Robust portfolio selection based on Gaussian rank correlation estimator

Mei Yan*, Xin Zhao
School of Mathematical Science, University of Electronic Science and Technology of China, Sichuan, China

Email: myanlxy920@hotmail.com

Abstract. Estimating the covariance matrix of stock returns is a fundamental question in empirical finance with implications for portfolio selection. In the presence of outliers, classical estimates of covariance and correlation matrices are not reliable. For this reason, many robust estimator methods are given. However, high dimensional data poses great challenges to the traditional estimation, the merit of robust estimator cannot retain and improved robust estimator is required. In this paper, we proposed a robust pairwise covariance matrix with Gaussian rank correlation, which is acting an input to Markowitz portfolio model. There are two main advantages of the proposed way. The first is that the final covariance estimator is positive semidefinite, we need not to transform the estimator. The second is that the covariance estimator can be computed fast because of the pairwise covariance and the Gaussian rank correlation estimation. Moreover, our numerical results on empirical data show that the proposed estimation is much better than the traditional method. Through empirical studies, we also found that estimation error in sample mean should be considered in Mean-variance portfolios.

1. Introduction

Markowitz's mean-variance analysis sets the basis for modern portfolio optimization theory. In the Markowitz portfolio selection model, the "return" on a portfolio is measured by the expected value of the random portfolio return, and the associated "risk" is quantified by the variance of the portfolio return[1]. This model showed that, given either an upper bound on the risk that the investor is willing to take or a lower bound on the return the investor is willing to accept, the optimal portfolio can be obtained by solving a convex quadratic problem. To implement these portfolios in practice, one has to estimate the mean and the covariance matrix of asset returns from historical data.

However, the classical sample covariance estimator is highly sensitive to outlying observations. For the aberrant observation, robust estimators of covariance matrices are desired. Robust estimation of covariance matrices have received much attention by statistician in the past. Classical robust covariance matrix estimators include $M$-estimators[2,3], minimum volume ellipsoid(MVE)[4] and minimum covariance determinant(MCD) estimators[5-7], $S$-estimators[8], and the Stahel-Donoho estimate (SDE) based on projection [9]. These estimators are almost specifically designed for data with very low dimensions and large sample sizes. Simultaneously, to overcome the sensitivity to outliers, robust correlation estimator is investigated widely by researchers. Such as the Spearman correlation[10], Kendall correlation[11], quadrant correlation[12], Gaussian rank correlation (GRCor)[13] and so on. Compared with the Pearson correlation, these proposed correlation have a lot of advantages. It is worth mentioning that correlation matrices constructed from the Spearman or Kendall correlation or GRCor are robust and computing very fast. Nevertheless, to obtain consistently, the first two need to transformate and are not guaranteed to be positive semidefinite anymore[14].

In recent years, both economic and financial applications have encountered very large data sets which high-dimensional variables[15-17]. For example, the World Bank has data for about two hundred countries over forty years; in portfolio allocation, the number of stocks can be in thousands.
and be large or of the same order of the sample size[18]. Analysis of high-dimensional data pose many challenges, for instance covariance estimation will face with burdensome computation. Clearly, all of these estimates mentioned above are facing a serious computational challenge. Realizing the limitation of the classical robust estimators in high-dimensional setting, some faster estimators with high breakdown points are researched and given in the past years[19,20].

Intuitively, it is simply to calculate a robust location estimate to each individual variable for multivariate location. In the case of multivariate scatter, we can similarly apply a robust covariance or correlation estimate to each pair of variables. Estimates of this way are called “coordinatewise” and “pairwise”[21-23]. The simplest pairwise methods are based on pairwise robust correlation or covariance estimates, for example the Spearman’s $\rho$, kendall’s $\tau$, the quadrant correlation and so on.

Unfortunately, those pairwise matrix estimates are not guaranteed to be positive definite. Recently, [24] proposed a new way to obtain positive-definite estimator based on pairwise robust method. In this article we will use the similar technique to solve the trouble of portfolio selection.

The remainder of this paper is structured as follows. Section 2 provides a brief review of the mean-variance portfolio selection model. The pairwise covariance matrix estimation and Gaussian rank correlation estimators are shown in section 3. In section 4 the results of simulation are given. In section 5, we compare the proposed estimator to the Classical and FMCD estimators under empirical asset returns. Finally, a brief discussion is given to conclude this work.

2. The mean-variance portfolio selection model

The objective of seminal work by Markowitz[1] is to seek a optimal portfolio that ensure the minimal variance with a given expected return. In this section, we review the mean-variance optimization problem. Given p stocks whose returns are distributed with the expected returns vector $\mu$ and covariance matrix $\Sigma$. We consider the mean-variance model defined as follows:

$$\begin{align*}
\min_{\omega} & \quad \omega \Sigma \omega \\
\text{st.} & \quad \omega e = 1, \\
& \quad \omega' \mu = q,
\end{align*}$$

(1)

where, $q$ is the expected rate of return which is required on the portfolio, $l=(1,1,1,\ldots,1)$ denotes a conformable vector of ones.

Solving Eq.(1), the well-known solution is:

$$\omega = \frac{C - qN}{MC - B^2} \hat{\Sigma}^{-1} + qM - N \hat{\Sigma}^{-1} \hat{\mu},$$

(2)

where $\hat{\Sigma}$ is the estimated covariance matrix of return, $\hat{\mu}$ is the estimated of expection vector. $M = 1' \hat{\Sigma}^{-1} 1'$, $N = 1' \hat{\Sigma}^{-1} \hat{\mu}$ and $C = \mu' \hat{\Sigma}^{-1} \hat{\mu}$. From the Eq.(2), it is clearly that the ideal portfolio weights depend on the inverse of the estimated covariance matrix and vector of expected returns, which are estimated from historical data available in the empirical implementation. In other words, if the covariance matrix estimator is not invertible or if it is numerically ill-conditioned, the model did not work at all. For this reason, the resulting estimator matrix $\hat{\Sigma}$ requires be positive semidefinite.

3. Robust covariance matrix estimation

Gnanadesikan and Kettenring[25] use a simple method for turning scale estimators into covariance estimators to obtain a robust, pairwise covariance estimate. A modified version of the Gnanadesikan-Kettenring robust covariance estimate was introduced by Ma and Zamar[20], which is to prominence in the context of robust estimation with shorter computer times. The idea is based on the identity,

$$\text{cov}(X,Y) = (4\alpha\beta)^{-1}[\text{Var}(\alpha X + \beta Y) - \text{Var}(\alpha X - \beta Y)],$$

(3)

where $X$ and $Y$ are random variable. In general, $X$ and $Y$ may have different scales, hence it is standard to let $\alpha = (\text{var}(X))^{\frac{1}{2}}$ and $\beta = (\text{var}(Y))^{\frac{1}{2}}$. A robust covariance estimator is found by replacing the variance in Eq. (3) with (squared) robust scale estimator. With identity Eq. (3), a symmetric matrix full
of pairwise covariances can be constructed simply, however the changed estimator is not necessarily positive semidefinite.

Following the similar spirit of Ma and Zamar[20], Öllerer and Croux[24] studied a modification that using the Gaussian rank correlation to estimate the pairwise covariance matrix. This approach has two virtue. On the one hand, it compute fast and easily. On the other hand, the resulting covariance matrix is positive definiteness, thus we do not need to require a transformation to ensure the result be positive semidefinite. To achieve this goal, tow elements-correlation and scale matrix estimators are considered in the following.

It is well known that the classical Pearson correlation estimator is sensitive to the outliers, even only a small amount of atypical observations may seriously affect it. In view of the outliers, many proposals for the robust estimation correlation of variables were given. For example, the Spearman and Kendall correlation estimators[10,11] which are based on ranks; the quadrant correlation estimator based on winsorizing data. All of those estimators are robust and efficient to compute, even in high dimensions. Unfortunately, the resulting correlation matrix is no longer guaranteed to be positive semidefinite. To avoid the problem, Boudt et al. [13] proposed the Gaussian rank correlation $GR(\cdot)$, which is defined as the sample correlation estimator from the Van Der Waerden scores (or normal scores):

$$GR_{cor} = \sum_{i}^{n} \Phi^{-1} \left( \frac{R(x_i)}{n+1} \right) \sum_{j}^{n} \Phi^{-1} \left( \frac{R(y_j)}{n+1} \right) \Phi^{-1} \left( \frac{R(x_i)}{n+1} \right) \Phi^{-1} \left( \frac{R(y_j)}{n+1} \right)$$

where $R(x_i)$ and $R(y_j)$ denote the rank of $x_i$ and $y_j$, respectively, for $1 \leq i, j \leq n$. Specifically, the corresponding Gaussian (or normal) scores are obtained by plugging these ranks in the quantile function $\Phi^{-1}$ of the standard normal distribution.

Consequently, we use the $GR(\cdot)$ as a valuable robust correlation estimator to estimate the pairwise covariance matrix $(s_{jk})_{j, k=1, \ldots, p} \in \mathbb{R}^{p \times p}$. Concretely, we use the following equation:

$$\sigma_{jk} = scale(X^i)scale(X^j) r(X^i, X^j), j, k = 1, \ldots, p$$

from the data $X = (X^1, \ldots, X^p) \in \mathbb{R}^{n \times p}$, where scale $(\cdot)$ is replaced by a robust scale estimation, such as the $\text{Q}_n$ estimator or the median absolute deviation. Both the two mentioned scale estimators are equally robust with a breakdown point of 50%. To simplify, we opt for the median absolute deviation as scale estimate.

It is worth mentioning the pairwise covariance estimator matrix $\hat{\Sigma}$ using Gaussian rank correlation is positive semidefinite. We study estimating a robust and positive semidefinite covariance matrix are performed by one step. Therefore, we do not need to apply Nearest positive definite matrix procedure(NPD) or Orthogonalised Gnanadesikan Kettenriing procedure(OGK) to obtain a positive semidefinite covariance estimate. In the other words, the proposed method not only saves computation time, but also simplifies the final precision matrix estimator.

4. Numerical simulation

4.1. The simulated data set

In this section we run a simulation to comparing the FMCD and the pairwise robust covariance (PRC) estimates. To evaluate these statistical performance, we need the situations in which the “true values” are known. Considering the heavy-tailed of the returns, the simulated data generated by a multivariate Student-T distribution. Concretely, the simulated data following a distribution is a multivariate Student-T distribution with the degrees freedom 3, and we let $\hat{\Sigma}$ be the covariance matrix process, where $\Sigma = a^{-H}$ characterizes the entry of in the row $i$ and column $j$, $a = 0.7$ in this simulation. For each sample $i$, the covariance is estimated by $\hat{\Theta}_i$, and the Mean Squared error(MSE) is computed as

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (\hat{\Theta}_i - \Theta_i)^2.$$
4.2. Discussion of the simulation

In our experiments, to study the effect of dimension on the performance of various estimators, we considered different cases \( n = 200, 600, 1000 \) with \( p \) taking the values 20, 60, 100, 150, 300 respectively. The MSE results are reported for different estimation methods in Table 1. In small \( p \) case, for \( p = 20 \) and \( p = 60 \), we find that the classical sample covariance (CSM) method substantially outperforms the PRC method, and markedly surpass the FMCD method. That means in relatively low dimension, the CSM method presents a promising alternative, the PRC is the second one. It can be found that the PRC estimator performs significantly better than the classical method when the value of \( p \) is large, see Table 1, for \( p = 300, p = 600 \). That means the PRC estimator is more robust than the CSM method in high dimension case. Note that the number of stocks are often large in practice, the high dimension often happen with the Age of Big Data. Focussing on the robust approach FMCD and PRC, we observed that FMCD perform better than the PRC in the case of large \( p \) and small \( n \). However, the procedure of FMCD is required huge values of subsample, this method cannot run when \( n \) is not large enough (see Table 1). Obviously, when \( n \) is not large we cannot get the estimator of FMCD. Moreover, from the simulation, we also find that the FMCD estimator is not ensure be positive definite in higher dimensions.

We also compared the running time of the two robust estimators in this simulation. Table 2 gives the simulation time in seconds. When \( n = 200 \), we let \( p = 20, 40, 200 \) and run separately. The time of FMCD are 11 and 25 times for the time of PRC.

### Table 1. Simulation results

| Estimate | \( p = 20 \) | \( p = 60 \) | \( p = 100 \) | \( p = 150 \) | \( p = 300 \) | \( p = 600 \) |
|----------|-------------|-------------|-------------|-------------|-------------|-------------|
| \( n = 200 \) |             |             |             |             |             |             |
| FMCD     | 0.024       | 0.029       | **0.018**   | ---         | ---         | ---         |
| PRC      | 0.012       | 0.027       | 0.039       | **0.052**   | **0.086**   | **0.156**   |
| Classical| 0.010       | 0.020       | 0.037       | 0.054       | 0.107       | 0.214       |
| \( n = 600 \) |             |             |             |             |             |             |
| FMCD     | 0.007       | 0.018       | 0.014       | ---         | ---         | ---         |
| PRC      | 0.003       | 0.008       | 0.011       | 0.014       | 0.022       | **0.036**   |
| Classical| 0.002       | 0.004       | 0.006       | 0.010       | 0.020       | 0.041       |
| \( n = 1000 \) |             |             |             |             |             |             |
| FMCD     | 0.004       | 0.020       | 0.025       | ---         | ---         | ---         |
| PRC      | 0.002       | 0.004       | 0.006       | 0.008       | 0.012       | **0.019**   |
| Classical| 0.001       | 0.002       | 0.003       | 0.005       | 0.009       | **0.019**   |

Simulated MSE (the Mean Squared error). ---represent simulation result NAN.

### Table 2. Times for simulation data

| Estimate | \( p = 20 \) | \( p = 40 \) | \( p = 100 \) |
|----------|-------------|-------------|--------------|
| FMCD     | 14.5        | 25.6        | 109          |
| PRC      | 0.13        | 0.7         | 4.4          |

Simulated time with \( n = 200 \).

5. Real data

5.1. The real data set

In this section we use the real data to show the superior performance of the proposed robust portfolio selection approach in this paper. This datasets containing 20 assets were extracted from Yahoo Web-Finance-Stocks data library database composing the S&P500 index. The data span from the sixth of December 2013 to the sixth of December 2016. Deleting the missing data in the sample, all same trading days are 163. This is that \( n = 163, p = 20 \) in the following.

5.2. Result
In order to detect whether the returns are normal distribution or not, we computed the Skewness and Kurtosis of each stock. For simplicity, we denoted the twenty stocks $S_i (i=1,2,\ldots,20)$ respectively. To simplify discussion, we selected the first 10 to study their distributions. It is clearly from Table 1 that each stock $S_i$’s ($i=1,2,\ldots,10$) Skewness and Kurtosis have different extent compared with standard normal distribution. Moreover, the Skewness and Kurtosis are differ considerably (see Table 3).

Another method to identify the difference between the return’s distribution and the normal ones is frequency chart. Figure 1 displays frequency chart of $S_6$. Finally, we also presented boxplots of some stock return (see figure 2). Similar to Table 3, Figure 1 provide that the distributions of returns of risky assets have fat tail nature and are noticeably leptokurtic. For the mean-variance model, the properties of normal distribution are important hypothesis. In the tradeoff between income and risk, the variance and the mean of the securities return are best statistics if the stock return data is normally distributed. Many empirical researches have shown that if the return of risky assets no longer accord with normal distribution, variance is not the best risk. Furthermore, the influence of outliers are shown in Figure 2. We selected six stocks from the twenty stocks stochastically, then given the boxplots of the six stocks. Apparently, the outliers have impact on each stock more or less in Figure 2.

To check the estimation error from sample mean estimator, we came up with the adjust pairwise robust covariance way, which is replace the sample mean with robust mean estimator from the FastMCD (FMCD) and used the robust covariance estimator from the pairwise covariance estimator based on Gaussian rank correlation. Next, we used the classical, FMCD, pairwise robust covariance (PRC) and adjust pairwise robust covariance (APRC) estimator as input into portfolio selection model (1). Figure 4 plots the Mean-Variance Efficient Frontier of the four estimators, the horizontal axis and vertical axis represents the standard deviation and expected return of portfolio selection separately. It is clearly that the FMCD, the PRC, the APRC are better than the classical estimator. With the same level of expected, the risk with robust estimation way is lower than the traditional way. Overall, both the adjust pairwise robust covariance way and the FMCD provide competitive performance relative to the pairwise robust covariance way. However, the FMCD has three drawbacks can not be neglect: (i) the FMCD requires substantial running time for large $p$, it means that if we need compute a few stocks, the FMCD is ineffective; (ii) it no longer retains its robust nature better anymore when $n$ is large, it means that no high breakdown point; (iii) the result covariance estimator cannot ensure positive semidefinite when $n$ is large, which is important in this paper. Note that the APRC estimation is better than the PRC, it verifies that sample mean estimator in mean-variance portfolios should be not overlooked. With the same robust covariance estimator, the robust mean estimator show a better quality than sample mean.

To compare the computing times of different estimates, the running times in seconds is given in Table 4. Clearly, pairwise robust covariance compute fast than the others.

### Table 3. Skewness and Kurtosis of the stock return

| Name | Skewness | Kurtosis | Name | Skewness | Kurtosis |
|------|----------|----------|------|----------|----------|
| S1   | -0.44    | 4.33     | S6   | -0.45    | 3.99     |
| S2   | 0.34     | 4.13     | S7   | 0.33     | 3.35     |
| S3   | 0.29     | 5.45     | S8   | 0.95     | 4.69     |
| S4   | 0.05     | 4.58     | S9   | 0.17     | 3.36     |
| S5   | 0.22     | 8.50     | S10  | 0.08     | 4.31     |

### Table 4. Time(seconds) for real data

| Estimate | Times |
|----------|-------|
| FMCD     | 11.7  |
| PRC      | 0.6   |
| APRC     | 10.2  |
Figure 1. Frequency chart of stock S6. Performance comparison normal distribution density curve.

Figure 2. The boxplots of the stocks for S2, S4, S6, S7 S9, S10.
Figure 3. The Efficient-Frontier of the four estimators.

6. Conclusion
Markowitz portfolio selection model has had a profound impact on the economic modeling of financial markets and the pricing of assets. It makes the optimal portfolio can be obtained by solving a convex quadratic programming problem. Consequently, we can study the optimal portfolio quantitatively and qualitatively. The curse of dimensionality and the influence of perturbations in practice make the classical estimation perform poorly. Various aspects of those phenomenon have been discussed extensively, generally can be divided into re-established model, renewed the estimation method.

In this paper we developed a robust estimation into a large dimensional. Motivated by the well-known solution Eq.(2), we consider a pairwise approach to covariance estimation, which is based on Gaussian rank correlation estimator. The proposed way have main advantage focus on the following:
- The estimator is robust. As we all know, the stock returns is non-normal distribution, and is fat-tail. The returns are sensitive to outliers, robust estimator can ease the disturbance from outliers.
- The final covariance estimator is straightforward positive semidefinite without transforming. From Eq.(2), the solution rely on the inverse of covariance matrix. The estimator we studied above-pairwise covariance and Gaussian rank correlation-are all positive semidefinite, therefore it is invertible. Furthermore, we need not special way to change the final estimator positively, which improves the efficiency in some degree.
- The estimation can be computed much faster than others. On account of the pairwise method and the rank correlation, the complexity of calculation is more simply, especially in the high dimension.

We compared the performance of the pairwise robust approach with the FMCD and classical non-robust method for the covariance estimation of stock returns. Overall, the PRA method performs better than the others from the simulation and real data set. The numerical results also show that the
improved PRA estimation perform well in the mean-variance model. Additional, we verified that
errors of mean sample estimation also manipulate the result of the portfolio selection in our method.

References
[1] Markowitz H 1952 Portfolio selection The journal of finance vol 7 pp 77-91
[2] Cressie N and Hawkins M 1980 Robust estimation of the variogram: I J. Int. Assoc. Math. Geol. vol 7 pp 115-125
[3] Hubert M, Rousseeuw P J and Van Aelst S 2008 High-breakdown robust multivariate methods Statist. Sci. pp 92-119
[4] Hubert M and Debruyne M 2010 Minimum covariance determinant Wiley Interdiscip. Rev. Comput. Stat. vol 2 pp 36-43
[5] Rousseeuw P J and Driessen K V 1999 A fast algorithm for the minimum covariance determinant estimator Technometrics vol 41 pp 212-223
[6] Hardin J and Rocke D M 2004 Outlier detection in the multiple cluster setting using the minimum covariance determinant estimator Comput. Statist. Data Anal. vol 44 pp 625-638
[7] Maronna R A and Yohai V J 2017 Robust and efficient estimation of multivariate scatter and location Comput. Statist. Data Anal. vol 109 pp 64-75
[8] Rocke D M 1996 Robustness properties of S-estimators of multivariate location and shape in high dimension Ann. Statist. pp 1327-1345
[9] Zuo Y, Cui H and He X 2004 On the Stahel-Donoho estimator and depth-weighted means of multivariate data Ann. Statist. pp 167-188
[10] Zar J H 1972 Significance testing of the Spearman rank correlation coefficient J. Amer. Statist. Assoc. vol 67 pp 578-580
[11] Kendall M G 1938 A new measure of rank correlation Biometrika vol 30 pp 81-93
[12] Devlin S J, Gnanadesikan R and Kettenring J R 1975 Robust estimation and outlier detection with correlation coefficients Biometrika vol 62 pp 531-545
[13] Boudt K, Cornelissen J and Croux C 2012 The Gaussian rank correlation estimator: robustness properties Stat. Comput. vol 22 pp 471-483
[14] Croux C and Dehon C 2010 Influence functions of the Spearman and Kendall correlation measures Stat. Methods Appl. vol 19 pp 497-515
[15] Oh D H and Patton A J 2017 Modeling dependence in high dimensions with factor copulas J. Bus. Econom. Statist. vol 35 pp 139-154
[16] Ledoit O and Wolf M 2004 A well-conditioned estimator for large-dimensional covariance matrices J. Multivariate Anal. vol 88 pp 365-411
[17] Díssmann J, Brechmann E C, Czado C and Kurowicka D 2013 Selecting and estimating regular vine copulae and application to financial returns Comput. Statist. Data Anal. vol 59 pp 52-69
[18] Fan J, Liao Y and Mincheva M 2011 High dimensional covariance matrix estimation in approximate factor models Ann. Statist. vol 39 pp 3320
[19] Fritz H, García-Escudero L A and Mayo-Iscar A 2013 A fast algorithm for robust constrained clustering Comput. Statist. Data Anal. vol 61 pp 124-136
[20] Maronna R A and Zamar R H 2002 Robust estimates of location and dispersion for high-dimensional datasets Technometrics vol 44 pp 307-317
[21] Maronna R 2005 Principal components and orthogonal regression based on robust scales Technometrics vol 47 pp 264-273
[22] Maronna R A and Yohai V J 2017 Robust and efficient estimation of multivariate scatter and location Comput. Statist. Data Anal. vol 109 pp 64-75
[23] Tarr G, Müller S and Weber N C 2016 Robust estimation of precision matrices under cellwise contamination Comput. Statist. Data Anal. vol 93 pp 404-420
[24] Öllerer V and Croux C 2015 Robust high-dimensional precision matrix estimation (Springer International Publishing) pp 325-350
[25] Gnanadesikan R and Kettenring J R 1972 Robust estimates, residuals, and outlier detection with multiresponse data Biometrics pp 81-124
Acknowledgments
This work was supported by a grant from the National Natural Science Foundation of China (No. G0511701069) and the Fundamental Research Funds for Central Universities. (No. ZYGX2015J100)