北京大学
PEKING UNIVERSITY

High-dimensional Portfolio Selection with Estimation on the Covariance and Precision Matrices

Name 杨宝旭 Department 光华管理学院

Name 刘锴昭 Department 数学科学学院

Abstract

This project studies a high-dimensional portfolio selection problem using new estimation methods for covariance and precision matrix when p>n. It considers a total 441-asset portfolio from the S&P500 index. It compares four estimation methods for covariance and precision matrix under high dimension including banding estimation, thresholding estimation, banding cholesky estiamtion, and graphical lasso estimation. We produce six out-of-sample predictions from 2015 to 2020, to compare the performance of the four methods, with a variation of time horizons, and balanced frequencies. The results show that graphical lasso method outperforms all other methods in most of the cases. We also discuss about how the assumptions of methods and the true characters of data can affect the results, by analyzing the performance of different methods.

Content

1 Introduction

1.1 Introduction

Investment is essential not only to all financial institutions but also to individuals. A portfolio is a key component and structure of organizing investments, and it consists of all the assets that the investors hold. Investment through portfolio organization is of paramount importance since it provides a structured way of asset allocation, and it enables measurements of its performance. Portfolio investment is important to groups with various backgrounds - hedge funds use it to construct multiple correlated assets to hedge risks; retirement funds use it to secure future financial security; individual investors use it for wealth management, etc. With the fast-growing investment universe, which not only contains stocks from companies after IPO, but also includes various commodities, fixed incomes, and ETFs, there are more choices for investors and investors tend to contain more assets in the portfolios compared to those from ten years ago. Nevertheless, the biggest issue that comes with it is that it is a non-trivial task to hold as many assets in a portfolio as one desires because the portfolio optimization strategy does not always work well with a high-dimensionality.

To be specific, the estimation of covariance matrix of assets return and its inverse play an important role in classical portfolio selection strategy. This limits the number of assets that can be considered, that is also, the dimension problem. However, with new methods under high dimension, we can consider portfolios constructed by more assets relative to the time interval, that is also, when $p > n$.

1.2 What We Do

Thus, in this project, We use several covariance and precision matrix estimation methods under high dimension (p>n), to address the above-mentioned problems. Firstly, we train the model using all the available historical daily price data from around 2014 till 2020, which include 1512 entries for one asset, and 441 assets from the S&P500 index, which are used to construct our portfolios. Secondly, for the methods, traditional methods estimate the covariance matrix and then take the inverse, thus failing to yield significant result in this setup (p>n). We loosely define high-dimension as the case when the number of assets is greater than the number of observations. The estimations of the covariance matrix become imprecise and when the assets are in high-dimension, usually greater than ten, and also the traditional inverse of covariance matrix does not exist. However, the proposed method Graphical Lasso by Friedman, Hastie, and Tibshirani (2008)[1] takes a shortcut and estimates the inverse covariance matrix with Lasso regularization, thus conveying significant results. More importantly, Graphical Lasso requires the regularization process and sparsity of inverse covariance matrix to enforce the estimation, which is the reason why it theoretically addresses the issues that other methods fail to achieve.

1.3 Structure of Report

The paper is organized as follows. Section 2 provides an overview of the literature on Markowitz's modern portfolio theory and the current strategy of sample covariance and precision matrix estimation. Section 3 presents our data, a basic analysis and shows how to organize them to build the portfolio and indicators of results for comparison. Section 4 introduces several methods of estimating covariance matrix and precision matrix under high dimension, and talks about some basic properties of these methods. Section 5 shows the results on real stock data with some analysis. Finally, we end with some discussion and conclusion.

2 Background and Problem statement

2.1 Background and Previous Work

Seventy years ago, Markowitz designed and developed mean-variance optimization which becomes the fundamental work for portfolio optimization. Mean-variance optimization is a theory supporting and advising risk-averse investors to make decisions on their investment portfolios. It considers the risk-return trade-off. Thus, it constructs a portfolio that maximizes the expected return based on a given market risk or minimizes the risk given an expected portfolio return. Harry Markowitz pioneered this theory in his famous and fundamental work "Portfolio Selection" (Markowitz, 1952)[2], which is published in the Journal of Finance. It is one of the most rigorous and popular ways of constructing investment portfolios once it came out. In the financial industry, portfolio managers usually compare their portfolio to a specific benchmark, from which they derive the expected return, and they can conclude that their portfolios outperform it if they have a lower risk. Similarly, they can also argue their portfolios outperforms the benchmark if they have the same risk, theirs have a higher return. Mean-variance optimization is a very well-designed algorithm that combines investors main concern, return and risk, and feed them into a quadratic optimization, which can be solved very fast.

The mean-variance optimization, as its name suggests, uses the vector of expected returns and the variance-covariance matrix as inputs. A key step is to estimate the inverse of the covariance matrix. It has always been the most challenging and trickiest part of optimization. A naive method that comes out originally is calculating the sample covariance matrix and take its inverse directly. It works well when there is a small number of assets, usually smaller than ten. Unfortunately, Jobson and Korkie (1980)[3] discovered and documented that the sample covariance matrix is estimated with lots of errors, especially when the number of assets is large compared to the number of observations, in which case we say it is high-dimensional. It implies that the most extreme coefficients in the matrix thus estimates tend to take extreme values not because this is "the truth" , but because they contain an extreme amount of error (Ledoit and Wolf, 2004)[4]. Michaud (1989)[5] calls this "error maximization" , which will be a terrible result for any portfolio manager who uses mean-variance optimization with the estimation of a sample covariance matrix. In this project we try to find an optimal way of estimating the covariance matrix and its inverse when there are a large number of assets.

Ledoit and Wolf (2004)[4] addresses this issue that lies within the sample covariance matrix for the purpose of portfolio optimization, which contains estimation error of the kind most likely to perturb a mean-variance optimizer. They suggest that nobody should be using the sample covariance matrix for the purpose of portfolio optimization. Instead, they propose linear shrinkage, which is a transformation to find the sample covariance matrix. They employed the shrinkage method on portfolios with the number of assets 30, 50, 100, 225, and 500 and it works decently well. They concluded that without changing any other step in the portfolio optimization process, they reduced tracking error relative to a benchmark index, and substantially increased the realized information ratio of the active portfolio manager. Ledoit, Wolf, et al. (2012)[6] further improves and expands their method 10 years later they proposed the linear shrinkage method and produce a non-linear shrinkage method. They successfully solve the issue that lies within the erroneous estimation of the sample covariance matrix. However, the close-form solution for mean-variance optimization does not require a covariance matrix as the ingredient. It only serves as a middle step to calculate the inverse of the covariance matrix. In a high-dimensional variance-covariance matrix, more computations lead to more inaccuracy. If we can find a method that bypasses the middle step of getting the covariance matrix, it can greatly increase the efficiency and accuracy of the inverse covariance matrix.

2.2 Problem Statement

The goal of portfolio selection (construction) is determining the weight of buying different assets. In Markowitz portfolio theory, there are two main elements being considered, which are payoff and risk. The expect payoff of portfolio is assumed to be the the weighted sum of historical returns, and the risk is measured by the variance, which is the quadratic form of weight and variance-covariance matrix. With the two elements, we can determine the weight in two ways, either minimizing the risk with payoff larger than certain value, or maximizing payoff with risk lower than a certain value. We consider the first way, in which the calculation of portfolio weights will be an optimization problem as following.

$$\begin{aligned}
&\min_{\omega} \omega^T \Sigma \omega \\
&\text{subject to } 1^T \omega = 1 \\
&\qquad\qquad \mu^T \omega \geq \mu_{target}
\end{aligned}$$

Where Σ is the variance-covariance matrix of assets return, μ is the expectation of assets return, and μ_{target} is the lowest return we assume the portfolio to gain. In this problem, what we need to provide are Σ, μ, and μ_{target}. We usually use the estimation of mean $\hat{\mu}$ and covariance matrix $\hat{\Sigma}$ of the sample assets return instead of Σ and μ, and this is also what we call "plug-in" method. μ_{target} is a value we can change, and usually the higher the assumed value, the better the new return of the portfolio is, but we also have to control it since it also affects the feasible minimum variance. The solution ω then determines the proportion of different assets we intend to buy, and the future pay-off time series could be calculated by $R\omega$, where R_{n*p} represents the future returns series of different assets.

Therefore We can firstly, estimate the covariance matrix $\hat{\Sigma}$, and then solve this optimization problem numerically to get the optimal weights. Moreover, this optimization problem has a closed form solution.

$$\omega^* = \Lambda_1 + \Lambda_2 \mu_{target},$$

with

$$\Lambda_1 = \frac{1}{D} \left[B \left(\Sigma^{-1} \iota \right) - A \left(\Sigma^{-1} \mu \right) \right],$$

and

$$\Lambda_2 = \frac{1}{D}\left[C\left(\Sigma^{-1}\mu\right) - A\left(\Sigma^{-1}\iota\right)\right],$$

where ι denotes an appropriately sized vector of ones and where

$$A = \iota^T\Sigma^{-1}\mu,$$
$$B = \mu^T\Sigma^{-1}\mu,$$
$$C = \iota^T\Sigma^{-1}\iota,$$

and

$$D = BC - A^2.$$

With this closed form solution, we can can actually skip the estimation of covariance matrix, and directly estimate the precision matrix Σ^{-1}, and then directly get the optimal weights by plug it in the closed form solution. In all, we can see that we have two main ways to get the optimal weights. The first one needs to estimate the covariance matrix, and the second one needs to estimate the precision matrix. Both require new methods when the number of assets p is larger than the number of observations n.

3 Data and Analysis Procedure

3.1 Data Description

The common elements to form a comprehensive portfolio for a investor is just the stocks. Firstly, the choice of equities involves individual stocks, indices, and ETFs. Ideally, we can choose as many assets as each individual wants to build the portfolios. For example, the benchmark, S&P500 consists of around 500 individual stocks. More assets will bring more noise in the estimation of the covariance matrix. We are trying to overcome this disadvantage by the new methods of estimation, making S&P500 a good choice as real data to be used.

What's more, the whole history data still needs cleaning and truncation due to the market changes. There are entries, exits, and splits of constituents of S&P500, and the stocks that play an important role in the market are not always included. For example, Facebook becomes a constituent of S&P500 in 2013, Tesla joins S&P500 at the end of 2020. These big changes will also affect the whole market, but it's easy to understand that the portfolio constructed by historical data highly depends on the approximate stationarity of return time series. Therefore, I finally chose to use the stocks that have always been on the market from 2014 to 2020. It has 441 stocks in total and includes 1512 entries for each asset (252 trading days a year). The calculation is by year, so that $p = 441$ is larger than $n = 252$, which is just the high dimension situation we want to discuss. It also covers a large variety of asset classes in equities, so we can see the interaction among them.

3.2 Summary Statistics

We take several famous stocks as a subset of the assets to firstly have a basic understanding of the stock returns. Firstly, we visualize the cumulative returns of these assets.

From the figure 1, we can see that the returns of different assets are variant. There are some with high cumulative return but the variance is also large. There are also some with returns always

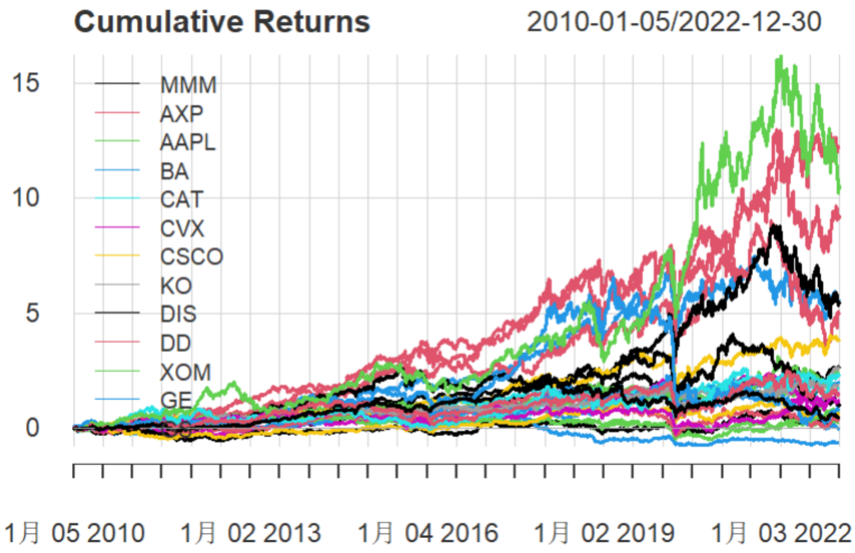

Figure 1: cumulative returns of different assets

close to zero, but the variance is also small. There is indeed a trade off between the return and the variance. This "trade-off" ensures that the basic idea of mean-variance portfolio makes sense.

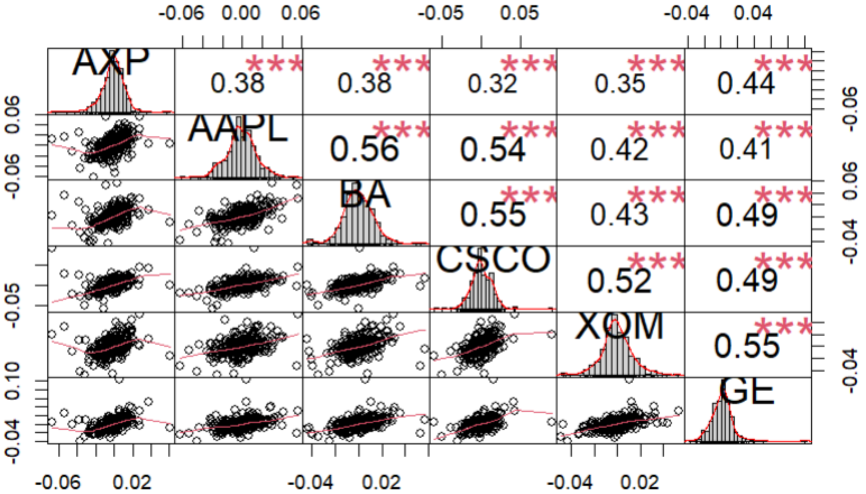

Figure 2: distribution and correlation

These characters of data needs to be further explored, due to estimation methods usually have some assumptions for data. In this correlation plot, we can see that the distribution of returns is close to normal distribution, but with heavy tails. The correlation shows that although the correlation may not be large, but all assets are still correlated to some degree. This may affect the subsequent methods to be introduced.

Finally, we give the summary statistics of the original return data in table 1.

3.3 Analysis Procedure

The data that we use is time-series data, and our goal of portfolio selection is to calculate the optimal weights that can get high returns in the future. Therefore, it is important to compare the

	AXP	AAPL	BA	CSCO	XOM	GE
Stand dev	0.020	0.019	0.027	0.017	0.017	0.024
Mean	0.000	0.001	0.000	0.000	-0.000	-0.001
n	1510.000	1510.000	1510.000	1510.000	1510.000	1510.000
Median	0.001	0.001	0.001	0.001	-0.000	-0.001
CoeffofVariation	84.781	16.862	63.309	38.223	-48.685	-45.528
Minimum	-0.160	-0.138	-0.272	-0.119	-0.130	-0.164
Maximun	0.198	0.113	0.218	0.126	0.119	0.137
Upper Quantile.90%	0.016	0.020	0.021	0.017	0.016	0.023
LowerQuartile.10%	-0.017	-0.019	-0.021	-0.016	-0.017	-0.024

Table 1: return data summary

out-of-sample predictability of the models. Thus, we split all the available sample data from the year 2014 to 2020 into training sets and testing sets. We use the data of last year to calculate the optimal weights for next year, and then test the real return of the constructed portfolio on next year data. Thus, we have six out-of-sample results in total, which we will dive deeper into in the result section. For example, if we use all the historical data from 2014/1/1 to 2014/12/31 as the training set, then the testing set will be 2015/1/1 to 2015/12/31. We mainly use the series of cumulative returns, the annualized average return and also the sharp ratio (defined as the ratio of average return and the standard deviation) as the indicators of comparison. Apart from comparison between different estimation methods, we will also compare the portfolios of each year with the S&P500 index of the year as a bench mark.

4 Main Estimation Methods

As discussed above, when the number of assets is very large and even larger than the number of observations, the estimation of the two core statistics, the covariance matrix and the precision matrix will be affected. Thus new estimation methods need to be introduced. Here, according to what we learn in class and further research, we mainly use four methods, with two estimation methods for each of the covariance matrix and precision matrix.

4.1 Banding Estimation

Bickel and Levina (2008a)[7] considered the following method. For any $p \times p$ matrix $M = (m_{ij})_{p\times p}$, the banding operator with bandwidth k is

$$B_k(M) = (m_{ij}1_{|i-j|\leq k})_{p\times p}$$

Suppose S_n is the sample covariance matrix. We estimate the covariance matrix by

$$\hat{\Sigma} = B_k(S_n)$$

It is noted that the original ordering of $X = (X_1, \cdots, X_p)$ may not admit a bandable covariance. We assume there is a permutation of the p-variate X whose covariance is bandable. The bandable class of covariance is defined as:

$$\begin{aligned}\mathfrak{U}(\alpha, C) = \Big\{\Sigma : \max_{l_2} \textstyle\sum_{|l_1-l_2|>q} |\sigma_{l_1 l_2}| \leq C q^{-\alpha} \text{ for all } q > 0 \\ \text{and } \ 0 < \nu^{-1} \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq \nu\}\end{aligned} \tag{1}$$

4.2 Threshold Estimation

Bickel and Levina (2008b)[8] proposed the threshold estimator under the following covariance class:

$$\mathfrak{V}\left(q, c_0(p), M\right) = \left\{\Sigma : \sigma_{l_1 l_1} \leq M, \sum_{l_2=1}^{p} |\sigma_{l_1 l_2}|^q \leq c_0(p), \quad \text{for all } l_1\right\}$$

For any $p \times p$ matrix $M = (m_{ij})_{p\times p}$, the threshold operator at s is

$$T_s(M) = (m_{ij} 1_{|m_{ij} \geq s|})_{p\times p}$$

Suppose S_n is the sample covariance matrix. We estimate the precision matrix by inverting $T_s(S_n)$ for a suitable s, where s is selected by cross validation.

4.3 Banding Cholesky Estimator

Bickel and Levina (2008a)[7] also considered the banding estimator for the precision matrix. The banding Cholesky estimator is based on the Cholesky decomposition. Without loss of generality, suppose $EX_i = 0$. Regress X_j on its $j-1$ predecessors $X_{j-1}, \cdots, X_1$ so that

$$X_j = \sum_{l=1}^{j-1} a_{jl} X_l + \epsilon_j \quad j = 1, \cdots, p$$

Let $a_j = (a_{j1}, \cdots, a_{j,j-1})^T$ which is the coefficient of the projection of X_j on the subspace generated by $Z_j = (X_1, \cdots, X_{j-1})^T$. Hence,

$$a_j = \frac{1}{\text{Var}(Z_j)} \text{Cov}(X_j, Z_j)$$

Let $d_j^2 = \text{Var}(\epsilon_j)$, $D = \text{diag}(d_1^2, \cdots, d_p^2)$ and $T = (T_{ij})_{p\times p}$ be a lower triangular matrix with ones on the diagonal and $T_{jl} = \begin{cases} -a_{jl} & l < j \\ 1 & l = j \\ 0 & l > j \end{cases}$. Then $T\Sigma T^T = D$.

The banding estimator of Σ^{-1} is to regress X_j on no more than k predecessors, so that

$$X_j = \sum_{l=j-k}^{j-1} a_{jl}^{(k)} X_l + \epsilon_j^{(k)} \quad j = 1, \cdots, p$$

where the bandwidth is selected by the bandwidth test suggested by An, Guo and Liu (2014)[9]. Let $T^{(k)} = (T_{ij}^{(k)})_{p\times p}$, $T_{jl}^{(k)} = \begin{cases} -a_{jl}^{(k)} & j-k \leq l < j \\ 1 & l = j \\ 0 & \text{else} \end{cases}$, and $D^{(k)} = \text{diag}(\text{Var}(\epsilon^{(k)}))$. Then the banding Cholesky estimator of Σ^{-1} is

$$\widehat{\Sigma^{-1}} = \hat{T}^{(k)T} (\hat{D}^{(k)})^{-1} \hat{T}^{(k)}$$

Suppose S_n is the sample covariance matrix. We estimate the precision matrix by inverting $T_s(S_n)$ for a suitable s, where s is selected by cross validation.

4.4 Graphical Lasso Estimation

Friedman J, Hastie T, Tibshirani R (2008)[1] proposed a method using graphical lasso. Suppose we have N multivariate normal observations of dimension p with mean μ and covariance Σ, the objective function of Graphical Lasso is given by

$$\hat{\Theta} = \arg\min_{\Theta} \operatorname{tr}(\hat{\Sigma}\Theta) - \log(\det(\Theta)) + \lambda \sum_{j \neq k} |\Theta_{j,k}| \tag{2}$$

where $\hat{\Theta}$ is an estimator of $\Theta = \Sigma^{-1}$ and $\lambda > 0$ is a regularization parameter, also called the penalty term of Lasso. The above equation is the Gaussian log-likelihood of the data partially maximized with respective to the mean parameter μ. By estimating Θ directly, we will make the estimation of optimal portfolio weight more accurate when we have high-dimensional assets. The regularization by the penalty and sparsity of the inverse covariance matrix enforces the estimation of Graphical Lasso. For the calculation and decision of the penalized term λ in the model selection, we use the Bayesian information criterion (BIC) as the model selection criterion to get the penalty term λ. Meantime, BIC also helps deciding the training set and testing set.

What's more there is also a method called constrained L1 estimation for precision matrix. The idea is to solve the optimal problem:

$$\min \|\boldsymbol{\Omega}\|_1 \text{ subject to: } |\boldsymbol{\Sigma}\boldsymbol{\Omega} - \boldsymbol{I}|_\infty \leq \lambda_n, \quad \boldsymbol{\Omega} \in \mathbb{R}^{p \times p} \tag{3}$$

The lasso problem can be similarly seen as

$$\min \|\boldsymbol{\Omega}\|_1 \text{ subject to: } |\boldsymbol{\Sigma}\boldsymbol{\Omega} - \boldsymbol{I}|_2 \leq \lambda_n, \quad \boldsymbol{\Omega} \in \mathbb{R}^{p \times p} \tag{4}$$

(the original object of lasso is $\min |\boldsymbol{\Sigma}\boldsymbol{\Omega} - \boldsymbol{I}|_2 + \lambda_n \|\boldsymbol{\Omega}\|_1$)

Thus it can be considered as a loose version of the constrained L1 estimation (CLIME). The computation time of graphical lasso estimation is much smaller than the constrained L1 estimation with large p as our data.

For these methods, some detailed properties will be covered in the final discussion section, based on the analysis of results on real data.

5 Results on Real Data and Comparison

As introduced previously, we have in total 6 years (2015-2020) for testing and comparison. Firstly we give the cumulative returns of different portfolios constructed by different estimation methods.

where from top to bottom are by order banding estimator for the covariance matrix, threshold estimator for the covariance matrix, graphical lasso estimator for the precision matrix, banding estimator for the precision matrix, and the bench mark index.

From the cumulative returns, we can see that the portfolio constructed by graphical lasso estimator for the precision matrix performs over all the best. However, only the portfolio constructed by banding estimator for the precision matrix is beaten by the bench mark index during a long time. The difference between the other two estimators of the covariance matrix is quite small.

We also have three indicators to compare between portfolios, which are shown in tables.

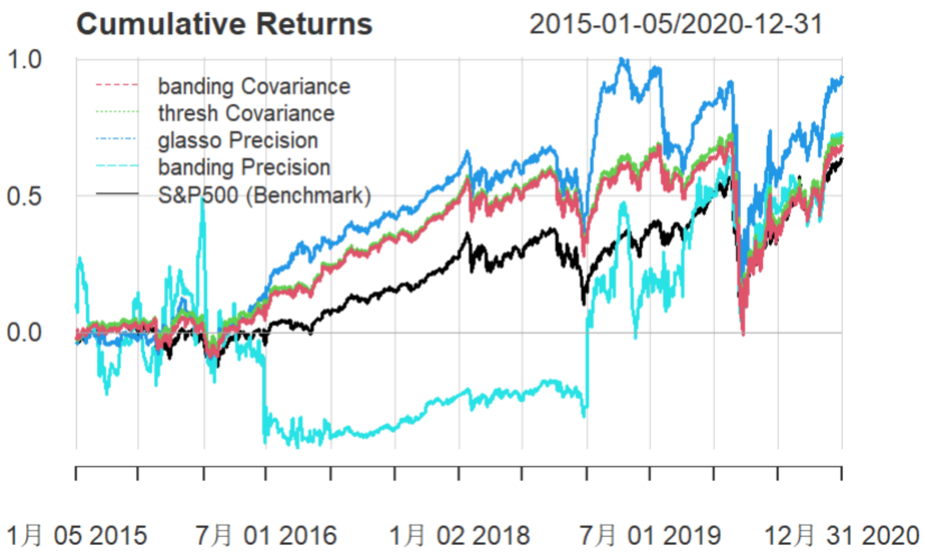

Figure 3: cumulative returns of different portfolios

	banding Covariance	thresh Covariance	glasso Precision	banding Precision	S&P500 (Benchmark)
2015	0.019	0.048	0.029	0.473	-0.019
2016	0.195	0.171	0.271	-0.582	0.086
2017	0.211	0.210	0.205	0.252	0.192
2018	-0.087	-0.087	-0.090	-0.052	-0.084
2019	0.220	0.233	0.283	1.084	0.278
2020	0.025	0.025	0.050	0.148	0.085

Table 2: annualized average return

From the table of annualized standard error, we can see that even though the two methods directly estimating the precision matrix, also both directly plug the precision matrix in the closed form solution. However, for the future standard error, the result value of banding estimator is always larger than the graphical lasso estimator, which shows that the banding estimation of precision matrix have some problems due to some reasons. What's more, the results of two covariance matrix estimators are quite similar. We will discuss these results combined with several properties of these methods.

6 Discussion and Conclusion

6.1 Discussion

Firstly, among all the estimation methods, the portfolio constructed by banding estimator for the precision matrix performs the worst. This may be because that Both the banding estimator require the variables in X having a natural ordering (at least for a permutation of them) such that the correlation decays as two variables are further apart.

From figure 4 we can see that actually the correlation between assets may not satisfy the assumption of banding estimator, which will affect the estimation.

	banding Covariance	thresh Covariance	glasso Precision	banding Precision	S&P500 (Benchmark)
2015	0.150	0.150	0.194	0.523	0.155
2016	0.144	0.145	0.174	0.586	0.131
2017	0.071	0.071	0.073	0.111	0.067
2018	0.156	0.156	0.157	0.170	0.171
2019	0.140	0.138	0.183	0.559	0.125
2020	0.383	0.383	0.381	0.404	0.347

Table 3: annualized standard error

	banding Covariance	thresh Covariance	glasso Precision	banding Precision	S&P500 (Benchmark)
2015	0.125	0.320	0.149	0.904	-0.121
2016	1.354	1.184	1.563	-0.993	0.656
2017	2.963	2.956	2.824	2.281	2.876
2018	-0.555	-0.556	-0.571	-0.307	-0.492
2019	1.569	1.685	1.546	1.938	2.225
2020	0.067	0.066	0.132	0.367	0.244

Table 4: annualized SharpeRatio

Besides, the bandwidth is also an important parameter that affects the estimation. However the larger the bandwidth is the more complicated the calculation is. As a consequence, we actually use bandwidth slightly smaller than the recommended value estimated by simulation to reduce the time of calculation. This may also affect the performance.

Secondly, the banding estimator for covariance matrix and threshold estimator for covariance matrix give very similar results. The convergence rate of this two estimators are given as follows.

For the banding estimator, if $k \sim \left(\frac{\log p}{n}\right)^{-\frac{1}{2\alpha+2}}$

$$\|B_k(S_n) - \Sigma\| = O_p\left\{\left(\frac{\log p}{n}\right)^{\frac{\alpha}{2\alpha+2}}\right\}$$

For the threshold estimator, if $(\log p)/n = o(1)$

$$\|T_{t_n}(S_n) - \Sigma\| = O_p\left\{c_0(p)(\log(p)/n)^{(1-q)/2}\right\}$$

In our problem, $n = 252$ and $p = 441$, and also, due to the nature of returns, the covariance could be bounded by a relatively small value. Thus, in this situation the two covariance matrix would be both very close to the real covariance, which may be the reason for the similar results.

Finally, we can see that the overall best method is directly estimating the precision matrix, instead of solving the problem numerically with covariance matrix estimator. By directly estimating the precision matrix and plug in the closed form solution, we actually skip a procedure of numerical calculation. Besides the consistency of the characters of data and assumptions of the graphical lasso method may also be an important reason for the good performance.

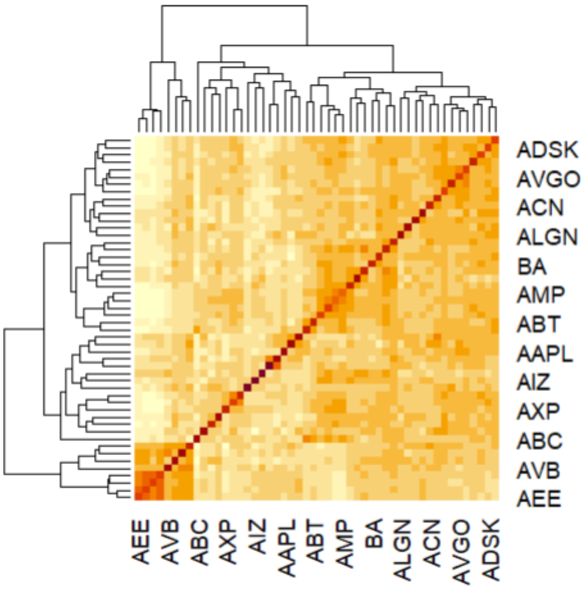

Figure 4: correlation of a sample subset of assets

6.2 Conclusion

In this project, we apply several estimation methods for covariance matrix or precision matrix to the interesting problem of portfolio selection under the situation of high dimension. We tried all these methods on the real data of S&P500 stocks data. Most methods perform very well and have a good results beating the market index, showing the advantage of these high dimension estimation methods. Besides, by comparison of results, we also discuss several problems to be notice using these methods, like the bandwidth selection and also the assumptions that the data must satisfy. In all, by solving this problem of portfolio construction, we see the challenge of high dimension with real world big data and thus also get familiar with new and effective estimation methods of high dimension covariance and precision matrix to deal with this challenge. This project can serve as a good supplement to the theory learned in class.

Reference

[1] Jerome H. Friedman, Trevor J. Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9 3:432–41, 2008.

[2] Harry Markowitz. Portfolio selection*. *The Journal of Finance*, 7(1):77–91, 1952.

[3] J. D. Jobson and Bob Korkie. Estimation for markowitz efficient portfolios. *Journal of the American Statistical Association*, 75:544–554, 1980.

[4] Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88:365–411, 2004.

[5] Richard O. Michaud. The markowitz optimization enigma: Is 'optimized' optimal? *Financial Analysts Journal*, 45(1):31–42, 1989.

[6] Olivier Ledoit and Michael Wolf. Nonlinear shrinkage estimation of large-dimensional covariance matrices. *The Annals of Statistics*, 40(2):1024 – 1060, 2012.

[7] Peter J. Bickel and Elizaveta Levina. Regularized estimation of large covariance matrices. *The Annals of Statistics*, 36(1):199 – 227, 2008.

[8] Peter J. Bickel and Elizaveta Levina. Covariance regularization by thresholding. *The Annals of Statistics*, 36(6):2577 – 2604, 2008.

[9] Baiguo An, Jianhua Guo, and Yufeng Liu. Hypothesis testing for band size detection of high-dimensional banded precision matrices. *Biometrika*, 101(2):477–483, 04 2014.