Applications of Mathematics

Yi Zhou; Bin Zhang

Ridge estimation of covariance matrix from data in two classes

Applications of Mathematics, Vol. 69 (2024), No. 2, 169–184

Persistent URL: http://dml.cz/dmlcz/152311

69 (2024) APPLICATIONS OF MATHEMATICS No. 2, 169–184

RIDGE ESTIMATION OF COVARIANCE MATRIX FROM DATA IN TWO CLASSES

Yi Zhou, Bin Zhang, Guilin

Received August 2, 2023. Published online February 22, 2024.

Abstract. This paper deals with the problem of estimating a covariance matrix from the data in two classes: (1) good data with the covariance matrix of interest and (2) contamination coming from a Gaussian distribution with a different covariance matrix. The ridge penalty is introduced to address the problem of high-dimensional challenges in estimating the covariance matrix from the two-class data model. A ridge estimator of the covariance matrix has a uniform expression and keeps positive-definite, whether the data size is larger or smaller than the data dimension. Furthermore, the ridge parameter is tuned through a cross-validation procedure. Lastly, the proposed ridge estimator is verified with better performance than the existing estimator from the data in two classes and the traditional ridge estimator only from the good data.

Keywords: covariance matrix; ridge estimation; two-class data; contamination

MSC 2020: 62H12, 62J07

1. Introduction

The covariance matrix is an imperative quantity for describing the dispersion of the data and measuring the linear correlation between each pair of variables during data processing. As is well known, it is an oracle estimator and always needs to be estimated from a finite sample in practice [12], [34], [32], [25]. Therefore, the problem of estimating a covariance matrix has aroused concern during the last two decades in a wide range of scientific fields, such as linear discriminant analysis [15], [19], high-dimensional regression [31], [14], large portfolio optimization[6], gene expression data analysis [16], and multi-source information fusion [21].

The research has been supported by the Guangxi Science and Technology Planning Project (Guike AD23026220) and the Science and Technology Project of Guangxi (Guike AD21220114).

DOI: 10.21136/AM.2024.0157-23

Numerous popular methods have been developed to estimate the covariance matrix. When the data is low-dimensional with a sufficient sample, the traditional sample covariance matrix is a widely adopted estimator. Under Gaussian assumption, it is the maximum likelihood estimator and enjoys many desired statistical properties. However, the sample eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_p$ follow the famous Marčenko-Pastur law when the dimension proportionally increases with the data size, which deviates from the population distribution of the true eigenvalues [9], [22]. It brings about the traditional sample covariance matrix being badly behaved when the data dimension is large compared to or larger than the data size and the concerned statistical methods and algorithms becoming invalid where a reliable or positive-definite estimator is indispensable [28], [35], [38]. Therefore, extensive attention has been attracted to finding a covariance matrix estimator for both the large-sample and the high-dimensional situations [18], [4], [29].

1.1. Good data case. Most existing estimators are developed from an identically distributed sample with the same covariance matrix, namely, the good data [33], [10], [5]. A direct strategy to produce an improved covariance matrix estimator is regularizing the sample eigenvalues, which can be traced back to Stein's estimation [8], [17], [37]. The classic ridge estimation is one of the representative methods based on this strategy [27], [36]. From an optimization perspective, the ridge estimation can be regarded as adding a penalty term to the loss function [13], [30]. Denote the sample covariance matrix as $\mathbf{S}_\mathbf{x}$ with the spectral decomposition $\mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$, where $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_p)$ is a diagonal matrix containing the sample eigenvalues and $\mathbf{U}$ is a unitary matrix consisting of the corresponding eigenvectors. To be specific, denote by $\mathcal{L}(\mathbf{\Sigma}_\mathbf{x}|\mathbf{X})$ the loss function of the data $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{n_1})$ with unknown parameter $\mathbf{\Sigma}_\mathbf{x}$, then the penalized loss function is

$$\widetilde{\mathcal{L}}(\mathbf{\Sigma}_\mathbf{x}|\mathbf{X}) = \mathcal{L}(\mathbf{\Sigma}_\mathbf{x}|\mathbf{X}) + \lambda g(\mathbf{\Sigma}_\mathbf{x}), \tag{1.1}$$

where $g(\mathbf{\Sigma}_\mathbf{x})$ is a given penalty. Under the Gaussian framework, the loss function is the negative log-likelihood function $\mathcal{L}(\mathbf{\Sigma}_\mathbf{x}|\mathbf{X}) = n_1(\log|\mathbf{\Sigma}_\mathbf{x}| + \mathrm{tr}(\mathbf{S}_\mathbf{x}\mathbf{\Sigma}_\mathbf{x}^{-1}))$. The penalty of the ridge estimation is $g(\mathbf{\Sigma}_\mathbf{x}) = \mathrm{tr}(\mathbf{\Sigma}_\mathbf{x}^{-1})$. Then the sample eigenvalues are regularized as $\psi(\mathbf{\Lambda}) = \mathbf{\Lambda} + \alpha\mathbf{I}$, and the ridge estimator is of the form $\widehat{\mathbf{\Sigma}}_\mathbf{x} = \mathbf{S}_\mathbf{x} + \alpha\mathbf{I}$, where α is a positive tuning parameter and $\mathbf{I}$ is the identity matrix. We can find that the ridge estimator enjoys an analytical expression and keeps positive-definite in both large-sample and high-dimensional situations. When the involved tuning parameter is optimally chosen under some criteria, the ridge estimator can largely improve the performance of the sample covariance matrix, especially in a high-dimensional situation.

1.2. Two-class data case. When the statistical model is specified, the performance of the estimators is limited by the finite information from the good data. The contamination can also contain the information of the covariance matrix of interest. Thus, taking advantage of the contamination is a sensible choice to improve the estimator's performance [7], [1], [11], [23], [24]. In [26], a two-class data model is suggested for improving the covariance matrix of the good data. However, their two-class data model requires the same covariance matrix and different means. Furthermore, in [2], another two-class data model is formulated by the good data and the contamination under Gaussian distribution with the same mean but different covariance matrix. With the aid of auxiliary information, the estimator from the data in two classes is shown to dominate the existing estimators from only the good data. As illustrated in Figure 1, the application scope of the existing estimator is limited to three situations:

1) the size of the good data is less than the dimension, but the size of the contamination should be larger than the dimension, namely $n_1 < p$, $n_2 \geqslant p$;

2) the size of the good data is larger than the dimension, but the size of contamination should be less than the dimension, namely $n_1 \geqslant p$, $n_2 < p$;

3) both the sizes of the good data and the contamination are less than the dimension, but the total size should be larger than the dimension, namely $n_1 < p$, $n_2 < p$, $n_1 + n_2 \geqslant p$.

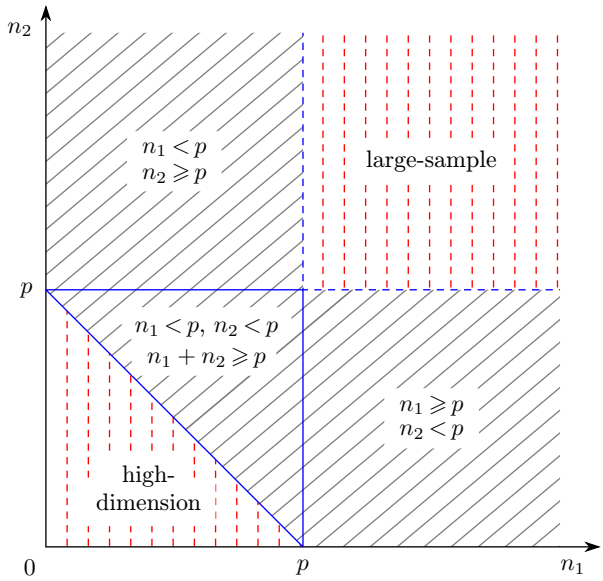

Figure 1. A variety of situations of estimating covariance matrix from the data in two classes.

In other words, the existing estimator is not applicable, neither in the large-sample situation where the good data size or the size of contamination is larger than the dimension nor in the high-dimensional situation where the total size is still less than the dimension. The deficiency in applicability makes the existing estimator from the data in two classes unable to take full advantage of the information on the contamination in many common situations. Therefore, developing a more applicable covariance matrix estimator from the data in two classes is necessary.

This paper employs the ridge method to estimate the covariance matrix from the data in two classes: (1) good data with the covariance matrix of interest and (2) contamination from a Gaussian distribution with a different covariance matrix. The main contribution is three-fold:

(1) The ridge penalty is employed in the negative log-likelihood function of the two-class data model to broaden the application scope of the estimator from the data in two classes.
(2) The ridge estimator of the covariance matrix from the data in two classes is obtained in closed form, where the concerned ridge parameter is tuned through the K-fold cross-validation procedure. The ridge estimator from the data in two classes is a generalized version of the traditional ridge estimator only from the good data. It keeps positive-definite and enjoys a uniform expression whether the dimension is less or greater than the data size.
(3) The proposed estimator is verified to perform better than the traditional ridge estimator and the existing estimator without ridge penalty.

The remainder of this paper is successively organized into four parts. Section 2 formulates the two-class data model and computes the log-likelihood function under Gaussian distribution. Section 3 joins the ridge penalty into the negative log-likelihood function, resulting in a closed-form covariance matrix estimator whenever the dimension is smaller or larger than the total size of the data in two classes. Section 4 verifies the numerical performance of the proposed estimators compared with some existing estimators of the same type. Section 5 concludes the main work of this paper.

2. Two-class data model

Let p-dimensional random vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{n_1}$ be an independent identically sample drawn from the Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_\mathbf{x})$. Besides, there exist p-dimensional data $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{n_2}$, which independently follow the p-dimensional Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_\mathbf{y})$. Further, we assume the two samples $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{n_1}$ and $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{n_2}$ are independent. Let the Cholesky decomposition of $\mathbf{\Sigma}_\mathbf{x}$ be $\mathbf{L}\mathbf{L}^\top$,

where $\mathbf{L}$ is a lower triangle matrix. We measure the geometric distance between $\mathbf{\Sigma_y}$ and $\mathbf{\Sigma_x}$ by

$$\text{dist}(\mathbf{\Sigma_y}, \mathbf{\Sigma_x}) = \sqrt{\sum_{k=1}^{p} \log^2 \lambda_k(\mathbf{L}^\top \mathbf{\Sigma_y}^{-1} \mathbf{L})}, \tag{2.1}$$

where $\lambda_k(\mathbf{A})$ represents the kth largest eigenvalue of matrix $\mathbf{A}$ (see [3]). In the Riemannian manifold consisting of all $p \times p$ real symmetric positive definite matrices, the distance $\text{dist}(\mathbf{\Sigma_y}, \mathbf{\Sigma_x})$ is the infimum length of geodesics between $\mathbf{\Sigma_x}$ and $\mathbf{\Sigma_y}$ (see [20]).

Let $\mathbf{W} = \mathbf{L}^\top \mathbf{\Sigma_y}^{-1} \mathbf{L}$, it is easy to verify that the following three propositions are equivalent to each other:

$$\text{dist}(\mathbf{\Sigma_y}, \mathbf{\Sigma_x}) = 0 \Leftrightarrow \mathbf{W} = \mathbf{I} \Leftrightarrow \mathbf{\Sigma_y} = \mathbf{\Sigma_x}. \tag{2.2}$$

When $\text{dist}(\mathbf{\Sigma_y}, \mathbf{\Sigma_x})$ is close to 0, $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{n_2}$ is considered the contamination relative to $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma_x})$. The matrix $\mathbf{W}$ is a key quantity for describing the distance of the contamination. As suggested in [2], we assume that $\mathbf{W}$ follows a Wishart distribution $\mathcal{W}(\nu, \mu^{-1}\mathbf{I})$, where ν is the degree of freedom. Moreover, we set $\mu = \nu - p - 1$ for meeting the mathematical assumption $\mathbb{E}(\mathbf{\Sigma_y}) = \mathbf{\Sigma_x}$.

This article aims to estimate $\mathbf{\Sigma_x}$ from the data $(\mathbf{X}, \mathbf{Y}) = \{\mathbf{x}_{i_1}, \mathbf{y}_{i_2} \mid i_1 = 1, 2, \ldots, n_1, i_2 = 1, 2, \ldots, n_2\}$. For given $\mathbf{\Sigma_x}$ and $\mathbf{W}$, we have $\mathbf{\Sigma_y} = \mathbf{L}\mathbf{W}^{-1}\mathbf{L}^\top$ and $|\mathbf{\Sigma_y}| = |\mathbf{W}^{-1}\mathbf{\Sigma_x}|$. Then, the conditional joint density function of the data $(\mathbf{X}, \mathbf{Y})$ is

$$\begin{aligned} &f(\mathbf{X}, \mathbf{Y} | \mathbf{\Sigma_x}, \mathbf{W}) \\ &\quad = (2\pi)^{-pn/2} |\mathbf{\Sigma_x}|^{-n/2} |\mathbf{W}|^{n_2/2} \text{etr}\Big\{-\frac{1}{2}\mathbf{X}^\top \mathbf{\Sigma_x}^{-1}\mathbf{X} - \frac{1}{2}\mathbf{Y}^\top \mathbf{L}^{-\top}\mathbf{W}\mathbf{L}^{-1}\mathbf{Y}\Big\}, \end{aligned} \tag{2.3}$$

where $n = n_1 + n_2$ is the total size of the data in two classes. Because $\mathbf{W}$ follows the Wishart distribution $\mathcal{W}(\nu, \mu^{-1}\mathbf{I})$, its density function is

$$f(\mathbf{W}) = \Big(\frac{\mu}{2}\Big)^{\nu p/2} \Gamma^{-1}\Big(\frac{\nu}{2}\Big) |\mathbf{W}|^{(\nu-p-1)/2} \text{etr}\Big\{-\frac{1}{2}\mu\mathbf{W}\Big\}. \tag{2.4}$$

Therefore, the conditional density function under $\mathbf{\Sigma_x}$ is

$$f(\mathbf{X}, \mathbf{Y} | \mathbf{\Sigma_x}) = \int_{\mathbf{W}>0} f(\mathbf{X}, \mathbf{Y} | \mathbf{\Sigma_x}, \mathbf{W}) f(\mathbf{W}) \, \mathrm{d}\mathbf{W}. \tag{2.5}$$

By equations (2.3) and (2.4), the conditional density function $f(\mathbf{X}, \mathbf{Y}|\boldsymbol{\Sigma}_\mathbf{x})$ becomes (2.6)

$$\begin{aligned} &f(\mathbf{X}, \mathbf{Y}|\boldsymbol{\Sigma}_\mathbf{x}) \\ &\quad = c_n|\boldsymbol{\Sigma}_\mathbf{x}|^{-n/2}\mathrm{etr}\Big\{-\frac{1}{2}\mathbf{X}^\top\boldsymbol{\Sigma}_\mathbf{x}^{-1}\mathbf{X}\Big\}\int_{\mathbf{W}>0}|\mathbf{W}|^{(\nu+n_2-p-1)/2} \\ &\qquad \times \mathrm{etr}\Big\{-\frac{1}{2}\mathbf{W}(\mu\mathbf{I}+\mathbf{L}^{-1}\mathbf{Y}\mathbf{Y}^\top\mathbf{L}^{-\top})\Big\}\,\mathrm{d}\mathbf{W} \\ &\quad = c_{n_1}c_{n_2}|\boldsymbol{\Sigma}_\mathbf{x}|^{-n_1/2}\mathrm{etr}\Big\{-\frac{1}{2}\mathbf{X}^\top\boldsymbol{\Sigma}_\mathbf{x}^{-1}\mathbf{X}\Big\}|\mu\boldsymbol{\Sigma}_\mathbf{x}|^{-n_2/2}|\mathbf{I}+\mathbf{Y}^\top(\mu\boldsymbol{\Sigma}_\mathbf{x})^{-1}\mathbf{Y}|^{-(\nu+n_2)/2}, \end{aligned}$$

where

$$c_{n_1} = (2\pi)^{-pn_1/2}, \quad c_{n_2} = \pi^{-pn_2/2}\Gamma\Big(\frac{\nu+n_2}{2}\Big)\Gamma^{-1}\Big(\frac{\nu}{2}\Big),$$

and

$$c_n = (2\pi)^{-pn/2}\mu^{\nu p/2}2^{-\nu p/2}\Gamma^{-1}\Big(\frac{\nu}{2}\Big).$$

Denote

$$f(\mathbf{X}|\boldsymbol{\Sigma}_\mathbf{x}) = c_{n_1}|\boldsymbol{\Sigma}_\mathbf{x}|^{-n_1/2}\mathrm{etr}\Big\{-\frac{1}{2}\mathbf{X}^\top\boldsymbol{\Sigma}_\mathbf{x}^{-1}\mathbf{X}\Big\}, \tag{2.7}$$

$$f(\mathbf{Y}|\boldsymbol{\Sigma}_\mathbf{x}) = c_{n_2}|\mu\boldsymbol{\Sigma}_\mathbf{x}|^{-n_2/2}|\mathbf{I}+\mathbf{Y}^\top(\mu\boldsymbol{\Sigma}_\mathbf{x})^{-1}\mathbf{Y}|^{-(\nu+n_2)/2}. \tag{2.8}$$

We have

$$f(\mathbf{X}, \mathbf{Y}|\boldsymbol{\Sigma}_\mathbf{x}) = f(\mathbf{X}|\boldsymbol{\Sigma}_\mathbf{x})f(\mathbf{Y}|\boldsymbol{\Sigma}_\mathbf{x}). \tag{2.9}$$

We can see that the contamination $\mathbf{Y}$ follows a matrix-variate Student distribution under the covariance matrix $\boldsymbol{\Sigma}_\mathbf{x}$ and keeps independent of the good data $\mathbf{X}$.

Let $\mathbf{S}_\mathbf{x} = n_1^{-1}\mathbf{X}\mathbf{X}^\top$ be the sample covariance matrices of $\mathbf{X}$, and $\mathbf{S}_\mathbf{y} = n_2^{-1}\mathbf{Y}\mathbf{Y}^\top$ be the one of $\mathbf{Y}$. Without regard to a constant term, we obtain that the negative log-likelihood function of $(\mathbf{X}, \mathbf{Y})$ is

$$\mathcal{L}(\mathbf{X}, \mathbf{Y}|\boldsymbol{\Sigma}_\mathbf{x}) = n\log|\boldsymbol{\Sigma}_\mathbf{x}| + n_1\mathrm{tr}(\mathbf{S}_\mathbf{x}\boldsymbol{\Sigma}_\mathbf{x}^{-1}) + (n_2+\nu)\log|\mathbf{I}+n_2\mu^{-1}\mathbf{S}_\mathbf{y}\boldsymbol{\Sigma}_\mathbf{x}^{-1}|. \tag{2.10}$$

R e m a r k 2.1. When $n_2 = 0$, we have $\mathcal{L}(\mathbf{X}, \mathbf{Y}|\boldsymbol{\Sigma}_\mathbf{x}) = n_1\log|\boldsymbol{\Sigma}_\mathbf{x}| + n_1\mathrm{tr}(\mathbf{S}_\mathbf{x}\boldsymbol{\Sigma}_\mathbf{x}^{-1})$. Then the negative log-likelihood function in (2.10) degenerates into the negative log-likelihood function of the good data $\mathbf{X}$.

R e m a r k 2.2. In some special applications, $\boldsymbol{\Sigma}_\mathbf{x}$ and $\boldsymbol{\Sigma}_\mathbf{y}$ may be proportional, namely $\boldsymbol{\Sigma}_\mathbf{x} = c\boldsymbol{\Sigma}_\mathbf{y}$. Then we have $\mathbf{W} = \mathbf{L}^\top\boldsymbol{\Sigma}_\mathbf{y}^{-1}\mathbf{L} = c\mathbf{I}$. Then the distance between $\boldsymbol{\Sigma}_\mathbf{x}$ and $\boldsymbol{\Sigma}_\mathbf{y}$ becomes

$$\mathrm{dist}(\boldsymbol{\Sigma}_\mathbf{y}, \boldsymbol{\Sigma}_\mathbf{x}) = \sqrt{\sum_{k=1}^{p}\log^2\lambda_k(\mathbf{W})} = \sqrt{\sum_{k=1}^{p}\log^2\lambda_k(c\mathbf{I})} = \sqrt{p}|\log c|.$$

Moreover, the conditional joint density function of $(\mathbf{X}, \mathbf{Y})$ can be simplified as

$$f(\mathbf{X}, \mathbf{Y}|\mathbf{\Sigma_x}) = (2\pi)^{-pn/2}|\mathbf{\Sigma_x}|^{-n/2}c^{pn_2/2}\mathrm{etr}\Big\{-\frac{1}{2}\mathbf{X}^\top\mathbf{\Sigma_x}^{-1}\mathbf{X} - \frac{c}{2}\mathbf{Y}^\top\mathbf{\Sigma_x}^{-1}\mathbf{Y}\Big\}. \tag{2.11}$$

The corresponding negative log-likelihood function is

$$\mathcal{L}(\mathbf{X}, \mathbf{Y}|\mathbf{\Sigma_x}) = n\log|\mathbf{\Sigma_x}| + \mathrm{tr}(n_1\mathbf{S_x}\mathbf{\Sigma_x}^{-1} + cn_2\mathbf{S_y}\mathbf{\Sigma_x}^{-1}). \tag{2.12}$$

The covariance matrix estimator turns out to be

$$\mathbf{\Sigma_x} = \frac{n_1}{n}\mathbf{S_x} + c\frac{n_2}{n}\mathbf{S_y}, \tag{2.13}$$

which is a weighted combination between the sample covariance matrices of $\mathbf{X}$ and $\mathbf{Y}$. Especially, when $c = 1$, $\mathbf{\Sigma_x} = \mathbf{\Sigma_y}$, then $\mathbf{X}$ and $\mathbf{Y}$ are from the same distribution. We have

$$\mathbf{\Sigma_x} = \frac{n_1}{n}\mathbf{S_x} + \frac{n_2}{n}\mathbf{S_y}, \tag{2.14}$$

which is the classic sample covariance matrix of $(\mathbf{X}, \mathbf{Y})$.

3. Covariance matrix estimator

3.1. Ridge estimation. The maximum likelihood estimator cannot be derived from the negative log-likelihood function given by (2.10) in many high-dimensional cases, such as $n_1 + n_2 < p$. Therefore, we consider the ridge penalty term in the negative log-likelihood function $\mathcal{L}(\mathbf{X}, \mathbf{Y}|\mathbf{\Sigma_x})$ to estimate the covariance matrix $\mathbf{\Sigma_x}$ from the data $(\mathbf{X}, \mathbf{Y})$ in high-dimensional situation. Generally, we assume $\mathbf{\Sigma_x} \neq \mathbf{\Sigma_y}$. The ridge penalized negative log-likelihood function is

$$\begin{aligned}\widetilde{\mathcal{L}}_\lambda(\mathbf{\Sigma_x}) &= n\log|\mathbf{\Sigma_x}| + n_1\mathrm{tr}(\mathbf{S_x}\mathbf{\Sigma_x}^{-1}) + (n_2+\nu)\log|\mathbf{I} + n_2\mu^{-1}\mathbf{S_y}\mathbf{\Sigma_x}^{-1}| + \lambda\mathrm{tr}(\mathbf{\Sigma_x}^{-1}) \\ &= (n_1-\nu)\log|\mathbf{\Sigma_x}| + n_1\mathrm{tr}(\mathbf{S_x}\mathbf{\Sigma_x}^{-1}) + (n_2+\nu)\log|\mathbf{\Sigma_x} + n_2\mu^{-1}\mathbf{S_y}| + \lambda\mathrm{tr}(\mathbf{\Sigma_x}^{-1}),\end{aligned} \tag{3.1}$$

where $\lambda > 0$ is the tuning parameter. Then the ridge-type estimator is

$$\begin{aligned}\widehat{\mathbf{\Sigma}}_\mathbf{x}(\lambda) = \arg\min_{\mathbf{\Sigma_x}\succ 0}\{&(n_1-\nu)\log|\mathbf{\Sigma_x}| + n_1\mathrm{tr}(\mathbf{S_x}\mathbf{\Sigma_x}^{-1}) \\ &+ (n_2+\nu)\log|\mathbf{\Sigma_x} + n_2\mu^{-1}\mathbf{S_y}| + \lambda\mathrm{tr}(\mathbf{\Sigma_x}^{-1})\}.\end{aligned} \tag{3.2}$$

Theorem 3.1. *For an arbitrary tuning parameter $\lambda > 0$, the ridge-type estimator $\widehat{\mathbf{\Sigma}}_\mathbf{x}(\lambda)$ holds positive definite whether $n_2 > p$ or $n_2 \leqslant p$.*

P r o o f. First of all, we develop the analytic expression of the ridge-type estimator $\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}}(\lambda)$. By taking the derivative of equation (3.1), we obtain the normal equation:

$$(n_1-\nu)\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}-\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}(n_1\mathbf{S}_{\mathbf{x}}+\lambda\mathbf{I})\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}+(n_2+\nu)(\boldsymbol{\Sigma}_{\mathbf{x}}+n_2\mu^{-1}\mathbf{S}_{\mathbf{y}})^{-1}=0. \tag{3.3}$$

By multiplying $(\boldsymbol{\Sigma}_{\mathbf{x}}+n_2\mu^{-1}\mathbf{S}_{\mathbf{y}})$ from the right, the normal equation (3.3) becomes

$$(n_1-\nu)\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}(\boldsymbol{\Sigma}_{\mathbf{x}}+n_2\mu^{-1}\mathbf{S}_{\mathbf{y}})-\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}(n_1\mathbf{S}_{\mathbf{x}}+\lambda\mathbf{I})\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}(\boldsymbol{\Sigma}_{\mathbf{x}}+n_2\mu^{-1}\mathbf{S}_{\mathbf{y}})+(n_2+\nu)\mathbf{I}=0. \tag{3.4}$$

Then we have

$$n\mathbf{I}+(n_1-\nu)n_2\mu^{-1}\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\mathbf{S}_{\mathbf{y}}-\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}(n_1\mathbf{S}_{\mathbf{x}}+\lambda\mathbf{I})-n_2\mu^{-1}\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}(n_1\mathbf{S}_{\mathbf{x}}+\lambda\mathbf{I})\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\mathbf{S}_{\mathbf{y}}=0. \tag{3.5}$$

By multiplying $\boldsymbol{\Sigma}_{\mathbf{x}}(n_1\mathbf{S}_{\mathbf{x}}+\lambda\mathbf{I})^{-1}\boldsymbol{\Sigma}_{\mathbf{x}}$ from the left, we obtain

$$n\boldsymbol{\Sigma}_{\mathbf{x}}(n_1\mathbf{S}_{\mathbf{x}}+\lambda\mathbf{I})^{-1}\boldsymbol{\Sigma}_{\mathbf{x}}+(n_1-\nu)n_2\mu^{-1}\boldsymbol{\Sigma}_{\mathbf{x}}(n_1\mathbf{S}_{\mathbf{x}}+\lambda\mathbf{I})^{-1}\mathbf{S}_{\mathbf{y}}-\boldsymbol{\Sigma}_{\mathbf{x}}-n_2\mu^{-1}\mathbf{S}_{\mathbf{y}}=0. \tag{3.6}$$

Therefore, we have

$$\boldsymbol{\Sigma}_{\mathbf{x}}(n_1\mathbf{S}_{\mathbf{x}}+\lambda\mathbf{I})^{-1}\boldsymbol{\Sigma}_{\mathbf{x}}-\boldsymbol{\Sigma}_{\mathbf{x}}\Big[\frac{(\nu-n_1)n_2}{\mu n}(n_1\mathbf{S}_{\mathbf{x}}+\lambda\mathbf{I})^{-1}\mathbf{S}_{\mathbf{y}}+\frac{1}{n}\mathbf{I}\Big]-\frac{n_2}{\mu n}\mathbf{S}_{\mathbf{y}}=0. \tag{3.7}$$

For an arbitrary tuning parameter $\lambda>0$, the matrix $n_1\mathbf{S}_{\mathbf{x}}+\lambda\mathbf{I}$ is positive definite whether $n\geqslant p$ or $n<p$. Denote by $\mathbf{U}$ the square root of $n_1\mathbf{S}_{\mathbf{x}}+\lambda\mathbf{I}$, we have $n_1\mathbf{S}_{\mathbf{x}}+\lambda\mathbf{I}=\mathbf{U}\mathbf{U}^{\top}$. Let $\widetilde{\boldsymbol{\Sigma}}_{\mathbf{x}}=\mathbf{U}^{-1}\boldsymbol{\Sigma}_{\mathbf{x}}\mathbf{U}^{-\top}$ and $\mathbf{S}_{\mathbf{xy}}=n_2\mathbf{U}^{-1}\mathbf{S}_{\mathbf{y}}\mathbf{U}^{-\top}$, we have

$$\widetilde{\boldsymbol{\Sigma}}_{\mathbf{x}}^2-\Big(\frac{\nu-n_1}{\mu n}\mathbf{S}_{\mathbf{xy}}+\frac{1}{n}\mathbf{I}\Big)\widetilde{\boldsymbol{\Sigma}}_{\mathbf{x}}-\frac{1}{\mu n}\mathbf{S}_{\mathbf{xy}}=0. \tag{3.8}$$

We can see that the matrix equation (3.8) is quadratic concerning $\boldsymbol{\Sigma}_{\mathbf{x}}$. Let q be an eigenvalue of $\widetilde{\boldsymbol{\Sigma}}_{\mathbf{x}}$ and $\mathbf{w}$ be the associated eigenvector. We can obtain

$$q^2\mathbf{w}-q\Big(\frac{\nu-n_1}{\mu n}\mathbf{S}_{\mathbf{xy}}+\frac{1}{n}\mathbf{I}\Big)\mathbf{w}-\frac{1}{\mu n}\mathbf{S}_{\mathbf{xy}}\mathbf{w}=0. \tag{3.9}$$

Therefore, we have

$$\Big(\frac{1}{\mu n}+\frac{q(\nu-n_1)}{\mu n}\Big)\mathbf{S}_{\mathbf{xy}}\mathbf{w}=q\Big(q-\frac{1}{n}\Big)\mathbf{w}. \tag{3.10}$$

The above equation (3.10) reveals $\mathbf{w}$ is also an eigenvector of $\mathbf{S}_{\mathbf{xy}}$. To explore the value of q, we discuss the eigenvalue problem of (3.10) in two cases.

Case 1: $n_2 > p$. We can see that $\mathbf{S_{xy}}$ is almost surely positive definite for an arbitrary $\lambda > 0$. Therefore, there is no zero eigenvalue in (3.10). Denote by η_j an eigenvalue of $\mathbf{S_{xy}}$, the eigenvalue of $\widetilde{\mathbf{\Sigma}}_{\mathbf{x}}$ can be obtained by taking the positive root of the following equation:

$$q^2 - q\Big(\frac{\eta_j(\nu - n_1)}{\mu n} + \frac{1}{n}\Big) - \frac{\eta_j}{\mu n} = 0, \quad j \in \{1, \ldots, p\}. \tag{3.11}$$

Note that $\mathbf{S_{xy}} = \mathbf{U}^{-1}\mathbf{Y}\mathbf{Y}^\top\mathbf{U}^{-\top}$. Let the singular value decomposition of $\mathbf{U}^{-1}\mathbf{Y}$ be $\mathbf{U}^{-1}\mathbf{Y} = \mathbf{A}\mathbf{\Theta}\mathbf{B}^\top$. Denote by θ_j the jth diagonal element of $\mathbf{\Theta}$, and by $\mathbf{a}_j$, $\mathbf{b}_j$ the jth columns of $\mathbf{A}$ and $\mathbf{B}$, respectively, we have $\mathbf{U}^{-1}\mathbf{Y} = \sum_{j=1}^{p} \theta_j \mathbf{a}_j \mathbf{b}_j^\top$ and $\eta_j = \theta_j^2$. Therefore, we have

$$\widetilde{\mathbf{\Sigma}}_{\mathbf{x}}(\lambda) = \sum_{j=1}^{p} q_j \mathbf{a}_j \mathbf{a}_j^\top, \tag{3.12}$$

where q_j is an eigenvalue of $\widetilde{\mathbf{\Sigma}}_{\mathbf{x}}$, which satisfies equation (3.11).

Case 2: $n_2 \leqslant p$. When $\mathbf{w}$ corresponds to the zero eigenvalue (with $p - n_2$ multiplicities), we have $q = 1/n$. When $\mathbf{w}$ corresponds to a nonzero eigenvalue η_j, we have

$$q^2 - q\Big(\frac{\eta_j(\nu - n_1)}{\mu n} + \frac{1}{n}\Big) - \frac{\eta_j}{\mu n} = 0, \quad j \in \{1, \ldots, n_2\}. \tag{3.13}$$

Then we can obtain the eigenvalues of $\widetilde{\mathbf{\Sigma}}_{\mathbf{x}}$ by searching for the positive root of equation (3.13). In the same manner, we have

$$\widetilde{\mathbf{\Sigma}}_{\mathbf{x}} = \sum_{j=1}^{p} q_j \mathbf{a}_j \mathbf{a}_j^\top = \sum_{j=1}^{n_2} q_j \mathbf{a}_j \mathbf{a}_j^\top + \frac{1}{n} \sum_{j=n_2+1}^{p} \mathbf{a}_j \mathbf{a}_j^\top, \tag{3.14}$$

where q_j are the positive roots of equation (3.13).

In summary, the ridge-type estimator of $\mathbf{\Sigma}_{\mathbf{x}}$ is

$$\widehat{\mathbf{\Sigma}}_{\mathbf{x}}(\lambda) = \sum_{j=1}^{p} q_j \mathbf{U}\mathbf{a}_j \mathbf{a}_j^\top \mathbf{U}^\top. \tag{3.15}$$

Next, we prove that $\widehat{\mathbf{\Sigma}}_{\mathbf{x}}(\lambda)$ is positive definite.

When $n_2 > p$, it is obvious that $\widehat{\mathbf{\Sigma}}_{\mathbf{x}}(\lambda)$ is positive definite because q_j are the positive roots of equation (3.11).

When $n_2 \leqslant p$, we have

$$\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}}(\lambda) = \sum_{j=1}^{n_2} q_j \mathbf{U}\mathbf{a}_j\mathbf{a}_j^\top\mathbf{U}^\top + \frac{1}{n}\sum_{j=n_2+1}^{p}\mathbf{U}\mathbf{a}_j\mathbf{a}_j^\top\mathbf{U}^\top. \tag{3.16}$$

As $q_j > 0$, $j = 1, 2, \ldots, n_2$, we have that the ridge estimator $\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}}(\lambda)$ is positive definite for arbitrary positive tuning parameter λ. The proof ends. □

R e m a r k 3.1. Since $\sum\limits_{j=1}^{p}\mathbf{a}_j\mathbf{a}_j^\top = \mathbf{I}$, we have

$$\begin{aligned}\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}}(\lambda) &= \sum_{j=1}^{n_2} q_j \mathbf{U}\mathbf{a}_j\mathbf{a}_j^\top\mathbf{U}^\top + \frac{1}{n}\sum_{j=n_2+1}^{p}\mathbf{U}\mathbf{a}_j\mathbf{a}_j^\top\mathbf{U}^\top \\ &= \sum_{j=1}^{n_2}\Big(q_j - \frac{1}{n}\Big)\mathbf{U}\mathbf{a}_j\mathbf{a}_j^\top\mathbf{U}^\top + \frac{1}{n}(n_1\mathbf{S}_{\mathbf{x}} + \lambda\mathbf{I}).\end{aligned} \tag{3.17}$$

Therefore, when $n_2 = 0$, the ridge estimator turns out to be $\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}}(\lambda) = \mathbf{S}_{\mathbf{x}} + \lambda\mathbf{I}/n_1$, which is the classic ridge estimator from the good data $\mathbf{X}$.

3.2. Tuning parameter. It is obvious that the penalized estimator $\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}}(\lambda)$ in (3.15) depends on the tuning parameter selection. In this subsection, we select the appropriate value of λ with the K-fold cross-validation procedure [30]. Firstly, we split data $\mathbf{X}$ and $\mathbf{Y}$ into K roughly equal groups. We reserve the kth group $(\mathbf{X}^k, \mathbf{Y}^k)$ as the test part, and use the other $K-1$ groups $(\mathbf{X}^{\setminus k}, \mathbf{Y}^{\setminus k})$ as the training part. Denote by $\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}}^{\setminus k}$ the ridge estimator from equation (3.15). For each $k \in \{1, \ldots, K\}$, denote by n_1^k and $\mathbf{S}_{\mathbf{x}}^k$ respectively the data size and sample covariance matrix of $\mathbf{X}^k$, by n_2^k and $\mathbf{S}_{\mathbf{y}}^k$ respectively the data size and sample covariance matrix of $\mathbf{Y}^k$. Then the total size of the test data $(\mathbf{X}^k, \mathbf{Y}^k)$ is $n^k = n_1^k + n_2^k$. Then we calculate the negative log-likelihood function:

$$\begin{aligned}\mathcal{L}(\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}}^{\setminus k}|\mathbf{X}^k, \mathbf{Y}^k) = n^k \log|\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}}^{\setminus k}| + n_1^k \mathrm{tr}(\mathbf{S}_{\mathbf{x}}^k(\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}}^{\setminus k})^{-1}) \\ + (n_2^k + \nu)\log|\mathbf{I} + n_2^k\mu^{-1}\mathbf{S}_{\mathbf{y}}^k(\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}}^{\setminus k})^{-1}|.\end{aligned} \tag{3.18}$$

Therefore, the optimal tuning parameter $\widehat{\lambda}$ can be estimated by minimizing the cross-validation likelihood function, namely,

$$\widehat{\lambda} = \arg\min_{\lambda}\sum_{k=1}^{K}\mathcal{L}(\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}}^{\setminus k}|\mathbf{X}^k, \mathbf{Y}^k). \tag{3.19}$$

4. Numerical simulation

This section verifies the performance of the proposed ridge estimator compared with its competitors. In the implementation, the good data $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{n_1})$ is drawn from the Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_\mathbf{x})$, where the population covariance matrix to be estimated is

$$\mathbf{\Sigma}_\mathbf{x} = (\sigma_{ij})_{p\times p} \text{ with } \sigma_{ij} = 0.5^{|i-j|}. \tag{4.1}$$

Besides, the contamination $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{n_2})$ is from the Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_\mathbf{y})$, where $\mathbf{\Sigma}_\mathbf{y}$ is generated by $\mathbf{\Sigma}_\mathbf{x}$ and a Wishart distributed random matrix $\mathbf{W}$ from $\mathcal{W}(\nu, \mu^{-1}\mathbf{I})$ with ν, μ being respectively $p+2$ and 1. We denote the proposed ridge estimator from the data in two classes developed in Section 3 as RPEe for the notation convenience. The competing estimators include the sample covariance matrix (SE) from the data $(\mathbf{X}, \mathbf{Y})$, the ridge estimator (RPE) from $\mathbf{X}$ in [30], and the non-penalty estimator (NPE) from $(\mathbf{X}, \mathbf{Y})$ in [2]. We investigate their ridge parameters and losses under various data sizes n_1, n_2 and dimension p.

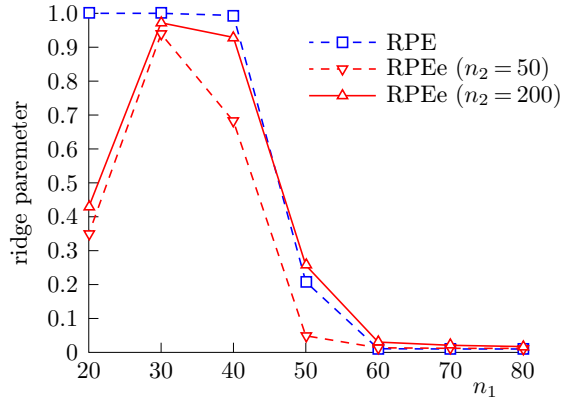

Figure 2. Ridge parameters in the estimator from the data in two classes and the one from the only good data when $p = 20$.

Figures 2 and 3 report the variation trends of the ridge parameters along with the data size n_1. The ridge parameters present significantly different variation trends in the cases of $p = 20$ and $p = 100$. When $p < n_1$, the ridge parameters decrease to 0 as the data size n_1 increases from 30 to 80. Furthermore, the ridge parameter in RPEe becomes smaller in the case of $n_2 = 50$ than in the case of $n_2 = 200$. It reveals that the estimator relies more heavily on good data in the large-sample setting. When $p > n_1$, the ridge parameters are close or increase to the bound 1 as n_1 increases. Besides, the ridge parameter in RPEe is smaller in the case of $n_2 = 200$ than in

the case of $n_2 = 50$, revealing that the ridge estimator relies more heavily on the likelihood function in the large-dimensional setting.

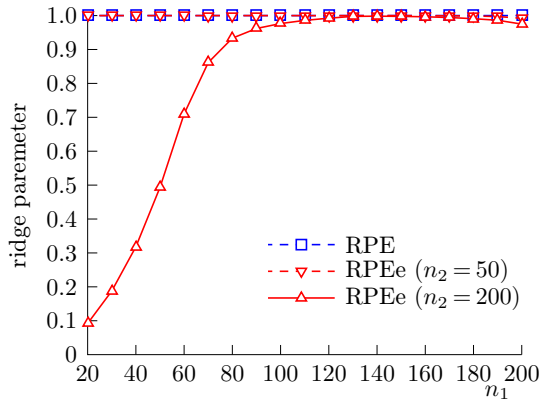

Figure 3. Ridge parameters in the estimator from the data in two classes and the one from the only good data when $p = 100$.

Next, we discover the estimation accuracy of the proposed ridge estimator from the data in two classes against its main competitors. For an intuitive comparison, we employ the loss function in [2], which is

$$l(\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}}) = 10 \log_{10} \left[\sum_{j=1}^{p} \log^2(\lambda_j(\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}} \boldsymbol{\Sigma}_{\mathbf{x}}^{-1})) \right]. \tag{4.2}$$

We compute the average losses of the covariance matrix estimators based on five thousand replications.

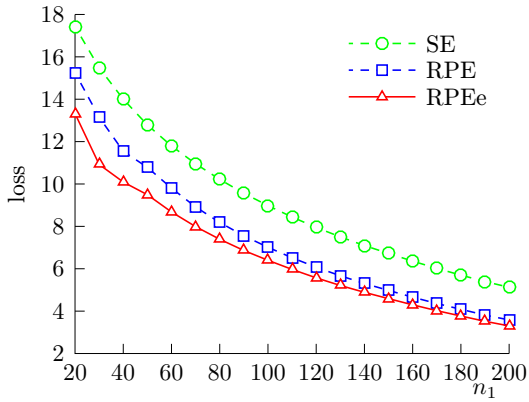

Figure 4. Loss comparison of the covariance estimator against its competitors when $p = 20$ and $n_2 = 50$.

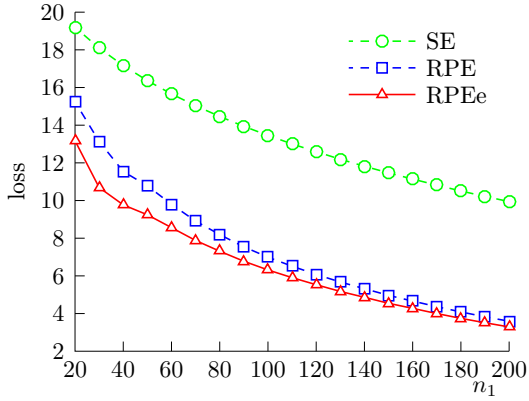

Figure 5. Loss comparison of the covariance estimator against its competitors when $p = 20$ and $n_2 = 200$.

Figures 4 and 5 reveal the losses of the ridge estimator RPEe and its competitors along with the data size n_1 when $p = 20$. The non-penalty estimator NPE does not work when $n_1 \geqslant p$, $n_2 > p$. We can see the losses of the other estimators decrease when n_1 gets larger. When n_2 increases from 50 to 200, the two ridge estimators RPE and RPEe have very similar performance, while the loss of SE becomes significantly larger. It reveals that the ridge estimators enjoy a robust performance relative to n_2. We can see that the proposed RPEe from $(\mathbf{X}, \mathbf{Y})$ enjoys lower loss than the existing ridge estimator RPE only from $\mathbf{X}$.

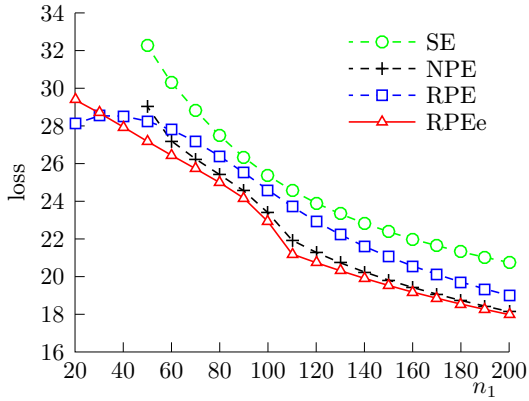

Figure 6. Loss comparison of the covariance estimator against its competitors when $p = 100$ and $n_2 = 50$.

Figures 6 and 7 report the losses of the four estimators relative to the data size n_1 when $p = 100$. The sample covariance matrix SE is only available when $n_1 + n_2 \geqslant p$ in Figure 6, and the estimator NPE from $(\mathbf{X}, \mathbf{Y})$ can work only when $n_1 > 50$ in Figure 6 and $n_1 < 100$ in Figure 7. We can see that the estimators NPE and RPEe,

obtained from the two-class data model, have lower losses than SE and RPE in most cases. Moreover, the proposed RPEe enjoys a lower loss than NPE when $n_2 = 50$, showing that the ridge penalty can further promote the estimator's performance in the high-dimensional setting. In the case of $n_2 = 200$, the proposed RPEe performs similarly to NPE as $n_1 < 100$ and performs well when NPE is unavailable.

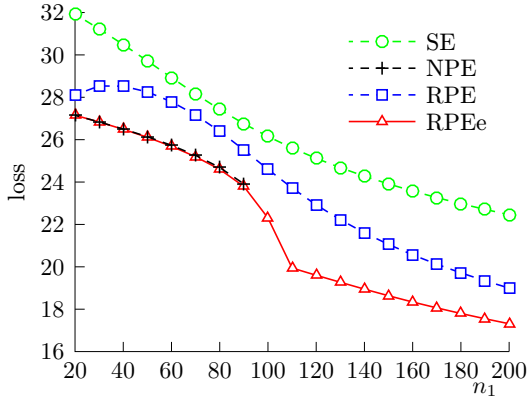

Figure 7. Loss comparison of the covariance estimator against its competitors when $p = 100$ and $n_2 = 200$.

Finally, we summarize the numerical performance of the proposed ridge estimator from the two-class data model as follows:

(1) The loss of the proposed RPEe significantly decreases as n_1 gets larger. In the large-sample setting, RPEe is robust relative to n_2. RPEe can reach lower losses in high-dimensional scenarios as n_2 gets large.

(2) The proposed RPEe from $(\mathbf{X}, \mathbf{Y})$ generally enjoys a lower loss than the traditional ridge estimator RPE only from $\mathbf{X}$.

(3) The proposed RPEe based on ridge penalty thoroughly addresses the usage limitations of existing NPE from $(\mathbf{X}, \mathbf{Y})$ and has obvious advantages on the loss.

5. Conclusions

This paper has studied the covariance matrix estimation from the data in two classes. The contamination was incorporated into the i.i.d. Gaussian sample to improve the traditional maximum likelihood estimation. The log-likelihood function was derived based on the data in two classes. The ridge penalty was considered in the two-class data model to handle estimating the covariance matrix in the high-dimensional situation, resulting in an analytic covariance matrix estimator. The proposed ridge estimator thoroughly addressed the limitations of the existing estimator from the data in two classes in theory. Numerical simulations verify that the

proposed ridge estimator from the data in two classes retains the advantages of the two-class data model. The ridge estimation often performs better than its competitors in large-sample or high-dimensional settings. Meanwhile, it is necessary to warn that the contamination may bring pessimism in some cases, such as when the size of good data is far less than the size of contamination in a high-dimensional setting.

References

[1] *M. Ahsanullah, V. B. Nevzorov*: Generalized spacings of order statistics from extended sample. J. Stat. Plann. Inference *85* (2000), 75–83. zbl MR doi

[2] *O. Besson*: Maximum likelihood covariance matrix estimation from two possibly mismatched data sets. Signal Process. *167* (2020), Article ID 107285, 9 pages. doi

[3] *R. Bhatia*: Positive Definite Matrices. Princeton Series in Applied Mathematics. Princeton University Press, Princeton, 2007. zbl MR doi

[4] *J. Bien, R. J. Tibshirani*: Sparse estimation of a covariance matrix. Biometrika *98* (2011), 807–820. zbl MR doi

[5] *O. Bodnar, T. Bodnar, N. Parolya*: Recent advances in shrinkage-based high-dimensional inference. J. Multivariate Anal. *188* (2022), Article ID 104826, 13 pages. zbl MR doi

[6] *S. Cho, S. Katayama, J. Lim, Y.-G. Choi*: Positive-definite modification of a covariance matrix by minimizing the matrix ℓ_∞ norm with applications to portfolio optimization. AStA, Adv. Stat. Anal. *105* (2021), 601–627. zbl MR doi

[7] *P. Danaher, P. Wang, D. M. Witten*: The joint graphical lasso for inverse covariance estimation across multiple classes. J. R. Stat. Soc., Ser. B, Stat. Methodol. *76* (2014), 373–397. zbl MR doi

[8] *T. J. Fisher, X. Sun*: Improved Stein-type shrinkage estimators for the high-dimensional multivariate normal covariance matrix. Comput. Stat. Data Anal. *55* (2011), 1909–1918. zbl MR doi

[9] *F. Götze, A. Tikhomirov*: Rate of convergence in probability to the Marchenko-Pastur law. Bernoulli *10* (2004), 503–548. zbl MR doi

[10] *A. Hannart, P. Naveau*: Estimating high dimensional covariance matrices: A new look at the Gaussian conjugate framework. J. Multivariate Anal. *131* (2014), 149–162. zbl MR doi

[11] *N. Hoshino, A. Takemura*: On reduction of finite-sample variance by extended Latin hypercube sampling. Bernoulli *6* (2000), 1035–1050. zbl MR doi

[12] *C. Huang, D. Farewell, J. Pan*: A calibration method for non-positive definite covariance matrix in multivariate data analysis. J. Multivariate Anal. *157* (2017), 45–52. zbl MR doi

[13] *J. Z. Huang, N. Liu, M. Pourahmadi, L. Liu*: Covariance matrix selection and estimation via penalised normal likelihood. Biometrika *93* (2006), 85–98. zbl MR doi

[14] *S. Jia, C. Zhang, H. Lu*: Covariance function versus covariance matrix estimation in efficient semi-parametric regression for longitudinal data analysis. J. Multivariate Anal. *187* (2022), Article ID 104900, 14 pages. zbl MR doi

[15] *J. Kalina, J. D. Tebbens*: Algorithms for regularized linear discriminant analysis. Proceedings of the International Conference on Bioinformatics Models, Methods and Algorithms. Scitepress, Setúbal, 2015, pp. 128–133. doi

[16] *N. Kochan, G. Y. Tütüncü, G. Giner*: A new local covariance matrix estimation for the classification of gene expression profiles in high dimensional RNA-Seq data. Expert Systems Appl. *167* (2021), Article ID 114200, 5 pages. doi

[17] *C. M. Le, K. Levin, P. J. Bickel, E. Levina*: Comment: Ridge regression and regularization of large matrices. Technometrics *62* (2020), 443–446. MR doi

[18] *O. Ledoit, M. Wolf*: A well-conditioned estimator for large-dimensional covariance matrices. J. Multivariate Anal. *88* (2004), 365–411. zbl MR doi

[19] *C.-N. Li, P.-W. Ren, Y.-R. Guo, Y.-F. Ye, Y.-H. Shao*: Regularized linear discriminant analysis based on generalized capped $\ell_{2,q}$-norm. To appear in Ann. Oper. Res. doi
[20] *L.-H. Lim, R. Sepulchre, K. Ye*: Geometric distance between positive definite matrices of different dimensions. IEEE Trans. Inf. Theory *65* (2019), 5401–5405. zbl MR doi
[21] *J. A. D. Massignan, J. B. A. London, M. Bessani, C. D. Maciel, R. Z. Fannucchi, V. Miranda*: Bayesian inference approach for information fusion in distribution system state estimation. IEEE Trans. Smart Grid *13* (2022), 526–540. doi
[22] *X. Mestre*: On the asymptotic behavior of the sample estimates of eigenvalues and eigenvectors of covariance matrices. IEEE Trans. Signal Process. *56* (2008), 5353–5368. zbl MR doi
[23] *E. Raninen, E. Ollila*: Coupled regularized sample covariance matrix estimator for multiple classes. IEEE Trans. Signal Process. *69* (2021), 5681–5692. MR doi
[24] *E. Raninen, D. E. Tyler, E. Ollila*: Linear pooling of sample covariance matrices. IEEE Trans. Signal Process. *70* (2022), 659–672. MR doi
[25] *C. Scheidegger, J. Hörrmann, P. Bühlmann*: The weighted generalised covariance measure. J. Mach. Learn. Res. *23* (2022), Article ID 273, 68 pages. MR
[26] *H. Tsukuma, T. Kubokawa*: Unified improvements in estimation of a normal covariance matrix in high and low dimensions. J. Multivariate Anal. *143* (2016), 233–248. zbl MR doi
[27] *W. N. van Wieringen, C. F. W. Peeters*: Ridge estimation of inverse covariance matrices from high-dimensional data. Comput. Stat. Data Anal. *103* (2016), 284–303. zbl MR doi
[28] *R. Vershynin*: How close is the sample covariance matrix to the actual covariance matrix? J. Theor. Probab. *25* (2012), 655–686. zbl MR doi
[29] *H. Wang, B. Peng, D. Li, C. Leng*: Nonparametric estimation of large covariance matrices with conditional sparsity. J. Econom. *223* (2021), 53–72. zbl MR doi
[30] *D. I. Warton*: Penalized normal likelihood and ridge regularization of correlation and covariance matrices. J. Am. Stat. Assoc. *103* (2008), 340–349. zbl MR doi
[31] *D. M. Witten, R. Tibshirani*: Covariance-regularized regression and classification for high dimensional problems. J. R. Stat. Soc., Ser. B, Stat. Methodol. *71* (2009), 615–636. zbl MR doi
[32] *B. Xi, J. Li, Y. Li, R. Song, D. Hong, J. Chanussot*: Few-shot learning with class-covariance metric for hyperspectral image classification. IEEE Trans. Image Process. *31* (2022), 5079–5092. doi
[33] *L. Xue, S. Ma, H. Zou*: Positive-definite ℓ_1-penalized estimation of large covariance matrices. J. Am. Stat. Assoc. *107* (2012), 1480–1491. zbl MR doi
[34] *Y. Yang, J. Zhou, J. Pan*: Estimation and optimal structure selection of high-dimensional Toeplitz covariance matrix. J. Multivariate Anal. *184* (2021), Article ID 104739, 17 pages. zbl MR doi
[35] *Y. Yin*: Spectral statistics of high dimensional sample covariance matrix with unbounded population spectral norm. Bernoulli *28* (2022), 1729–1756. zbl MR doi
[36] *R. Yuasa, T. Kubokawa*: Ridge-type linear shrinkage estimation of the mean matrix of a high-dimensional normal distribution. J. Multivariate Anal. *178* (2020), Article ID 104608, 18 pages. zbl MR doi
[37] *H. Zhang, J. Jia*: Elastic-net regularized high-dimensional negative binomial regression: Consistency and weak signals detection. Stat. Sin. *32* (2022), 181–207. zbl MR doi
[38] *Y. Zhang, Y. Zhou, X. Liu*: Applications on linear spectral statistics of high-dimensional sample covariance matrix with divergent spectrum. Comput. Stat. Data Anal. *178* (2023), Article ID 107617, 19 pages. zbl MR doi

Authors' address: *Yi Zhou*, *Bin Zhang* (corresponding author), School of Mathematics and Statistics, Guangxi Normal University, 15 Yucai Road, Qixing District, Guilin, Guangxi 541004, P. R. China, e-mail: `jyzyy98@163.com`, `binzhang@gxnu.edu.cn`.