High Dimensional Covariance Matrix Estimation via Bayesian Method

TANG Jie, HUANG Si-ming

Institute of Policy and Management, Chinese Academy of Sciences, Beijing 100190, China
tj19880912@163.com

Abstract - High-dimensional covariance matrix estimation and its applications in the portfolio selection are increasingly becoming an important topic. However, classical statistical methods used to estimate the sample covariance matrix will lead to inverse covariance matrix biased. Based on this, we propose the high dimensional covariance matrix estimation via Bayesian Method, to ensure that the result of the inverse covariance estimation is unbiased.

Index Terms - covariance matrix estimation, portfolios, bayes, unbiasedness

1. Introduction

The theory of portfolio selection was presented by Markowitz in 1952[1]. It tells that portfolio variance is related to the eigenvalues of a covariance matrix, and optimal portfolio allocation is related to eigenvectors. Therefore, high-dimensional covariance matrix estimation becomes an important topic in many areas, especially, in the financial research.

The mean-variance model raised by Markowitz, is expressed as

$$\min \sigma^2 = \frac{1}{2} x^T S x$$

$$s.t. \begin{cases} {e_n}^T x + x_0 = 1 \\ R^T x + R_0 x_0 = \overline{R} \end{cases} \tag{1}$$

where $R = (R_1, R_2, \cdots, R_n)^T$ is a $N \times 1$ random vector involving n expected yields of risk assets in securities; $\overline{R}$ is the given expectation of the investment portfolio yield; S is a covariance matrix of the expected yields of risk assets in securities; $x = (x_1, x_2, \cdots, x_n)^T$ is a $N \times 1$ random vector involving n risk asset ratios in securities; R_0 is the risk-free asset yield; x_0 is the risk-free asset ratio

If x_0 is eliminated from (1), then (1) can be transformed into (2), that is

$$\min \sigma^2 = \frac{1}{2} x^T S x$$

$$s.t. \ (R - R_0 e_n)^T x = \overline{R} - R_0 \tag{2}$$

According to the Lagrange multiplier method, the first-order partial derivatives of L with respect to x is zero, that means $\partial L / \partial x = 0$,

if

$$L = \frac{1}{2} x^T S x + \lambda[\overline{R} - R_0 - (R - R_0 e_n)^T x].$$

So

$$x = \lambda S^{-1}(R - R_0 e_n).$$

Then put the constraints (2) into it, and assume that

$$Y = R - R_0 e_n$$

is excess yield of risk assets in securities, we can have optimal investment ratio for all risk assets in securities, that is

$$x = \frac{(\overline{R} - R_0) S^{-1} Y}{Y^T S^{-1} Y} \tag{3}$$

From (3), if Y, R_0, $\overline{R}$ have been given or can be obtained based on the sample data by simple addition and subtraction, then the optimal composition of the assets of a proportion of investment portfolios can be simply solved after having got the inverse covariance matrix.

However, the classical statistical methods used to estimate the population covariance matrix use the sample covariance matrix. Although this estimation has more advantages, for example, one can ensure that there is no bias, but it will lead to inverse covariance matrix estimation biased. In addition, more and more domestic and foreign scholars attempt to estimate high-dimensional covariance matrix using various methods and techniques, which could solve the curse of dimensionality, but it cannot meet the requirement of unbiased. In recent years, many methods mentioned are mainly two ideas: one is to improve the sample covariance matrix, such as the use of convergence, eigenvalue decomposition method so as to construct a new covariance matrix, for instance, Guangzhi Cao, Charles A. Bouman have proposed in their paper[2]; the other is by strengthening the data structure to achieve the effect of dimensionality reduction, such as sparsity, compound symmetry, regression models, for example, Ming Yuan[3], Jianqing Fan, Yingying Fan, Jinchi Lv[4] have mentioned in their papers, and so on. Until

Bayesian estimator and empirical Bayesian estimator were proposed by Jushan Bai, Shuzhong Shi[2], methods mentioned above can't guarantee the inverse covariance matrix estimation is unbiased. But unfortunately, some of the conclusions are not very appropriate. For example, Jushan Bai, Shuzhong Shi [2] thought that the posterior distribution of general covariance matrix in the prior distribution is assumed $\Sigma \mid (T-1)S \sim W_N^{-1}(V+T+N,((T-1)S+V\Omega)^{-1})$. However, we verify it and find that this posterior distribution should be $\Sigma \mid (T-1)S \sim W_N^{-1}(V+T+N,(T-1)S+V\Omega)$. And think that the estimator $\hat{\Sigma} = \frac{(T-1)S+V\Omega}{V+T-1}$ is a expectation, not the mode of maximum likelihood estimates based on the posterior distribution, which is proposed by Jushan Bai, Shuzhong Shi[2].

Therefore, this paper will represent high dimensional covariance matrix estimation via Bayesian method, then gives relevant evidence.

2. Sample Covariance Estimation

Sample covariance matrix is the unbiased estimator of general covariance matrix in theory with good property, so it's often used to estimate the general covariance matrix directly.

Let $X_t = \left(X_{1t}, X_{2t}, \cdots, X_{Nt}\right)'$ be an $N \times 1$ -dimensional random vector and assumed to represent the general mean and general covariance. Σ is assumed to be of full rank, and the rank is N. So sample mean and sample covariance can be defined as

$$\bar{X} = \frac{1}{T}\sum_{t=1}^{T} X_t, \quad S = \frac{1}{T-1}\sum_{t=1}^{T}\left(X_t - \bar{X}\right)\left(X_t - \bar{X}\right)'$$

Sample mean and sample covariance are respectively the unbiased estimators of general mean and general covariance, that is

$$E\left(\bar{X}\right) = \mu, \quad E\left(S\right) = \Sigma$$

Although sample covariance is the unbiased estimator of general covariance in theory, meaning that its expected value is equal to the true covariance matrix, there are many shortcomings using sample covariance to estimate the general covariance. When the number of samples are less than the number of cases (the number of indicators), the sample covariance matrix won't be of full rank. At this point, the inverse sample covariance matrix does not exist, and this will bring inconvenience to the application of requiring inverse covariance matrix to estimate. Even if the number of samples is more than the number of cases (the number of indicators) with the sample covariance inverse matrix existing, the estimation of S^{-1} is Biased estimation of Σ^{-1}, that is

$$E(S^{-1}) = \frac{T-1}{T-N-2}\Sigma^{-1}$$

For the reasons above, in order to make the estimation of the inverse sample covariance matrix be the unbiased estimation of the inverse general covariance matrix, general covariance matrix will be estimated with Bayesian statistical method in the following.

3. Bayesian Statistical Method

Bayesian statistics[3,4] is one of statistical inference theories. Different from classical statistics, the theoretical principles of Bayesian statistics are prior probability and posterior distribution. Bayesian statistics make use of both posterior information provided by the sample and prior information given by human's subjective judgments to estimate the probability of an event. However, it uses only information provided by the sample (namely posterior information) to estimate an event's probability in classical statistics[5].

In section 2, we have proven that the estimation for inverse covariance matrix is biased when using the classical statistical method to estimate the high dimensional covariance matrix. For the differences between Bayesian statistics and classical statistics, we take Bayesian statistics method to estimate the population covariance matrix in this section, and prove that the estimate for inverse covariance matrix is unbiased.

We use the following notation.

$X_t = \left(X_{1t}, X_{2t}, \cdots, X_{Nt}\right)'$ is a $N \times 1$ random vector and satisfies multivariate normal distribution. The population mean and covariance matrix are given by $E\left(X_t\right) = \mu$, and $E\left[\left(X_t - \mu\right)\left(X_t - \mu\right)'\right] = \Sigma$.

According to the characters of Wishart distribution[6], if

$$y_1, y_2, \cdots, y_{n+1} \sim N_m(\mu, \sum)$$

and

$$\tilde{A} = \sum_{i=1}^{n+1}(y_i - \bar{y})(y_i - \bar{y})',$$

Then

$$\tilde{A} \sim W_m(n, \sum).$$

Since

$$S = \frac{1}{T-1}\sum_{i=0}^{T-1}(y_i - \bar{y})(y_i - \bar{y})',$$

we conclude that $(T-1)S \sim W_N(T-1, \Sigma)$. Therefore, we have the following density function:

$$p((T-1)S \mid \Sigma) = \frac{etr(-\frac{1}{2}\sum^{-1}(T-1)S)(\det(T-1)S)^{(T-N-2)/2}}{2^{N(T-1)/2}\Gamma_N(\frac{T-1}{2})(\det\sum)^{\frac{T-1}{2}}}$$

For the convenience of calculation, we assume the prior distribution of the population mean, μ , satisfies normal distribution, and the prior distribution of the population inverse covariance matrix, Σ^{-1} , satisfies *Wishart* distribution[2].

We denote

$$\Sigma^{-1} \sim W_N((v\Omega)^{-1}, v),$$

then the prior distribution of Σ satisfies

$$\Sigma \sim W_N{}^{-1}(N+V+1, V\Omega),$$

and its density function is:

$$\pi(\Sigma) = \frac{etr(-\frac{1}{2}\Sigma^{-1}V\Omega)(\det V\Omega)^{(N+V+1-N-1)/2}}{2^{(N+V+1-N-1)N/2}\Gamma_N(\frac{(N+V+1-N-1)}{2})(\det\Sigma)^{\frac{N+V+1}{2}}}$$

$$= \frac{etr(-\frac{1}{2}\Sigma^{-1}V\Omega)(\det V\Omega)^{V/2}}{2^{VN/2}\Gamma_N(\frac{V}{2})(\det\Sigma)^{\frac{N+V+1}{2}}}$$

Applying Bayesian formula to estimate the posterior of Σ, we have

$$p(\Sigma \mid (T-1)S) = \frac{p((T-1) \mid S)\pi(\Sigma)}{P((T-1)S)} \tag{4}$$

The denominator of expression (4) is an integration，and the numerator is the product of two density functions. From the exponential degree, we have

$$tr(-\frac{1}{2}\Sigma^{-1}(T-1)S) + tr(-\frac{1}{2}\Sigma^{-1}V\Omega) = tr(-\frac{1}{2}\Sigma^{-1}((T-1)S+V\Omega))$$

From the part of $\det\Sigma$, we have

$$\frac{1}{2}((N+V+1)+(T-1)) = \frac{1}{2}(N+V+T)$$

Then

$$\Sigma \mid (T-1)S \sim W_N{}^{-1}(V+T+N, (T-1)S+V\Omega)$$

and

$$\Sigma^{-1} \mid ((T-1)S)^{-1} \sim W_N(V+T+N-N-1, ((T-1)S+V\Omega))^{-1})$$

which can be simplified as

$$\Sigma^{-1} \mid ((T-1)S)^{-1} \sim W_N(V+T-1, ((T-1)S+V\Omega))^{-1})$$

Therefore, *Wishart* distribution ensures the estimate for inverse covariance matrix is unbiased，namely

$$E(\Sigma^{-1} \mid ((T-1)S)^{-1}) = (V+T-1)\cdot((T-1)S+V\Omega))^{-1}) = \hat{\Omega}^{-1} \tag{5}$$

From the expression (5), we can gain the Bayesian estimates for the high dimensional covariance matrix and inverse covariance matrix, which are expressed as follows:

$$\hat{\Omega} = \frac{(T-1)S+V\Omega}{V+T-1} \tag{6}$$

$$\hat{\Omega}^{-1} = (V+T-1)\cdot((T-1)S+V\Omega))^{-1}) \tag{7}$$

Chen uses EM (Expectation Maximization) algorithm[7-8] to simplify the calculation. If we express the factor structure of Ω as $\Omega = \beta\beta' + \Delta$, where Δ is a diagonal matrix, and express the parameters as $\theta = (\beta, \Delta)$, the number of parameters cuts down sharply and its degree decreases from $n(n+1)/2$ of Ω to n of θ. When using simulation, such as MCMC (Markov Chain Monte Carlo), to estimate θ, denoted as θ^*, the estimate for covariance matrix is:

$$\hat{\Sigma} = \frac{(T-1)\cdot S + V^*\cdot\Omega(\theta^*)}{V^*+T-1} \tag{8}$$

So we can use the express (8) to estimate the population covariance matrix.

4. Conclusions

This paper corrects the improper conclusion given by Jushan Bai, Shuzhong Shi [2] with a rigorous prove. Using Bayesian methods, we not only gain the estimators for the high dimensional covariance matrix and inverse covariance matrix, but also prove this method ensures the unbiased estimators. Bayesian statistics method can be viewed as a special case of compression methods. From the expression (8) in section 3, Bayesian estimator $\hat{\Sigma}$ is a compression estimate to target $\Omega(\theta^*)$. The Bayesian statistics method present in this paper can be applied in the research on security markets. After getting the estimator of the inverse covariance matrix through Bayesian statistics method and taking the estimator into expression (3) in section 1, we'll gain the best proportion of composition of assets for some portfolio selection easily.

However, despite the rich theoretical researches, there are still rare empirical studies in domestic securities markets, for obtaining prior distribution (prior information) has much trouble. The Bayesian method presented in this paper can overcome the difficulties when the sample covariance matrix was not full rank, and ensure the estimator is unbiased meanwhile. However, there are still rare studies on the test of the Bayesian estimator and how to solve the unstable problem in out-of-sample testing. Hence the questions above are the important research areas in the future.

References

[1] Markowitz H M. Portfolio selection. *Journal of Finance*, 1952, 7: 77-91.

[2] Guangzhi Cao, Charles A.Bouman. Covariance Estimation for High Dimensional Data Vectors Using the Sparse Matrix Transform. *Electrical and Computer Engineering*, 2008.

[3] Ming Yuan. High Dimensional Inverse Covariance Matrix Estimation via Linear Programming. *Journal of Machine Learning Research*, 2010, 11:2261-2286.

[4] Jianqing Fan, Yingying Fan, Jinchi Lv. High Dimensional Covariance Matrix Estimation Using a Factor Model. *Journal of Econometrics*, 2008, 147:185-197.

[5] Jushan Bai, Shuzhong Shi. Estimating High Dimensional Covariance Matrices and Its Applications. *Annals of Economics & Finance*, 2011, 12(2).

[6] Zellner, A. *An introduction to Bayesian inference in econometrics*. Wiley, New York, 1971.

[7] Kotz M, Wu Xizhi. *Modern Bayesian statistics*. Beijing, China Statistics Press, 2000:174.

[8] Liu Jishan. *Introduction to Wishart Distribution*. Beijing, Science Press, 2005.

[9] Chen, C.F.. Bayesian Inference for a Normal Dispersion Matrix and its Application to Stochastic Multiple Regression Analysis. *Journal of the Royal Statistical Society*. 1979, B41: 235-248.

[10] Chen, N., R. Roll, and S. Ross. Economic Forces and the Stock Market. *Journal of Business*. 1986, 59:383-403

Appendix

Proof of the inverse covariance estimation is biased using the inverse sample covariance matrix.

In the normal distribution assumption, the expectation of inverse covariance matrix S^{-1} can be calculated as follows:

If

$$y_1, y_2, \cdots, y_{n+1} \sim N_m(\mu, \sum) ,$$

and

$$\tilde{A} = \sum_{i=1}^{n+1}(y_i - \bar{y})(y_i - \bar{y})' ,$$

then

$$\tilde{A} \sim W_m(n, \sum) .$$

If

$$S = \frac{1}{T-1}\sum_{i=0}^{T-1}(y_i - \bar{y})(y_i - \bar{y})' ,$$

Then

$$(T-1)S \sim W_N(T-1, \Sigma) \text{ [6]} ,$$

and

$$((T-1)S)^{-1} \sim W_N^{-1}(T+N, \Sigma^{-1}) .$$

So the expectation of $((T-1)S)^{-1}$ is

$$E(((T-1)S)^{-1}) = \frac{\Sigma^{-1}}{T+N-2N-2} = \frac{\Sigma^{-1}}{T-N-2} ,$$

that is

$$E(S^{-1}) = \frac{T-1}{T-N-2}\Sigma^{-1} .$$

Hence the inverse covariance estimation is biased using the inverse sample covariance matrix.