CIRJE-F-906

Optimal Ridge-type Estimators of Covariance Matrix in High Dimension

Tatsuya Kubokawa
The University of Tokyo

Muni S. Srivastava
University of Toronto

October 2013

Optimal Ridge-type Estimators of Covariance Matrix in High Dimension

Tatsuya Kubokawa[*]and Muni S. Srivastava[†]

University of Tokyo and University of Toronto

October 12, 2013

Abstract

The problem of estimating the covariance matrix of normal and non-normal distributions is addressed when both the sample size and the dimension of covariance matrix tend to infinity. In this paper, we consider a class of ridge-type estimators which are linear combinations of the unbiased estimator and the identity matrix multiplied by a scalor statistic, and we derive a leading term of their risk functions relative to a quadratic loss function. Within this class, we obtain the optimal ridge-type estimator by minimizing the leading term in the risk approximation. It is interesting to note that the optimal weight is based on a statistic for testing sphericity of the covariance matrix.

Key words and phrases: Covariance matrix, high dimension, non-normal distribution, normal distribution, ridge-type estimator, risk function.

1 Introduction

Many applied problems in multivariate analysis require estimates of a covariance matrix and/or of its inverse. For example, the inverse of estimators of the covarinace matrix is used in the Fisher linear discriminant analysis, confidence intervals based on the Mahalanobis distance and weighted least squares estimator in multivariate linear regression models. However, the unbiased estimator based on the sample covariance matrix is not invertible when the dimension p of the variables is larger than the sample size N. When p is large and close to N, the inverse of the unbiased estimator may be ill-conditioned even if $N > p$. Thus, an estimator for the covariance matrix is required to be both invertible

[*] Faculty of Economics, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, JAPAN. E-Mail: tatsuya@e.u-tokyo.ac.jp

[†] Department of Statistics, University of Toronto, 100 St George Street, Toronto, Ontario, CANADA M5S 3G3, E-Mail: srivasta@utstat.toronto.edu

and well-conditioned. The estimation procedures satisfying these properties have been studied in a lot of articles. For instance, see Daniels and Kass (2001), Ledoit and Wolf (2003, 2004), Srivastava and Kubokawa (2007), Konno (2009), Fisher and Sun (2011) and Bai and Shi (2011).

To specify the problem considered here, consider p-dimensional random vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ which are mutually independently and identically distributed with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Then, $\boldsymbol{\Sigma}$ is estimated unbiasedly by $\boldsymbol{S} = n^{-1}\boldsymbol{V}$, where

$$\boldsymbol{V} = \sum_{j=1}^{N} (\boldsymbol{x}_j - \overline{\boldsymbol{x}})(\boldsymbol{x}_j - \overline{\boldsymbol{x}})^t, \quad n = N - 1, \tag{1.1}$$

for $\overline{\boldsymbol{x}} = N^{-1}\sum_{j=1}^{N} \boldsymbol{x}_j$. Ledoit and Wolf (2004) considered a class of a convex combination of $\boldsymbol{S}$ and $a_1\boldsymbol{I}_p$, namely

$$\boldsymbol{S}_w^* = w\boldsymbol{S} + (1 - w)a_1\boldsymbol{I}_p,$$

for $a_1 = \mathrm{tr}\,[\boldsymbol{\Sigma}]/p$, and showed that the optimal weight w in the sense of minimizing the risk function $E[\mathrm{tr}\,[(\boldsymbol{S}_w^* - \boldsymbol{\Sigma})^2]]/p$ is given by

$$w(\boldsymbol{\Sigma}) = \mathrm{tr}\,[(\boldsymbol{\Sigma} - a_1\boldsymbol{I}_p)^2]/E[\mathrm{tr}\,[(\boldsymbol{S} - a_1\boldsymbol{I}_p)^2]].$$

Ledoit and Wolf (2004) provided an estimator $\hat{w}$ of w and suggested the use of a plug-in estimator. In the case of a normal distribution, Fisher and Sun (2011) showed that $w(\boldsymbol{\Sigma})$ is described as

$$w(\boldsymbol{\Sigma}) = \frac{n(a_2 - a_1^2)}{n(a_2 - a_1^2) + pa_1^2 + a_2},$$

for $a_2 = \mathrm{tr}\,[\boldsymbol{\Sigma}^2]/p$, and suggested a simple estimator of $w(\boldsymbol{\Sigma})$, given by

$$\hat{w} = \frac{n(\hat{a}_2 - \hat{a}_1^2)}{n(\hat{a}_1 - \hat{a}_1^2) + p\hat{a}_1^2 + \hat{a}_2} = \frac{nT}{nT + p + T + 1}, \tag{1.2}$$

where $T = \hat{a}_2/\hat{a}_1^2 - 1$ for

$$\hat{a}_1 = \mathrm{tr}\,[\boldsymbol{S}]/p = \mathrm{tr}\,[\boldsymbol{V}]/(np), \tag{1.3}$$

$$\hat{a}_2 = \frac{1}{(n-1)(n+2)p}\Big[\mathrm{tr}\,[\boldsymbol{V}^2] - \frac{1}{n}(\mathrm{tr}\,[\boldsymbol{V}])^2\Big]. \tag{1.4}$$

The resulting plug-in estimator is given by $\hat{w}\boldsymbol{S} + (1 - \hat{w})\hat{a}_1\boldsymbol{I}_p$. Although it seems reasonable, it is natural to raise the following queries.

(I) The weighted estimator $\boldsymbol{S}_w^*$ considered above is a convex combination of the statistic $\boldsymbol{S}$ and the parameter $a_1\boldsymbol{I}_p$. Rather than this combination, it is more natural to consider a convex combination of $\boldsymbol{S}$ and the statistic $\hat{a}_1\boldsymbol{I}_p$. Are there any difference between these two estimators ?

(II) Although the estimator $w(\boldsymbol{\Sigma})\boldsymbol{S} + (1 - w(\boldsymbol{\Sigma}))a_1\boldsymbol{I}_p$ with the optimal weight $w(\boldsymbol{\Sigma})$ is optimal in the sense of minimizing the risk, this fact does not necessarily guarantee the optimality of the plug-in estimator $\hat{w}\boldsymbol{S} + (1-\hat{w})\hat{a}_1\boldsymbol{I}_p$, since the estimated weight $\hat{w}$ has correlations with $\boldsymbol{S}$ and $\hat{a}_1$. Can we establish any optimality property of the plug-in estimator ?

(III) In Fisher and Sun (2011), the estimators $\hat{a}_1$ and $\hat{a}_2$ were used for a_1 and a_2 without any justification in cases of non-normal distributions. In the normal distribution, Srivastava (2005) showed that these are unbiased and that $\hat{a}_1 - a_1 = O_p((np)^{-1/2})$ and $\hat{a}_2 - a_2 = O_p((np)^{-1/2}) + O_p(n^{-1})$. However, these properties for $\hat{a}_2$ can not be necessarily established. For instance, $\hat{a}_2$ is not an unbiased estimator of a_2 in non-normal cases as shown in Srivastava, Kollo and von Rosen (2011) and Srivastava, Yanagihara and Kubokawa (2013). Can we derive order of $\hat{a}_2 - a_2$ and show the consistency of $\hat{a}_2$?

In this paper, we try to answer these questions. In Section 2, we provide the optimal weights in linear combinations $c_1\boldsymbol{S}+c_2\hat{a}_1\boldsymbol{I}_p$ in normal and non-normal distributions, which corresponds to (I). Based on approximation of the optimal weights, we can suggest the plug-in estimator

$$\widehat{\boldsymbol{\Sigma}}_T = \frac{nT}{nT+p}\boldsymbol{S} + \Big(1 - \frac{nT}{nT+p}\Big)\hat{a}_1\boldsymbol{I}_p,$$

where $T = \hat{a}_2/\hat{a}_1^2 - 1$. It is noted that T is used as a test statistic for testing the sphericity as shown in Srivastava (2005). Namely, the hypothesis of the sphericity is accepted if T is small, but it is rejected otherwise. This tells us that the ridge-type estimator $\widehat{\boldsymbol{\Sigma}}_T$ has a reasonable form. To approximate the optimal weights and to evaluate the risk functions of the ridge-type estimators, we need to obtain the order of $\hat{a}_1 - a_1$ and $\hat{a}_2 - a_2$ as described in (III). In fact, some properties for moments of $\hat{a}_1$ and $\hat{a}_2$ are derived in Section 2, where the order of $\hat{a}_2 - a_2$ is harder to evaluate and the proof is given in Section 4.

In Section 3, we consider a class of estimators $\hat{w}_1\boldsymbol{S} + \hat{w}_2\hat{a}_1\boldsymbol{I}_p$ with random weights $\hat{w}_1$ and $\hat{w}_2$ assuming some conditions, and we show that within this class, the estimator $\widehat{\boldsymbol{\Sigma}}_T$ is the best in the sense of minimizing the leading term of the risk function. This is in reply to the query (II). The normal case is treated in Section 3.1 and an extension to the non-normal cases is given in Section 3.2.

Concerning the numerical performance of multivariate procedures based on optimal ridge-type estimator of the covariance, see, for example, Srivastava and Kubokawa (2007), Hyodo, Yamada, Himeno and Seo (2012) and Kubokawa, Hyodo and Srivastava (2013) who showed that the linear classification rule based on the optimal ridge-type estimator improves on that based on the Moore-Penrose inverse in terms of the error misclassification rates.

2 Optimal Weights and Their Estimators

Consider p-dimensional random vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ which are mutually independently and identically distributed with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{F}\boldsymbol{F}$, where $\boldsymbol{F}$ is the unique factorization of $\boldsymbol{\Sigma}$. Assume that the observation vectors $\boldsymbol{x}_j$ are generated as

$$\boldsymbol{x}_j = \boldsymbol{\mu} + \boldsymbol{F}\boldsymbol{u}_j, \quad j = 1, \ldots, N, \tag{2.1}$$

with

$$E(\boldsymbol{u}_j) = \mathbf{0}, \ \ \mathbf{Cov}\,(\boldsymbol{u}_j) = \boldsymbol{I}_p, \tag{2.2}$$

and for integers $\gamma_1, \ldots, \gamma_k$ satisfying $0 \le \sum_{k=1}^p \gamma_k \le 4$,

$$E\left[\prod_{k=1}^p u_{jk}^{\gamma_k}\right] = \prod_{k=1}^p E(u_{jk}^{\gamma_k}), \quad j = 1, \ldots, N, \tag{2.3}$$

where u_{jk} is the k^{th} component of the vector $\boldsymbol{u}_j = (u_{j1}, \ldots, u_{jk}, \ldots, u_{jp})^t$. We shall write the third and fourth moments of u_{jk} as $E[u_{jk}^3] = K_3$ and $E[u_{jk}^4] = K_4 + 3$. In the case of a normal distribution, we have $K_3 = K_4 = 0$.

In this paper, we use the notation $a_i = \mathrm{tr}\,[\boldsymbol{\Sigma}^i]/p$ for integer $i \ge 1$, and we assume the following conditions:

(A1) Both n and p tend to infinity where the relation between them is given as follows:
(A1-1) $n = O(p^\delta)$ for $0 < \delta \le 1$ in the case of $p \ge n$,
(A1-2) $p = O(n^\delta)$ for $0 < \delta \le 1$ in the case of $n > p$.

(A2) a_i converges to a positive constant for $i = 1, 2, 3, 4$. Also, $a_2 - a_1^2$ converges to a positive constant.

(A3) Let $a_{20} = \sum_{i=1}^p \sigma_{ii}^2/p$ for $\boldsymbol{\Sigma} = (\sigma_{ij})$. Then a_{20} converges to a positive constant.

Let $\hat{a}_1$ and $\hat{a}_2$ be defined in (1.3) and (1.4). In the case of normal distributions, Stivastava (2005) showed that

$$\begin{pmatrix} \hat{a}_1 \\ \hat{a}_2 \end{pmatrix} \sim \mathcal{N}_2\Big(\begin{pmatrix} a_1 \\ a_2 \end{pmatrix}, (np)^{-1}\begin{pmatrix} 2a_2 & 4a_3 \\ 4a_3 & 8a_4 + 4(p/n)a_2^2 \end{pmatrix}\Big), \tag{2.4}$$

which imples that $\hat{a}_1 - a_1 = O_p((np)^{-1/2})$ and $\hat{a}_2 - a_2 = O_p((np)^{-1/2}) + O_p(n^{-1})$ under the conditions (A1) and (A2). However, similar results are harder to derive for non-normal distributions.

Theorem 2.1 *Assume the model described in* (2.1)-(2.3) *for non-normal distributions, and the conditions* (A1)-(A3)*. Then,* $E[\hat{a}_1] = a_1$ *and*

$$Var(\hat{a}_1) = \frac{1}{Np} K_4 a_{20} + \frac{2}{np} a_2, \tag{2.5}$$

which means that $\hat{a}_1 - a_1 = O_p((np)^{-1/2})$*. Also,*

$$E[\hat{a}_2] = \frac{n}{N(N+1)} K_4 a_{20} + a_2, \tag{2.6}$$

and

$$\begin{aligned} \hat{a}_2 - a_2 =& O_p(n^{-1}p^{1/2}) + O_p((np)^{-1/2}) \\ =& O_p(n^{-1}p^{1/2}) I(p \ge n) \\ &+ \big\{O_p(n^{-1}p^{1/2}) I(\delta \ge 1/2) + O_p((np)^{-1/2}) I(\delta < 1/2)\big\} I(n > p), \end{aligned} \tag{2.7}$$

where in the case of $p \ge n$ *we need to assume that* $\delta > 1/2$.

It is hard to establish (2.7), and the proof of Theorem 2.1 is long. Thus, the proof is deferred to Section 4. As seen from (2.6), $\hat{a}_2$ is not unbiased in non-normal cases. This fact was pointed out in Srivastava, *et al.* (2011), and the same formula as in (2.6) was derived by Srivastava, *et al.* (2013). Although $\hat{a}_2$ is an unbiased estimator of a_2 in the normal case since $K_4 = 0$, in non-nomal cases $\hat{a}_2$ is not unbiased. It follows from (2.7) that $\hat{a}_2 - a_2$ satisfies

$$\hat{a}_2 - a_2 = O_p(n^{-1}p^{1/2})I(p \geq n) + O_p(n^{-1/2})I(n > p),$$

which gives a weaker and simpler order in the case of $n > p$.

We now consider the problem of estimating $\boldsymbol{\Sigma}$ by an estimator $\widehat{\boldsymbol{\Sigma}}$ relative to the quadratic loss function $L_q(\widehat{\boldsymbol{\Sigma}}, \boldsymbol{\Sigma}) = \mathrm{tr}\,[(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})^2]/p$. The risk function of $\widehat{\boldsymbol{\Sigma}}$ is given by $R(\boldsymbol{\Sigma}, \widehat{\boldsymbol{\Sigma}}) = E[\mathrm{tr}\,[(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})^2]]/p$. An unbiased estimator of $\boldsymbol{\Sigma}$ is $\boldsymbol{S} = n^{-1}\boldsymbol{V}$, where $\boldsymbol{V}$ is given in (1.1). Since $\boldsymbol{S}$ is not invertible in the case of $p > n$ nor well-conditioned in the case that p is close to n even if $n > p$, it is reasonable to consider convex combinations of $\boldsymbol{S}$ and a positive definite matrix based on $\boldsymbol{S}$. We here treat a class of linear combinations

$$\widehat{\boldsymbol{\Sigma}}_{c_1,c_2} = c_1\boldsymbol{S} + c_2\hat{a}_1\boldsymbol{I}_p,$$

where c_1 and c_2 are constants, and $\hat{a}_1 = \mathrm{tr}\,[\boldsymbol{S}]/p$. Then, the risk function is written as

$$\begin{aligned} R(\boldsymbol{\Sigma}, \widehat{\boldsymbol{\Sigma}}_{c_1,c_2}) =& E[\mathrm{tr}\,[\{c_1\boldsymbol{S} + c_2\hat{a}_1\boldsymbol{I}_p - \boldsymbol{\Sigma}\}^2]]/p \\ =& c_1^2E[\mathrm{tr}\,[\boldsymbol{S}^2]/p] + c_2^2E[\hat{a}_1^2] + a_2 + 2c_1c_2E[\hat{a}_1^2] - 2c_1a_2 - 2c_2a_1^2. \end{aligned} \tag{2.8}$$

By differentiating the risk function with respect to c_1 and c_2, it can be seen that the optimal c_1 and c_2 are given as solutions of the equations

$$\begin{aligned} &c_1E[\mathrm{tr}\,[\boldsymbol{S}^2]/p] + c_2E[\hat{a}_1^2] = a_2, \\ &c_1E[\hat{a}_1^2] + c_2E[\hat{a}_1^2] = a_1^2. \end{aligned}$$

Thus, the optimal c_1 and c_2 are

$$\begin{aligned} c_1 =& \frac{a_2 - a_1^2}{E[\mathrm{tr}\,[\boldsymbol{S}^2]/p - \hat{a}_1^2]}, \\ c_2 =& \frac{(a_1^2/E[\hat{a}_1^2])E[\mathrm{tr}\,[\boldsymbol{S}^2]/p] - a_2}{E[\mathrm{tr}\,[\boldsymbol{S}^2]/p - \hat{a}_1^2]}. \end{aligned}$$

Since $\hat{a}_2$ is defined in (1.4), $n\mathrm{tr}\,[\boldsymbol{S}^2]/p$ is written as

$$n\mathrm{tr}\,[\boldsymbol{S}^2]/p = n\hat{a}_2 + p\hat{a}_1^2 + (1 - 2/n)\hat{a}_2, \tag{2.9}$$

so that the optimal c_1 and c_2 are rewritten as

$$\begin{aligned} c_1 =& \frac{n(a_2 - a_1^2)}{E[n(\hat{a}_2 - \hat{a}_1^2) + p\hat{a}_1^2 + (1 - 2/n)\hat{a}_2]}, \\ c_2 =& \frac{(a_1^2/E[\hat{a}_1^2])E[n\hat{a}_2 + p\hat{a}_1^2 + (1 - 2/n)\hat{a}_2] - na_2}{E[n(\hat{a}_2 - \hat{a}_1^2) + p\hat{a}_1^2 + (1 - 2/n)\hat{a}_2]}. \end{aligned} \tag{2.10}$$

It follows from (2.5) that $E[\hat{a}_1^2] = a_1^2 + (Np)^{-1}K_4a_{20} + 2(np)^{-1}a_2$. Using this expectation and (2.6), we can evaluate the expectation in the denominator of c_1 and c_2 as

$$\begin{aligned} E[n(\hat{a}_2 - \hat{a}_1^2) + p\hat{a}_1^2 + (1 - 2/n)\hat{a}_2] =& n(a_2 - a_1^2) + pa_1^2 + \frac{p-2}{p}a_2 + \frac{n(p-1)}{Np}K_4a_{20} \\ =& n(a_2 - a_1^2) + pa_1^2 + O(1). \end{aligned}$$

Similarly, the expectation in the numerator of c_2 can be approximated as

$$\begin{aligned} \frac{a_1^2}{E[\hat{a}_1^2]}E[n\hat{a}_2 + p\hat{a}_1^2 + (1 - 2/n)\hat{a}_2] - na_2 =& \frac{a_1^2}{a_1^2 + O((np)^{-1})}\{na_2 + pa_1^2 + O(1)\} - na_2 \\ =& pa_1^2 + O(1). \end{aligned}$$

Hence, the optimal c_1 and c_2 can be approximated as

$$\begin{aligned} c_1 =& \frac{n\tau}{n\tau + p} + O(p^{-1})I(p \geq n) + O(n^{-1})I(n > p), \\ c_2 =& \frac{p}{n\tau + p} + O(p^{-1})I(p \geq n) + O(n^{-1})I(n > p), \end{aligned} \tag{2.11}$$

where $\tau = a_2/a_1^2 - 1$. Using these approximations, we can get the ridge-type estimator of the form

$$\widehat{\boldsymbol{\Sigma}}_\tau = \frac{n\tau}{n\tau + p}\boldsymbol{S} + \Big(1 - \frac{n\tau}{n\tau + p}\Big)\hat{a}_1\boldsymbol{I}_p. \tag{2.12}$$

Since τ is unknown, it is natural to estimate it by

$$T = \hat{a}_2/\hat{a}_1^2 - 1,$$

which suggests the estimator

$$\widehat{\boldsymbol{\Sigma}}_T = \frac{nT}{nT + p}\boldsymbol{S} + \Big(1 - \frac{nT}{nT + p}\Big)\hat{a}_1\boldsymbol{I}_p. \tag{2.13}$$

It is interesting to note that T is used by Srivastava (2005) as a test statistics for the sphericity test. Ledoit and Wolf (2004), Fisher and Sun (2011) and Hyodo, *et al.* (2012) suggested such weighted estimators. But the dominance property of these kind of estimators has not been established. In the next section, we establish this property.

Remark 2.1 As described in (4.22) in Section 4, $\hat{a}_2 - a_2$ is approximated as

$$\begin{aligned} \hat{a}_2 - a_2 =& -\frac{2(n^2+1)}{(n-1)(n+2)nN^2p}\sum_{i=1}^{N}\sum_{j\neq k}\boldsymbol{x}_i^t\boldsymbol{x}_j\boldsymbol{x}_i^t\boldsymbol{x}_k \\ &+ \frac{2}{(n-1)(n+2)nNp}\sum_{i=1}^{N}\sum_{j\neq k}\boldsymbol{x}_i^t\boldsymbol{x}_i\boldsymbol{x}_j^t\boldsymbol{x}_k \\ &+ \frac{1}{(n+2)nN^2p}\sum_{A}\boldsymbol{x}_a^t\boldsymbol{x}_b\boldsymbol{x}_c^t\boldsymbol{x}_d + O_p((np)^{-1/2})) + O_p(n^{-1}), \end{aligned} \tag{2.14}$$

where $\sum_A$ is defined below (4.4). The remainder term $O_p((np)^{-1/2})) + O_p(n^{-1})$ corresponds to the order of $\hat{a}_2 - a_2$ in the normal distribution as indicated in (2.4). If the first three terms in (2.14) were of the same order, we could establish that $\hat{a}_2 - a_2 = O_p((np)^{-1/2})) + O_p(n^{-1})$.

3 Dominance Property of Ridge-type Estimators

The ridge-type estimator $\widehat{\mathbf{\Sigma}}_T$ given in (2.13) is a plug-in estimator, and no dominance property of $\widehat{\mathbf{\Sigma}}_T$ has been known. In this paper, we show that the estimator $\widehat{\mathbf{\Sigma}}_T$ is the best of a class of weighted estimators in light of minimizing a leading term in the approximation of the risk function.

3.1 Case of normal distributions

We first assume normality for the distributions of $\boldsymbol{x}_i$'s. As a general class of weighted linear combinations of $\boldsymbol{S}$ and $\hat{a}_1\boldsymbol{I}_p$, we consider a class of the weighting functions $w_1(\boldsymbol{X})$ and $w_2(\boldsymbol{X})$ which satisfy the properties

$$\begin{aligned} w_i(\boldsymbol{X}) =& O_p(1)I(p \geq n) + O_p(1)I(n > p), \\ w_i(\boldsymbol{X}) - w_{i0} =& O_p(1)I(p \geq n) + O_p((p/n)^{1/2})I(n > p), \end{aligned} \tag{3.1}$$

for $i = 1, 2$, where $w_{i0} = w_{i0}(\mathbf{\Sigma})$ is a nonnegative function of $\mathbf{\Sigma}$. The corresponding class of estimators is given by

$$\widehat{\mathbf{\Sigma}}_{w_1,w_2} = w_1(\boldsymbol{X})\boldsymbol{S} + w_2(\boldsymbol{X})\hat{a}_1\boldsymbol{I}_p. \tag{3.2}$$

Theorem 3.1 *Assume the conditions* (A1) *and* (A2)*. Also assume that $\boldsymbol{x}_i$'s have normal distributions. Then, for weighting functions $w_1(\boldsymbol{X})$ and $w_2(\boldsymbol{x})$ satisfying* (3.1)*, the risk function of $\widehat{\mathbf{\Sigma}}_{w_1,w_2}$ is approximated as*

$$R(\mathbf{\Sigma}, \widehat{\mathbf{\Sigma}}_{w_1,w_2}) = R_0(\mathbf{\Sigma}, \widehat{\mathbf{\Sigma}}_{w_1,w_2}) + O(n^{-1})I(p \geq n) + O(n^{-1})I(n > p), \tag{3.3}$$

where the leading term $R_0(\mathbf{\Sigma}, \widehat{\mathbf{\Sigma}}_{w_1,w_2})$ is expressed as

$$\begin{aligned} R_0(\mathbf{\Sigma}, \widehat{\mathbf{\Sigma}}_{w_1,w_2}) =& E\Big[(\hat{a}_1)^2\Big\{(T + 1 + p/n)w_1^2(\boldsymbol{X}) + w_2^2(\boldsymbol{X}) + 2w_1(\boldsymbol{X})w_2(\boldsymbol{X}) \\ & - 2(T + 1)w_1(\boldsymbol{X}) - 2w_2(\boldsymbol{X}) + T + 1\Big\}\Big]. \end{aligned} \tag{3.4}$$

Proof. The risk function of $\widehat{\mathbf{\Sigma}}_{w_1,w_2}$ is written as

$$\begin{aligned} R(\mathbf{\Sigma}, \widehat{\mathbf{\Sigma}}_{w_1,w_2}) =& E\Big[\mathrm{tr}\left\{w_1(\boldsymbol{X})\boldsymbol{V}/n + w_2(\boldsymbol{X})\hat{a}_1\boldsymbol{I}_p - \mathbf{\Sigma}\right\}^2\Big]/p \\ =& E\Big[w_1^2(\boldsymbol{X})\frac{\mathrm{tr}\,[\boldsymbol{V}^2]}{n^2p} + w_2^2(\boldsymbol{X})\hat{a}_1^2 + 2w_1(\boldsymbol{X})w_2(\boldsymbol{X})\hat{a}_1^2 \\ & - 2w_1(\boldsymbol{X})\frac{\mathrm{tr}\,[\boldsymbol{V}\mathbf{\Sigma}]}{np} - 2w_2(\boldsymbol{X})\hat{a}_1a_1\Big] + a_2. \end{aligned}$$

Since $\mathrm{tr}\,[\boldsymbol{V}^2]/(n^2p) = \{(n-1)(n+2)/n^2\}\hat{a}_2 + (p/n)\hat{a}_1^2 = \hat{a}_2 + (p/n)\hat{a}_1^2 + \{(n-2)/n^2\}\hat{a}_2$, $E[\hat{a}_2] = a_2$ and $\hat{a}_2/\hat{a}_1^2 = T+1$, the risk is rewritten as

$$\begin{aligned} R(\boldsymbol{\Sigma}, \widehat{\boldsymbol{\Sigma}}_{w_1,w_2}) =& E\Big[\hat{a}_1^2\Big\{w_1^2(\boldsymbol{X})\big[T+1+\frac{p}{n}\big] + w_2^2(\boldsymbol{X}) + 2w_1(\boldsymbol{X})w_2(\boldsymbol{X}) + T + 1\Big\}\Big] \\ &+ \frac{n-2}{n^2}E[\hat{a}_1^2(T+1)w_1^2(\boldsymbol{X})] - 2E\Big[w_1(\boldsymbol{X})\frac{\mathrm{tr}\,[\boldsymbol{V\Sigma}]}{np} + w_2(\boldsymbol{X})\hat{a}_1a_1\Big]. \end{aligned} \tag{3.5}$$

It is here observed that

$$\begin{aligned} w_1(\boldsymbol{X})\frac{\mathrm{tr}\,[\boldsymbol{V\Sigma}]}{np} =& w_1(\boldsymbol{X})\Big(a_2 + \mathrm{tr}\,[(\boldsymbol{V}/n - \boldsymbol{\Sigma})\boldsymbol{\Sigma}]/p\Big) \\ =& w_1(\boldsymbol{X})\hat{a}_2 + (w_1(\boldsymbol{X}) - w_{10})\big\{-(\hat{a}_2 - a_2) + \mathrm{tr}\,[(\boldsymbol{V}/n - \boldsymbol{\Sigma})\boldsymbol{\Sigma}]/p\big\} \\ &+ w_{10}\big\{-(\hat{a}_2 - a_2) + \mathrm{tr}\,[(\boldsymbol{V}/n - \boldsymbol{\Sigma})\boldsymbol{\Sigma}]/p\big\}, \end{aligned}$$

which implies that

$$\begin{aligned} E\Big[w_1(\boldsymbol{X})\frac{\mathrm{tr}\,[\boldsymbol{V\Sigma}]}{np}\Big] =& E[w_1(\boldsymbol{X})\hat{a}_2] \\ &+ E\big[(w_1(\boldsymbol{X}) - w_{10})\big\{-(\hat{a}_2 - a_2) + \mathrm{tr}\,[(\boldsymbol{V}/n - \boldsymbol{\Sigma})\boldsymbol{\Sigma}]/p\big\}\big]. \end{aligned} \tag{3.6}$$

Similarly, $w_2(\boldsymbol{X})\hat{a}_1a_1 = w_2(\boldsymbol{X})(\hat{a}_1)^2 - w_2(\boldsymbol{X})(\hat{a}_1 - a_1)^2 - (w_2(\boldsymbol{X}) - w_{20})(\hat{a}_1 - a_1)a_1 - w_{20}(\hat{a}_1 - a_1)a_1$, so that

$$E[w_2(\boldsymbol{X})\hat{a}_1a_1] = E[w_2(\boldsymbol{X})(\hat{a}_1)^2] - E[w_2(\boldsymbol{X})(\hat{a}_1 - a_1)^2 + (w_2(\boldsymbol{X}) - w_{20})(\hat{a}_1 - a_1)a_1]. \tag{3.7}$$

Also,

$$\hat{a}_1^2w^2(\boldsymbol{X})\frac{(n-1)(n+2)}{n^2}(T+1) = \hat{a}_1^2w^2(\boldsymbol{X})(T+1) + \frac{n-2}{n^2}\hat{a}_1^2w^2(\boldsymbol{X})(T+1). \tag{3.8}$$

Combining (3.5)-(3.8) gives the expression

$$\begin{aligned} R(\boldsymbol{\Sigma}, \widehat{\boldsymbol{\Sigma}}_{w_1,w_2}) =& E\big[(\hat{a}_1)^2\big\{(T+1+p/n)w_1^2(\boldsymbol{X}) + w_2^2(\boldsymbol{X}) + 2w_1(\boldsymbol{X})w_2(\boldsymbol{X}) \\ &\qquad - 2(T+1)w_1(\boldsymbol{X}) - 2w_2(\boldsymbol{X}) + T + 1\big\}\big] \\ &+ 2E\big[(w_1(\boldsymbol{X}) - w_{10})\big\{(\hat{a}_2 - a_2) - \mathrm{tr}\,[(\boldsymbol{V}/n - \boldsymbol{\Sigma})\boldsymbol{\Sigma}]/p\big\}\big] \\ &- 2E[(w_2(\boldsymbol{X}) - w_{20})(\hat{a}_1 - a_1)a_1] + 2E[w_2(\boldsymbol{X})(\hat{a}_1 - a_1)^2] \\ &+ \frac{n-2}{n^2}E[\hat{a}_1^2w_1^2(\boldsymbol{X})(T+1)]. \end{aligned} \tag{3.9}$$

It can be demonstrated that

$$E[\big\{\mathrm{tr}\,[(\boldsymbol{V}/n - \boldsymbol{\Sigma})\boldsymbol{\Sigma}]/p\big\}^2] = 2a_4/(np),$$

so that $\mathrm{tr}\,[(\boldsymbol{V}/n - \boldsymbol{\Sigma})\boldsymbol{\Sigma}]/p = O_p((np)^{-1/2})$. It follows from (2.4) that $\hat{a}_1 - a_1 = O_p((np)^{-1/2})$ and

$$\hat{a}_2 - a_2 = O_p(n^{-1})I(p \geq n) + O_p((np)^{-1/2})I(n > p).$$

Since $w_i(\boldsymbol{X}) - w_{i0} = O_p(1)I(p \geq n) + O_p((p/n)^{1/2})I(n > p)$, $i = 1, 2$, from (3.1), it is seen that

$$E\big[(w_1(\boldsymbol{X}) - w_{10})\big\{(\hat{a}_2 - a_2) - \mathrm{tr}\,[(\boldsymbol{V}/n - \boldsymbol{\Sigma})\boldsymbol{\Sigma}]/p\big\}\big]$$
$$= O(n^{-1})I(p \geq n) + O(n^{-1})I(n > p).$$

Similarly, $E[(w_2(\boldsymbol{X}) - w_{20})(\hat{a}_1 - a_1)a_1] = O(n^{-1})I(p \geq n) + O(n^{-1})I(n > p)$. Also, from (3.1), it can be seen that $E[w_2(\boldsymbol{X})(\hat{a}_1 - a_1)^2] = O((np)^{-1})$ and

$$\frac{n-2}{n^2}E[\hat{a}_1^2 w_1^2(\boldsymbol{X})(T+1)] = O(n^{-1})I(p \geq n) + O(n^{-1})I(n > p).$$

Hence, we get the approximation (3.3). □

The leading term $R_0(\boldsymbol{\Sigma}, \widehat{\boldsymbol{\Sigma}}_{w_1,w_2})$ in (3.4) is minimized at $w_1(\boldsymbol{X}) = w^*(T)$ and $w_2(\boldsymbol{X}) = 1 - w^*(T)$ for

$$w^*(T) = \frac{nT}{nT + p}, \tag{3.10}$$

since

$$\begin{aligned}
&(T + 1 + p/n)w_1^2(\boldsymbol{X}) + w_2^2(\boldsymbol{X}) + 2w_1(\boldsymbol{X})w_2(\boldsymbol{X}) - 2(T+1)w_1(\boldsymbol{X}) - 2w_2(\boldsymbol{X}) + T + 1\\
&\quad= \begin{pmatrix} w_1(\boldsymbol{X}) - w^*(T) \\ w_2(\boldsymbol{X}) - 1 + w^*(T) \end{pmatrix}^t \begin{pmatrix} T + p/n + 1 + 1 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} w_1(\boldsymbol{X}) - w^*(T) \\ w_2(\boldsymbol{X}) - 1 + w^*(T) \end{pmatrix}\\
&\qquad+ \frac{pT}{nT + p}.
\end{aligned}$$

The weight $w^*(T)$ minimizing the risk in the leading term of (3.3) corresponds to the weight in the estimator $\widehat{\boldsymbol{\Sigma}}_T$ given in (2.13). It is noted that

$$\begin{aligned}
\frac{nT}{nT+p} &= O_p(n/p)I(p \geq n) + O_p(1)I(n > p),\\
1 - \frac{nT}{nT+p} &= O_p(1)I(p \geq n) + O_p(p/n)I(n > p).
\end{aligned} \tag{3.11}$$

Also, it can be seen that

$$\begin{aligned}
\frac{nT}{nT+p} - \frac{n\tau}{n\tau+p} &= \frac{np(T-\tau)}{(nT+p)(n\tau+p)}\\
&= \frac{np}{(nT+p)(n\tau+p)} \frac{a_1^2(\hat{a}_2 - a_2) - a_2(\hat{a}_1 - a_1)(\hat{a}_1 + a_1)}{\hat{a}_1^2 a_1^2}\\
&= O_p(p^{-1})I(p \geq n) + O_p(p^{1/2}n^{-3/2})I(n > p).
\end{aligned} \tag{3.12}$$

Thus, the optimal weight $w^*(T)$ satisfies the conditions in (3.1). Hence from Theorem 3.1, we get the following result.

Theorem 3.2 *Assume the conditions* (A1) *and* (A2) *with the normality. Among estimators with weighting functions* $w_1(\boldsymbol{X})$ *and* $w_2(\boldsymbol{X})$ *satisfying* (3.1)*, the risk function of* $\widehat{\boldsymbol{\Sigma}}_{w_1,w_2}$ *is improved on by the ridge estimator* $\widehat{\boldsymbol{\Sigma}}_T$ *given in* (2.13) *in terms of minimizing the leading term* $R_0(\boldsymbol{\Sigma}, \widehat{\boldsymbol{\Sigma}}_{w_1,w_2})$ *given in* (3.4)*. Also, the risk of* $\widehat{\boldsymbol{\Sigma}}_T$ *with the best weight* $w^*(T)$ *is evaluated as* $R(\boldsymbol{\Sigma}, \boldsymbol{\Sigma}_T) = E[pT/(nT+p)] + O(n^{-1})I(p \geq n) + O(n^{-1})I(n > p)$*, where* $E[pT/(nT+p)] = O(1)I(p \geq n) + O(p/n)I(n > p)$*.*

Estimators suggested by Fisher and Sun (2011) and Hyodo, *et al.* (2012) are slightly different from $\widehat{\boldsymbol{\Sigma}}_T$, but they have the same leading term. Thus, it may be guessed that their estimators has the same leading term in risk. In general, we consider the weighting functions $w_1(\boldsymbol{X})$ and $w_2(\boldsymbol{X})$ which can be approximated as

$$\begin{aligned} w_1(\boldsymbol{X}) =& \frac{nT}{nT+p} + O_p(p^{-1})I(p \geq n) + O_p(n^{-1})I(n > p), \\ w_2(\boldsymbol{X}) =& \frac{p}{nT+p} + O_p(n^{-1}). \end{aligned} \tag{3.13}$$

Since these functions clearly satisfy the conditions in (3.1), we can show that $R_0(\boldsymbol{\Sigma}, \widehat{\boldsymbol{\Sigma}}_{w_1,w_2})$ given in (3.4) is evaluated as

$$\begin{aligned} R_0(\boldsymbol{\Sigma}, \widehat{\boldsymbol{\Sigma}}_{w_1,w_2}) =& E[pT/(nT+p)] + \big\{O(p^{-1}) + O(n^{-1})\big\} I(p \geq n) + O(n^{-1})I(n > p) \\ =& E[pT/(nT+p)] + O(n^{-1})I(p \geq n) + O(n^{-1})I(n > p). \end{aligned}$$

Theorem 3.3 *Assume the conditions* (A1) *and* (A2) *with the normality. If the weights* $w_1(\boldsymbol{X})$ *and* $w_2(\boldsymbol{X})$ *satisfy the conditions in* (3.13)*, then the ridge-type estimators* $\widehat{\boldsymbol{\Sigma}}_{w_1,w_2}$ *and* $\widehat{\boldsymbol{\Sigma}}_T$ *have the same risk approximation, namely,* $R(\boldsymbol{\Sigma}, \widehat{\boldsymbol{\Sigma}}_{w_1,w_2}) = E[pT/(nT+p)] + O(n^{-1})I(p \geq n) + O(n^{-1})I(n > p)$*.*

For example, for the estimator of Fisher and Sun (2011), the weight function (1.2) is rewritten as

$$w_1(\boldsymbol{X}) = \hat{w} = \frac{n(\hat{a}_2 - \hat{a}_1^2)}{n(\hat{a}_2 - \hat{a}_1^2) + p\hat{a}_1^2 + \hat{a}_2} = \frac{nT}{nT + p + T + 1},$$

which is approximated as $nT/(nT+p) + O_p(np^{-2})I(p \geq n) + O_p(n^{-1})I(n > p)$. Also, it can be verified that

$$w_2(\boldsymbol{X}) = 1 - \hat{w} = \frac{p + T + 1}{nT + p + T + 1} = \frac{p}{nT+p} + O_p(p^{-1})I(p \geq n) + O_p(n^{-1})I(n > p).$$

Hence from Theorem 3.3, it follows that the estimator $\widehat{\boldsymbol{\Sigma}}_T$ and the estimator of Fisher and Sun (2011) have the same risk with respect to the leading term of risk.

Example 3.1 For example, a crude estimator of $\boldsymbol{\Sigma}$ is $\widehat{\boldsymbol{\Sigma}}_0 = \boldsymbol{S}$, and it belongs to the class (3.2) with $w_1(\boldsymbol{X}) = 1$ and $w_2(\boldsymbol{X}) = 0$. Since these weights clearly satisfy the conditions (3.1), Theorem 3.2 means that $\boldsymbol{S}$ is improved on by $\widehat{\boldsymbol{\Sigma}}_T$ under the conditions (A1) and (A2). □

Example 3.2 The optimal multiple among estimators $c\boldsymbol{S}$ can be derived from (2.8) and (2.9) by

$$c = \frac{n(\tau+1)}{(n+1-2/n)(\tau+1)+p},$$

which suggests the plug-in estimator

$$\widehat{\boldsymbol{\Sigma}}_{0B} = \frac{n(T+1-2/n)}{(n+1)(T+1)+p}\boldsymbol{S}.$$

This belongs to the class (3.2) with $w_1(\boldsymbol{X}) = n(T+1)/\{(n+1-2/n)(T+1)+p\}$ and $w_2(\boldsymbol{X}) = 0$. Similarly to (3.11) and (3.12), it can be seen that these weighst satisfy the conditions in (3.1). Thus, from Theorem 3.2, $\widehat{\boldsymbol{\Sigma}}_{0B}$ is improved on by $\widehat{\boldsymbol{\Sigma}}_T$ under the conditions (A1) and (A2). □

3.2 Case of non-normal distributions

We next extend the dominance result in the previous section to the non-normal models. In this case, we need to consider a more restricted class than (3.1), since $\hat{a}_2 - a_2 \neq O_p(n^{-1}) + O_p((np)^{-1/2})$ as shown in Theorem 2.1. Namely, we treat a class of weighting functions $w_1(\boldsymbol{X})$ and $w_2(\boldsymbol{X})$ satisfying the properties

$$\begin{aligned} w_i(\boldsymbol{X}) =& O_p(1)I(p \geq n) + O_p(1)I(n > p), \\ w_i(\boldsymbol{X}) - w_{i0} =& O_p(p^{-1/2})I(p \geq n) + O_p(n^{-1/2})I(n > p), \end{aligned} \tag{3.14}$$

for $i = 1, 2$, where $w_{i0} = w_{i0}(\boldsymbol{\Sigma})$ is a nonnegative function of $\boldsymbol{\Sigma}$. The corresponding class of estimators is given by

$$\widehat{\boldsymbol{\Sigma}}_{w_1,w_2} = w_1(\boldsymbol{X})\boldsymbol{S} + w_2(\boldsymbol{X})\hat{a}_1\boldsymbol{I}_p. \tag{3.15}$$

To establish a dominance result, we need to show the following lemma, which will be proved in Section 4.

Lemma 3.1 *Assume the conditions* (A1)-(A3)*, and the following condition:*

(A4) $\sum_{i=1}^p \{(\boldsymbol{\Sigma}^2)_{ii}\}^2/p$ *converges to a positive constant, where* $(\boldsymbol{\Sigma}^2)_{ii}$ *is the* (i,i) *element of* $\boldsymbol{\Sigma}^2$*.*

Then,

$$E[\big\{\mathrm{tr}\,[\boldsymbol{V}\boldsymbol{\Sigma}]/(np) - a_2\big\}^2] = \frac{1}{Np}K_4\sum_{i=1}^p \{(\boldsymbol{\Sigma}^2)_{ii}\}^2/p + \frac{2}{np}a_4,$$

which is of $O((np)^{-1})$*.*

Theorem 3.4 *Assume the conditions* (A1)-(A4)*. Also, it is assumed that* $\delta > 1/2$ *in the case of* $p \geq n$*. For weighting functions* $w_1(\boldsymbol{X})$ *and* $w_2(\boldsymbol{X})$ *satisfying* (3.14)*, the risk function of* $\widehat{\boldsymbol{\Sigma}}_{w_1,w_2}$ *is approximated as*

$$R(\boldsymbol{\Sigma}, \widehat{\boldsymbol{\Sigma}}_{w_1,w_2}) = R_0(\boldsymbol{\Sigma}, \widehat{\boldsymbol{\Sigma}}_{w_1,w_2}) + O(n^{-1})I(p \geq n) + O(n^{-1})I(n > p), \tag{3.16}$$

where the leading term $R_0(\boldsymbol{\Sigma}, \widehat{\boldsymbol{\Sigma}}_{w_1,w_2})$ *is given in* (3.4)*.*

Proof. Recall the risk expression given in (3.9), namely,

$$\begin{aligned}R(\boldsymbol{\Sigma}, \widehat{\boldsymbol{\Sigma}}_{w_1,w_2}) =& R_0(\boldsymbol{\Sigma}, \widehat{\boldsymbol{\Sigma}}_{w_1,w_2}) + \frac{n-2}{n^2}E[\hat{a}_1^2 w_1^2(\boldsymbol{X})(T+1)] \\ &+ 2E\big[(w_1(\boldsymbol{X}) - w_{10})\big\{(\hat{a}_2 - a_2) - \mathrm{tr}\,[(\boldsymbol{V}/n - \boldsymbol{\Sigma})\boldsymbol{\Sigma}]/p\big\}\big] \\ &- 2E[(w_2(\boldsymbol{X}) - w_{20})(\hat{a}_1 - a_1)a_1] + 2E[w_2(\boldsymbol{X})(\hat{a}_1 - a_1)^2]. \qquad (3.17)\end{aligned}$$

It follows from (2.7) and Lemma 3.1 that

$$\begin{aligned}&\hat{a}_2 - a_2 - \mathrm{tr}\,[(\boldsymbol{V}/n - \boldsymbol{\Sigma})\boldsymbol{\Sigma}]/p \\ =& O_p(n^{-1}p^{1/2})I(p \geq n) \\ &+ \{O_p(n^{-1}p^{1/2})I(\delta \geq 1/2) + O_p((np)^{-1/2})I(\delta < 1/2)\}I(n > p),\end{aligned}$$

so that from (3.14),

$$\begin{aligned}E\big[(w_1(\boldsymbol{X}) - w_{10})\big\{(\hat{a}_2 - a_2) - \mathrm{tr}\,[(\boldsymbol{V}/n - \boldsymbol{\Sigma})\boldsymbol{\Sigma}]/p\big\}\big]& \\ = O(n^{-1})I(p \geq n) + O(n^{-1})I(n > p).&\end{aligned}$$

Similarly, $E[(w_2(\boldsymbol{X}) - w_{20})(\hat{a}_1 - a_1)a_1] = O(n^{-1})I(p \geq n) + O(n^{-1})I(n > p)$. Also, from (3.14), it can be seen that $E[w_2(\boldsymbol{X})(\hat{a}_1 - a_1)^2] = O((np)^{-1})$ and

$$\frac{n-2}{n^2}E[\hat{a}_1^2 w_1^2(\boldsymbol{X})(T+1)] = O(n^{-1})I(p \geq n) + O(n^{-1})I(n > p).$$

Hence, we get the approximation (3.16). □

When we consider the weight $w^*(T) = nT/(nT + p)$ given in (3.10), it follows from (3.11) that $w^*(T)$ and $1 - w^*(T)$ satisfy the first condition in (3.1). Similarly to (3.12),

$$\begin{aligned}\frac{nT}{nT+p} - \frac{n\tau}{n\tau+p} =& \frac{np}{(nT+p)(n\tau+p)} \frac{a_1^2(\hat{a}_2 - a_2) - a_2(\hat{a}_1 - a_1)(\hat{a}_1 + a_1)}{\hat{a}_1^2 a_1^2} \\ =& O_p(p^{-1/2})I(p \geq n) + O_p(n^{-1/2})I(n > p),\end{aligned}$$

which satisfies the second condition in (3.14).

Theorem 3.5 *Assume the conditions* (A1)-(A4) *and that* $\delta > 1/2$ *in the case of* $p \geq n$. *Among estimators with weighting functions* $w_1(\boldsymbol{X})$ *and* $w_2(\boldsymbol{X})$ *satisfying* (3.14), *the risk function of* $\widehat{\boldsymbol{\Sigma}}_{w_1,w_2}$ *is improved on by the ridge estimator* $\widehat{\boldsymbol{\Sigma}}_T$ *given in* (2.13) *in terms of minimizing the leading term* $R_0(\boldsymbol{\Sigma}, \widehat{\boldsymbol{\Sigma}}_{w_1,w_2})$ *given in* (3.4). *Also, the risk of* $\widehat{\boldsymbol{\Sigma}}_T$ *with the best weight* $w^*(T)$ *is evaluated as* $R(\boldsymbol{\Sigma}, \boldsymbol{\Sigma}_T) = E[pT/(nT + p)] + O(n^{-1})I(p \geq n) + O(n^{-1})I(n > p)$, *where* $E[pT/(nT + p)] = O(1)I(p \geq n) + O(p/n)I(n > p)$.

Combining Theorems 3.4 and 3.5, we can get a similar result as in Theorem 3.3 although the detail is omitted here.

4 Proofs

In this section, we shall prove Theorem 2.1 and Lemma 3.1. Especially, we give a proof of (2.7), since (2.5) and (2.6) can be verified similarly to (2.7).

[1] Decomposition of $\hat{a}_2$. Note that $\boldsymbol{V}$ and $\boldsymbol{V}^2$ are written as $\boldsymbol{V} = \sum_{i=1}^N \boldsymbol{x}_i\boldsymbol{x}_i^t - N\overline{\boldsymbol{x}}\overline{\boldsymbol{x}}^t$ and $\boldsymbol{V}^2 = (\sum_{i=1}^N \boldsymbol{x}_i\boldsymbol{x}_i^t)^2 + N^2(\overline{\boldsymbol{x}}\overline{\boldsymbol{x}}^t)^2 - N\overline{\boldsymbol{x}}\overline{\boldsymbol{x}}^t\sum_{i=1}^N \boldsymbol{x}_i\boldsymbol{x}_i^t - N\sum_{i=1}^N \boldsymbol{x}_i\boldsymbol{x}_i^t\overline{\boldsymbol{x}}\overline{\boldsymbol{x}}^t$. Then,

$$\begin{aligned}\mathrm{tr}\,[\boldsymbol{V}^2] =& \sum_{j,k} \boldsymbol{x}_j^t\boldsymbol{x}_k\boldsymbol{x}_k^t\boldsymbol{x}_j + N^2(\overline{\boldsymbol{x}}^t\overline{\boldsymbol{x}})^2 - 2N\overline{\boldsymbol{x}}^t\sum_{i=1}^N \boldsymbol{x}_i\boldsymbol{x}_i^t\overline{\boldsymbol{x}},\\ (\mathrm{tr}\,[\boldsymbol{V}])^2 =& (\sum_{i=1}^N \boldsymbol{x}_i^t\boldsymbol{x}_i)^2 + N^2(\overline{\boldsymbol{x}}^t\overline{\boldsymbol{x}})^2 - 2N\overline{\boldsymbol{x}}^t\overline{\boldsymbol{x}}\sum_{i=1}^N \boldsymbol{x}_i^t\boldsymbol{x}_i,\end{aligned} \tag{4.1}$$

which yields that

$$\begin{aligned}\hat{a}_2 =& \frac{1}{(n-1)(n+2)p}\Big[\mathrm{tr}\,[\boldsymbol{V}^2] - \frac{1}{n}(\mathrm{tr}\,[\boldsymbol{V}])^2\Big]\\ =& \frac{1}{(n-1)(n+2)p}\Big[\sum_{i=1}^N(\boldsymbol{x}_i^t\boldsymbol{x}_i)^2 + \sum_{j\neq k}\boldsymbol{x}_j^t\boldsymbol{x}_k\boldsymbol{x}_k^t\boldsymbol{x}_j + N^2(\overline{\boldsymbol{x}}^t\overline{\boldsymbol{x}})^2 - 2N\overline{\boldsymbol{x}}^t\sum_{i=1}^N \boldsymbol{x}_i\boldsymbol{x}_i^t\overline{\boldsymbol{x}}\\ &- \frac{1}{n}\sum_{i=1}^N(\boldsymbol{x}_i^t\boldsymbol{x}_i)^2 - \frac{1}{n}\sum_{j\neq k}\boldsymbol{x}_j^t\boldsymbol{x}_j\boldsymbol{x}_k^t\boldsymbol{x}_k - \frac{N^2}{n}(\overline{\boldsymbol{x}}^t\overline{\boldsymbol{x}})^2 + 2\frac{N}{n}\overline{\boldsymbol{x}}^t\overline{\boldsymbol{x}}\sum_{i=1}^N\boldsymbol{x}_i^t\boldsymbol{x}_i\Big].\end{aligned} \tag{4.2}$$

Here, it is observed that

$$\begin{aligned}\overline{\boldsymbol{x}}^t\overline{\boldsymbol{x}}\sum_{i=1}^N\boldsymbol{x}_i^t\boldsymbol{x}_i =& \frac{1}{N^2}\sum_{i,j,k}\boldsymbol{x}_i^t\boldsymbol{x}_i\boldsymbol{x}_j^t\boldsymbol{x}_k = \frac{1}{N^2}\sum_{i,j}\boldsymbol{x}_i^t\boldsymbol{x}_i\boldsymbol{x}_j^t\boldsymbol{x}_j + \frac{1}{N^2}\sum_i\boldsymbol{x}_i^t\boldsymbol{x}_i\sum_{j\neq k}\boldsymbol{x}_j^t\boldsymbol{x}_k\\ =& \frac{1}{N^2}\sum_{i=1}^N(\boldsymbol{x}_i^t\boldsymbol{x}_i)^2 + \frac{1}{N^2}\sum_{j\neq k}\boldsymbol{x}_j^t\boldsymbol{x}_j\boldsymbol{x}_k^t\boldsymbol{x}_k + \frac{1}{N^2}\sum_i\boldsymbol{x}_i^t\boldsymbol{x}_i\sum_{j\neq k}\boldsymbol{x}_j^t\boldsymbol{x}_k,\end{aligned} \tag{4.3}$$

$$\begin{aligned}\overline{\boldsymbol{x}}^t\sum_{i=1}^N\boldsymbol{x}_i\boldsymbol{x}_i^t\overline{\boldsymbol{x}} =& \frac{1}{N^2}\sum_{i,j,k}\boldsymbol{x}_i^t\boldsymbol{x}_j\boldsymbol{x}_i^t\boldsymbol{x}_k = \frac{1}{N^2}\sum_{i,j}(\boldsymbol{x}_i^t\boldsymbol{x}_j)^2 + \frac{1}{N^2}\sum_i\sum_{j\neq k}\boldsymbol{x}_i^t\boldsymbol{x}_j\boldsymbol{x}_i^t\boldsymbol{x}_k\\ =& \frac{1}{N^2}\sum_{i=1}^N(\boldsymbol{x}_i^t\boldsymbol{x}_i)^2 + \frac{1}{N^2}\sum_{j\neq k}(\boldsymbol{x}_j^t\boldsymbol{x}_k)^2 + \frac{1}{N^2}\sum_i\sum_{j\neq k}\boldsymbol{x}_i^t\boldsymbol{x}_j\boldsymbol{x}_i^t\boldsymbol{x}_k.\end{aligned} \tag{4.4}$$

To evaluate $N^4(\overline{\boldsymbol{x}}^t\overline{\boldsymbol{x}})^2 = \sum_{a,b,c,d}\boldsymbol{x}_a^t\boldsymbol{x}_b\boldsymbol{x}_c^t\boldsymbol{x}_d$, note that $\sum_{a,b,c,d} = \sum_{a=b=c=d} + \sum_{a=c,b=d,a\neq b} + \sum_{a=c,b\neq d} + \sum_{a\neq c,b=d} + \sum_{a\neq c,b\neq d,a=b,c=d} + \sum_{a\neq c,b\neq d,a=d,b=c} + \sum_A$, where $A = \{(a,b,c,d)|a\neq c, b\neq d\}\cap\{(a,b,c,d)|a=b,c=d\}^c\cap\{(a,b,c,d)|a=d,b=c\}^c$. Then,

$$\begin{aligned}N^4(\overline{\boldsymbol{x}}^t\overline{\boldsymbol{x}})^2 =& \sum_{i=1}^N(\boldsymbol{x}_i^t\boldsymbol{x}_i)^2 + 2\sum_{j\neq k}(\boldsymbol{x}_j^t\boldsymbol{x}_k)^2 + 2\sum_{i=1}^N\sum_{j\neq k}\boldsymbol{x}_i^t\boldsymbol{x}_j\boldsymbol{x}_i^t\boldsymbol{x}_k\\ &+ \sum_{j\neq k}\boldsymbol{x}_j^t\boldsymbol{x}_j\boldsymbol{x}_k^t\boldsymbol{x}_k + \sum_A\boldsymbol{x}_a^t\boldsymbol{x}_b\boldsymbol{x}_c^t\boldsymbol{x}_d.\end{aligned} \tag{4.5}$$

The last term can be further rewritten as

$$\sum_A \boldsymbol{x}_a^t\boldsymbol{x}_b\boldsymbol{x}_c^t\boldsymbol{x}_d = 4\sum_{i\neq j\neq k}(\boldsymbol{x}_i^t\boldsymbol{x}_i\boldsymbol{x}_j^t\boldsymbol{x}_k + \boldsymbol{x}_i^t\boldsymbol{x}_j\boldsymbol{x}_i^t\boldsymbol{x}_k) + 4\sum_D \boldsymbol{x}_a^t\boldsymbol{x}_b\boldsymbol{x}_c^t\boldsymbol{x}_d, \tag{4.6}$$

where $\sum_{i\neq j\neq k}$ means that the summation is taken for mutually different i, j, k, and $\sum_D$ means that the summation is taken for mutually different a, b, c, d.

Substituting these expressions (4.3)-(4.5) into (4.2), we get the expression

$$\hat{a}_2 = [(n-1)(n+2)p]^{-1}\{I_1 + I_2\},$$

where

$$\begin{aligned} I_1 =& \Big(1 - \frac{1}{n} - \frac{2}{N} + \frac{2}{nN} + \frac{1}{N^2} - \frac{1}{nN^2}\Big)\sum_{i=1}^N (\boldsymbol{x}_i^t\boldsymbol{x}_i)^2 \\ &+ \Big(-\frac{1}{n} + \frac{2}{nN} + \frac{1}{N^2} - \frac{1}{nN^2}\Big)\sum_{j\neq k}\boldsymbol{x}_j^t\boldsymbol{x}_j\boldsymbol{x}_k^t\boldsymbol{x}_k, \end{aligned} \tag{4.7}$$

$$\begin{aligned} I_2 =& \Big(1 - \frac{2}{N} + \frac{2}{N^2} - \frac{2}{nN^2}\Big)\sum_{j\neq k}(\boldsymbol{x}_j^t\boldsymbol{x}_k)^2 + \Big(-\frac{2}{n} + \frac{2}{N^2} - \frac{2}{nN^2}\Big)\sum_{i=1}^N\sum_{j\neq k}\boldsymbol{x}_i^t\boldsymbol{x}_j\boldsymbol{x}_i^t\boldsymbol{x}_k \\ &+ \frac{2}{nN}\sum_{i=1}^N\sum_{j\neq k}\boldsymbol{x}_i^t\boldsymbol{x}_i\boldsymbol{x}_j^t\boldsymbol{x}_k + \Big(\frac{1}{N^2} - \frac{1}{nN^2}\Big)\sum_A \boldsymbol{x}_a^t\boldsymbol{x}_b\boldsymbol{x}_c^t\boldsymbol{x}_d. \end{aligned} \tag{4.8}$$

Noting that $\sum_{j\neq k}\boldsymbol{x}_j^t\boldsymbol{x}_j\boldsymbol{x}_k^t\boldsymbol{x}_k = (\sum_{i=1}^N \boldsymbol{x}_i^t\boldsymbol{x}_i)^2 - \sum_{i=1}^N(\boldsymbol{x}_i^t\boldsymbol{x}_i)^2$, we can get a simple expression of I_1 as

$$I_1 = \frac{n-1}{N}\Big\{\sum_{i=1}^N(\boldsymbol{x}_i^t\boldsymbol{x}_i)^2 - \frac{1}{N}\Big(\sum_{i=1}^N \boldsymbol{x}_i^t\boldsymbol{x}_i\Big)^2\Big\} = \frac{n-1}{N}\sum_{i=1}^N\Big\{\boldsymbol{x}_i^t\boldsymbol{x}_i - N^{-1}\sum_{j=1}^N \boldsymbol{x}_j^t\boldsymbol{x}_j\Big\}^2. \tag{4.9}$$

Also, I_2 can be rewritten as

$$\begin{aligned} I_2 =& \frac{n-1}{N}(1 + \frac{2}{nN})\sum_{j\neq k}(\boldsymbol{x}_j^t\boldsymbol{x}_k)^2 - 2\frac{n^2+1}{nN^2}\sum_{i=1}^N\sum_{j\neq k}\boldsymbol{x}_i^t\boldsymbol{x}_j\boldsymbol{x}_i^t\boldsymbol{x}_k \\ &+ \frac{2}{nN}\sum_{i=1}^N\sum_{j\neq k}\boldsymbol{x}_i^t\boldsymbol{x}_i\boldsymbol{x}_j^t\boldsymbol{x}_k + \frac{n-1}{nN^2}\sum_A \boldsymbol{x}_a^t\boldsymbol{x}_b\boldsymbol{x}_c^t\boldsymbol{x}_d. \end{aligned} \tag{4.10}$$

The expressions (4.9) and (4.10) are used to evaluate the moments of $\hat{a}_2 = [(n-1)(n+2)p]^{-1}\{I_1 + I_2\}$.

[2] Expectation of $\hat{a}_2$. For the proofs of (2.6) and (2.7), we use the following moment given in Srivastava, *et al.* (2013):

$$E[\boldsymbol{z}^t\boldsymbol{A}\boldsymbol{z}\boldsymbol{z}^t\boldsymbol{B}\boldsymbol{z}] = K_4\sum_{i=1} a_{ii}b_{ii} + 2\mathrm{tr}\,[\boldsymbol{A}\boldsymbol{B}] + \mathrm{tr}\,[\boldsymbol{A}]\mathrm{tr}\,[\boldsymbol{B}], \tag{4.11}$$

where $\boldsymbol{z} = (z_1, \ldots, z_p)^t$ is a random vector with $E[\boldsymbol{z}] = \mathbf{0}$, $\mathbf{Cov}\,(\boldsymbol{z}) = \boldsymbol{I}_p$ and $E[z_i^4] = K_4 + 3$, and $\boldsymbol{A} = (a_{ij})$ and $\boldsymbol{B} = (b_{ij})$ are $p \times p$ matrices of constants.

To evaluate the terms in (4.9) and (4.10), it can be demonstrated that

$$\begin{aligned} \sum_{i=1}^{N} E[(\boldsymbol{x}_i^t \boldsymbol{x}_i)^2] =& N E[(\boldsymbol{x}_1^t \boldsymbol{x}_1)^2] = N\{K_4 p a_{20} + 2\mathrm{tr}\,[\boldsymbol{\Sigma}^2] + (\mathrm{tr}\,[\boldsymbol{\Sigma}])^2\} \equiv N\beta, & (4.12) \\ E[(\sum_{i=1}^{N} \boldsymbol{x}_i^t \boldsymbol{x}_i)^2] =& E[\sum_{i=1}^{N} (\boldsymbol{x}_i^t \boldsymbol{x}_i)^2 + \sum_{j\neq k} \boldsymbol{x}_j^t \boldsymbol{x}_j \boldsymbol{x}_k^t \boldsymbol{x}_k] = N\beta + Nn(\mathrm{tr}\,[\boldsymbol{\Sigma}])^2, & (4.13) \\ \sum_{j\neq k} E[(\boldsymbol{x}_j^t \boldsymbol{x}_k)^2] =& NnE[\boldsymbol{x}_1^t \boldsymbol{x}_2 \boldsymbol{x}_2^t \boldsymbol{x}_1] = Nn\mathrm{tr}\,[\boldsymbol{\Sigma}^2]. & (4.14) \end{aligned}$$

It can be also seen that $\sum_{i=1}^{N} \sum_{j\neq k} E[\boldsymbol{x}_i^t \boldsymbol{x}_j \boldsymbol{x}_i^t \boldsymbol{x}_k] = 0$, $\sum_{i=1}^{N} \sum_{j\neq k} E[\boldsymbol{x}_i^t \boldsymbol{x}_i \boldsymbol{x}_j^t \boldsymbol{x}_k] = 0$ and $\sum_{A} E[\boldsymbol{x}_a^t \boldsymbol{x}_b \boldsymbol{x}_c^t \boldsymbol{x}_d] = 0$ from (4.6). Using these observations, we can evaluate the expectaions $E[I_1]$ and $E[I_2]$ as

$$\begin{aligned} E[I_1] =& (n-1)N^{-1}\{N\beta - \beta - n(\mathrm{tr}\,[\boldsymbol{\Sigma}])^2\} = (n-1)nN^{-1}\{\beta - (\mathrm{tr}\,[\boldsymbol{\Sigma}])^2\} \\ =& (n-1)nN^{-1}p(K_4 a_{20} + 2a_2), & (4.15) \\ E[I_2] =& (n-1)N^{-1}(nN+2)\mathrm{tr}\,[\boldsymbol{\Sigma}^2] = (n-1)N^{-1}(nN+2)pa_2, \end{aligned}$$

which implies that

$$E[\hat{a}_2] = \frac{n}{(n+2)N} K_4 a_{20} + a_2. \tag{4.16}$$

This was shown in Srivastava, *et al.* (2013).

[3] Evaluation of I_1. We next investigate the order of $\hat{a}_2 - a_2$. For I_1, it is noted from (4.9) that $I_1 > 0$. Also from (4.15), $E[I_1] = O(np)$. Then from the Markov inequality, we see that for any $\varepsilon > 0$,

$$P(I_1 > \varepsilon) < E[I_1]/\varepsilon = O(np), \tag{4.17}$$

which means that $I_1 = O_p(np)$, namely,

$$\frac{1}{(n+2)Np}\Big\{\sum_{i=1}^{N} (\boldsymbol{x}_i^t \boldsymbol{x}_i)^2 - \frac{1}{N}\big(\sum_{i=1}^{N} \boldsymbol{x}_i^t \boldsymbol{x}_i\big)^2\Big\} = O_p(n^{-1}). \tag{4.18}$$

[4] Evaluation of $(Nnp)^{-1}\sum_{j\neq k}(\boldsymbol{x}_j^t \boldsymbol{x}_k)^2 - a_2$. Similarly to (4.17), it can be shown that $\sum_{j\neq k}(\boldsymbol{x}_j^t \boldsymbol{x}_k)^2 = O_p(Nnp)$, so that

$$\frac{1}{N(n+2)p}(1 + \frac{2}{nN}) \sum_{j\neq k} (\boldsymbol{x}_j^t \boldsymbol{x}_k)^2 - a_2 = \frac{1}{Nnp} \sum_{j\neq k} (\boldsymbol{x}_j^t \boldsymbol{x}_k)^2 - a_2 + O_p(n^{-1}). \tag{4.19}$$

Thus, we shall show that

$$(Nnp)^{-1} \sum_{j\neq k} (\boldsymbol{x}_j^t \boldsymbol{x}_k)^2 - a_2 = O_p(n^{-1}).$$

Since $E[(Nnp)^{-1}\sum_{j\neq k}(\boldsymbol{x}_j^t\boldsymbol{x}_k)^2]=a_2$, it is observed that

$$E\Big[\Big\{(Nnp)^{-1}\sum_{j\neq k}(\boldsymbol{x}_j^t\boldsymbol{x}_k)^2-a_2\Big\}^2\Big]=(Nnp)^{-2}E\Big[\Big\{\sum_{j\neq k}(\boldsymbol{x}_j^t\boldsymbol{x}_k)^2\Big\}^2\Big]-a_2^2. \tag{4.20}$$

It is here note that

$$\sum_{a\neq b}\sum_{c\neq d}(\boldsymbol{x}_a^t\boldsymbol{x}_b)^2(\boldsymbol{x}_c\boldsymbol{x}_d)^2=2\sum_{a\neq b}(\boldsymbol{x}_a^t\boldsymbol{x}_b)^4+4\sum_{a\neq b\neq c}(\boldsymbol{x}_a^t\boldsymbol{x}_b)^2(\boldsymbol{x}_c^t\boldsymbol{x}_b)^2+\sum_D(\boldsymbol{x}_a^t\boldsymbol{x}_b)^2(\boldsymbol{x}_c^t\boldsymbol{x}_d)^2,$$

where $\sum_{a\neq b\neq c}$ means that the summation is taken for mutually different a,b,c, and $\sum_D$ means that the summation is taken for mutually different a,b,c,d. Thus,

$$\begin{aligned}E[\frac{1}{(Nnp)^2}\sum_{a\neq b}(\boldsymbol{x}_a^t\boldsymbol{x}_b)^4]=&\frac{1}{Nnp^2}E[(\boldsymbol{x}_1^t\boldsymbol{x}_2\boldsymbol{x}_2^t\boldsymbol{x}_1)(\boldsymbol{x}_1^t\boldsymbol{x}_2\boldsymbol{x}_2^t\boldsymbol{x}_1)]\\=&\frac{1}{Nnp^2}E[K_4\sum_{i=1}^p\{(\boldsymbol{\Sigma}^{1/2}\boldsymbol{x}_2\boldsymbol{x}_2^t\boldsymbol{\Sigma}^{1/2})_{ii}\}^2+3(\boldsymbol{x}_2^t\boldsymbol{\Sigma}\boldsymbol{x}_2)^2]\\\leq&\frac{1}{Nn}\Big\{K_4E[(\boldsymbol{x}_2^t\boldsymbol{\Sigma}\boldsymbol{x}_2)^2]/p^2+3K_4\sum_{i=1}^N\{(\boldsymbol{\Sigma}^2)_{ii}\}^2/p^2+6a_4/p+3a_2^2\Big\}\\=&O(n^{-2}),\end{aligned}$$

since $\sum_{i=1}^p\{(\boldsymbol{\Sigma}^{1/2}\boldsymbol{x}_2\boldsymbol{x}_2^t\boldsymbol{\Sigma}^{1/2})_{ii}\}^2\leq\mathrm{tr}\,[(\boldsymbol{\Sigma}^{1/2}\boldsymbol{x}_2\boldsymbol{x}_2^t\boldsymbol{\Sigma}^{1/2})^2]=(\boldsymbol{x}_2\boldsymbol{\Sigma}\boldsymbol{x}_2)^2$. Also,

$$\begin{aligned}E[\frac{1}{(Nnp)^2}\sum_{a\neq b\neq c}(\boldsymbol{x}_a^t\boldsymbol{x}_b)^2(\boldsymbol{x}_c^t\boldsymbol{x}_b)^2]=&\frac{1}{Np^2}E[(\boldsymbol{x}_2^t\boldsymbol{x}_1\boldsymbol{x}_1^t\boldsymbol{x}_2)(\boldsymbol{x}_2^t\boldsymbol{x}_1\boldsymbol{x}_1^t\boldsymbol{x}_2)]=\frac{1}{Np^2}E[(\boldsymbol{x}_2^t\boldsymbol{\Sigma}\boldsymbol{x}_2)^2]\\=&\frac{1}{n}\Big\{K_4\sum_{i=1}^p\{(\boldsymbol{\Sigma}^2)_{ii}\}^2/p^2+2a_4/p+a_2^2\Big\}.\end{aligned}$$

Finally,

$$\begin{aligned}E[\frac{1}{(Nnp)^2}\sum_D(\boldsymbol{x}_a^t\boldsymbol{x}_b)^2(\boldsymbol{x}_c^t\boldsymbol{x}_d)^2]=&\frac{N(N-1)(N-2)(N-3)}{(Nnp)^2}E[(\boldsymbol{x}_1^t\boldsymbol{x}_2\boldsymbol{x}_2^t\boldsymbol{x}_1)(\boldsymbol{x}_3^t\boldsymbol{x}_4\boldsymbol{x}_4^t\boldsymbol{x}_3)]\\=&\frac{(N-2)(N-3)}{Nnp^2}E[\boldsymbol{x}_1^t\boldsymbol{\Sigma}\boldsymbol{x}_1]E[\boldsymbol{x}_3^t\boldsymbol{\Sigma}\boldsymbol{x}_3]\\=&\frac{(N-2)(N-3)}{Nn}a_2^2=a_2^2-\frac{4}{n}a_2^2+O(n^{-2}).\end{aligned}$$

Combining these observations gives that

$$\begin{aligned}(Nnp)^{-2}E\Big[\Big\{\sum_{j\neq k}(\boldsymbol{x}_j^t\boldsymbol{x}_k)^2\Big\}^2\Big]-a_2^2=&\frac{1}{n}\Big\{K_4\sum_{i=1}^p\{(\boldsymbol{\Sigma}^2)_{ii}\}^2/p^2+2a_4/p\Big\}+O(n^{-2})\\=&O((np)^{-1})+O(n^{-2}).\end{aligned}$$

Then from (4.19),

$$\frac{1}{N(n+2)p}(1+\frac{2}{nN})\sum_{j\neq k}(\boldsymbol{x}_j^t\boldsymbol{x}_k)^2 - a_2 = O_p((np)^{-1/2})) + O_p(n^{-1}). \tag{4.21}$$

[5] Evaluation of $\hat{a}_2 - a_2$. It follows from (4.10), (4.18) and (4.21) that

$$\begin{aligned}\hat{a}_2 - a_2 =& -\frac{2(n^2+1)}{(n-1)(n+2)nN^2p}\sum_{i=1}^N\sum_{j\neq k}\boldsymbol{x}_i^t\boldsymbol{x}_j\boldsymbol{x}_i^t\boldsymbol{x}_k\\ &+\frac{2}{(n-1)(n+2)nNp}\sum_{i=1}^N\sum_{j\neq k}\boldsymbol{x}_i^t\boldsymbol{x}_i\boldsymbol{x}_j^t\boldsymbol{x}_k\\ &+\frac{1}{(n+2)nN^2p}\sum_A\boldsymbol{x}_a^t\boldsymbol{x}_b\boldsymbol{x}_c^t\boldsymbol{x}_d + O_p((np)^{-1/2})) + O_p(n^{-1}).\end{aligned} \tag{4.22}$$

In (4.22), we shall evaluate the term $\sum_{i=1}^N\sum_{j\neq k}\boldsymbol{x}_i^t\boldsymbol{x}_i\boldsymbol{x}_j^t\boldsymbol{x}_k$, which is

$$\sum_{i=1}^N\boldsymbol{x}_i^t\boldsymbol{x}_i\sum_{j\neq k}\boldsymbol{x}_j^t\boldsymbol{x}_k = 2\sum_{j\neq k}\boldsymbol{x}_j^t\boldsymbol{x}_j\boldsymbol{x}_j^t\boldsymbol{x}_k + \sum_{i\neq j\neq k}\boldsymbol{x}_i^t\boldsymbol{x}_i\boldsymbol{x}_j^t\boldsymbol{x}_k. \tag{4.23}$$

Since $E[\sum_{i=1}^N\boldsymbol{x}_i^t\boldsymbol{x}_i] = Npa_1$, it follows that $\sum_{i=1}^N\boldsymbol{x}_i^t\boldsymbol{x}_i = O_p(np)$. Noting that

$$E[\{\sum_{j\neq k}\boldsymbol{x}_j^t\boldsymbol{x}_k\}^2] = 2E[\sum_{j\neq k}(\boldsymbol{x}_j^t\boldsymbol{x}_k)^2] = 2Nnpa_2 = O(n^2p),$$

we can see that

$$\sum_{i=1}^N\boldsymbol{x}_i^t\boldsymbol{x}_i\sum_{j\neq k}\boldsymbol{x}_j^t\boldsymbol{x}_k = O_p(n^2p^{3/2}), \tag{4.24}$$

which gives that

$$\frac{2}{(n-1)(n+2)nNp}\sum_{i=1}^N\sum_{j\neq k}\boldsymbol{x}_i^t\boldsymbol{x}_i\boldsymbol{x}_j^t\boldsymbol{x}_k = O_p(n^{-2}p^{1/2}). \tag{4.25}$$

On the other hand,

$$\begin{aligned}E\Big[\Big\{\sum_{i\neq j\neq k}\boldsymbol{x}_i^t\boldsymbol{x}_i\boldsymbol{x}_j^t\boldsymbol{x}_k\Big\}^2\Big] =&2\sum_{i\neq j\neq k}E[\{\boldsymbol{x}_i^t\boldsymbol{x}_i\boldsymbol{x}_j^t\boldsymbol{x}_k\}^2]\\ =&2N(N-1)(N-2)E[(\boldsymbol{x}_1^t\boldsymbol{x}_1)^2]E[(\boldsymbol{x}_2^t\boldsymbol{x}_3)^2]\\ =&2N(N-1)(N-2)\beta\mathrm{tr}\,[\boldsymbol{\Sigma}^2] = O(n^3p^3),\end{aligned}$$

which shows that $\sum_{i\neq j\neq k}\boldsymbol{x}_i^t\boldsymbol{x}_i\boldsymbol{x}_j^t\boldsymbol{x}_k = O_p((np)^{3/2})$. Hence, from (4.23),

$$\sum_{j\neq k}\boldsymbol{x}_j^t\boldsymbol{x}_j\boldsymbol{x}_j^t\boldsymbol{x}_k = O_p(n^2p^{3/2}) - O_p((np)^{3/2}) = O_p(n^2p^{3/2}). \tag{4.26}$$

Using (4.26), we evaluate the term $\sum_{i=1}^N\sum_{j\neq k}\boldsymbol{x}_i^t\boldsymbol{x}_j\boldsymbol{x}_i^t\boldsymbol{x}_k$ in (4.22), which is

$$\sum_{i=1}^N\sum_{j\neq k}\boldsymbol{x}_i^t\boldsymbol{x}_j\boldsymbol{x}_i^t\boldsymbol{x}_k = 2\sum_{j\neq k}\boldsymbol{x}_j^t\boldsymbol{x}_j\boldsymbol{x}_j^t\boldsymbol{x}_k + \sum_{i\neq j\neq k}\boldsymbol{x}_i^t\boldsymbol{x}_j\boldsymbol{x}_i^t\boldsymbol{x}_k.$$

Since it can be verified that $\sum_{i\neq j\neq k}\boldsymbol{x}_i^t\boldsymbol{x}_j\boldsymbol{x}_i^t\boldsymbol{x}_k = O_p((np)^{3/2})$, from (4.26), it is observed that

$$\sum_{i=1}^N\sum_{j\neq k}\boldsymbol{x}_i^t\boldsymbol{x}_j\boldsymbol{x}_i^t\boldsymbol{x}_k = O_p(n^2p^{3/2}), \tag{4.27}$$

which leads to

$$\frac{2(n^2+1)}{(n-1)(n+2)nN^2p}\sum_{i=1}^N\sum_{j\neq k}\boldsymbol{x}_i^t\boldsymbol{x}_j\boldsymbol{x}_i^t\boldsymbol{x}_k = O_p(n^{-1}p^{1/2}). \tag{4.28}$$

Finally, from (4.6),

$$\begin{aligned}\sum_A\boldsymbol{x}_a^t\boldsymbol{x}_b\boldsymbol{x}_c^t\boldsymbol{x}_d =&4\sum_{i\neq j\neq k}(\boldsymbol{x}_i^t\boldsymbol{x}_i\boldsymbol{x}_j^t\boldsymbol{x}_k + \boldsymbol{x}_i^t\boldsymbol{x}_j\boldsymbol{x}_i^t\boldsymbol{x}_k) + 4\sum_D\boldsymbol{x}_a^t\boldsymbol{x}_b\boldsymbol{x}_c^t\boldsymbol{x}_d\\ =&O_p((np)^{3/2}) + O_p(n^2p),\end{aligned}$$

which implies that

$$\frac{1}{(n+2)nN^2p}\sum_A\boldsymbol{x}_a^t\boldsymbol{x}_b\boldsymbol{x}_c^t\boldsymbol{x}_d = O_p(n^{-5/2}p^{1/2}) + O_p(n^{-2}). \tag{4.29}$$

Thus, combining (4.22), (4.21), (4.25), (4.28) and (4.29), we can see that

$$\hat{a}_2 - a_2 = O_p(n^{-1}p^{1/2}) + O_p((np)^{-1/2}). \tag{4.30}$$

[6] Variance of $\hat{a}_1$. Concerning the variance of $\hat{a}_1$, from (4.1),

$$E[(\hat{a}_1-a_1)^2] = \frac{1}{(np)^2}E\Big[(\sum_{i=1}^N\boldsymbol{x}_i^t\boldsymbol{x}_i)^2 + N^2(\overline{\boldsymbol{x}}^t\overline{\boldsymbol{x}})^2 - 2N\overline{\boldsymbol{x}}^t\overline{\boldsymbol{x}}\sum_{i=1}^N\boldsymbol{x}_i^t\boldsymbol{x}_i\Big] - a_1^2.$$

We can use (4.3), (4.5), (4.12), (4.13) and (4.14) to evaluate the moments of the variance, and we get

$$\begin{aligned}E[(\hat{a}_1-a_1)^2] =&\frac{1}{(np)^2}\Big\{\frac{n^2}{N}\beta + \frac{2n}{N}pa_2(Nn + \frac{n}{N} - 2n)p^2a_1^2\Big\} - a_1^2\\ =&\frac{1}{Np}K_4a_{20} + \frac{2}{np}a_2,\end{aligned}$$

which is order $O((np)^{-1})$.

[7] Proof of Lemma 3.1. It is observed that

$$E[\{\mathrm{tr}\,[\boldsymbol{V\Sigma}]/(np) - a_2\}^2] = E[\{\mathrm{tr}\,[\boldsymbol{V\Sigma}]\}^2]/(np)^2 - a_2^2,$$

and that $\{\mathrm{tr}\,[\boldsymbol{V\Sigma}]\}^2 = \left\{\sum_{i=1}^N \boldsymbol{x}_i^t\boldsymbol{\Sigma}\boldsymbol{x}_i\right\}^2 - 2N\overline{\boldsymbol{x}}^t\boldsymbol{\Sigma}\overline{\boldsymbol{x}}\sum_{i=1}^N \boldsymbol{x}_i^t\boldsymbol{\Sigma}\boldsymbol{x}_i + N^2(\overline{\boldsymbol{x}}^t\boldsymbol{\Sigma}\overline{\boldsymbol{x}})^2$. To calculate the expectation, we can use similar decompositions and evaluations as in (4.3), (4.4), (4.5), (4.11), (4.12), (4.13) and (4.14). Then, we can prove Lemma 3.1.

5 Concluding Remarks

As invertible and well-conditioned estimators of a large covariance matrix, the plug-in estimators based on the optimal convex combination of $\boldsymbol{S}$ and $a_1\boldsymbol{I}_p$ have been suggested in the literature. However, the plug-in estimators can not necessarily be guaranteed to be optimal because there exist correlations between the random weights and $\boldsymbol{S}$ and $\hat{a}_1$. In this paper, we have shown that the plug-in estimators are optimal properties within a class of estimators with random convex combinations in the sense of minimizing the leading term of the risk approximation. This dominance property has been established not only for a normal distribution, but also for non-normal distributions. In the case of non-normal distributions, we have obtained the order of $\hat{a}_2 - a_2$, which is necessary for proving the dominance property.

Through normal and non-normal distributions, it is seen that $\hat{a}_1 - a_1 = O_p((np)^{-1/2})$. However, it is harder to evaluate $\hat{a}_2 - a_2$ for non-normal distributions. As shown in Srivastava (2005), $\hat{a}_2 - a_2 = O_p((np)^{-1/2}) + O_p(n^{-1})$ for a normal distribution. In the case of non-normal distributions, however, $\hat{a}_2 - a_2 = O_p(n^{-1}p^{1/2}) + O_p((np)^{-1/2})$ as shown in Theorem 2.1. Especially, we need to assume that $\delta > 1/2$ for consistency of $\hat{a}_2$ in the case of $p \geq n$ and $n = O(p^\delta)$ for $0 < \delta \leq 1$. As described in Remark 2.1, we hope that the order of $\hat{a}_2 - a_2$ will be improved on in a future.

Acknowledgments.

Research of the first author was supported in part by Grant-in-Aid for Scientific Research (21540114 and 23243039) from Japan Society for the Promotion of Science.

References

[1] Bai, J., and Shi, S. (2011). Estimating high dimensional covariance matrices and its applications. *Ann. Economics and finance*, **12**, 199-215.

[2] Daniels, M., and Kass, R.E. (2001). Shrinkage estimators for covariance matrices. *Biometrics*, **57**, 1173-1184.

[3] Fan, J., and Fan, Y. (2008). High-dimensional classification using features annealed independence rules. *Ann. Statist.*, **36**, 2605-2637.

[4] Fan, J., Fan, Y. and Lv, J. (2008). High dimensional covariance matrix estimation using a factor model. *J. Econometrics*, **147**, 186-197.

[5] Fisher, T.J., and Sun, X. (2011). Improved Stein-type shrinkage estimators for the high-dimensional multivariate normal covariance matrix. *Comp. Statist. Data Analysis*, **55**, 1909-1918.

[6] Hyodo, M., Yamada, T., Himeno, T., and Seo, T. (2012). A modified discriminant analysis for high-dimensional data. *Hiroshima Math. J.*, **42**, 209-231.

[7] Konno, Y. (2009). Shrinkage estimators for large covariance matrices in multivariate real and complex normal distributions under an invariant quadratic loss. *J. Multivariate Analysis*, **100**, 2237-2253.

[8] Kubokawa, T., Hyodo, M, and Srivastava, M.S. (2013). Asymptotic expansion and estimation of EPMC for linear classification rules in high dimension. *J. Multivariate Analysis*, **115**, 496-515.

[9] Ledoit, O., and Wolf, M. (2003). Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *J. Empirical Finance*, **10**, 603-621.

[10] Ledoit, O., and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *J. Multivariate Analysis*, **88**, 365-411.

[11] Srivastava, M.S. (2005). Some tests concerning the covariance matrix in high dimensional data. *J. Japan Statist. Soc.*, **35**, 251-272.

[12] Srivastava, M.S., Kollo, T., and von Rosen, D. (2011). Some tests for the covariance matrix with fewer observations than the dimension under non-normality. *J. Multivariate Analysis*, **102**, 1090-1103.

[13] Srivastava, M.S., and Kubokawa, T. (2007). Comparison of discrimination methods for high dimensional data. *J. Japan Statist. Soc.*, **37**, 123-134.

[14] Srivastava, M.S., Yanagihara, H., and Kubokawa, T. (2013). Tests for covariance matrices in high dimension with less sample size. Unpublished manuscript.