-0.00045	0.00328	0.53901	1.45527	-0.00068	0.00318	-3.80231	1.28830
T = 100	Classic	0.00030	0.00223	1.58302	0.07236	0.00033	0.00226	1.64966	0.06994	0.00024	0.00199	0.10471	0.01050
	RMT	0.00030	0.00223	1.59425	0.07074	0.00034	0.00226	1.66608	0.05837	0.00024	0.00199	0.09566	0.01217
	Single linkage	0.00012	0.00213	0.62113	0.26152	0.00014	0.00219	0.55639	0.25039	0.00009	0.00187	-0.46180	0.15097
	Average linkage	0.00022	0.00218	1.08339	0.16704	0.00024	0.00222	1.11380	0.15679	0.00017	0.00194	-0.33963	0.08565
T = 500	Classic	-0.00030	0.00070	-1.11800	0.38939	-0.00029	0.00070	-1.05358	0.36550	-0.00024	0.00062	-0.69456	0.38257
	RMT	-0.00030	0.00070	-1.11424	0.38804	-0.00029	0.00070	-1.05042	0.36223	-0.00024	0.00061	-0.69082	0.38035
	Single linkage	-0.00032	0.00069	-1.05861	0.37295	-0.00031	0.00070	-1.03993	0.34634	-0.00027	0.00061	-0.67158	0.36608
	Average linkage	-0.00030	0.00068	-1.08412	0.36562	-0.00029	0.00069	-1.03587	0.35142	-0.00024	0.00059	-0.69585	0.35530

Table 1

Bootstrap experiments using 50 random samples for each value of T when the return is the mean of the average expected return of the portfolio and the maximal expected return over all stocks