Statistica Sinica **35** (2025), 1991-2012
doi:https://doi.org/10.5705/ss.202022.0380

HIGH-DIMENSIONAL SCALE INVARIANT DISCRIMINANT ANALYSIS

Ming Li, Cheng Wang, Yanqing Yin and Shurong Zheng*

Shandong Technology and Business University, Shanghai Jiao Tong University, Nanjing Audit University and Northeast Normal University

Abstract: In this paper, we propose a scale invariant linear discriminant analysis classifier for high-dimensional data with dense signals. The method is valid for both cases that the data dimension is smaller or greater than the sample size. Based on recent advances of the sample correlation matrix in random matrix theory, we derive the asymptotic limits of the error rate which characterizes the influences of the data dimension and the tuning parameter. The major advantage of our proposed classifier is scale invariant and it is applicable to any variances of the feature. Several numerical studies are investigated and our proposed classifier performs favorably in comparison to some existing methods.

Key words and phrases: Dimension effect, discriminant analysis, random matrix theory, sample correlation matrix, scale invariant.

1. Introduction

Linear discriminant analysis (LDA), which can be dated back to Fisher (1936), is a fundamental problem in multivariate statistical analysis (Anderson, 2003, Chap. 6). From the perspective of methodology, LDA is closely related to many other important statistical methods such as principal component analysis, analysis of variance and regression analysis etc. In real problems, LDA usually has a reliable performance and can compete with many sophisticated methods such as neural networks and support vector machines (Hand, 2006).

In the era of high-dimensional data, many improved methods have been proposed for linear discriminant analysis. For example, to address the singularity of the sample covariance matrix, Dudoit, Fridlyand and Speed (2002) proposed a diagonal linear discriminant analysis (DLDA) which is valid for the situation that the data dimension p is greater than the sample size n. To reduce the dispersion of the eigenvalues of the sample covariance matrix, Friedman (1989) conducted a regularized linear discriminant analysis (RLDA), and see also Guo, Hastie and Tibshirani (2007) for RLDA. Moreover, various other improved LDA methods were studied under some certain assumptions. One may refer to the papers (Shao et al., 2011; Cai and Liu, 2011; Mai, Zou and Yuan, 2012; Fan, Feng and Tong, 2012; Fan, Jin and Yao, 2013; Hao, Dong and Fan, 2015) and references therein

*Corresponding author. E-mail: zhengsr@nenu.edu.cn

for high-dimensional discriminant analysis under sparse assumption. Chen and Tang (2021) considered the classification of high-dimensional data with spiked covariance matrix structure while Jiang et al. (2021) studied high-dimensional classification with mixed variables. Auguin, Morales-Jimenez and McKay (2021) proposed a robust classifier when there are outlying samples in the training data and Park, Baek and Park (2022) used non-parametric methods for high-dimensional discriminant analysis. Cai and Zhang (2019) considered linear discriminant analysis with missing data.

To better understand the performance of LDA for high-dimensional data, one may focus on the effect of a diverging dimension p. For instance, Bickel and Levina (2004) found that the well-known linear discriminant analysis is no better than random guessing as $p/n \to \infty$. Shao et al. (2011) further showed that the empirical LDA is consistent if and only if $p/n \to 0$. For the asymptotic regime that the data dimension p is comparable to the sample size n, Dobriban and Wager (2018) studied the misclassification rate of RLDA, and Wang and Jiang (2018) further extended the result to general settings. The technical analysis builds on results of the sample covariance matrix in the random matrix theory.

In specific, considering two classes $\Pi_1 : N(\mathbf{u}_1, \mathbf{\Sigma})$ and $\Pi_2 : N(\mathbf{u}_2, \mathbf{\Sigma})$, we have independent and identically distributed samples $\{\mathbf{x}_{i,j} : \ i = 1, 2, \ j = 1, \ldots, n_i\}$ where $\mathbf{x}_{1,1}, \ldots, \mathbf{x}_{1,n_1} \sim N(\mathbf{u}_1, \mathbf{\Sigma})$ and $\mathbf{x}_{2,1}, \ldots, \mathbf{x}_{2,n_2} \sim N(\mathbf{u}_2, \mathbf{\Sigma})$. Here, $\mathbf{u}_1, \ \mathbf{u}_2 \in \mathbb{R}^p$ are the two population means and $\mathbf{\Sigma} \in \mathbb{R}^{p \times p}$ is the covariance matrix. Based on the samples, we define the pooled sample covariance matrix

$$\mathbf{S}_n = (n_1 + n_2 - 2)^{-1} \sum_{i=1}^{2} \sum_{j=1}^{n_i} (\mathbf{x}_{i,j} - \bar{\mathbf{x}}_i)(\mathbf{x}_{i,j} - \bar{\mathbf{x}}_i)^{\mathrm{T}}, \tag{1.1}$$

where $\bar{\mathbf{x}}_1 = n_1^{-1} \sum_{j=1}^{n_1} \mathbf{x}_{1,j}$ and $\bar{\mathbf{x}}_2 = n_2^{-1} \sum_{j=1}^{n_2} \mathbf{x}_{2,j}$ are the sample means. In this work, we study a Scale-Invariant Discriminant Analysis (SIDA)

$$\text{SIDA} : \ D_s(\mathbf{x}) = \delta \left\{ \left(\mathbf{x} - \frac{\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2}{2} \right)^{\mathrm{T}} \left[\mathbf{S}_n + \lambda \cdot \mathrm{diag}(\mathbf{S}_n) \right]^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) > 0 \right\}, \tag{1.2}$$

where $\mathbf{x} \in \mathbb{R}^p$ is a new observation, $\lambda \geq 0$ is a tuning parameter. Here, $\delta(\cdot)$ is an indicator function and $\mathrm{diag}(\mathbf{S}_n)$ being a diagonal matrix from the diagonal elements of a matrix $\mathbf{S}_n$. Discriminant rule is as follows

- If $D_s(\mathbf{x}) = 1$, then $\mathbf{x}$ is regarded as being from the population Π_1;
- If $D_s(\mathbf{x}) = 0$, then $\mathbf{x}$ is regarded as being from the population Π_2.

When $\lambda = 0$ and the sample covariance matrix is invertible, the SIDA will be reduced to the classical Fisher's LDA. If the tuning parameter tends to infinity, the SIDA will comes to DLDA (Dudoit, Fridlyand and Speed, 2002). If $\mathrm{diag}(\mathbf{S}_n)$ is replaced by an p-dimension identity matrix $\mathbf{I}_p$, we have the RLDA (Friedman,

Table 1. Percentages of misclassification rates of RLDA.

	p=100			p=150			p=200		
n	40	120	200	60	180	300	80	240	400
RLDA_1	31.64	32.02	21.52	29.16	27.25	20.25	**24.26**	23.73	**13.35**
RLDA_2	44.89	35.87	29.26	43.86	36.02	27.83	**40.67**	29.05	**26.19**

1989; Guo, Hastie and Tibshirani, 2007), i.e.,

$$\text{RLDA}: \ D_r(\mathbf{x}) = \delta\left[\left(\mathbf{x} - \frac{\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2}{2}\right)^{\mathrm{T}} (\mathbf{S}_n + \lambda \mathbf{I}_p)^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) > 0\right]. \tag{1.3}$$

One obvious advantage of SIDA over RLDA is that SIDA is scale-invariant. That is, for any diagonal matrix $\mathbf{\Lambda} = \text{diag}\{\sigma_1, \ldots, \sigma_p\}$, the performance of SIDA is invariant if we make the transformations $\mathbf{x}_{i,j} \to \mathbf{\Lambda}\mathbf{x}_{i,j}$ and $\mathbf{x} \to \mathbf{\Lambda}\mathbf{x}$. Dobriban and Wager (2018) and Wang and Jiang (2018) analyzed the misclassification rates of RLDA. In this work, we study the performance of SIDA under the regime that the data dimension is comparable with the sample size.

In analysis of real data, the normalization is a usual procedure. Given the training data $\{\mathbf{x}_{i,j} : \ i = 1, 2, \ j = 1, \ldots, n_i\}$ and the test data $\mathbf{x}$, we can re-scale the data as follows:

$$\mathbf{y}_{i,j} = [\text{diag}(\mathbf{S}_n)]^{-1/2} \mathbf{x}_{i,j}, \ \mathbf{y} = [\text{diag}(\mathbf{S}_n)]^{-1/2} \mathbf{x}$$

For the normalized data $\{\mathbf{y}_{i,j} : \ i = 1, 2, \ j = 1, \ldots, n_i\}$ and $\mathbf{y}$, the RLDA is equivalent to our SIDA. In other works, if we make the data normalization and then apply RLDA to a real data, we are actually using the SIDA method. Thus, studying SIDA is more realistic to the real applications.

Furthermore, from the perspective of statistical methods, the scale-invariant property is attractive and can bring the improvement of the method. To illustrate this issue of RLDA and SIDA, we conduct a toy experiment. In details, we simulate the data as follows.

- The data dimension p is taken as $p = 100, 150, 200$ and $p/n = 5/2, 5/6, 1/2$, where $n = n_1 + n_2$. Set $\lambda = 0.1$;
- For RLDA_1, the data $\{\mathbf{x}_{i,j} : \ i = 1, 2, \ j = 1, \ldots, n_i\}$ are an i.i.d. sample from a p-dimensional population with mean vector $\mathbf{u}_i$ and covariance matrix $\mathbf{\Sigma}$ for $i = 1, 2$ with the elements of $\mathbf{u}_1$ being i.i.d. from the standard uniform distribution, $\mathbf{u}_2 = (0, 0, \ldots, 0)^{\mathrm{T}}$ and $\mathbf{\Sigma} = \text{diag}(5, 10, \ldots, 5, 10)$.
- For RLDA_2, the data are $\{\mathbf{\Lambda}\mathbf{x}_{i,j} : \ i = 1, 2, \ j = 1, \ldots, n_i\}$ with $\mathbf{\Lambda} = \text{diag}(0.001, 1000, \ldots, 0.001, 1000)$.

From Table 1, we can see that the data $\{\mathbf{x}_{i,j} : \ i = 1, 2, \ j = 1, \ldots, n_i\}$ and $\{\mathbf{\Lambda x}_{i,j} : \ i = 1, 2, \ j = 1, \ldots, n_i\}$ are different only in the scales. But the misclassification rates of RLDA_1 and RLDA_2 are very different, for example, **24.26** and **40.67**. Thus, the target of this paper is to propose a scale-invariant discriminant analysis classifier which is valid for both the low-dimensional situation and the high-dimensional situation.

The present work is motivated by the recent interest in ridge regression (Dobriban and Liu, 2019; Liu and Dobriban, 2020; Kobak, Lomond and Sanchez, 2020; Dobriban and Sheng, 2021; Hastie et al., 2022). The ridge regression is a bias-variance tradeoff. Under high-dimensional setting that the number of unknown parameters p is of the same order as the number of samples n, several interesting phenomena are observed for the classical ridge regression. For example, Hastie et al. (2022) recovered the "double descent" behavior in the simple linear model. Liu and Dobriban (2020) proposed a bias-correction to the tuning parameter of the well-known $K-$fold cross-validation procedure and Kobak, Lomond and Sanchez (2020) provided a situation where the optimal value of the ridge penalty can be negative. The regularized LDA connects to the ridge regression and intuitively, we expect that these interesting phenomena also hold for RLDA. Furthermore, from the perspective of statistical methods, the scale-invariant property can provide the classification improvement. To verify these intuitions, this paper studies the properties of the scale-invariant discriminant analysis classifier and we leave the demonstration of other phenomena as future works.

The remainder of this paper is organized as follows. In Section 2, we present the assumptions and the main results. Specially, we derive the explicit limits of the misclassification error rates and propose a bias correction to the SIDA. Several interesting examples are also discussed in this part. To demonstrate the performance of SIDA, we conduct several simulation studies in Section 3 and we conclude the paper in Section 4 with discussions. The proofs of main results and technical details are relegated to the Appendix.

2. Scale Invariant Discriminant Analysis

Given the classifier

$$\text{SIDA}: \ D_s(\mathbf{x}) = \delta\left\{\left(\mathbf{x} - \frac{\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2}{2}\right)^{\mathrm{T}} \left[\mathbf{S}_n + \lambda \cdot \operatorname{diag}(\mathbf{S}_n)\right]^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) > 0\right\},$$

and assuming $\mathbf{x} \sim N(\mathbf{u}_1, \mathbf{\Sigma})$ or $\mathbf{x} \sim N(\mathbf{u}_2, \mathbf{\Sigma})$ with equal prior probability, we can get the misclassification rate of SIDA conditional on the samples

$$R_n = \frac{1}{2} P\left[D_s(\mathbf{x}) = 0 | \mathbf{x} \sim N(\mathbf{u}_1, \mathbf{\Sigma})\right] + \frac{1}{2} P\left[D_s(\mathbf{x}) = 1 | \mathbf{x} \sim N(\mathbf{u}_2, \mathbf{\Sigma})\right] \tag{2.1}$$

$$= \frac{1}{2}\sum_{i=1}^{2}\Phi\left\{\frac{(-1)^i\left(2\mathbf{u}_i-\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2\right)^{\mathrm{T}}\left[\mathbf{S}_n+\lambda\cdot\mathrm{diag}(\mathbf{S}_n)\right]^{-1}\left(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2\right)}{2\sqrt{(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)^{\mathrm{T}}\left[\mathbf{S}_n+\lambda\cdot\mathrm{diag}(\mathbf{S}_n)\right]^{-1}\boldsymbol{\Sigma}\left[\mathbf{S}_n+\lambda\cdot\mathrm{diag}(\mathbf{S}_n)\right]^{-1}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)}}\right\},$$

where $\Phi(\cdot)$ is the distribution function of the standard normal population. The main contribution of this work is to derive the asymptotic properties of R_n from which we can gain insights of how the increasing dimension and the tuning parameter λ affect the classification rate.

First of all, we review some key basics and notations in random matrix theory (RMT). Suppose $\mathbf{D}_m$ is an $m\times m$ Hermitian matrix with real eigenvalues $\lambda_j, j=1,\ldots,m$. The empirical spectral distribution (ESD) of the matrix $\mathbf{D}_m$ is defined as

$$F^{\mathbf{D}_m}(x)=\frac{1}{m}\sum_{j=1}^{m}\mathbf{I}(\lambda_j\le x).$$

One of the important problems in RMT is to investigate the convergence of the sequence of empirical spectral distributions $F^{\mathbf{D}_m}$ for a given sequence of random matrices $\{\mathbf{D}_m\}$. The limit of $F^{\mathbf{D}_m}$, if it exists and is non-random, is called the limiting spectral distribution (LSD) of the sequence $\mathbf{D}_m$. In RMT, the Stieltjes transform of a function $F(x)$ bounded variation on the real line is defined by

$$m_F(z)=\int\frac{1}{t-z}dF(t), z\in\mathbb{C}^+,$$

where $\mathbb{C}^+=\{z\in\mathbb{C}:\Im z>0\}$ and $\Im z$ is the imaginary part of z. For any continuity points $a<b$ of F, one could get

$$F\{[a,b]\}=\lim_{\varepsilon\to0^+}\frac{1}{\pi}\int_a^b\Im m_F(x+i\varepsilon)dx,$$

which is the famous Inversion formula. The Inversion formula shows a one-to-one correspondence between the distribution functions and their Stieltjes transforms.

Noting $\bar{\mathbf{x}}_1,\ \bar{\mathbf{x}}_2$ and $\mathbf{S}_n$ are independent, we can write

$$\bar{\mathbf{x}}_1=\mathbf{u}_1+\frac{1}{\sqrt{n_1}}\boldsymbol{\Sigma}^{1/2}\mathbf{z}_1,\ \bar{\mathbf{x}}_2=\mathbf{u}_2+\frac{1}{\sqrt{n_2}}\boldsymbol{\Sigma}^{1/2}\mathbf{z}_2$$

where $\mathbf{z}_1,\ \mathbf{z}_2\sim N(\mathbf{0},\mathbf{I})$ are independent with $\mathbf{S}_n$. Then, the numerator parts of the error rate (2.1) are

$$\begin{aligned}&-\left(2\mathbf{u}_1-\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2\right)^{\mathrm{T}}\left[\mathbf{S}_n+\lambda\cdot\mathrm{diag}(\mathbf{S}_n)\right]^{-1}\left(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2\right)\\&=-\tilde{\mathbf{u}}^{\mathrm{T}}\mathbf{A}_n\tilde{\mathbf{u}}+\left(\frac{1}{n_1}-\frac{1}{n_2}\right)\mathrm{tr}\left(\mathbf{A}_n\right)+o_p(1),\\&\left(2\mathbf{u}_2-\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2\right)^{\mathrm{T}}\left[\mathbf{S}_n+\lambda\cdot\mathrm{diag}(\mathbf{S}_n)\right]^{-1}\left(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2\right)\end{aligned}$$

$$= -\tilde{\mathbf{u}}^{\mathrm{T}}\mathbf{A}_n\tilde{\mathbf{u}} - \left(\frac{1}{n_1} - \frac{1}{n_2}\right)\mathrm{tr}\left(\mathbf{A}_n\right) + o_p(1),$$

and the denominator part is

$$\begin{aligned}&\left(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2\right)^{\mathrm{T}}\left[\mathbf{S}_n + \lambda\cdot\mathrm{diag}(\mathbf{S}_n)\right]^{-1}\boldsymbol{\Sigma}\left[\mathbf{S}_n + \lambda\cdot\mathrm{diag}(\mathbf{S}_n)\right]^{-1}\left(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2\right)\\ &= \tilde{\mathbf{u}}^{\mathrm{T}}\mathbf{A}_n^2\tilde{\mathbf{u}} + \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\mathrm{tr}\left(\mathbf{A}_n^2\right) + o_p(1),\end{aligned}$$

where

$$\tilde{\mathbf{u}} = \boldsymbol{\Sigma}^{-1/2}\mathbf{u}_1 - \boldsymbol{\Sigma}^{-1/2}\mathbf{u}_2,\ \mathbf{A}_n = \boldsymbol{\Sigma}^{1/2}\left[\mathbf{S}_n + \lambda\cdot\mathrm{diag}(\mathbf{S}_n)\right]^{-1}\boldsymbol{\Sigma}^{1/2}.$$

To derive the limits of these terms, we need to study the eigenvalues of $\mathbf{A}_n$ and the quadratic forms involving $\mathbf{A}_n$ (Karoui and Kösters, 2011). The technical challenge is to deal with two random matrices $\mathbf{S}_n$ and $\mathrm{diag}(\mathbf{S}_n)$. Some knowing results in random matrix theory about the sample covariance matrix (e.g., Karoui and Kösters, 2011; Ledoit and Péché, 2011) are not applicable. Noting

$$\mathbf{A}_n = \boldsymbol{\Sigma}^{1/2}\mathrm{diag}(\mathbf{S}_n)^{-1/2}(\mathbf{R}_n + \lambda\mathbf{I}_p)^{-1}\mathrm{diag}(\mathbf{S}_n)^{-1/2}\boldsymbol{\Sigma}^{1/2},$$

where $\mathbf{R}_n$ is the sample correlation matrix, we need to study asymptotic properties of the correlation matrix (El Karoui, 2009; Yin, Zheng and Zou, 2023). The key element of the argument is to consider another random matrix

$$\mathbf{B}_n = \boldsymbol{\Sigma}^{1/2}\left[\mathbf{S}_n + \lambda\cdot\mathrm{diag}(\boldsymbol{\Sigma})\right]^{-1}\boldsymbol{\Sigma}^{1/2},$$

and we bound the difference between $\mathbf{A}_n$ and $\mathbf{B}_n$. We will present the details in next section.

2.1. Main results

Without loss of generality, we let $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\mathbf{R}\boldsymbol{\Lambda}$ where $\boldsymbol{\Lambda} = \mathrm{diag}(\sigma_1, \ldots, \sigma_p)$, $\sigma_i > 0$ and $\mathbf{R}$ is a correlation matrix. We first present the assumptions on the populations means $\mathbf{u}_1, \mathbf{u}_2$ and the population covariance matrix $\boldsymbol{\Sigma}$.

(C1) $p/n_1 \to y_1 \in (0,\infty)$, $p/n_2 \to y_2 \in (0,\infty)$ and denote $y = y_1y_2/(y_1+y_2)$.

(C2) For the population correlation matrix $\mathbf{R}$, the eigenvalues of $\mathbf{R}$ are bounded, i.e., $1/c \le \lambda(\mathbf{R}) \le c$ for some $c > 1$. The empirical spectral distribution of the eigenvalues converges to a distribution function $G(\cdot)$ in distribution.

(C3) For any $t \ge 0$ and $p \to +\infty$,

$$\begin{aligned}&\Delta^{-2}\left(\boldsymbol{\Lambda}^{-1}\mathbf{u}_1 - \boldsymbol{\Lambda}^{-1}\mathbf{u}_2\right)^{\mathrm{T}}\left(\mathbf{R} + t\mathbf{I}_p\right)^{-1}\left(\boldsymbol{\Lambda}^{-1}\mathbf{u}_1 - \boldsymbol{\Lambda}^{-1}\mathbf{u}_2\right) \to f_1(t),\\ &\Delta^{-2}\left(\boldsymbol{\Lambda}^{-1}\mathbf{u}_1 - \boldsymbol{\Lambda}^{-1}\mathbf{u}_2\right)^{\mathrm{T}}\left(\mathbf{R} + t\mathbf{I}_p\right)^{-2}\left(\boldsymbol{\Lambda}^{-1}\mathbf{u}_1 - \boldsymbol{\Lambda}^{-1}\mathbf{u}_2\right) \to f_2(t).\end{aligned}$$

Assumptions (C1)–(C2) are two common conditions in random matrix theory. Assumption (C3) could be regarded as a technical assumption to express explicit limits. For brevity, we let $\Delta = \sqrt{(\mathbf{u}_1 - \mathbf{u}_2)^{\mathrm{T}} \mathbf{\Sigma}^{-1} (\mathbf{u}_1 - \mathbf{u}_2)}$ be a constant. Compared with the conditions of Wang and Jiang (2018) and Dobriban and Wager (2018), the main difference is that our assumption is on the correlation matrix $\mathbf{R}$ and we do not impose any conditions on the variances σ_i^2. Thus, our proposed SIDA is scale invariant and can handle the data with very different scales.

Let $m_0(z)$ be the Stieltjes transform of the limiting spectral distribution with $G(\cdot)$, i.e., $m_0(z)$ is the unique solution to the Marčenko–Pastur equation

$$m_0(z) = \int \frac{1}{t\left[1 - y - yzm_0(z)\right] - z} dG(t). \tag{2.2}$$

Theorem 1. *Under Conditions* (C1)–(C2)*, for any* $\lambda > 0$*, as* $n \to \infty$*, we have*

$$p^{-1}\mathrm{tr}\left(\mathbf{A}_n\right) \xrightarrow{a.s.} h_1(\lambda), \tag{2.3}$$

$$p^{-1}\mathrm{tr}\left(\mathbf{A}_n^2\right) \xrightarrow{a.s.} h_2(\lambda), \tag{2.4}$$

where

$$h_1(\lambda) = \frac{1 - \lambda m_0(-\lambda)}{1 - y\left[1 - \lambda m_0(-\lambda)\right]},$$

$$h_2(\lambda) = \frac{1 - \lambda m_0(-\lambda)}{\left\{1 - y\left[1 - \lambda m_0(-\lambda)\right]\right\}^3} - \frac{\lambda m_0(-\lambda) - \lambda^2 m_0'(-\lambda)}{\left\{1 - y\left[1 - \lambda m_0(-\lambda)\right]\right\}^4},$$

and $m_0'(-\lambda)$ *is the derivative of the Stieltjes transform* $m_0(z)$ *evaluated at* $z = -\lambda$*.*

The condition (C3) is used to derive the limits of the quadratic forms involving $\mathbf{A}_n$. In summary, we can get the limit of the misclassification error rate.

Theorem 2. *Under Conditions* (C1)–(C3)*, we have*

$$R_n \xrightarrow{p} \frac{1}{2} \sum_{i=1}^{2} \Phi\left[-\frac{H_1(\lambda)\Delta^2 + (-1)^i (y_1 - y_2) h_1(\lambda)}{2\sqrt{H_2(\lambda)\Delta^2 + (y_1 + y_2) h_2(\lambda)}}\right], \tag{2.5}$$

where

$$H_1(\lambda) = \frac{1}{1 - y\left[1 - \lambda m_0(-\lambda)\right]} f_1\left\{\frac{\lambda}{1 - y\left[1 - \lambda m_0(-\lambda)\right]}\right\},$$

$$H_2(\lambda) = \left\{\left[1 + y h_1(\lambda)\right]^2 + y h_2(\lambda)\right\} f_2\left\{\frac{\lambda}{1 - y\left[1 - \lambda m_0(-\lambda)\right]}\right\}.$$

Theorem 2 provides the explicit effects of the ratios p/n_1, p/n_2 and the tuning parameter λ. Before presenting the illustration of the results, we turn to

a bias correction for unequal sample sizes.

2.2. Bias correction

From (2.5), we can see that the misclassification rates are different for the two classes, i.e.,

$$\Phi\left[-\frac{H_1(\lambda)\Delta^2+(-1)(y_1-y_2)h_1(\lambda)}{2\sqrt{H_2(\lambda)\Delta^2+(y_1+y_2)h_2(\lambda)}}\right]\neq\Phi\left[-\frac{H_1(\lambda)\Delta^2+(y_1-y_2)h_1(\lambda)}{2\sqrt{H_2(\lambda)\Delta^2+(y_1+y_2)h_2(\lambda)}}\right]$$

as $y_1 \neq y_2$. That is, the unequal sample sizes lead to different misclassification rates. This is due to the estimation bias of intercept part in SIDA. As we know, $\Phi(x)$ is strictly convex for $x < 0$, which implies that the misclassification rate can be improved if we can remove the unnecessary term $(y_1 - y_2)h_1(\lambda)$.

To reduce the bias of SIDA brought by the different sample sizes, following the bias correction in Wang and Jiang (2018), we consider the following classifier,

$$D(\mathbf{x}) = \delta\{\mathbf{x}^T\left[\mathbf{S}_n + \lambda\mathrm{diag}(\mathbf{S}_n)\right]^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) + \alpha > 0\}. \tag{2.6}$$

By the Proposition 2 in Mai, Zou and Yuan (2012), when the classification direction is $\left[\mathbf{S}_n + \lambda\mathrm{diag}(\mathbf{S}_n)\right]^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$, the optimal intercept corresponding to minimum misclassification rate is

$$\alpha_o = -\frac{1}{2}(\mathbf{u}_1+\mathbf{u}_2)^T\left[\mathbf{S}_n + \lambda\mathrm{diag}(\mathbf{S}_n)\right]^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2),$$

while for SIDA the intercept is set to be

$$\alpha_s = -\frac{1}{2}(\bar{\mathbf{x}}_1+\bar{\mathbf{x}}_2)^T\left[\mathbf{S}_n + \lambda\mathrm{diag}(\mathbf{S}_n)\right]^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2).$$

Then,

$$\begin{aligned}
\alpha_s - \alpha_o &= -\frac{1}{2n_1}\mathbf{z}_1^T\boldsymbol{\Sigma}^{1/2}\left[\mathbf{S}_n + \lambda\mathrm{diag}(\mathbf{S}_n)\right]^{-1}\boldsymbol{\Sigma}^{1/2}\mathbf{z}_1 \\
&\quad + \frac{1}{2n_2}\left(\mathbf{z}_2\right)^T\boldsymbol{\Sigma}^{1/2}\left[\mathbf{S}_n + \lambda\mathrm{diag}(\mathbf{S}_n)\right]^{-1}\boldsymbol{\Sigma}^{1/2}\mathbf{z}_2 \\
&\quad + \frac{1}{2}\left(\frac{1}{\sqrt{n_1}}\boldsymbol{\Sigma}^{1/2}\mathbf{z}_1 + \frac{1}{\sqrt{n_2}}\boldsymbol{\Sigma}^{1/2}\mathbf{z}_2\right)^T\left[\mathbf{S}_n + \lambda\mathrm{diag}(\mathbf{S}_n)\right]^{-1}(\mathbf{u}_1 - \mathbf{u}_2) \\
&= -\frac{1}{2n_1}\mathbf{z}_1^T\boldsymbol{\Sigma}^{1/2}\left[\mathbf{S}_n + \lambda\mathrm{diag}(\mathbf{S}_n)\right]^{-1}\boldsymbol{\Sigma}^{1/2}\mathbf{z}_1 \\
&\quad + \frac{1}{2n_2}\mathbf{z}_2^T\boldsymbol{\Sigma}^{1/2}\left[\mathbf{S}_n + \lambda\mathrm{diag}(\mathbf{S}_n)\right]^{-1}\boldsymbol{\Sigma}^{1/2}\mathbf{z}_2 + o_p(1) \\
&\triangleq -\alpha + o_p(1),
\end{aligned}$$

As can be noticed, the α depends on the population covariance matrix $\mathbf{\Sigma}$ which is unknown in practice, we need to find a consistent estimator of it. Based on the formula of $h_1(\lambda)$, we propose the corrected SIDA

$$D_s^c(\mathbf{x}) = \delta\left\{\left(\mathbf{x} - \frac{\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2}{2}\right)^{\mathrm{T}} \left[\mathbf{S}_n + \lambda \cdot \mathrm{diag}(\mathbf{S}_n)\right]^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) + \widehat{\alpha} > 0\right\}, \quad (2.7)$$

with

$$\widehat{\alpha} = \left(\frac{p}{2n_1} - \frac{p}{2n_2}\right) \frac{1 - (1/p)\mathrm{tr}\left\{(1/\lambda)\mathbf{R}_n + \mathbf{I}_p\right\}^{-1}}{1 - \{p/(n-2)\} + \{1/(n-2)\}\mathrm{tr}\left\{(1/\lambda)\mathbf{R}_n + \mathbf{I}_p\right\}^{-1}}$$

and $n = n_1 + n_2$. About the conduction of $\widehat{\alpha}$, one could refer to Chen et al. (2011) for more details.

Let R_n^c be the misclassification rate of SIDA after bias correction. Then, the misclassification rate of corrected SIDA is

$$R_n^c = \quad (2.8)$$
$$\frac{1}{2}\sum_{i=1}^{2} \Phi\left\{\frac{(-1)^i \left[(2\mathbf{u}_i - \bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^{\mathrm{T}} (\mathbf{S}_n + \lambda \cdot \mathrm{diag}(\mathbf{S}_n))^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) + 2\widehat{\alpha}\right]}{2\sqrt{(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^{\mathrm{T}} \left[\mathbf{S}_n + \lambda \cdot \mathrm{diag}(\mathbf{S}_n)\right]^{-1} \mathbf{\Sigma} \left[\mathbf{S}_n + \lambda \cdot \mathrm{diag}(\mathbf{S}_n)\right]^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)}}\right\}.$$

The dimension effect of the bias-corrected SIDA is described in Proposition 1 below.

Proposition 1. *Under the conditions of Theorem* 2*, for the corrected SIDA, we have*

$$R_n^c \xrightarrow{p} \Phi\left[-\frac{H_1(\lambda)\Delta^2}{2\sqrt{H_2(\lambda)\Delta^2 + (y_1 + y_2)h_2(\lambda)}}\right]. \quad (2.9)$$

As $\Phi(x)$ is strictly convex in $(-\infty, 0)$, which implies that the asymptotic misclassification rate of the bias-corrected SIDA given in (2.9) is smaller than that of SIDA given in (2.5).

2.3. Examples

In this part, we use several examples to illustrate our Theorem 2 and we assume $n_1 = n_2$ for brevity. Specially, we consider $\mathbf{\Sigma} = \sigma^2\mathbf{I}_p$ and $\mathbf{\Sigma} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_p^2)$. Then, the true population correlation matrix is an identity matrix, which means $f_1(t) = (1+t)^{-1}$ and $f_2(t) = (1+t)^{-2}$.

Example 1. Assuming $\mathbf{\Sigma} = \sigma^2\mathbf{I}_p$, we have

$$m_0(-\lambda) = \frac{\sqrt{(1-y+\lambda)^2 + 4y\lambda} - (1-y+\lambda)}{2y\lambda},$$
$$m_0'(-\lambda) = \frac{\lambda(1+y) + (1-y)^2}{2y\lambda^2\sqrt{(1-y+\lambda)^2 + 4y\lambda}} - \frac{1-y}{2y\lambda^2},$$

$$h_1(\lambda) = m_0(-\lambda), h_2(\lambda) = m_0'(-\lambda),$$
$$H_1(\lambda) = \frac{1}{1 - y\left[1 - \lambda m_0(-\lambda)\right] + \lambda},$$
$$H_2(\lambda) = \frac{\left\{\left[1 + y m_0(-\lambda)\right]^2 + y m_0'(-\lambda)\right\}\left\{1 - y\left[1 - \lambda m_0(-\lambda)\right]\right\}^2}{\left\{1 - y\left[1 - \lambda m_0(-\lambda)\right] + \lambda\right\}^2}.$$

Then the misclassification rate of SIDA is

$$\Phi\left[-\frac{\Delta^2}{2\sqrt{\Delta^2 + y_1 + y_2}}\sqrt{\frac{2\sqrt{(1+y+\lambda)^2 - 4y}}{\sqrt{(1+y+\lambda)^2 - 4y} + (1+y+\lambda)}}\right]. \tag{2.10}$$

As a comparison, the misclassification rate of RLDA is provided by Wang and Jiang (2018) as follows

$$\Phi\left[-\frac{\Delta^2}{2\sqrt{\Delta^2 + y_1 + y_2}}\sqrt{\frac{2\sqrt{(1+y+\lambda\sigma^{-2})^2 - 4y}}{\sqrt{(1+y+\lambda\sigma^{-2})^2 - 4y} + (1+y+\lambda\sigma^{-2})}}\right]. \tag{2.11}$$

From the error rates (2.10) and (2.11), we can see that the SIDA is not influenced by the scale σ. In other words, if we set $\lambda_1 = \sigma^2\lambda$ for RLDA, our SIDA have the exact performance as RLDA which means inducting $\mathrm{diag}(\mathbf{S}_n)$ do not loss any information even if we know all the variances are the same.

Example 2. Assuming $\mathbf{\Sigma} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_p^2)$, we have

$$m_0(-\lambda) = \frac{\sqrt{(1-y+\lambda)^2 + 4y\lambda} - (1-y+\lambda)}{2y\lambda},$$
$$m_0'(-\lambda) = \frac{\lambda(1+y) + (1-y)^2}{2y\lambda^2\sqrt{(1-y+\lambda)^2 + 4y\lambda}} - \frac{1-y}{2y\lambda^2},$$
$$h_1(\lambda) = m_0(-\lambda), h_2(\lambda) = m_0'(-\lambda).$$

Then the misclassification rate of SIDA is

$$\Phi\left[-\frac{\Delta^2}{2\sqrt{\Delta^2 + y_1 + y_2}}\sqrt{\frac{2\sqrt{(1+y+\lambda)^2 - 4y}}{\sqrt{(1+y+\lambda)^2 - 4y} + (1+y+\lambda)}}\right]. \tag{2.12}$$

From (2.12), we can see again that SIDA is not influenced by the variances $\sigma_1^2, \ldots, \sigma_p^2$. For RLDA, we can not derive the explicit result and more importantly, some additional conditions on σ_i^2 are needed to guarantee the limits.

3. Simulation

In this section, we conduct several simulations to show the performance of our proposed SIDA. For comparison, we also include the diagonal LDA ("DLDA") and the regularized LDA ("RLDA"). Here, we focus on the dense case

(Dobriban and Wager, 2018) where many features contribute to the classification rule and therefore we do not compare SIDA with many sparse methods.

We generate the training samples $\mathbf{x}_{1,1}, \ldots, \mathbf{x}_{1,n_1} \sim N(\mathbf{u}_1, \mathbf{\Sigma})$ and $\mathbf{x}_{2,1}, \ldots, \mathbf{x}_{2,n_2} \sim N(\mathbf{u}_2, \mathbf{\Sigma})$. Here the population covariance matrix and the population means are defined as follows.

- To compare these methods for different covariance structures, we consider five cases. For the first four cases, we set $\mathbf{\Sigma} = \mathbf{\Lambda R \Lambda}$ where $\mathbf{\Lambda} = \mathrm{diag}(\sigma_1, \ldots, \sigma_p)$ and $\mathbf{R} = (0.5^{|i-j|})_{p\times p}$. The variances are generated as follows:

  Case 1: $\sigma_1 = \cdots = \sigma_p = 1$;

  Case 2: $\sigma_1 = \cdots = \sigma_{p/2} = 1,\ \sigma_{p/2+1} = \cdots = \sigma_p = 10$;

  Case 3: $\sigma_1, \ldots, \sigma_p$ are i.i.d from $|Z|$ where $Z \sim \mathrm{Cauchy}(0, 1)$;

  Case 4: $\sigma_1, \ldots, \sigma_p$ are i.i.d from $|Z|$ where $Z \sim \mathrm{Cauchy}(5, 10)$.

  To evaluate the methods on strong correction structure, we also include the model

  Case 5: $\mathbf{\Sigma} = (\sigma_{ij})_{p\times p},\ \sigma_{ij} = \delta(i = j) + 0.5 \cdot \delta(i \neq j)$.

- The elements of $\mathbf{u}_1$ are independent and identically distributed from $N(0, 1)$ and $\mathbf{u}_2 = \mathbf{0_p}$. As a benchmark, $\mathbf{u}_1$ will be re-scaled to let $\mathbf{u}_1^{\mathrm{T}} \mathbf{\Sigma}^{-1} \mathbf{u}_1 = 6.57$ and the true Bayes error is 10%.

3.1. Performance of classifiers

In this part, we compare the misclassification rates of SIDA, DLDA and RLDA under cases 1–5. The tuning parameter λ is selected by 5-fold cross-validation. For each case, we fix $p = 100$ or $n = 200$. Figures 1 and 2 show the empirical misclassification rates based on 1,000 replications. The vertical axis is the percentage of empirical misclassification rates and the horizontal axis is the training sample size n or the dimension p. From these simulation results, we can see that the proposed SIDA has better performance than RLDA under cases 2–4 where the variances are heterogeneous. For case 1 and case 5 with homogeneous variances, SIDA and RLDA have similar performance. For all the cases, SIDA can achieve smaller error rates than DLDA since DLDA dose not utilize the correlation information of the covariance structure. In other words, DLDA can be considered as a special case of the SIDA method with $\lambda = \infty$. Consequently, SIDA consistently outperforms DLDA when an appropriate λ value is chosen. The performance of RLDA and DLDA varies depending on cases, as RLDA is sensitive to different variances and DLDA ignores the correlation structure. For example, RLDA can achieve a better performance on case 1 and case 5 where the variances are homogeneous and RLDA also includes the correlation

information. Conversely, for case 3 and case 4 where the variances are generated from Cauchy distributions, DLDA outperforms RLDA since DLDA is also scale-invariant. For case 2, the variances are different, but the difference in scale is not so significant. Interestingly, a phase transition can be observed, where RLDA perform better for small p/n ratios and DLDA has advantages for large p/n ratios. In summary, our proposed SIDA method is both scale-invariant and leverages correlation information.

3.2. Theoretical and empirical dimension effect of SIDA

We will compare the theoretical dimension effect in (2.5) and the empirical dimension effect of SIDA. For convenience, we set $\mathbf{\Sigma} = \mathbf{I}_p$ and $\lambda = 0.1$. The data dimension p ranges from 20 to 200 and the ratio p/n is fixed as $0.5, 1, 2$. Figure 3 presents the box plot of the error rate based on 1,000 replications. The horizontal axis is the dimension p and the vertical axis is the error rate. From Figure 3, we observe that the empirical error rates R_n converges to the theoretical results (2.10) which is consistent with the conclusion of our Theorem 2.

3.3. Bias correction of SIDA

In this subsection, we will compare "SIDA" and the corrected "SIDA" in (2.7) denoted by "C-SIDA" in Case 1. The simulation times are 1,000. As a benchmark, the linear classifier with optimal constant $\alpha_o = -(1/2)(\mathbf{u}_1 + \mathbf{u}_2)^T \left[\mathbf{S}_n + \lambda \text{diag}(\mathbf{S}_n)\right]^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$ in Mai, Zou and Yuan (2012) is also included and it is written as "O-SIDA". The total sample size is $n = n_1 + n_2 = 200$ and n_1 ranges from 30 to 170. The testing sample size is set as 100 and the data dimension p is set as $100, 200, 400$. λ is set to be $0, 0.1, 0.5$ for $p < n$ and $1, 0.5, 0.1$ for $p \geq n$. Figure 4 shows that "C-SIDA, O-SIDA" have lower empirical misclassification rates than SIDA and they have lower empirical misclassification rates when $n_1 = n_2$.

3.4. Real data analysis

In this section, we utilize two real datasets to demonstrate the performance of our proposed classifier SIDA. The first dataset is the central nervous system embryonal tumor data which was analyzed by Pomeroy et al. (2002). This dataset consists of 60 individuals, each with 7,128 gene data. The individuals are categorized into two classes: 21 in the "died" class and 39 in the "survived" class. The second dataset is the classical breast cancer data which was analyzed by Gravier et al. (2010). There are 168 individuals, with each individual having 2,905 gene data. The 168 individuals are from two classes: 111 in the "good" class and 57 in the "poor" class. Since these datasets are high-dimensional, we use a screening procedure (Fan and Lv, 2008) to select features of order $O(n)$. Specially, we employ a two-sample t test to select 50, 100, 150 or 200 important

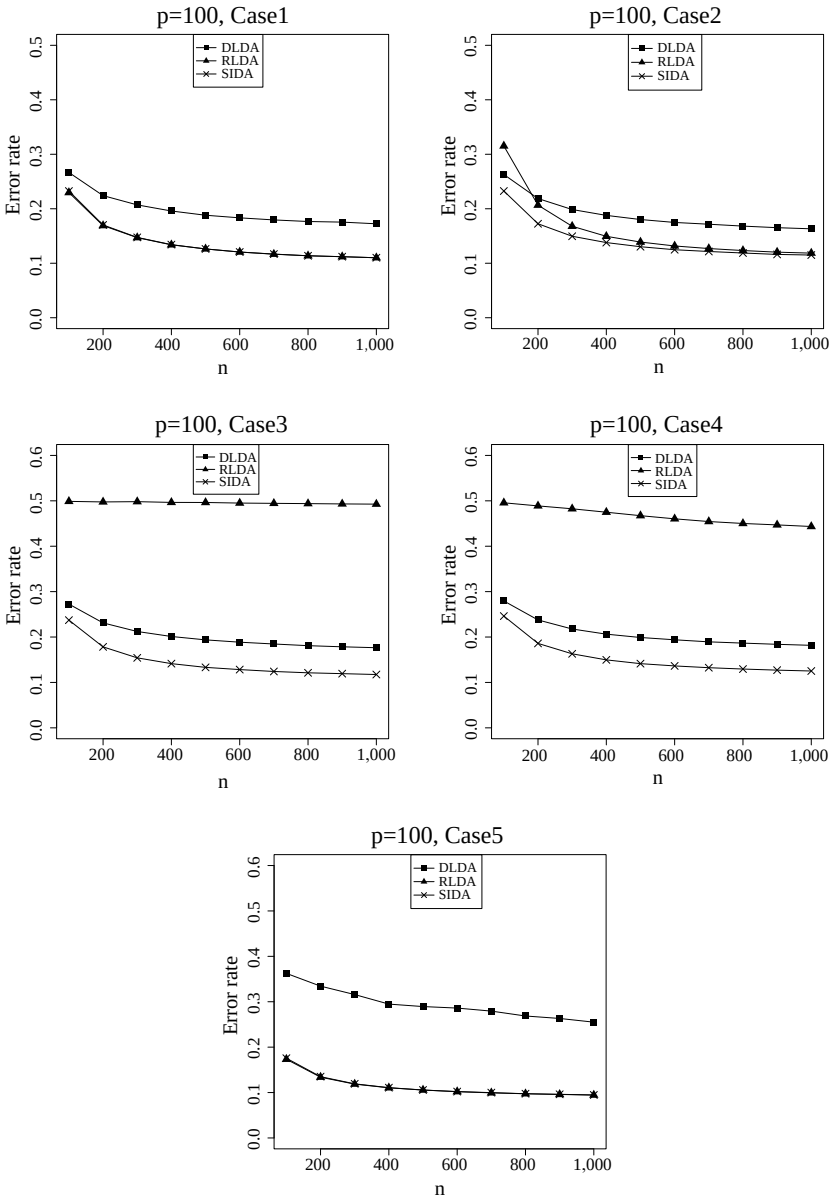


Figure 1. Misclassification rates of DLDA, RLDA and SIDA where p is fixed.

features which are comparable to the sample size. The tuning parameter λ is selected using a five-fold cross-validation approach.

Tables 2 and 3 show the empirical misclassification error rates for the embryonal tumor dataset and the breast cancer dataset, respectively. Each dataset is evaluated using two different methods: leave-one-out cross-validation (LOOCV) and random sampling, where approximately 80% of the individuals are

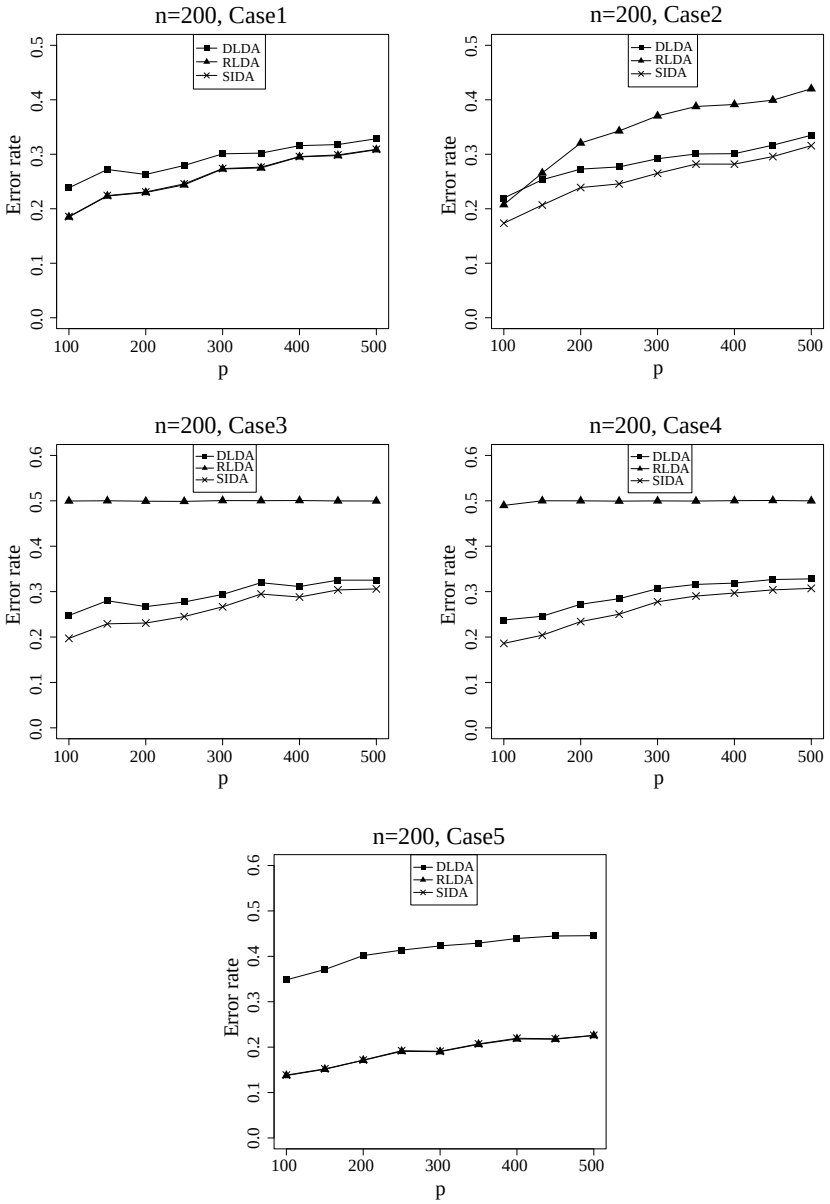


Figure 2. Misclassification rates of DLDA, RLDA and SIDA where n is fixed.

randomly selected as training data and the remaining 20% are used as test data. From Tables 2 and 3, we can see that RLDA and SIDA consistently outperform DLDA in most cases. Specifically, SIDA demonstrates superior performance on the embryonal tumor dataset, while for the breast cancer dataset, SIDA and RLDA exhibit similar error rates.

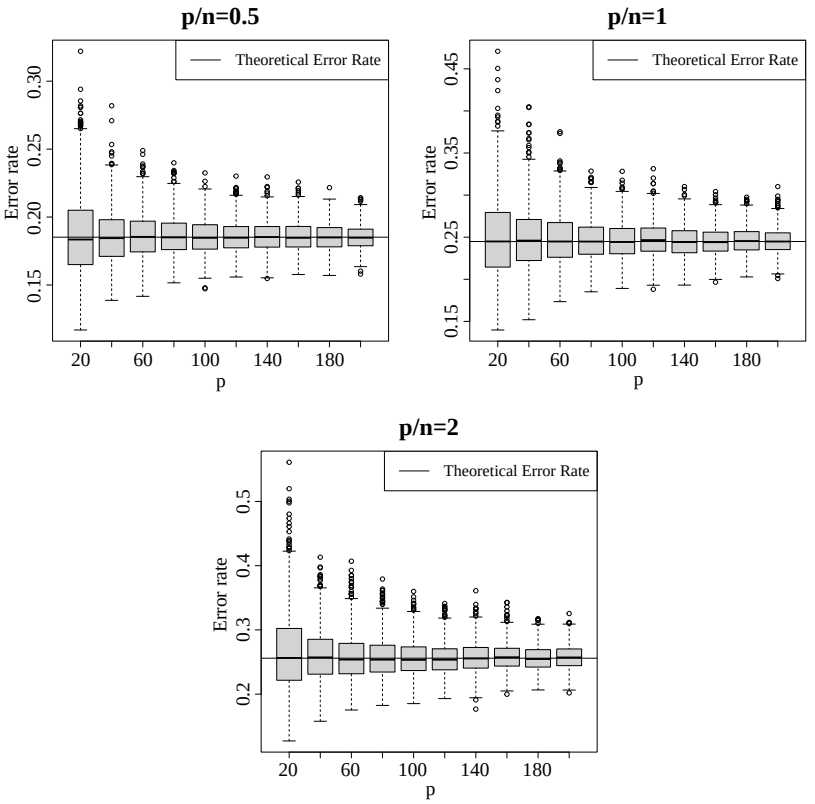


Figure 3. Consistency of theoretical and empirical dimension effects of SIDA.

Figure 5 further show the variances of the selected 200 important features. The variances of the embryonal tumor dataset show more heterogeneity, whereas the variances of the breast cancer dataset are more homogeneous. All these results demonstrate the advantage of SIDA on data with heterogeneous variances.

4. Discussions

In this paper, we study a scale-invariant discriminant analysis classifier

$$\left(\mathbf{x}-\frac{\bar{\mathbf{x}}_1+\bar{\mathbf{x}}_2}{2}\right)^{\mathrm{T}}\left[\mathbf{S}_n+\lambda\cdot\mathrm{diag}(\mathbf{S}_n)\right]^{-1}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)$$
$$=\left\{\left[\mathrm{diag}(\mathbf{S}_n)\right]^{-1/2}\left(\mathbf{x}-\frac{\bar{\mathbf{x}}_1+\bar{\mathbf{x}}_2}{2}\right)\right\}^{\mathrm{T}}\left(\mathbf{R}_n+\lambda\mathbf{I}_p\right)^{-1}\left[\mathrm{diag}(\mathbf{S}_n)\right]^{-1/2}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2).$$

In specific, we derive the limits of the misclassification error rates under the regime that the data dimension p and the sample sizes n_1, n_2 both tend to be infinite with fixed ratios. Technically, we relax the assumptions on the variance of each component, which is an improvement of Dobriban and Wager (2018) and

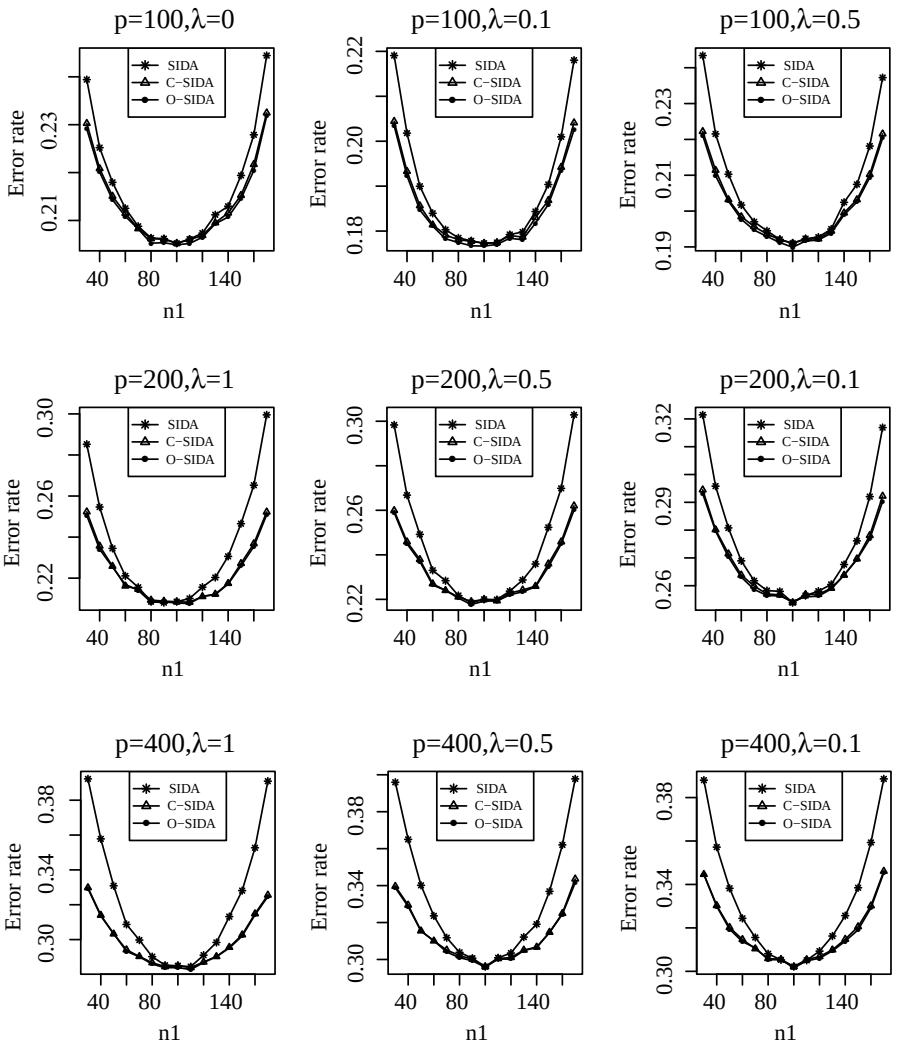


Figure 4. Empirical misclassification rates for SIDA, C-SIDA and O-SIDA for $n_1 + n_2 = 200$.

Wang and Jiang (2018). The results are based on random matrix theory of the sample correlation matrix, which is more challenging than those based on the sample covariance matrix. Moreover, simulation study and real data analysis show the advantages of the proposed SIDA.

This paper establishes the theoretical properties of SIDA under Gaussian settings where the misclassification error rate has a closed form on population means and covariance matrices. A natural extension is to consider more general population distributions. In multivariate statistical analysis (Anderson, 2003), we can consider the the elliptical distribution for which Hu et al. (2019) recently established random matrix theory for the sample covariance matrix. Another

Table 2. Misclassification error rates for embryonal tumor dataset.

p	DLDA	RLDA	SIDA	DLDA	RLDA	SIDA
	LOOCV($n = 59$)			random sample($n = 48$)		
50	15.25%	18.64%	6.78%	15.73%	16.86%	11.04%
100	18.64%	22.03%	13.56%	17.88%	25.08%	13.05%
150	20.34%	22.03%	3.39%	18.91%	26.87%	8.08%
200	20.34%	25.42%	1.69%	20.27%	29.01%	6.06%

Table 3. Misclassification error rates for breast cancer dataset.

p	DLDA	RLDA	SIDA	DLDA	RLDA	SIDA
	LOOCV($n = 167$)			random sample($n = 134$)		
50	20.36%	19.16%	17.96%	21.16%	18.15%	17.45%
100	17.96%	16.17%	17.96%	19.20%	17.92%	17.78%
150	20.36%	14.97%	14.97%	19.79%	17.32%	16.74%
200	18.56%	15.57%	14.97%	18.94%	16.14%	15.39%



Figure 5. The variances of top 200 features for embryonal tumor dataset and breast cancer dataset.

direction is to study SIDA for non-paranormal distributions (Liu, Lafferty and Wasserman, 2009; Mai, He and Zou, 2022) where a normalization procedure can be applied to the rank statistics. For these complicated distributions, studying SIDA is important from both empirical and theoretical aspects. All these extensions are interesting future work.

Acknowledgments

We are grateful to the Editor, the Associate Editor and the two referees for their constructive comments, which helped us to improve the manuscript. Li and Zheng's research is supported by NSFC Grant Nos. 12071066, 12231011, and

12326606, and by KLAS. Wang's research is supported by NSFC 12031005, NSF of Shanghai 21ZR1432900 and the fundamental research funds for the central universities. Yin's research is supported by NSFC 12271065.

Appendix

The key step of the technical analysis is to bound $\mathrm{diag}(\mathbf{S}_n) - \mathrm{diag}(\mathbf{\Sigma})$. Let $\|\cdot\|$ denote the spectral norm of matrices. Specially, we have the following lemma.

Lemma 1. *Under Assumptions* (C1)–(C3)*, we have*

$$\left\|\mathbf{\Sigma}^{-1/2}\mathrm{diag}(\mathbf{S}_n - \mathbf{\Sigma})\mathbf{\Sigma}^{-1/2}\right\| \xrightarrow{a.s.} 0, \quad \|\mathbf{A}_n - \mathbf{B}_n\| \xrightarrow{a.s.} 0.$$

proof. Let $\mathbf{z}_1, \mathbf{z}_2, \ldots, \overset{i.i.d.}{\sim} N(\mathbf{0}, \mathbf{I})$, we know

$$\mathbf{S}_n \overset{d}{=} \mathbf{\Sigma}^{1/2}\left(\frac{1}{n_1+n_2-2}\sum_{i=1}^{n_1+n_2-2}\mathbf{z}_i\mathbf{z}_i^{\mathrm{T}}\right)\mathbf{\Sigma}^{1/2}.$$

Since $\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{R}\mathbf{\Lambda}$, we can set $\mathbf{\Sigma}^{1/2} = \mathbf{\Lambda}\mathbf{R}^{1/2}$ and then

$$\mathbf{\Sigma}^{-1/2}\mathrm{diag}(\mathbf{S}_n - \mathbf{\Sigma})\mathbf{\Sigma}^{-1/2} = \mathbf{R}^{-1/2}\mathrm{diag}(\mathbf{\Lambda}^{-1}\mathbf{S}_n\mathbf{\Lambda}^{-1} - \mathbf{R})\mathbf{R}^{-1/2}.$$

Note $\mathbf{\Lambda}^{-1}\mathbf{S}_n\mathbf{\Lambda}^{-1} \overset{d}{=} \mathbf{R}^{1/2}\{\sum_{i=1}^{n_1+n_2-2}\mathbf{z}_i\mathbf{z}_i^{\mathrm{T}}/(n_1+n_2-2)\}\mathbf{R}^{1/2}$, which can be regarded as a sample covariance matrix with the true population covariance matrix $\mathbf{R}$. By Lemma 4 of El Karoui (2009), we have

$$\left\|\mathrm{diag}(\mathbf{\Lambda}^{-1}\mathbf{S}_n\mathbf{\Lambda}^{-1} - \mathbf{R})\right\| \xrightarrow{a.s.} 0.$$

Thus $\left\|\mathbf{\Sigma}^{-1/2}\mathrm{diag}(\mathbf{S}_n - \mathbf{\Sigma})\mathbf{\Sigma}^{-1/2}\right\| \leq \|\mathbf{R}\|\,\|\mathrm{diag}(\mathbf{\Lambda}^{-1}\mathbf{S}_n\mathbf{\Lambda}^{-1} - \mathbf{R})\| \xrightarrow{a.s.} 0$. For the second conclusion, by Weyl's inequality

$$\begin{aligned}
&\lambda_{min}\left[\mathbf{\Sigma}^{-1/2}\mathrm{diag}(\mathbf{S}_n)\mathbf{\Sigma}^{-1/2}\right] \\
&\geq \lambda_{min}\left[\mathbf{\Sigma}^{-1/2}\mathrm{diag}(\mathbf{\Sigma})\mathbf{\Sigma}^{-1/2}\right] - \left\|\mathbf{\Sigma}^{-1/2}\mathrm{diag}(\mathbf{S}_n - \mathbf{\Sigma})\mathbf{\Sigma}^{-1/2}\right\| \\
&= \|\mathbf{R}\|^{-1} - \left\|\mathbf{\Sigma}^{-1/2}\mathrm{diag}(\mathbf{S}_n - \mathbf{\Sigma})\mathbf{\Sigma}^{-1/2}\right\|,
\end{aligned}$$

together with the assumption $1/c \leq \lambda(\mathbf{R}) \leq c$, with probability 1, we have $\lambda_{min}\left[\mathbf{\Sigma}^{-1/2}\mathrm{diag}(\mathbf{S}_n)\mathbf{\Sigma}^{-1/2})\right] \geq 1/2c$. Noting

$$\begin{aligned}
\mathbf{A}_n =& \mathbf{\Sigma}^{1/2}\left[\mathbf{S}_n + \lambda\cdot\mathrm{diag}(\mathbf{S}_n)\right]^{-1}\mathbf{\Sigma}^{1/2} \\
=& \left[\mathbf{\Sigma}^{-1/2}\mathbf{S}_n\mathbf{\Sigma}^{-1/2} + \lambda\cdot\mathbf{\Sigma}^{-1/2}\mathrm{diag}(\mathbf{S}_n)\mathbf{\Sigma}^{-1/2}\right]^{-1},
\end{aligned}$$

we can get

$$\|\mathbf{A}_n\| \leq \frac{1}{\lambda_{min}\left[\boldsymbol{\Sigma}^{-1/2}\text{diag}(\mathbf{S}_n)\boldsymbol{\Sigma}^{-1/2}\right]} \leq 2c, \text{ a.s..}$$

For $\mathbf{B}_n = \boldsymbol{\Sigma}^{1/2}\left[\mathbf{S}_n + \lambda \cdot \text{diag}(\boldsymbol{\Sigma})\right]^{-1}\boldsymbol{\Sigma}^{1/2}$, we have

$$\|\mathbf{B}_n\| \leq \frac{1}{\lambda_{min}\left[\boldsymbol{\Sigma}^{-1/2}\text{diag}(\boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1/2}\right]} = \|\mathbf{R}\| \leq c.$$

Since

$$\begin{aligned}
&\mathbf{A}_n - \mathbf{B}_n \\
&= \boldsymbol{\Sigma}^{1/2}\left[\mathbf{S}_n + \lambda \cdot \text{diag}(\mathbf{S}_n)\right]^{-1}\boldsymbol{\Sigma}^{1/2} - \boldsymbol{\Sigma}^{1/2}\left[\mathbf{S}_n + \lambda \cdot \text{diag}(\boldsymbol{\Sigma})\right]^{-1}\boldsymbol{\Sigma}^{1/2} \\
&= \lambda\boldsymbol{\Sigma}^{1/2}\left[\mathbf{S}_n + \lambda \cdot \text{diag}(\mathbf{S}_n)\right]^{-1}\left[\text{diag}(\boldsymbol{\Sigma}) - \text{diag}(\mathbf{S}_n)\right]\left[\mathbf{S}_n + \lambda \cdot \text{diag}(\boldsymbol{\Sigma})\right]^{-1}\boldsymbol{\Sigma}^{1/2} \\
&= \lambda\mathbf{A}_n\boldsymbol{\Sigma}^{-1/2}\text{diag}(\boldsymbol{\Sigma} - \mathbf{S}_n)\boldsymbol{\Sigma}^{-1/2}\mathbf{B}_n,
\end{aligned}$$

thus we can show $\|\mathbf{A}_n - \mathbf{B}_n\| \leq \lambda\|\mathbf{A}_n\|\|\mathbf{B}_n\|\left\|\boldsymbol{\Sigma}^{-1/2}\text{diag}(\mathbf{S}_n - \boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1/2}\right\| \xrightarrow{a.s.} 0$.

A.1. Proofs of Theorems 1 and 2

Noting $\bar{\mathbf{x}}_1,\ \bar{\mathbf{x}}_2$ and $\mathbf{S}_n$ are independent, we can write

$$\bar{\mathbf{x}}_1 = \mathbf{u}_1 + \frac{1}{\sqrt{n_1}}\boldsymbol{\Sigma}^{1/2}\mathbf{z}_1,\ \bar{\mathbf{x}}_2 = \mathbf{u}_2 + \frac{1}{\sqrt{n_2}}\boldsymbol{\Sigma}^{1/2}\mathbf{z}_2$$

where $\mathbf{z}_1,\ \mathbf{z}_2 \sim N(\mathbf{0}, \mathbf{I})$ are independent with $\mathbf{S}_n$. Then,

$$\begin{aligned}
&\left(2\mathbf{u}_1 - \bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2\right)^{\mathrm{T}}\left[\mathbf{S}_n + \lambda \cdot \text{diag}(\mathbf{S}_n)\right]^{-1}\left(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2\right) \\
&= \tilde{\mathbf{u}}^{\mathrm{T}}\mathbf{A}_n\tilde{\mathbf{u}} - \frac{1}{n_1}\mathbf{z}_1^{\mathrm{T}}\mathbf{A}_n\mathbf{z}_1 + \frac{1}{n_2}\mathbf{z}_2^{\mathrm{T}}\mathbf{A}_n\mathbf{z}_2 - \frac{2}{\sqrt{n_2}}\mathbf{z}_2^{\mathrm{T}}\mathbf{A}_n\tilde{\mathbf{u}}
\end{aligned}$$

where $\tilde{\mathbf{u}} = \boldsymbol{\Sigma}^{-1/2}\mathbf{u}_1 - \boldsymbol{\Sigma}^{-1/2}\mathbf{u}_2$. Let $\|\cdot\|_2$ denote the Euclidean norm of vectors. For each part, it is trivial to show

$$\begin{aligned}
\left|\tilde{\mathbf{u}}^{\mathrm{T}}\mathbf{A}_n\tilde{\mathbf{u}} - \tilde{\mathbf{u}}^{\mathrm{T}}\mathbf{B}_n\tilde{\mathbf{u}}\right| &\leq \|\tilde{\mathbf{u}}\|_2^2\|\mathbf{A}_n - \mathbf{B}_n\| \xrightarrow{a.s.} 0; \\
\left|\frac{1}{n_1}\mathbf{z}_1^{\mathrm{T}}\mathbf{A}_n\mathbf{z}_1 - \frac{1}{n_1}\mathbf{z}_1^{\mathrm{T}}\mathbf{B}_n\mathbf{z}_1\right| &\leq \frac{p}{n_1}\frac{\|\mathbf{z}_1\|_2^2}{p}\|\mathbf{A}_n - \mathbf{B}_n\| \xrightarrow{a.s.} 0; \\
\left|\frac{1}{n_2}\mathbf{z}_2^{\mathrm{T}}\mathbf{A}_n\mathbf{z}_2 - \frac{1}{n_2}\mathbf{z}_2^{\mathrm{T}}\mathbf{B}_n\mathbf{z}_2\right| &\leq \frac{p}{n_2}\frac{\|\mathbf{z}_2\|_2^2}{p}\|\mathbf{A}_n - \mathbf{B}_n\| \xrightarrow{a.s.} 0; \\
\left|\frac{2}{\sqrt{n_2}}\mathbf{z}_2^{\mathrm{T}}\mathbf{A}_n\tilde{\mathbf{u}} - \frac{2}{\sqrt{n_2}}\mathbf{z}_2^{\mathrm{T}}\mathbf{B}_n\tilde{\mathbf{u}}\right| &\leq \left\|\frac{2}{\sqrt{n_2}}\mathbf{z}_2\right\|_2\|\mathbf{A}_n\tilde{\mathbf{u}} - \mathbf{B}_n\tilde{\mathbf{u}}\|_2 \xrightarrow{a.s.} 0.
\end{aligned}$$

Thus, we have

$$\begin{aligned}
&\left(2\mathbf{u}_1 - \bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2\right)^{\mathrm{T}}\left[\mathbf{S}_n + \lambda \cdot \text{diag}(\mathbf{S}_n)\right]^{-1}\left(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2\right) \\
&- \left(2\mathbf{u}_1 - \bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2\right)^{\mathrm{T}}\left[\mathbf{S}_n + \lambda \cdot \text{diag}(\boldsymbol{\Sigma})\right]^{-1}\left(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2\right) \xrightarrow{a.s.} 0. \qquad \text{(A.1)}
\end{aligned}$$

and similarly

$$\begin{aligned}&\left(2\mathbf{u}_2-\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2\right)^{\mathrm{T}}\left[\mathbf{S}_n+\lambda\cdot\mathrm{diag}(\mathbf{S}_n)\right]^{-1}\left(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2\right)\\&-\left(2\mathbf{u}_2-\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2\right)^{\mathrm{T}}\left[\mathbf{S}_n+\lambda\cdot\mathrm{diag}(\mathbf{\Sigma})\right]^{-1}\left(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2\right)\xrightarrow{a.s.}0.\end{aligned}\tag{A.2}$$

For the denominator

$$\begin{aligned}&\left(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2\right)^{\mathrm{T}}\left[\mathbf{S}_n+\lambda\cdot\mathrm{diag}(\mathbf{S}_n)\right]^{-1}\mathbf{\Sigma}\left[\mathbf{S}_n+\lambda\cdot\mathrm{diag}(\mathbf{S}_n)\right]^{-1}\left(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2\right)\\&=\left(\mathbf{\Sigma}^{-1/2}\bar{\mathbf{x}}_1-\mathbf{\Sigma}^{-1/2}\bar{\mathbf{x}}_2\right)^{\mathrm{T}}\mathbf{A}_n^2\left(\mathbf{\Sigma}^{-1/2}\bar{\mathbf{x}}_1-\mathbf{\Sigma}^{-1/2}\bar{\mathbf{x}}_2\right)\\&=\left(\tilde{\mathbf{u}}+\frac{1}{\sqrt{n_1}}\mathbf{z}_1-\frac{1}{\sqrt{n_2}}\mathbf{z}_2\right)^{\mathrm{T}}\mathbf{A}_n^2\left(\tilde{\mathbf{u}}+\frac{1}{\sqrt{n_1}}\mathbf{z}_1-\frac{1}{\sqrt{n_2}}\mathbf{z}_2\right)\end{aligned}$$

and noting

$$\left\|\mathbf{A}_n^2-\mathbf{B}_n^2\right\|\leq\left(\left\|\mathbf{A}_n\right\|+\left\|\mathbf{B}_n\right\|\right)\left\|\mathbf{A}_n-\mathbf{B}_n\right\|\xrightarrow{a.s.}0,$$

we can show

$$\begin{aligned}&\left(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2\right)^{\mathrm{T}}\left[\mathbf{S}_n+\lambda\cdot\mathrm{diag}(\mathbf{S}_n)\right]^{-1}\mathbf{\Sigma}\left[\mathbf{S}_n+\lambda\cdot\mathrm{diag}(\mathbf{S}_n)\right]^{-1}\left(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2\right)\\&-\left(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2\right)^{\mathrm{T}}\left[\mathbf{S}_n+\lambda\cdot\mathrm{diag}(\mathbf{\Sigma})\right]^{-1}\mathbf{\Sigma}\left[\mathbf{S}_n+\lambda\cdot\mathrm{diag}(\mathbf{\Sigma})\right]^{-1}\left(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2\right)\xrightarrow{a.s.}0.\end{aligned}$$

Finally, we know $\left[\mathbf{S}_n+\lambda\cdot\mathrm{diag}(\mathbf{\Sigma})\right]^{-1}=\mathbf{\Lambda}^{-1}\left(\mathbf{\Lambda}^{-1}\mathbf{S}_n\mathbf{\Lambda}^{-1}+\lambda\mathbf{I}_p\right)^{-1}\mathbf{\Lambda}^{-1}$ which yields that we can standardize the sample means $\bar{\mathbf{x}}_1,\bar{\mathbf{x}}_2$ and the sample covariance matrix $\mathbf{S}_n$ with the true standard deviations $\mathbf{\Lambda}$. To study the asymptotic performances of our SIDA, it is reduced to the case of Wang and Jiang (2018) with $\mathbf{\Sigma}=\mathbf{R}$. The proof is completed.

References

Anderson, T. (2003). *An Introduction to Multivariate Statistical Analysis.* 3rd Edition. Wiley, New York.

Auguin, N., Morales-Jimenez, D. and McKay, M. R. (2021). Large-dimensional characterization of robust linear discriminant analysis. *IEEE Transactions on Signal Processing* **69**, 2625–2638.

Bickel, P. J. and Levina, E. (2004). Some theory for Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations. *Bernoulli* **10**, 989–1010.

Cai, T. and Liu, W. (2011). A direct estimation approach to sparse linear discriminant analysis. *Journal of the American Statistical Association* **106**, 1566–1577.

Cai, T. T. and Zhang, L. (2019). High dimensional linear discriminant analysis: Optimality, adaptive algorithm and missing data. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **81**, 675–705.

Chen, L. S., Paul, D., Prentice, R. L. and Wang, P. (2011). A regularized Hotelling's T^2 test for pathway analysis in proteomic studies. *Journal of the American Statistical Association* **106**, 1345–1360.

Chen, Y.-J. and Tang, M. (2021). Classification of high-dimensional data with spiked covariance matrix structure. *arXiv:2110.01950*.

Dobriban, E. and Liu, S. (2019). Asymptotics for sketching in least squares regression. *Advances in Neural Information Processing Systems* **32**, 3675–3685.

Dobriban, E. and Sheng, Y. (2021). Distributed linear regression by averaging. *The Annals of Statistics* **49**, 918–943.

Dobriban, E. and Wager, S. (2018). High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics* **46**, 247–279.

Dudoit, S., Fridlyand, J. and Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association* **97**, 77–87.

El Karoui, N. (2009). Concentration of measure and spectra of random matrices: Applications to correlation matrices, elliptical distributions and beyond. *The Annals of Applied Probability* **19**, 2362–2405.

Fan, J., Feng, Y. and Tong, X. (2012). A road to classification in high dimensional space: The regularized optimal affine discriminant. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **74**, 745–771.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **70**, 849–911.

Fan, Y., Jin, J. and Yao, Z. (2013). Optimal classification in sparse Gaussian graphic model. *The Annals of Statistics* **41**, 2537–2571.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics* **7**, 179–188.

Friedman, J. H. (1989). Regularized discriminant analysis. *Journal of the American Statistical Association* **84**, 165–175.

Gravier, E., Pierron, G., Vincent-Salomon, A., Gruel, N., Raynal, V., Savignoni, A. et al. (2010). A prognostic DNA signature for T1T2 node-negative breast cancer patients. *Genes, Chromosomes and Cancer* **49**, 1125–1134.

Guo, Y., Hastie, T. and Tibshirani, R. (2007). Regularized linear discriminant analysis and its application in microarrays. *Biostatistics* **8**, 86–100.

Hand, D. J. (2006). Classifier technology and the illusion of progress. *Statistical Science* **21**, 1–14.

Hao, N., Dong, B. and Fan, J. (2015). Sparsifying the Fisher linear discriminant by rotation. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **77**, 827–851.

Hastie, T., Montanari, A., Rosset, S. and Tibshirani, R. J. (2022). Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics* **50**, 949–986.

Hu, J., Li, W., Liu, Z. and Zhou, W. (2019). High-dimensional covariance matrices in elliptical distributions with application to spherical test. *The Annals of Statistics* **47**, 527–555.

Jiang, B., Leng, C., Wang, C. and Yang, Z. (2021). Linear discriminant analysis with high-dimensional mixed variables. *arXiv:2112.07145*.

Karoui, N. E. and Kösters, H. (2011). Geometric sensitivity of random matrix results: Consequences for shrinkage estimators of covariance and related statistical methods. *arXiv:1105.1404*.

Kobak, D., Lomond, J. and Sanchez, B. (2020). The optimal ridge penalty for real-world high-dimensional data can be zero or negative due to the implicit ridge regularization. *Journal of Machine Learning Research* **21**, 6863–6878.

Ledoit, O. and Péché, S. (2011). Eigenvectors of some large sample covariance matrix ensembles. *Probability Theory and Related Fields* **151**, 233–264.

Liu, H., Lafferty, J. and Wasserman, L. (2009). The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *Journal of Machine Learning Research* **10**, 2295–2328.

Liu, S. and Dobriban, E. (2020). Ridge regression: Structure, cross-validation, and sketching. In *International Conference on Learning Representations*. *arXiv:1910.02373*.

Mai, Q., He, D. and Zou, H. (2022). Coordinatewise Gaussianization: Theories and applications. *Journal of the American Statistical Association* **118**, 2329–2343.

Mai, Q., Zou, H. and Yuan, M. (2012). A direct approach to sparse discriminant analysis in ultra-high dimensions. *Biometrika* **99**, 29–42.

Park, H., Baek, S. and Park, J. (2022). High-dimensional linear discriminant analysis using nonparametric methods. *Journal of Multivariate Analysis* **188**, 104836.

Pomeroy, S. L., Tamayo, P., Gaasenbeek, M., Sturla, L. M., Angelo, M., McLaughlin, M. E. et al. (2002). Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature* **415**, 436–442.

Shao, J., Wang, Y., Deng, X. and Wang, S. (2011). Sparse linear discriminant analysis by thresholding for high dimensional data. *The Annals of Statistics* **39**, 1241–1265.

Wang, C. and Jiang, B. (2018). On the dimension effect of regularized linear discriminant analysis. *Electronic Journal of Statistics* **12**, 2709–2742.

Yin, Y., Zheng, S. and Zou, T. (2023). Central limit theorem of linear spectral statistics of high-dimensional sample correlation matrices. *Bernoulli* **29**, 984–1006.

(Received November 2022; accepted September 2023)