Keio University

CALAMVS GLADIO FORTIOR 1858

Statistical Estimation of High-Dimensional Portfolio

Hiroyuki Oka*, Hiroshi Shiraishi

Keio University Graduate School of Science and Technology

Abstract

We introduce a Markowitz's mean-variance optimal portfolio estimator from $d \times n$ data matrix under high dimensional setting where d is the number of assets and n is the sample size. When d/n converges in $(0, 1)$, we show inconsistency of the traditional estimator and propose a consistent estimator.

Background

When an investor invest in d financial products, he consider how maximize the portfolio return for a given level of risk, defined as variance. Define optimal portfolio as follows.

Let $\boldsymbol{X}$ be a asset returns (r.v.), and $\boldsymbol{w}$ be portfolio weights. Then, optimal portfolio weights are the solution of following optimization problem.

$$\begin{cases} \max_{\boldsymbol{w} \in \mathbb{R}^d} u(\boldsymbol{w}) = \mathsf{E}[\boldsymbol{w}^\top \boldsymbol{X}] - \frac{1}{2\gamma} \mathsf{Var}(\boldsymbol{w}^\top \boldsymbol{X}) \\ \text{subject to } \boldsymbol{w}^\top \mathbf{1}_d = 1 \end{cases}$$

Here, γ is a positive constant depending on individual investor.

And then, using expressions $\mathsf{E}[\boldsymbol{X}] = \boldsymbol{\mu}$ and $\mathsf{Var}(\boldsymbol{X}) = \boldsymbol{\Sigma}$, expected value and variance of optimal portfolio are expressed as follows.

Expected value μ_{opt} and variance σ^2_{opt} of optimal portfolio return are expressed as follows.

$$\mu_{\text{opt}}(\gamma; \boldsymbol{\theta}) = \gamma \left(\theta_1 - \frac{\theta_2^2}{\theta_3} \right) + \frac{\theta_2}{\theta_3}$$

$$\sigma^2_{\text{opt}}(\gamma; \boldsymbol{\theta}) = \gamma^2 \left(\theta_1 - \frac{\theta_2^2}{\theta_3} \right) + \frac{1}{\theta_3}$$

A set of $(\sigma_{\text{opt}}, \mu_{\text{opt}})$ is called "*efficient frontier*". Here, $\boldsymbol{\theta}$ is

$$\boldsymbol{\theta} = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{pmatrix} = \begin{pmatrix} \boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} \\ \mathbf{1}_d^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} \\ \mathbf{1}_d^\top \boldsymbol{\Sigma}^{-1} \mathbf{1}_d \end{pmatrix}$$

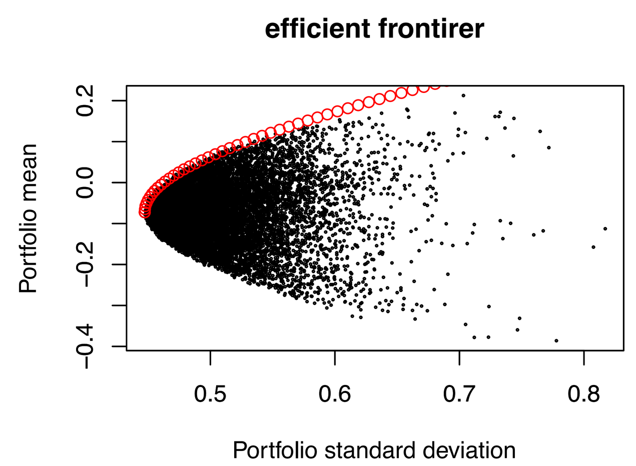

Fig.1 Efficient frontier : Following figure1 shows portfolio plot. Black points $(\cdot)$ are the portofolios in feasible area which can obtain by changing value of $\boldsymbol{w}$. And red circles $(\circ)$ shows optimal portfolios which can obtain by changing value of γ. This figure shows that rational investor prefer lower risk when mean is same value, and higher return when risk is same value.

Objective

Because $\mathsf{E}[\boldsymbol{X}] = \boldsymbol{\mu}$ and $\mathsf{Var}(\boldsymbol{X}) = \boldsymbol{\Sigma}$ is generally unknown, it is considered that optimal portofolio should be estimated by $d \times n$ data matrix $(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$. We estimate $\boldsymbol{\mu}$ by sample mean vector $\bar{\boldsymbol{X}}$, and $\boldsymbol{\Sigma}$ by sample covariance matrix $\boldsymbol{S}$.

$$\bar{\boldsymbol{X}} = \frac{1}{n} \sum_{t=1}^{n} \boldsymbol{X}_t, \ \boldsymbol{S} = \frac{1}{n} \sum_{t=1}^{n} (\boldsymbol{X}_t - \bar{\boldsymbol{X}})(\boldsymbol{X}_t - \bar{\boldsymbol{X}})^\top$$

In these days, because of expansion of market scale, the number of assets d grows bigger. But, it is known that the bigger dimension size d grows, the worse estimator $\boldsymbol{S}^{-1}$ becomes.

- d : fix, $n \to \infty \Rightarrow \boldsymbol{S}^{-1}$ is consistent
- $n, d \to \infty \Rightarrow \boldsymbol{S}^{-1}$ is inconsistent

So, we would like to derive asymptotic property of optimal portfolio on the following assumption.

$$n \to \infty, \quad d \to \infty, \quad \frac{d}{n} \to \rho \in (0, 1)$$

Purpose : estimate efficient frontier in high dimension

Methods

To estimate optimal portfolio, we should estimate optimal portfolio parameter $\boldsymbol{\theta}$. It is assumed that n data vectors $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ is following unknown distribution which has mean vector $\boldsymbol{\mu}$, and covariance matrix $\boldsymbol{\Sigma}$.
Then, $\tilde{\boldsymbol{\theta}}$, estimator of parmetor $\boldsymbol{\theta}$, is defined as follows.

$$\tilde{\boldsymbol{\theta}} = \begin{pmatrix} \tilde{\theta}_1 \\ \tilde{\theta}_2 \\ \tilde{\theta}_3 \end{pmatrix} = \begin{pmatrix} \bar{\boldsymbol{X}}^\top \boldsymbol{S}^{-1} \bar{\boldsymbol{X}} \\ \mathbf{1}_d{}^\top \boldsymbol{S}^{-1} \bar{\boldsymbol{X}} \\ \mathbf{1}_d{}^\top \boldsymbol{S}^{-1} \mathbf{1}_d \end{pmatrix}$$

For some mathematical argument, we put some assumptions.

Assumption1 Let $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n \overset{i.i.d.}{\sim} (\mathbf{0}, \boldsymbol{I}_d)$. Assume that entries of $\boldsymbol{Z}_t$ are independent with $4 + \epsilon$ moment.

Assumption2 Then, data vectors $\boldsymbol{X}_t$ can be expressed as $\boldsymbol{X}_t = \boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{Z}_t + \boldsymbol{\mu}$. $(\boldsymbol{\mu} \in \mathbb{R}^d, \boldsymbol{\Sigma} > \mathbf{0})$

Assumption3 d is expressed with n, and $d/n \to \rho \in (0, 1)$ $(n \to \infty)$ is satisfied.
We call this limit operation "(n, d)-*asymptotic*".

Assumption4 Assume that $\boldsymbol{\theta}$ converge the following constants $\alpha_1, \alpha_2, \alpha_3$.

$$\theta_1 \to \alpha_1, \ \theta_2 \to \alpha_2, \ \theta_3 \to \alpha_3 \quad (d \to \infty)$$

However, it is satisfied that $\alpha_1, \alpha_3 > 0, \alpha_2 \in \mathbb{R}, \alpha_1 \alpha_3 - {\alpha_2}^2 > 0$.

Results

We introduce (n, d)-*asymptotic* properties of estimators of optimal portfolio parameter $\boldsymbol{\theta}$. It is known that when $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \overset{i.i.d.}{\sim} (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and satisfy previous assumption 1∼4, $\tilde{\boldsymbol{\theta}}$ converges following value as n goes to infinity.

$$\tilde{\theta}_1 \overset{a.s.}{\to} \frac{1}{1-\rho} \alpha_1 + \frac{\rho}{1-\rho}, \quad \tilde{\theta}_2 \overset{a.s.}{\to} \frac{1}{1-\rho} \alpha_2, \quad \tilde{\theta}_3 \overset{a.s.}{\to} \frac{1}{1-\rho} \alpha_3$$

This shows that estimator using $\bar{\boldsymbol{X}}$ and $\boldsymbol{S}$ is overestimated. So, we need to correct estimators. We propose the following estimator $\hat{\boldsymbol{\theta}}$.

Result1 Define estimator $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3)^\top$ as following expressions.

$$\hat{\theta}_1 = \left(1 - \frac{d}{n} \right) \tilde{\theta}_1 - \frac{d}{n}, \quad \hat{\theta}_2 = \left(1 - \frac{d}{n} \right) \tilde{\theta}_2, \quad \hat{\theta}_3 = \left(1 - \frac{d}{n} \right) \tilde{\theta}_3$$

Then, $\hat{\boldsymbol{\theta}}$ is consistent estimator of $\boldsymbol{\theta}$.

This estimator $\hat{\boldsymbol{\theta}}$ has asymptotic normality.

Result2 $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \overset{i.i.d.}{\sim} (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ satisfy previous assumption 1∼4. Then, $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ converges to normal distribution as n goes to infinity.

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \overset{\mathbb{D}}{\to} N_3(\mathbf{0}, \boldsymbol{\Omega}) \quad ((n, d)\text{-}\textit{asymptotic})$$

In this, $\boldsymbol{\Omega}$ is following matrix.

$$\boldsymbol{\Omega} = \frac{1}{1-\rho} \begin{pmatrix} 2\alpha_1^2 + 4\alpha_1 + 2\rho & * & * \\ 2\alpha_1 \alpha_2 & \alpha_2^2 + \alpha_1 \alpha_3 + \alpha_3 & * \\ 2\alpha_1 \alpha_3 & 2\alpha_2 \alpha_3 & 2\alpha_3^2 \end{pmatrix}$$

Using this estimator $\hat{\boldsymbol{\theta}}$, we make efficient frontier estimators $\hat{\mu}_{\text{opt}}$ and $\hat{\sigma}^2_{\text{opt}}$ as follows.

$$\hat{\mu}_{\text{opt}}(\gamma; \boldsymbol{\theta}) = \gamma \left(\hat{\theta}_1 - \frac{\hat{\theta}_2^2}{\hat{\theta}_3} \right) + \frac{\hat{\theta}_2}{\hat{\theta}_3}, \ \hat{\sigma}^2_{\text{opt}}(\gamma; \boldsymbol{\theta}) = \gamma^2 \left(\hat{\theta}_1 - \frac{\hat{\theta}_2^2}{\hat{\theta}_3} \right) + \frac{1}{\hat{\theta}_3}$$

These efficient frontier estimators $\hat{\mu}_{\text{opt}}$ and $\hat{\sigma}^2_{\text{opt}}$ has consistency.

Result3 $\hat{\mu}_{\text{opt}}$ and $\hat{\sigma}^2_{\text{opt}}$ satisfy previous assumption 1∼4. Then, $\hat{\mu}_{\text{opt}}$ and $\hat{\sigma}^2_{\text{opt}}$ converge to following values as n goes to infinity.

$$\hat{\mu}_{\text{opt}} \overset{a.s.}{\to} \mu_{\text{opt}}, \ \hat{\sigma}^2_{\text{opt}} \overset{a.s.}{\to} \sigma^2_{\text{opt}} \quad ((n, d)\text{-}\textit{asymptotic})$$

Simulation Study

- fix $d/n = 0.8$
- increase $n = 100, 500, 1000$, and assume $\boldsymbol{X}_t \overset{i.i.d.}{\sim} N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
- components of $\boldsymbol{\mu}$ are devided $[-1, 1]$ into d equal parts
- $\boldsymbol{\Sigma}$ has diagonal components 1、and the others 0.5

In this condition, generate $\tilde{\theta}_1$ and $\hat{\theta}_1$ in 10000 times and confirm theoretical results.

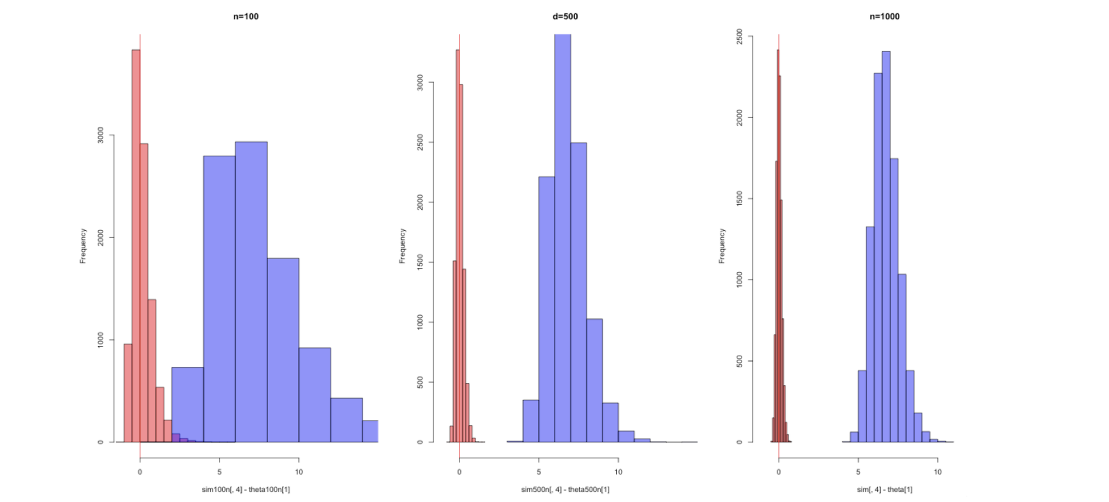

Fig.2 Consistency of estimators : Fig.2 shows histgrams of generated $\tilde{\theta}_1$ and $\hat{\theta}_1$. Red part is histgram of $\hat{\theta}_1$, and blue part is histgram of $\tilde{\theta}_1$. In order of $n = 100, 500, 1000$ from the left, and this shows that $\hat{\theta}_1$ converge to true value θ_1

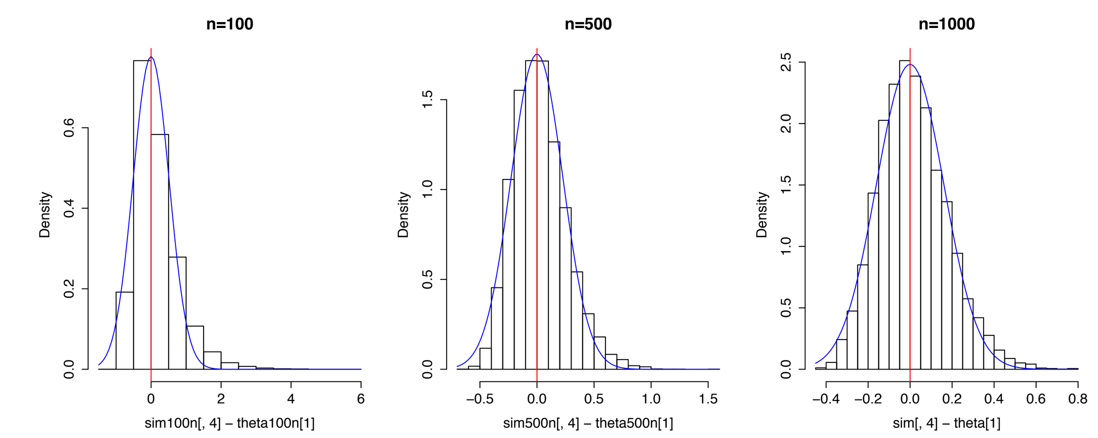

Fig.3 Asymptotic normality of $\sqrt{n}(\hat{\theta}_1 - \theta_1)$: Fig.3 shows histgram of $\sqrt{n}(\hat{\theta}_1 - \theta_1)$ and blue curve line of asymptotic normal distribution. In order of $n = 100, 500, 1000$ from the left, and this shows that $\sqrt{n}(\hat{\theta}_1 - \theta_1)$ converge to objective normal distribution.

Future Work

- To derive asymptotic nomality of $\hat{\mu}_{\text{opt}}$ and $\hat{\sigma}^2_{\text{opt}}$
- To derive confidential interval and test of efficient frontier
- To analyze efficient frontier from actual stock price data

References

Bai, Z., Liu, H., and Wong, W.-K. (2009). Enhancement of the applicability of markowitz's portfolio optimization by utilizing random matrix theory. Mathematical Finance, 19(4):639–667.

El Karoui, N. et al. (2010). High-dimensionality effects in the markowitz problem and other quadratic programs with linear constraints: Risk underestimation. The Annals of Statistics, 38(6):3487–3566.

Bodnar, T., Parolya, N., and Schmid, W. (2014). Estimation of the global minimum variance portfolio in high dimensions. arXiv preprint arXiv:1406.0437.

Fujikoshi, Y., Ulyanov, V. V., and Shimizu, R. (2010). Multivariate Statistics: High-Dimensional and Large-Sample Approximations (Wiley Series in Probabilityand Statistics). Wiley, 1 edition.

Glombek, K. (2012). High-Dimensionality in Statistics and Portfolio Optimization. Josef Eul Verlag Gmbh.