On the dimension effect of regularized linear discriminant analysis

Cheng Wang[1] **and Binyan Jiang**[2]

[1] *School of Mathematical Sciences,*
Shanghai Jiao Tong University, Shanghai, 200240, China.
e-mail: chengwang@sjtu.edu.cn

[2] *Department of Applied Mathematics,*
The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong.
e-mail: by.jiang@polyu.edu.hk

Abstract: This paper studies the dimension effect of the linear discriminant analysis (LDA) and the regularized linear discriminant analysis (RLDA) classifiers for large dimensional data where the observation dimension p is of the same order as the sample size n. More specifically, built on properties of the Wishart distribution and recent results in random matrix theory, we derive explicit expressions for the asymptotic misclassification errors of LDA and RLDA respectively, from which we gain insights of how dimension affects the performance of classification and in what sense. Motivated by these results, we propose adjusted classifiers by correcting the bias brought by the unequal sample sizes. The bias-corrected LDA and RLDA classifiers are shown to have smaller misclassification rates than LDA and RLDA respectively. Several interesting examples are discussed in detail and the theoretical results on dimension effect are illustrated via extensive simulation studies.

Keywords and phrases: Dimension effect, linear discriminant analysis, random matrix theory, regularized linear discriminant analysis.

Contents

1. Introduction

Discriminant analysis that aims to allocate objects into one of the predefined classes has been an important topic in statistical learning and data analysis. Modern data is frequently featured by complex structures such as high dimensionality. In the past decades, extensive research has been done to address new challenges of high dimensionality in discriminant analysis. In particular, due to its simplicity, optimality (Anderson, 2003), and promising performance in real data analysis (Hand, 2006), the linear discriminant analysis (LDA) classifier has received more and more attention for classifying data of very large dimension under certain sparsity assumptions (Cai and Liu, 2011; Mai, Zou and Yuan, 2012; Fan, Feng and Tong, 2012). In this paper, we will focus on the general case where no sparsity assumption is assumed and study the effect of dimensionality on the performance of LDA and the regularized LDA (Friedman, 1989). For simplicity, we shall focus on the two-class classification problem.

The issue of dimension effect, or more specifically, the effect that a diverging dimension p would brought to classical statistical inference (Bai, Liu and Wong, 2009), has been studied in several statistical problems in the literature. Bai and Saranadasa (1996) first studied the dimension effect of Hotelling's T^2 test by deriving the explicit power under local alternatives. Bai, Liu and Wong (2009) and El Karoui (2010) considered the dimension effect of Markowitz's portfolio. For LDA, when the dimension p is assumed to be fixed, many works have been done to study the asymptotic of the misclassification rate by studying the properties of Wishart distribution. For a diverging p, a seminal work by Bickel and Levina (2004) showed that when $p/n \to \infty$, LDA tends to random guessing. Shao et al. (2011) proved that when $p\sqrt{\log p}/\sqrt{n} \to 0$, LDA is consistent in that the error rate of sample LDA tends to the oracle Bayes error rate. Other related works that handle variations of LDA (for example, quadratic discriminant analysis etc.) can also be found in Saranadasa (1993), Cheng (2004) and Li and Yao (2016), among others. On the other hand, when p is large, regularization method is commonly used to reduce the dispersion of the sample covariance matrix; see

for example Chen et al. (2011) for the study of the regularized Hotelling's T^2 test and Ledoit and Wolf (2004) for regularized estimation in Markowitz's portfolio. Regularization procedure has also been widely applied in many other statistical analysis in current high dimensional literature; see for example Cai, Liu and Luo (2011), Bühlmann (2013) and Wang and Leng (2016). For discriminant analysis, the regularized LDA (RLDA) was proposed and studied in Friedman (1989) and Guo, Hastie and Tibshirani (2007). Recently, Zollanvari and Dougherty (2013, 2015) studied the misclassification error rate of RLDA with the motivation to estimate this error rate and moreover, they did not provide explicit expressions for the rate. Dobriban and Wager (2018) analyzed the predictive risk of ridge regression and regularized discriminant analysis and derived some explicit formulas. However, the results were based on the random effects hypothesis which is a strong assumption.

Among these existing literatures, the problem of how the high dimensionality would affect the classification accuracy of LDA and RLDA when the observation dimension p is of the same order as the sample size n is still not well understood and in this work we provide a comprehensive analysis of the misclassification error rates under mild conditions. Built on properties of the Wishart distribution, we will first study the dimension effect of LDA (Section 2) under the assumption that $p/n \to y \in (0,1)$. In the general case where $p/n \to y \in (0,\infty)$, using recent results in random matrix theory, we provide a systematic study on the dimension effect of the RLDA in Section 3. Overall, this paper aims at providing theoretical studies on the dimension effect of LDA and RLDA, from which we gain insights of how the increasing dimension affects the classification accuracy of LDA and RLDA. Following is a summary of the contributions of our work:

- For LDA, we study the effect of the sample means and the sample covariance matrix respectively. Interestingly, our results show that the dimension effect brought by the sample means would introduce some bias to LDA, which would be further amplified by the effect brought by the sample covariance matrix. Motivated by these observations, a bias correction is proposed to improve the classification accuracy.

- For RLDA, we derive the closed form of the error rates under general cases. Dobriban and Wager (2018) studied a similar question under random effect which is a strong condition. We relax this condition and consider a general case which includes random effect as a special case. Similar to LDA, a bias corrected RLDA is proposed to reduce the bias brought by the dimension effect that improves the classification performance.

- To study the dimension effect of RLDA, the key question is to study the asymptotic limits of the moments of a class of random matrices and their random quadratic forms involved of the population means. The former one is studied by Ledoit and Péché (2011), Chen et al. (2011) and our previous work Wang et al. (2015). When the population means satisfy some special structures such as the eigenvector structures in Bai, Miao and Pan (2007) or the random effect of Dobriban and Wager (2018), the limits of quadratic forms are the same as the moments. By studying the limits of these mo-

ments, Dobriban and Wager (2018) derived the dimension effect of RLDA under the random effect assumption. For general cases, the asymptotic of the random quadratic forms are much more challenging. Bai, Liu and Wong (2011) derived results for an identity matrix. El Karoui and Holger (2011) provided a tool to study the limits of the quadratic forms. Based on the methods of El Karoui and Holger (2011), we develop explicit results for the dimension effect of RLDA under mild conditions.

The remainder of this paper is organized as follows. We study the dimension effect of LDA when $p/n \to y \in (0,1)$ in Section 2. Dimension effect of RLDA is provided in Section 3. To better understand the dimension effect of RLDA and verify our theoretical results, we conduct several interesting examples in Section 4 and simulations in Section 5. All technical details are relegated to the Appendix.

2. Linear discriminant analysis

Let X be a p-dimensional normal random vector belonging to class k if $X \sim N(\mu_k, \Sigma)$, $k = 1, 2$, where $\mu_1 \neq \mu_2$ are the population means and Σ is the population covariance matrix. For simplicity, in this paper we consider the case where equal weights are used for the two types of misclassification error. If μ_1, μ_2 and Σ are known, the Bayes' classification rule is

$$\text{Bayes' rule:} \quad \delta_0(X) = \mathbb{I}\left\{\left(X - \frac{\mu_1 + \mu_2}{2}\right)^{\mathrm{T}} \Sigma^{-1}(\mu_1 - \mu_2) > 0\right\}, \qquad (2.1)$$

where $\mathbb{I}(\cdot)$ is the indicator function which assigns X to class 1 if and only if $\delta_0(X) = 1$. Bayes' rule is known to be optimal in that it has the minimum misclassification rate among all classifiers. Specifically, the optimal Bayes error rate is

$$\begin{aligned}
R_0 &= \frac{1}{2}\text{pr}\{\delta_0(X) = 0 | X \sim N(\mu_1, \Sigma)\} + \frac{1}{2}\text{pr}\{\delta_0(X) = 1 | X \sim N(\mu_2, \Sigma)\} \\
&= \Phi\left(-\frac{\Delta}{2}\right),
\end{aligned} \qquad (2.2)$$

where $\Delta = \sqrt{(\mu_1 - \mu_2)^{\mathrm{T}} \Sigma^{-1}(\mu_1 - \mu_2)}$ and $\Phi(\cdot)$ is the standard normal distribution function.

Let $\{X_{1,j}, j = 1, \cdots, n_1\}$ and $\{X_{2,j}, j = 1, \cdots, n_2\}$ be independent and identically distributed random samples from $N_p(\mu_1, \Sigma)$ and $N_p(\mu_2, \Sigma)$, respectively. We can estimate μ_1, μ_2 and Σ by their sample analogs,

$$\bar{X}_k = \frac{1}{n_k}\sum_{j=1}^{n_k} X_{k,j}, \ k = 1, 2, \ S_n = \frac{1}{n-2}\sum_{k=1}^{2}\sum_{j=1}^{n_k}(X_{k,j} - \bar{X}_k)(X_{k,j} - \bar{X}_k)^{\mathrm{T}},$$

where $n = n_1 + n_2$. Plugging the estimation into Bayes' rule (2.1) we obtain the classical LDA classifier

$$\delta_{\text{LDA}}(X) = \mathbb{I}\left\{\left(X - \frac{\bar{X}_1 + \bar{X}_2}{2}\right)^{\mathrm{T}} S_n^{-1}(\bar{X}_1 - \bar{X}_2) > 0\right\}. \qquad (2.3)$$

After some simple calculation we can obtain that conditional on the samples the misclassification rate of LDA is

$$\mathrm{R}_{\mathrm{LDA}} = \frac{1}{2}\sum_{j=1}^{2} \Phi\left(\frac{(-1)^j(\mu_j - \frac{\bar{X}_1 + \bar{X}_2}{2})^{\mathrm{T}} S_n^{-1}(\bar{X}_1 - \bar{X}_2)}{\sqrt{(\bar{X}_1 - \bar{X}_2)^{\mathrm{T}} S_n^{-1} \Sigma S_n^{-1}(\bar{X}_1 - \bar{X}_2)}}\right).$$

To investigate the dimension effect of the sample means and the sample co-variance matrix separately, we will first study two simple cases to gain some insights:

(i) Assuming that Σ is known, what is the effect brought by the sample mean?
(ii) Assuming that μ_1, μ_2 are known, what is the effect brought by the sample covariance matrix?

For convenience, we assume that Δ is a constant throughout the paper. As pointed out by Cai and Liu (2011), the case where $\Delta \to 0$ would indicate that no classifier is better than random guessing. Similarly, the case where $\Delta \to \infty$ would imply that the two classes are so separated that classification is trivial such that even naive Bayes would asymptotically have zero misclassification rate (Aoshima and Yata, 2014).

2.1. Effect of the sample mean

To investigate the effect brought by the sample mean, we first assume Σ is known and consider the following linear classifier

$$\delta_{\mathrm{LDA1}}(X) = \mathbb{I}\left\{\left(X - \frac{\bar{X}_1 + \bar{X}_2}{2}\right)^{\mathrm{T}} \Sigma^{-1}(\bar{X}_1 - \bar{X}_2) > 0\right\},$$

which is obtained by replacing S_n in the LDA classifier (2.3) by the true Σ. The misclassification rate of the above classifier can be computed as:

$$\mathrm{R}_{\mathrm{LDA1}} = \frac{1}{2}\sum_{j=1}^{2} \Phi\left(\frac{(-1)^j\left(\mu_j - \frac{\bar{X}_1 + \bar{X}_2}{2}\right)^{\mathrm{T}} \Sigma^{-1}(\bar{X}_1 - \bar{X}_2)}{\sqrt{(\bar{X}_1 - \bar{X}_2)^{\mathrm{T}} \Sigma^{-1}(\bar{X}_1 - \bar{X}_2)}}\right).$$

The dimension effect brought by the sample means can then be characterized by the asymptotic misclassification rate given in the following theorem.

Theorem 2.1. *Assuming $p/n_j \to y_j \in (0, \infty)$, $j = 1, 2$, we have,*

$$R_{LDA1} \xrightarrow{p} \frac{1}{2}\Phi\left(-\frac{\Delta^2 + y_2 - y_1}{2\sqrt{\Delta^2 + y_1 + y_2}}\right) + \frac{1}{2}\Phi\left(-\frac{\Delta^2 + y_1 - y_2}{2\sqrt{\Delta^2 + y_1 + y_2}}\right). \quad (2.4)$$

By comparing (2.4) with the optimal Bayes error (2.2), we can see that the dimension effect brought by the sample means depends on y_1, y_2 only. From

Theorem 2.1 we know that even if Σ is known, when y_1 or y_2 is very large which indicates that the dimension far more exceeds the sample size, the misclassification rate would tend to 50%. The conclusion is consistent with the one in the seminal paper Bickel and Levina (2004) which studied the dimension effect from a minimax point of view under the setting $p/n \to \infty$.

2.2. Effect of the sample covariance matrix

Similarly, we study the effect brought by the sample covariance matrix. Assuming the population means μ_1, μ_2 are known, we consider the classifier

$$\delta_{\mathrm{LDA2}}(X) = \mathbb{I}\left\{ \left(X - \frac{\mu_1 + \mu_2}{2} \right)^{\mathrm{T}} S_n^{-1}(\mu_1 - \mu_2) > 0 \right\}.$$

The corresponding misclassification rate is

$$\mathrm{R_{LDA2}} = \Phi\left(\frac{-(\mu_1 - \mu_2)^{\mathrm{T}} S_n^{-1}(\mu_1 - \mu_2)}{2\sqrt{(\mu_1 - \mu_2)^{\mathrm{T}} S_n^{-1} \Sigma S_n^{-1}(\mu_1 - \mu_2)}} \right).$$

Theorem 2.2. *Assuming $p/(n_1 + n_2) \to y \in (0, 1)$, we have,*

$$R_{LDA2} \xrightarrow{p} \Phi\left(-\frac{\Delta}{2}\sqrt{1-y} \right). \tag{2.5}$$

Comparing with the optimal Bayes error (2.2) we can see that the term $\sqrt{1-y}$ in (2.5) is exactly the price we pay by using the sample covariance matrix in LDA. Since the results only depend on Wishart distribution, it would be possible to built up more precise results under weaker conditions such as $n > p + 2$ and $p/n \to y \in (0, 1]$ (Jiang and Yang, 2013) and we leave it to future work. By the properties of Wishart distribution (Cook and Forzani, 2011), we have

$$ES_n^{-1} = \frac{n-2}{n-p-3}\Sigma^{-1}, \ ES_n^{-1}\Sigma S_n^{-1} = \frac{(n-2)^2(n-3)}{(n-p-2)(n-p-3)(n-p-5)}\Sigma^{-1}.$$

Therefore, the term $\sqrt{1-y}$ is actually introduced by the inverse Wishart distribution. We note $\sqrt{1-y}$ also arises in the dimension effect for Hotelling's test (Pan and Zhou, 2011) or the Markowitz portfolio (Bai, Liu and Wong, 2009; El Karoui, 2010) which both involve the inverse of the sample covariance matrix.

2.3. Dimension effect of LDA

Theorems 2.1 and 2.2 show the explicit results for the sample means and the sample covariance from which we can see the explicit price we pay for the estimation. Now, we present the result for LDA as follows.

Theorem 2.3. *Assuming* $p/n_j \to y_j \in (0,\infty)$, $j = 1,2$ *and* $y \stackrel{def}{=} y_1 y_2/(y_1 + y_2) < 1$, *we have*

$$R_{LDA} \xrightarrow{p} \frac{1}{2} \sum_{j=1}^{2} \Phi\left(-\frac{\Delta^2 + (-1)^j(y_1 - y_2)}{2\sqrt{\Delta^2 + y_1 + y_2}}\sqrt{1-y}\right). \qquad (2.6)$$

Note that the term on the right hand side of (2.6) correspond to the misclassification rate within class 1 and class 2 respectively. LDA has different classification performance on the two classes when $y_1 \neq y_2$. From the proof of Theorem 2.1 we can see that this is due to the estimation bias of intercept part in LDA. On the other hand, from (2.5) we learn that the effect from the sample covariance matrix would result in a multiplication factor $\sqrt{1-y}$. Interestingly, we show that when both sample means and sample covariance matrix are used as in LDA (2.3), the bias introduced by the sample means was further amplified by the multiplication factor $\sqrt{1-y}$.

To investigate the bias of LDA, we consider the classifier,

$$\mathbb{I}\left\{X^{\mathrm{T}}S_n^{-1}(\bar{X}_1 - \bar{X}_2) + \alpha > 0\right\}. \qquad (2.7)$$

By Proposition 2 of Mai, Zou and Yuan (2012), given the classification direction $S_n^{-1}(\bar{X}_1 - \bar{X}_2)$ in (2.7), the optimal intercept corresponding to minimum misclassification rate is given as,

$$\alpha_0 = -\frac{1}{2}(\mu_1 + \mu_2)^{\mathrm{T}}S_n^{-1}(\bar{X}_1 - \bar{X}_2).$$

One the other hand, note that (2.7) would reduce to the LDA classifier (2.3) if the intercept α is set to be $-\frac{1}{2}(\bar{X}_1 + \bar{X}_2)^{\mathrm{T}}S_n^{-1}(\bar{X}_1 - \bar{X}_2)$. However, direct calculation shows that,

$$-\frac{1}{2}E(\bar{X}_1 + \bar{X}_2)^{\mathrm{T}}S_n^{-1}(\bar{X}_1 - \bar{X}_2) = E\alpha_0 + \frac{n-2}{n-p-3}\left(\frac{p}{2n_2} - \frac{p}{2n_1}\right),$$

which indicates that the expected bias between the intercept of LDA and the optimal intercept α_0 is exactly $\frac{n-2}{n-p-3}\left(\frac{p}{2n_2} - \frac{p}{2n_1}\right)$. We thus can make a bias correction to LDA as follow,

$$\delta_{\mathrm{LDA}}^c(X) = \mathbb{I}\left\{\left(X - \frac{\bar{X}_1 + \bar{X}_2}{2}\right)^{\mathrm{T}}S_n^{-1}(\bar{X}_1 - \bar{X}_2) + \frac{n-2}{n-p-3}\left(\frac{p}{2n_1} - \frac{p}{2n_2}\right) > 0\right\}. \qquad (2.8)$$

Let R_{LDA}^c be the misclassification rate of the bias-corrected classifier (2.8). The following proposition gives the asymptotic misclassification rate of the above bias-corrected classifier.

Proposition 2.1. *Under the conditions of Theorem 2.3,*

$$R_{LDA}^c \xrightarrow{p} \Phi\left(-\frac{\Delta^2}{2\sqrt{\Delta^2 + y_1 + y_2}}\sqrt{1-y}\right). \qquad (2.9)$$

By noticing that $\Phi(x)$ is strictly convex in $x \in (-\infty, 0)$, we immediately have that the asymptotic misclassification rate of the bias corrected LDA given in (2.9) is smaller than the asymptotic misclassification rate of LDA given in (2.6). In addition, considering the linear classifier with optimal intercept,

$$\mathbb{I}\left\{X^{\mathrm{T}}S_n^{-1}(\bar{X}_1 - \bar{X}_2) - \frac{1}{2}(\mu_1 + \mu_2)^{\mathrm{T}}S_n^{-1}(\bar{X}_1 - \bar{X}_2) > 0\right\}.$$

It can be shown that the result given in (2.9) would also hold for this oracle classifier, indicating that our bias correction procedure does eliminate the bias introduced by the unequal sample sizes. Finally, we remark that bias issues in binary classification have received much attention in early literatures; see for example Chan and Hall (2009) and Huang, Tong and Zhao (2010) and the references therein. For LDA, the bias-corrected classifier (2.8) happens to be identical to formula (8) of Moran and Murphy (1979). Our motivation and interpretation provide a different view on (2.8) and theoretical justifications provided in Proposition 2.1 are also new.

In literature, to handle high dimension data, naive Bayes (Dudoit, Fridlyand and Speed, 2002; Bickel and Levina, 2004) and sparse LDA (Cai and Liu, 2011; Mai, Zou and Yuan, 2012; Fan, Feng and Tong, 2012) were proposed. For naive Bayes, the correlation between the covariates is ignored and only the diagonal elements of the sample covariance matrix are used. Sparse LDA methods assumed that the discriminant direction $\beta = \Sigma^{-1}(\mu_1 - \mu_2)$ is sparse in that only several elements in β are non-zero and others are either zero or close to zero. In the following remarks we provide some comparison of the LDA rule to the naive Bayes rule and sparse LDA approaches.

Remark 2.1. *For simplicity, we consider the homogeneous case where the diagonal elements of Σ equal to each other and assume that $n_1 = n_2$. Consider the following naive Bayes classification rule,*

$$\delta_{NB}(X) = \mathbb{I}\left\{\left(X - \frac{\bar{X}_1 + \bar{X}_2}{2}\right)^T (\bar{X}_1 - \bar{X}_2) > 0\right\}.$$

The conditional misclassification rate of $\delta_{NB}(X)$ is

$$R_{NB} = \frac{1}{2}\sum_{j=1}^{2}\Phi\left(\frac{(-1)^j\left(\mu_j - \frac{\bar{X}_1 + \bar{X}_2}{2}\right)^T(\bar{X}_1 - \bar{X}_2)}{\sqrt{(\bar{X}_1 - \bar{X}_2)^T\Sigma(\bar{X}_1 - \bar{X}_2)}}\right).$$

By a similar analysis as our proofs for Theorem 2.3, we have

$$R_{NB} = \Phi\left(-\frac{(\mu_1 - \mu_2)^T(\mu_1 - \mu_2) + o_p(1)}{2\sqrt{(\mu_1 - \mu_2)^T\Sigma(\mu_1 - \mu_2) + \frac{4}{n}tr(\Sigma^2) + o_p(1)}}\right).$$

We can see that the performance of naive Bayes depends on the structure of Σ. When $\Sigma = \sigma^2 I_p$ for some constant $\sigma^2 > 0$, comparing with (2.6), the asymptotic

error rate of the naive Bayes rule is the same as that of LDA except the term $\sqrt{1-y}$, indicating that naive Bayes is always better than LDA. When the off diagonal elements of Σ are nonzero, we would need to pay some price for wrongly ignoring the correlations. A simulation study will be conducted in Section 5, providing further numerical comparison between naive Bayes and LDA.

Remark 2.2. *For sparse LDA, we take the linear programming discriminant (LPD) rule as an example which was proposed by Cai and Liu (2011). The LPD rule estimates the discriminant direction by solving the following optimization problem,*

$$\hat{\beta} = \arg\min |\beta|_1, \ subject \ to \ |S_n\beta - (\bar{X}_1 - \bar{X}_2)|_\infty \leq t_n,$$

where t_n is a tuning parameter. Under regular conditions, the misclassification rate R_{LPD} of LPD (Cai and Liu, 2011, Theorem 3) satisfies,

$$\frac{R_{LPD}}{\Phi(-\Delta/2)} - 1 = O_p\left(\Delta^2 s\sqrt{\frac{\log p}{n}}\right).$$

where s is the number of non-zero elements of $\Sigma^{-1}(\mu_1 - \mu_2)$. Unlike LDA, the misclassification rate of LPD not only depends on p and n, but also hinge on the size of s. When s is small, LPD will achieve a good performance and even close to the Bayes rule while the dimension p is allowed to grow exponentially in n. However, when the sparsity assumption is violated, it is expected that LPD will perform poor if s is large. Further numerical comparison will be provided in Section 5.

3. Regularized linear discriminant analysis

RLDA (Friedman, 1989; Guo, Hastie and Tibshirani, 2007) was proposed to reduce the dispersion of eigenvalues of the sample covariance matrix when p is large and overcome the singularity issue when $p > n - 2$. For a given tuning parameter $\lambda > 0$, the RLDA classifier is given as,

$$\delta_{\mathrm{RLDA}}(X) = \mathbb{I}\left\{\left(X - \frac{\bar{X}_1 + \bar{X}_2}{2}\right)^\mathrm{T}(S_n + \lambda I_p)^{-1}(\bar{X}_1 - \bar{X}_2) > 0\right\}.$$

When S_n is invertible, RLDA would reduce to LDA if λ is set to be zero. With one more degree of freedom to adjust the estimation of the covariance matrix, it is well known that RLDA generally poses better performance than LDA, providing that λ is chosen properly. The misclassification rate of RLDA is

$$\mathrm{R}_{\mathrm{RLDA}}(\lambda) = \frac{1}{2}\sum_{j=1}^2 \Phi\left(\frac{(-1)^j\left(\mu_j - \frac{\bar{X}_1+\bar{X}_2}{2}\right)^\mathrm{T}(S_n + \lambda I_p)^{-1}(\bar{X}_1 - \bar{X}_2)}{\sqrt{(\bar{X}_1 - \bar{X}_2)^\mathrm{T}(S_n + \lambda I_p)^{-1}\Sigma(S_n + \lambda I_p)^{-1}(\bar{X}_1 - \bar{X}_2)}}\right).$$

To explore the dimension effect on RLDA, we need to study the asymptotic properties of $R_{RLDA}(\lambda)$. Denote

$$\mu = \Sigma^{-\frac{1}{2}}(\mu_1 - \mu_2), \ B_n(\lambda) = \Sigma^{\frac{1}{2}}(S_n + \lambda I)^{-1}\Sigma^{\frac{1}{2}}. \tag{3.1}$$

By noticing that

$$\bar{X}_1 \overset{d}{=} \frac{1}{\sqrt{n_1}}\Sigma^{\frac{1}{2}}Y_1 + \mu_1, \ \bar{X}_2 \overset{d}{=} \frac{1}{\sqrt{n_2}}\Sigma^{\frac{1}{2}}Y_2 + \mu_2,$$

where $Y_1, Y_2 \sim N(0, I_p)$ and Y_1, Y_2, S_n are independent, we can see that the asymptotic properties of $R_{RLDA}(\lambda)$ would rely on the asymptotic properties of

$$\frac{1}{p}Y_1^{\mathrm{T}}\mathrm{tr}B_n(\lambda)Y_1, \ \ \frac{1}{p}Y_1^{\mathrm{T}}\mathrm{tr}B_n^2(\lambda)Y_1, \tag{3.2}$$

and the random quadratic forms

$$\mu^{\mathrm{T}}B_n(\lambda)\mu, \ \ \mu^{\mathrm{T}}B_n^2(\lambda)\mu. \tag{3.3}$$

Under mild conditions, it can be shown that

$$\frac{1}{p}Y_1^{\mathrm{T}}\mathrm{tr}B_n^k(\lambda)Y_1 - \frac{1}{p}\mathrm{tr}B_n^k(\lambda) = o_p(1), \ k = 1, 2. \tag{3.4}$$

Thus, finding the limits of the statistics (3.2) can be reduced to finding the moments of $\mathrm{tr}B_n^k(\lambda)/p$. Related studies on the moments of $\mathrm{tr}B_n^k(\lambda)/p$ can be found in Ledoit and Péché (2011), Chen et al. (2011) and Wang et al. (2015). Unlike the statistics (3.2) whose limits only depend on the eigenvalues of $B_n(\lambda)$, the asymptotic of the quadratic forms (3.3) generally depend on the entry-wise properties of $B_n(\lambda)$ (El Karoui and Holger, 2011). Under special case such as μ or $\mu_1 - \mu_2$ itself is generated by i.i.d. entries, the quadratic forms can be expressed by the trace of a random matrix. For example, Dobriban and Wager (2018) studied RLDA by setting $\mu_1 - \mu_2$ is i.i.d. where the statistics (3.4) can be approximated by $\mathrm{tr}(S_n + \lambda I_p)^{-k}/p, \ k = 1, 2$, respectively. Here, we study the problem under more general settings where no structure on μ_1, μ_2 is assumed. Before proceed to the theoretical results, we introduce some technical assumptions.

(C1): $p/n_j \to y_j \in (0, \infty), \ j = 1, 2$ and denote $y = y_1y_2/(y_1 + y_2)$.
(C2): The eigenvalues of Σ are uniformly bounded and the empirical spectral distribution F^Σ converges to a nonrandom distribution function H as $p \to \infty$.
(C3): For $t \geq 0$,

$$\Delta^{-2}\mu^{\mathrm{T}}(I_p + t\Sigma^{-1})^{-1}\mu \to h_1(t),$$
$$\Delta^{-2}\mu^{\mathrm{T}}(I_p + t\Sigma^{-1})^{-2}\mu \to h_2(t).$$

We remark that (C1) and (C2) are common conditions in random matrix theory and (C3) is a technical assumption to express the explicit limits of the quadratic forms (3.3). For a given λ, (C3) can be relaxed to some specific t which is a function of λ and here for brevity we make the assumption for general t. Under the conditions (C1) and (C2), for any $\lambda > 0$, it can be shown that,

$$\frac{1}{p}\text{tr}\{(S_n + \lambda I_p)^{-1}\} \xrightarrow{a.s.} m_0(-\lambda),$$

$$\frac{1}{p}\text{tr}\{(S_n + \lambda I_p)^{-2}\} \xrightarrow{a.s.} m_0'(-\lambda) = \frac{dm_0(z)}{dz}\Big|_{z=-\lambda},$$

where $m_0(-\lambda)$ is the unique solution of the Marčenko-Pastur equation (Marčenko and Pastur, 1967; El Karoui, 2008),

$$m(-\lambda) = \int \frac{dH(t)}{t(1 - y + y\lambda m(-\lambda)) + \lambda},$$

under the condition $1 - y + y\lambda m(-\lambda) \geq 0$. More details can be found in Wang et al. (2015).

Following is our main results.

3.1. Dimension effect of RLDA

Theorem 3.1. *Under the conditions (C1)-(C3), for any $\lambda > 0$,*

$$R_{RLDA}(\lambda) \xrightarrow{p} \frac{1}{2} \sum_{j=1}^{2} \Phi\left(-\frac{H_1(\lambda)\Delta^2 + (-1)^j(y_1 - y_2)R_1(\lambda)}{2\sqrt{H_2(\lambda)\Delta^2 + (y_1 + y_2)R_2(\lambda)}}\right), \qquad (3.5)$$

where

$$R_1(\lambda) = \frac{1 - \lambda m_0(-\lambda)}{1 - y[1 - \lambda m_0(-\lambda)]},$$

$$R_2(\lambda) = \frac{1 - \lambda m_0(-\lambda)}{\{1 - y[1 - \lambda m_0(-\lambda)]\}^3} - \frac{\lambda m_0(-\lambda) - \lambda^2 m_0'(-\lambda)}{\{1 - y[1 - \lambda m_0(-\lambda)]\}^4},$$

$$H_1(\lambda) = \frac{1}{1 - y[1 - \lambda m_0(-\lambda)]} h_1\left(\frac{\lambda}{1 - y[1 - \lambda m_0(-\lambda)]}\right),$$

$$H_2(\lambda) = \{(1 + yR_1(\lambda))^2 + yR_2(\lambda)\} h_2\left(\frac{\lambda}{1 - y[1 - \lambda m_0(-\lambda)]}\right).$$

Theorem 3.1 provides an explicit expression for the asymptotic misclassification rate of RLDA, which characterizes the effect of the data dimension in terms of y_1, y_2 and the functions on the regularization parameter λ defined above.

Remark 3.1. *From the proof of Theorem 3.1, we note that when $y < 1$ and $\lambda = 0$, (3.5) reduces to (2.6), implying that Theorem 2.3 could be regarded as a special case of Theorem 3.1. We remark that the proof of Theorem 2.3 is built on properties of normal and Wishart distributions only while Theorem 3.1 is built on the more complicated random matrix theory.*

By noticing that if

$$h_1(t) = \int \frac{x}{x+t} dH(x), \ h_2(t) = \int \frac{x^2}{(x+t)^2} dH(x), \qquad (3.6)$$

we would have $H_1(\lambda) = R_1(\lambda), H_2(\lambda) = R_2(\lambda)$ and hence the asymptotic mis-classification rate (3.5) can be further simplified. This is summarized in the following proposition.

Proposition 3.1. *Under the conditions of Theorem 3.1 and assuming that* (3.6) *holds, we have*

$$R_{RLDA}(\lambda) \xrightarrow{p} \frac{1}{2} \sum_{j=1}^{2} \Phi \left(-\frac{\Delta^2 + (-1)^j (y_1 - y_2)}{2\sqrt{\Delta^2 + y_1 + y_2}} \frac{R_1(\lambda)}{\sqrt{R_2(\lambda)}} \right).$$

Note that condition (3.6) would hold when $\mu_1 - \mu_2$ and Σ satisfy the following structures,

$$\mu^{\mathrm{T}}(I_p + t\Sigma^{-1})^{-k}\mu = \|\mu\|^2 \frac{1}{p} \mathrm{tr}(I_p + t\Sigma^{-1})^{-k}, \ k = 1, 2,$$

A trivial example satisfies the above structure is $\Sigma = I_p$ and an interesting isotropic example will also be discussed in Section 4.

3.2. Bias correction for RLDA

To reduce the bias brought by the unequal sample sizes, following the same idea of bias correction for LDA as in Section 2.3, we consider the following class of classifiers,

$$\mathbb{I}\left\{ X^{\mathrm{T}}(S_n + \lambda I_p)^{-1}(\bar{X}_1 - \bar{X}_2) + \alpha > 0 \right\}.$$

Given the discriminant direction $(S_n + \lambda I_p)^{-1}(\bar{X}_1 - \bar{X}_2)$, the optimal intercept corresponding to minimum misclassification rate is given as,

$$\alpha_r = -\frac{1}{2}(\mu_1 + \mu_2)^{\mathrm{T}}(S_n + \lambda I_p)^{-1}(\bar{X}_1 - \bar{X}_2),$$

while for RLDA the intercept is $-\frac{1}{2}(\bar{X}_1 + \bar{X}_2)^{\mathrm{T}}(S_n + \lambda I_p)^{-1}(\bar{X}_1 - \bar{X}_2)$. On the other hand, from the proof of Theorem 3.1 we have

$$-\frac{1}{2}E(\bar{X}_1 + \bar{X}_2)^{\mathrm{T}}(S_n + \lambda I_p)^{-1}(\bar{X}_1 - \bar{X}_2)$$

$$=E\alpha_r + E\frac{1}{2n_1}Y_1^{\mathrm{T}}B_n(\lambda)Y_1 - E\frac{1}{2n_2}Y_2^{\mathrm{T}}B_n(\lambda)Y_2$$

$$=E\alpha_r - \left(\frac{p}{2n_1} - \frac{p}{2n_2} \right) \frac{E\mathrm{tr}(B_n(\lambda))}{p}.$$

Therefore, we can correct the bias as follow,

$$
\left(X - \frac{\bar{X}_1 + \bar{X}_2}{2} \right)^{\mathrm{T}} (S_n + \lambda I_p)^{-1} (\bar{X}_1 - \bar{X}_2) + \left(\frac{p}{2n_1} - \frac{p}{2n_2} \right) \frac{E\mathrm{tr}(B_n(\lambda))}{p} > 0.
$$

The matrix $B_n(\lambda)$ depends on the true population covariance matrix Σ and the regularized term $(S_n + \lambda I_p)^{-1}$. It is complicated to derive the explicit results for $E\mathrm{tr}(B_n(\lambda))$. By noticing that $\mathrm{tr}(B_n(\lambda))/p$ converges to $R_1(\lambda)$ almost surely (Wang et al., 2015), we consider estimating $E\mathrm{tr}(B_n(\lambda))/p$ by the consistent estimator given in (Wang et al., 2015) and propose the following bias-corrected RLDA classifier,

$$
\begin{aligned}
\delta_{\mathrm{RLDA}}^c(X) = \mathbb{I} \Bigg\{ & \left(X - \frac{\bar{X}_1 + \bar{X}_2}{2} \right)^{\mathrm{T}} (S_n + \lambda I_p)^{-1} (\bar{X}_1 - \bar{X}_2) \\
& + \left(\frac{p}{2n_1} - \frac{p}{2n_2} \right) \frac{1 - \frac{1}{p} tr(\frac{1}{\lambda} S_n + I_p)^{-1}}{1 - \frac{p}{n-2} + \frac{1}{n-2} tr(\frac{1}{\lambda} S_n + I_p)^{-1}} > 0 \Bigg\}.
\end{aligned} \quad (3.7)
$$

Let $\mathrm{R}_{\mathrm{RLDA}}^c(\lambda)$ be the misclassification rate of the bias-corrected classifier (3.7).

Proposition 3.2. *Under the conditions of Theorem 3.1,*

$$
R_{RLDA}^c(\lambda) \xrightarrow{p} \Phi \left(- \frac{H_1(\lambda)\Delta^2}{2\sqrt{H_2(\lambda)\Delta^2 + (y_1 + y_2)R_2(\lambda)}} \right). \quad (3.8)
$$

Again by the convexity of $\Phi(x)$ in $x \in (-\infty, 0)$, we conclude that the asymptotic misclassification rate of the bias corrected RLDA given in (3.8) is smaller than the asymptotic misclassification rate of RLDA given in (3.5). Similarly to the bias correction for LDA, we can claim the result of (3.8) is also the asymptotic error rate of RLDA with optimal intercept.

3.3. Selection of λ

From the perspective of applications, the RLDA can avoid the singularity problem of S_n when $p > n$ and provides a better performance than LDA even if $p < n$. However, the performance of RLDA relies on the choice of the regularization parameter λ. Empirically, we can use re-sampling procedures such as cross validation to choose λ. However, there is little work on the theoretical analysis of these re-sampling procedures, and cross validation can be computationally expensive when both p and n are large. In this section, we discuss the effect of λ on the misclassification error rate and provide a direct estimation of the optimal λ for the special cases.

By Proposition 3.2, the optimal λ with minimum error rate is

$$
\lambda_0 \in \arg\max \frac{H_1(\lambda)\Delta^2}{\sqrt{H_2(\lambda)\Delta^2 + (y_1 + y_2)R_2(\lambda)}}. \quad (3.9)
$$

From the definitions of $H_1(\lambda)$ and $H_2(\lambda)$, these parameters depend on the true structures of Σ and μ and it is not easy to construct estimations for them. However, under the assumption of Proposition 3.1, the optimal λ has a simple form

$$\lambda_0 \in \arg\max \frac{R_1^2(\lambda)}{R_2(\lambda)}.$$

This motivates us to select λ_0 by maximizing an estimate of $R_1^2(\lambda)/R_2(\lambda)$. As discussed before, $R_1(\lambda)$ and $R_2(\lambda)$ only depends on the eigenvalues of Σ and could be estimated by the eigenvalues of S_n. Considering a shrinkage estimation $\alpha(S_n + \lambda I_p)^{-1}$ (Kubokawa and Srivastava, 2008; Wang et al., 2015) for the precision matrix Σ^{-1}, we have

$$\min_\alpha \frac{1}{p}\mathrm{tr}(\alpha(S_n + \lambda I_p)^{-1}\Sigma - I_p)^2 \xrightarrow{p} 1 - \frac{R_1^2(\lambda)}{R_2(\lambda)}.$$

Therefore, similar to Wang et al. (2015), we estimate the tuning parameter by

$$\hat{\lambda}_0 = \arg\min\left\{1 - \frac{(\hat{R}_1(\lambda))^2}{\hat{R}_2(\lambda)}\right\},$$

where

$$\hat{y} = \frac{p}{n-2}, \; a_1(\lambda) = 1 - \frac{1}{p}tr\left(\frac{1}{\lambda}S_n + I_p\right)^{-1},$$
$$a_2(\lambda) = \frac{1}{p}tr\left(\frac{1}{\lambda}S_n + I_p\right)^{-1} - \frac{1}{p}tr\left(\frac{1}{\lambda}S_n + I_p\right)^{-2},$$
$$\hat{R}_1(\lambda) = \frac{a_1(\lambda)}{1 - \hat{y}a_1(\lambda)}, \; \hat{R}_2(\lambda) = \frac{a_1(\lambda)}{(1 - \hat{y}a_1(\lambda))^3} - \frac{a_2(\lambda)}{(1 - \hat{y}a_1(\lambda))^4}.$$

Following our previous work Wang et al. (2015), it could be possible to establish some theoretical analysis of the estimation $\hat{\lambda}_0$ under the conditions given in the formula (3.6). For general Σ, μ_1, μ_2, theoretically it is hard to conduct such an effective estimation. Therefore, we skip the theoretical analysis of the proposed tuning parameter estimation and conduct simulations in Section 5, to compare the $\hat{\lambda}_0$ proposed above with the one obtained by cross validations.

4. Examples

In this section, we use several examples to illustrate the theoretical results obtained in Section 3. For simplicity, in all numerical studies, n_1 and n_2 are set to be equal and hence the bias-corrected RLDA is identical to RLDA. From Theorem 3.1, given the ratios y_1, y_2, we know the misclassification error rate depends on $R_k(\lambda), k = 1, 2$ and $H_k(\lambda), k = 1, 2$ where $R_k(\lambda)$ is determined by the population covariance Σ and $H_k(\lambda)$ involves the structure between $\mu_1 - \mu_2$

and Σ. Write the spectral decomposition of Σ as

$$\Sigma = \sum_{i=1}^{p} \lambda_i v_i v_i^{\mathrm{T}},$$

where $\lambda_1 \geq \cdots \geq \lambda_p > 0$ are the eigenvalues of Σ and $v_1, \cdots, v_p$ are the corresponding eigenvectors. To be specific, $R_k(\lambda)$ is determined by the eigenvalues $\lambda_1, \cdots, \lambda_p$ while $H_k(\lambda)$ involves the eigenvalues as well as the inner products $v_k^{\mathrm{T}}(\mu_1 - \mu_2), k = 1, \cdots, p$. Using the spectral decomposition of Σ, we have

$$(\mu_1 - \mu_2)^{\mathrm{T}}\Sigma^{-1/2}(I_p + t\Sigma^{-1})^{-1}\Sigma^{-1/2}(\mu_1 - \mu_2) = \sum_{i=1}^{p} \frac{1}{\lambda_i + t}(v_i^{\mathrm{T}}(\mu_1 - \mu_2))^2,$$

$$(\mu_1 - \mu_2)^{\mathrm{T}}\Sigma^{-1/2}(I_p + t\Sigma^{-1})^{-2}\Sigma^{-1/2}(\mu_1 - \mu_2) = \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i + t)^2}(v_i^{\mathrm{T}}(\mu_1 - \mu_2))^2.$$

In what following, we consider three cases. Firstly, we consider $\Sigma = \sigma^2 I_p$ where all the eigenvalues are identical, and the results would depends on $\sum_{i=1}^{p}(v_i^{\mathrm{T}}(\mu_1 - \mu_2))^2 = (\mu_1 - \mu_2)^{\mathrm{T}}(\mu_1 - \mu_2)$ only. Secondly, we consider three isotropic cases where $\mu_1 - \mu_2$, $\Sigma^{-1/2}(\mu_1 - \mu_2)$ or $\Sigma^{-1}(\mu_1 - \mu_2)$ has nonzero projections on all the eigenvectors of Σ. Consequently, we would have $v_i^{\mathrm{T}}(\mu_1 - \mu_2) \neq 0$ for all i. Lastly, we consider the case where $\mu_1 - \mu_2$ is parallel to one of the eigenvectors of Σ, and hence only one of the $v_i^{\mathrm{T}}(\mu_1 - \mu_2)$, $i = 1, \cdots, p$ is nonzero.

4.1. $\Sigma = \sigma^2 I_p$

We first study the case where Σ has a simple form $\Sigma = \sigma^2 I_p$. It can be easily seen that (C2) and (C3) hold with H being a degenerated distribution function at σ^2 and

$$h_1(t) = \frac{\sigma^2}{\sigma^2 + t}, \quad h_2(t) = \frac{\sigma^4}{(\sigma^2 + t)^2}.$$

By the Stieltjes transformation of the MP law (see Bai and Silverstein, 2010, Section 3.3), we have,

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

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

Together with the fact that $R_1(\lambda) = \sigma^2 m_0(-\lambda)$ and $R_2(\lambda) = \sigma^4 m_0'(-\lambda)$, we would have

$$\mathrm{R}_{\mathrm{RLDA}}^c(\lambda) \xrightarrow{p}$$

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

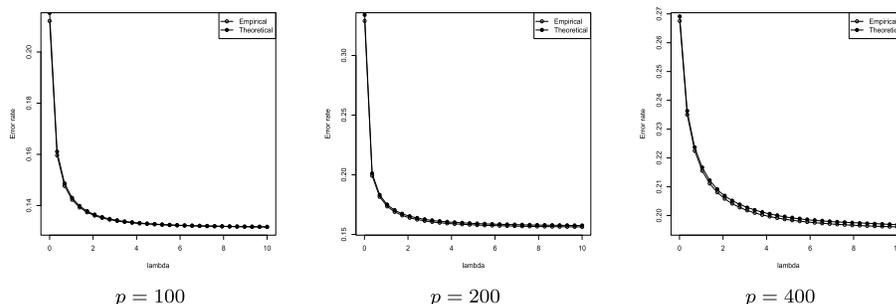$p = 100$ $p = 200$ $p = 400$

FIG 1. *Simulations for independent cases where $\Sigma = I_p$, $n_1 = n_2 = 100$, $\mu_2 = (0, \cdots, 0)$ and $\mu_1 = (2.56, 0, \cdots, 0)$. The Bayes misclassification error rate is 10%.*

We remark that Corollary 3.3 of Dobriban and Wager (2018) presents the same result for a special case where $y = 1$ and $\sigma^2 = 1$. Here in our results we explicitly describe the effect of dimension and the tuning parameter λ for general y and σ^2. Figure 1 show the simulation results for $y = 0.5, y = 1$ and $y = 2$ from which we can see the empirical results and the theoretical conclusions are consistent.

4.2. Isotropic case

Given the eigenvectors $v_1, \cdots, v_p$, we consider three different isotropic cases

$$\mu_1 - \mu_2 \propto \frac{1}{\sqrt{p}}(v_1 + \cdots + v_p), \tag{4.1}$$

$$\Sigma^{-1/2}(\mu_1 - \mu_2) \propto \frac{1}{\sqrt{p}}(v_1 + \cdots + v_p), \tag{4.2}$$

$$\Sigma^{-1}(\mu_1 - \mu_2) \propto \frac{1}{\sqrt{p}}(v_1 + \cdots + v_p). \tag{4.3}$$

In statistics, these cases have different implications. (4.1) implies that the direct differences between the two population means, (4.2) indicates the differences after covariance standardization, and (4.3) implies that the true classification weights. Similar mean-covariance structures also arise in Cai, Liu and Xia (2014) for the testing problem $H_0 : \mu_1 = \mu_2$. In random matrix theory, Bai, Miao and Pan (2007) considered (4.1) when studying the asymptotic property of the eigenvectors and more details can be found in their Remark 1.

To fix the Bayes error rate, we set $\Delta^2 = (\mu_1 - \mu_2)^{\mathrm{T}}\Sigma^{-1}(\mu_1 - \mu_2)$ to be a constant and for each case, we list the results as follows.

- Case (4.1):

$$\mu_1 - \mu_2 = \frac{\Delta}{\sqrt{\sum_{i=1}^{p} \lambda_i^{-1}}}(v_1 + \cdots + v_p),$$

$$h_1(t) = \left\{ \int \frac{1}{x} dH(x) \right\}^{-1} \int \frac{1}{x+t} dH(x),$$

$$h_2(t) = \left\{ \int \frac{1}{x} dH(x) \right\}^{-1} \int \frac{x}{(x+t)^2} dH(x),$$

$$H_1(\lambda) = \left\{ \int \frac{1}{x} dH(x) \right\}^{-1} m_0(-\lambda),$$

$$H_2(\lambda) = \left\{ \int \frac{1}{x} dH(x) \right\}^{-1} \frac{m_0(-\lambda) - \lambda m_0'(-\lambda)}{\{1 - y[1 - \lambda m_0(-\lambda)]\}^2}.$$

- Case (4.2):

$$\mu_1 - \mu_2 = \frac{\Delta}{\sqrt{p}} (\sqrt{\lambda_1} v_1 + \cdots + \sqrt{\lambda_p} v_p),$$

$$h_1(t) = \int \frac{x}{x+t} dH(x), \ h_2(t) = \int \frac{x^2}{(x+t)^2} dH(x),$$

$$H_1(\lambda) = R_1(\lambda), \ H_2(\lambda) = R_2(\lambda).$$

- Case (4.3):

$$\mu_1 - \mu_2 = \frac{\Delta}{\sqrt{\sum_{i=1}^p \lambda_i}} (\lambda_1 v_1 + \cdots + \lambda_p v_p),$$

$$h_1(t) = \left\{ \int x dH(x) \right\}^{-1} \int \frac{x^2}{x+t} dH(x),$$

$$h_2(t) = \left\{ \int x dH(x) \right\}^{-1} \int \frac{x^3}{(x+t)^2} dH(x),$$

$$H_1(\lambda) = \frac{1}{1 - y[1 - \lambda m_0(-\lambda)]} - \left\{ \int x dH(x) \right\}^{-1} \frac{\lambda - \lambda^2 m_0(-\lambda)}{\{1 - y[1 - \lambda m_0(-\lambda)]\}^2},$$

$$H_2(\lambda) = (1 + R_1(\lambda))^2 \left\{ 1 - \frac{\lambda R_1(\lambda)}{\int x dH(x)} \right\} + \left\{ y - \frac{\lambda + 2y\lambda R_1(\lambda)}{\int x dH(x)} \right\} R_2(\lambda).$$

From the Bayesian perspective, we can use random effects that assume the true parameters are random to describe isotropic cases. For example, Dobriban and Wager (2018) considered random weights:

$$\mu_1 - \mu_2 \overset{d}{=} (Z_1, \cdots, Z_p)^{\mathrm{T}},$$

where $Z_1, \cdots, Z_p$ are i.i.d random variables. Under their setting,

$$(\mu_1 - \mu_2)^{\mathrm{T}} \Sigma^{-1/2} (I_p + t\Sigma^{-1})^{-1} \Sigma^{-1/2} (\mu_1 - \mu_2) \overset{a.s.}{\to} c_1 \int \frac{1}{x+t} dH(x),$$

$$(\mu_1 - \mu_2)^{\mathrm{T}} \Sigma^{-1/2} (I_p + t\Sigma^{-1})^{-2} \Sigma^{-1/2} (\mu_1 - \mu_2) \overset{a.s.}{\to} c_2 \int \frac{x}{(x+t)^2} dH(x),$$

and then $H_k(\lambda)$ have similar forms as our case (4.1). Similarly, we can also consider the random effects

$$\Sigma^{-1/2}(\mu_1 - \mu_2) \overset{d}{=} (Z_1, \cdots, Z_p)^{\mathrm{T}}, \text{ or } \Sigma^{-1}(\mu_1 - \mu_2) \overset{d}{=} (Z_1, \cdots, Z_p)^{\mathrm{T}},$$

and the results would exactly correspond to the cases (4.2) and (4.3), respectively.

4.3. Sparse case

In this subsection, we consider the case where

$$\mu_1 - \mu_2 \propto v_k.$$

We view this case as a sparse case since $\mu_1 - \mu_2 \propto (0, \cdots, 0, 1, 0, \cdots, 0)$ under the basis vectors $\{v_1, \cdots, v_p\}$. When $\mu_1 - \mu_2 \propto v_k$, we have $\Sigma^{-1}(\mu_1 - \mu_2) \propto v_k$, indicating that the optimal linear classification direction is parallel to the eigenvector v_k. We consider $\mu_1 - \mu_2 = \Delta\sqrt{\lambda_i}v_i$. Then

$$\Delta^{-2}(\mu_1 - \mu_2)^{\mathrm{T}}\Sigma^{-1/2}(I_p + t\Sigma^{-1})^{-1}\Sigma^{-1/2}(\mu_1 - \mu_2) = \frac{\lambda_i}{\lambda_i + t},$$

$$\Delta^{-2}(\mu_1 - \mu_2)^{\mathrm{T}}\Sigma^{-1/2}(I_p + t\Sigma^{-1})^{-2}\Sigma^{-1/2}(\mu_1 - \mu_2) = \frac{\lambda_i^2}{(\lambda_i + t)^2}.$$

When λ_i has a limit which we still denote as λ_i, we have

$$h_1(t) = \frac{\lambda_i}{\lambda_i + t}, h_2(t) = \frac{\lambda_i^2}{(\lambda_i + t)^2},$$

and

$$H_1(\lambda) = \frac{1}{1 - y + y\lambda m_0(-\lambda) + \lambda/\lambda_i},$$

$$H_2(\lambda) = \frac{1}{\{1 - y + y\lambda m_0(-\lambda) + \lambda/\lambda_i\}^2} \left\{ \frac{1}{1 - y(1 - \lambda m_0(-\lambda))} - \frac{y(\lambda m_0(-\lambda) - \lambda^2 m_0'(-\lambda))}{(1 - y(1 - \lambda m_0(-\lambda)))^2} \right\}.$$

To illustrate the results, we consider a toy example $\Sigma = (\rho^{|i-j|})_{p \times p}$ with $|\rho| < 1$. The covariance matrix Σ is related to a stationary AR(1) process and is also used in Bickel and Levina (2004) for LDA. By the Szegö theorem, we have,

$$\lambda_k \approx \frac{1 - \rho^2}{1 + \rho^2 - 2\rho \cos\frac{k\pi}{p+1}}.$$

Thus, $\lambda_1 \to (1+\rho)/(1-\rho)$, $\lambda_p \to (1-\rho)/(1+\rho)$ and $\lambda_{[p/2]} \to (1-\rho^2)/(1+\rho^2)$. Figure 2 show the corresponding eigenvectors and Figure 3 presents the results for $\mu_1 - \mu_2 \propto v_k$, $k = 1, 50, 100$. We can see from Figure 3 that the dimension effects have different performances for each v_k and the tuning parameter λ brings in different effects on the misclassification error rate. For all the simulations, the theoretical misclassification rates are consistent with the empirical ones.
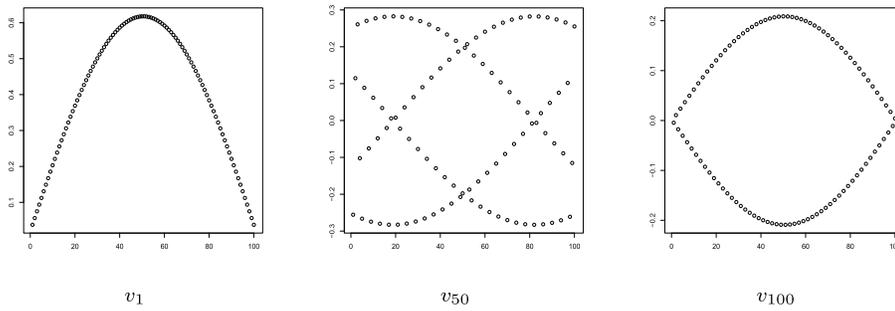
v_1 v_{50} v_{100}

FIG 2. *The eigenvectors of $\Sigma = (0.5^{|i-j|})_{100 \times 100}$ corresponding to λ_1, λ_{50} and λ_{100}.*
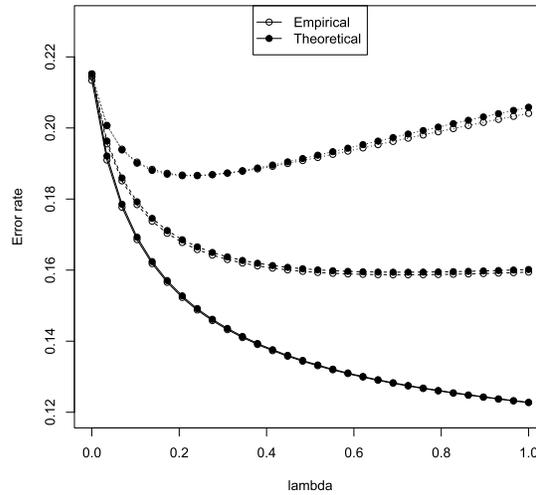

FIG 3. *The empirical and theoretical misclassification rates for $\Sigma = (0.5^{|i-j|})_{100 \times 100}$ and $n_1 = n_2 = 100$. The line stands for $\mu_1 - \mu_2 = 4.435 \cdot v_1$; the dashed line is the results for $\mu_1 - \mu_2 = 2.005 \cdot v_{50}$ and the dotted line is the one for $\mu_1 - \mu_2 = 1.480 \cdot v_{100}$. For all the cases, the Bayes error rate is 10%.*

5. Simulations

In this section, we conduct several simulations to illustrate the results. We consider the covariance matrix structures as follows:

$$\Sigma = (\rho^{|i-j|})_{p \times p},$$

where ρ reflects the correlations between the covariates and $|\rho| < 1$. For all of our simulations, we let $\mu_2 = 0$ and rescale μ_1 to control the Bayes error rate as 10%. Here we fix the sample size $n = 200$ and all the results are based on 100 replications.

5.1. Bias correction for LDA and RLDA

When $n_1 \neq n_2$, we propose a bias correction procedure to the intercept part for LDA and RLDA. Specially, we have

$$\delta_{\mathrm{LDA}}^c(X) = \mathbb{I}\left\{\left(X - \frac{\bar{X}_1 + \bar{X}_2}{2}\right)^{\mathrm{T}} S_n^{-1}(\bar{X}_1 - \bar{X}_2) + \frac{n-2}{n-p-3}\left(\frac{p}{2n_1} - \frac{p}{2n_2}\right) > 0\right\},$$

and

$$\delta_{\mathrm{RLDA}}^c(X) = \mathbb{I}\Bigg\{\left(X - \frac{\bar{X}_1 + \bar{X}_2}{2}\right)^{\mathrm{T}} (S_n + \lambda I_p)^{-1}(\bar{X}_1 - \bar{X}_2)$$
$$+ \left(\frac{p}{2n_1} - \frac{p}{2n_2}\right)\frac{1 - \frac{1}{p}tr(\frac{1}{\lambda}S_n + I_p)^{-1}}{1 - \frac{p}{n-2} + \frac{1}{n-2}tr(\frac{1}{\lambda}S_n + I_p)^{-1}} > 0\Bigg\}, \ \lambda > 0.$$

Our first simulation is to show the performance of the bias-corrected LDA and RLDA. We fix $n = 200$ and let n_1 rang from 10 to 190. We set λ as 0 (which corresponds to LDA), 0.1 and 0.5 for $p < n$ and 0.1, 0.5 and 1 when $p \geq n$. The data dimension p is set to be 100, 200 or 400. As a benchmark, we also include the linear classifier with optimal constant

$$\alpha_r = -\frac{1}{2}(\mu_1 + \mu_2)^{\mathrm{T}}(S_n + \lambda I_p)^{-1}(\bar{X}_1 - \bar{X}_2),$$

which is denoted by O-RLDA. We set $\Sigma = (0.5^{|i-j|})_{p \times p}$ and $\mu_1 \propto (1, 0, \cdots, 0)$, and all the simulation results are plotted in Figure 4. We can observe that the bias corrected RLDA (C-RLDA) achieves less misclassification rate than the original RLDA and also the performance of C-RLDA is quite close the one of RLDA with optimal constants. Specially, the first figure of Figure 4 is the results for LDA where $p = 100$, $n = 200$. We also conduct simulations for more covariance and mean structures and the results follow similar patterns.

5.2. Dimension effect of LDA, naive Bayes and RLDA

In this work, our main focus is to explicitly derive the dimension effects for LDA and RLDA. In detail, the asymptotic misclassification rate of LDA depends on the ratios $p/n_1, p/n_2$ and the one of RLDA also involves the structure of the covariance the means. More details can be found in Proposition 2.1 and 3.2. In Remark 2.1, we derive the dimension effect for naive Bayes. Here, we conduct a comprehensive comparison for LDA, naive Bayes and RLDA. For μ_1, we consider three scenarios:

    Case 1: $\mu_1 \propto (1, 0, \cdots, 0)$;
    Case 2: 10 % elements of μ_1 are from $N(0, 1)$;
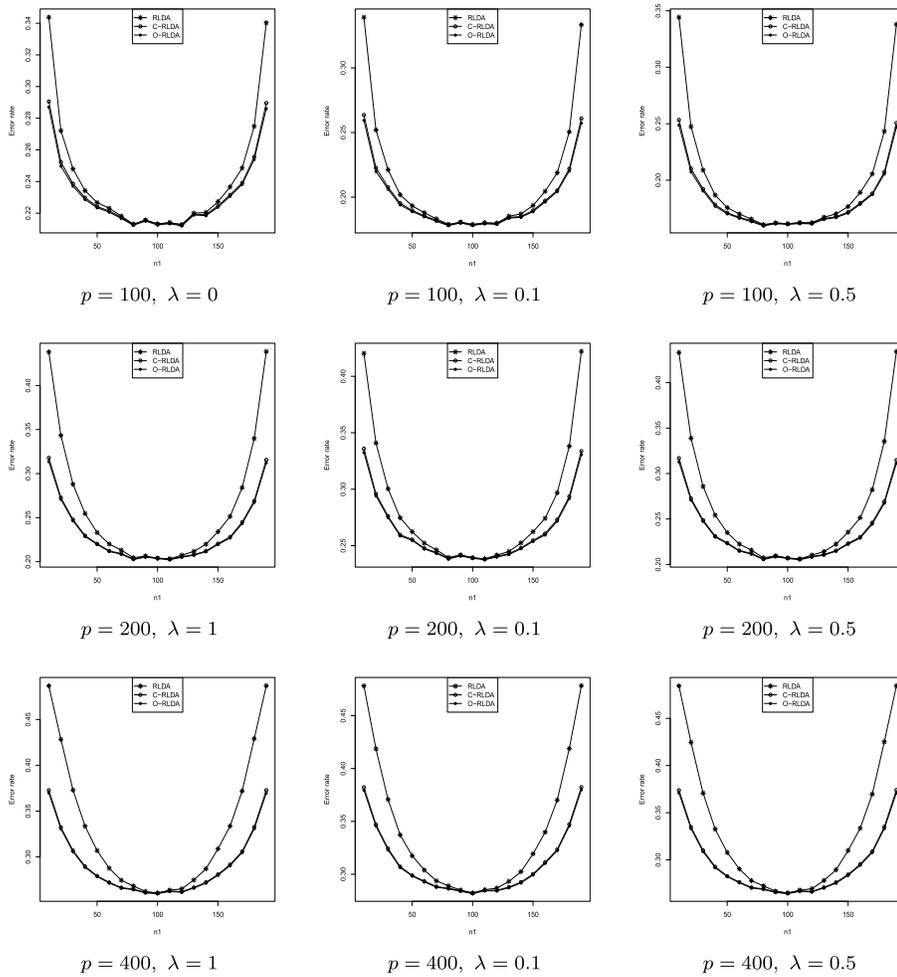   Case 3: all the elements of μ_1 are from $N(0, 1)$;

FIG 4. *Simulations for RLDA, bias corrected RLDA (C-RLDA) and RLDA with optimal constant (O-RLDA). The Bayes misclassification error rate is* 10%.

and for each case we still rescale μ_1 to control the Bayes error rate to be 10%. All the results are presented in Figure 5. For each heat map of the error rate, the horizontal is the data dimension p ranging from 10 to 190 for LDA or 400 for naive Bayes and RLDA. The vertical is the correlation ρ ranging from -0.9 to 0.9. When the data dimension p is increased, the misclassification rates for all the classifiers also increases which is well understood. When the correlations $|\rho|$ ranges from 0 to 0.9, the performance of LDA almost has no changes and Naive Bayes gets worse and worse. For RLDA, the performance is affected by $|\rho|$, but is less sensitive than naive Bayes.
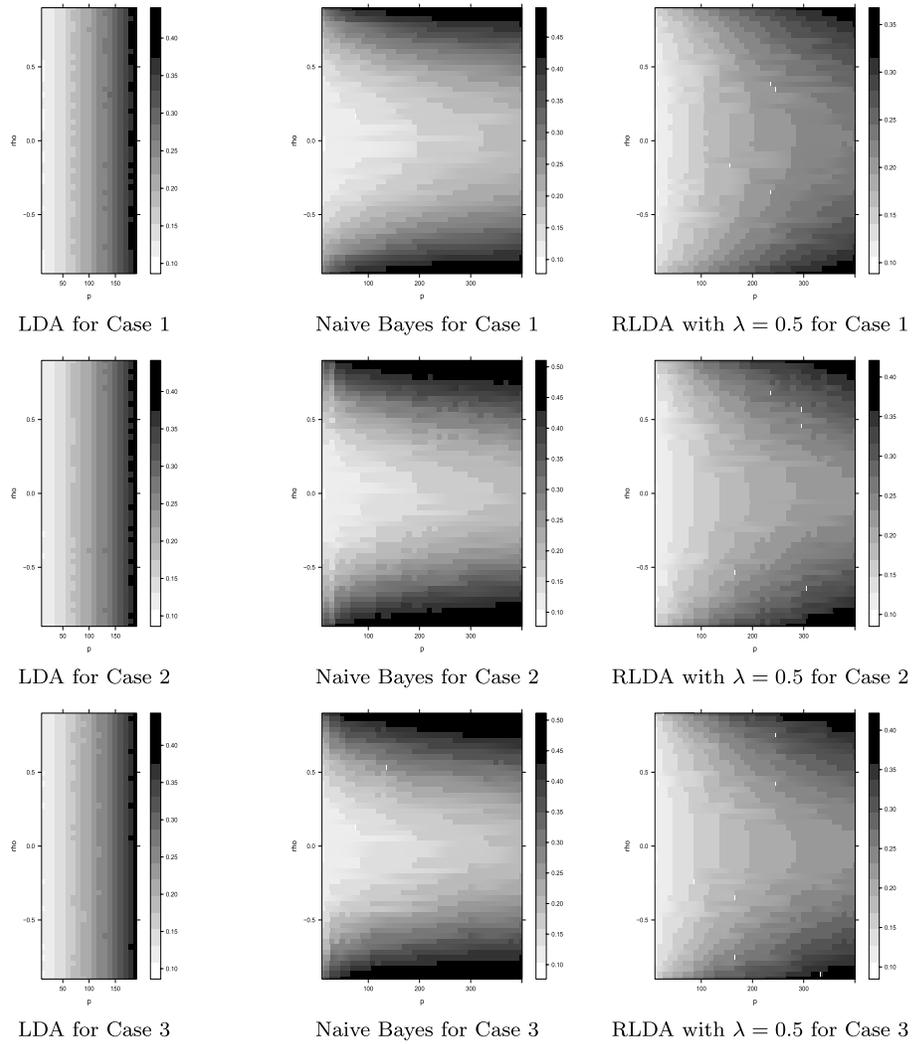
LDA for Case 1   Naive Bayes for Case 1   RLDA with $\lambda = 0.5$ for Case 1

LDA for Case 2   Naive Bayes for Case 2   RLDA with $\lambda = 0.5$ for Case 2

LDA for Case 3   Naive Bayes for Case 3   RLDA with $\lambda = 0.5$ for Case 3

FIG 5. *Heat maps of the error rates for LDA, Naive Bayes and RLDA with $\lambda = 0.5$ where the horizontal is the data dimension p ranging from 10 to 190 or 400 and the vertical is the correlation ρ ranging from -0.9 to 0.9. The Bayes misclassification error rate is* 10%.

5.3. RLDA and sparse LDA

In this part, we compare RLDA with the LPD (Cai and Liu, 2011) which is one of sparse LDA methods. The tuning parameters are chosen by 5-folds cross validation. We also include the RLDA where λ_n is estimated by our method in Section 3.3 and the naive Bayes. In summary, we have four methods NB, LPD, RLDA-CV and RLDA. We control the sparsity level as

$$\Sigma^{-1}(\mu_1 - \mu_2) \propto (\text{rnorm(s)}, 0, \cdots, 0);$$
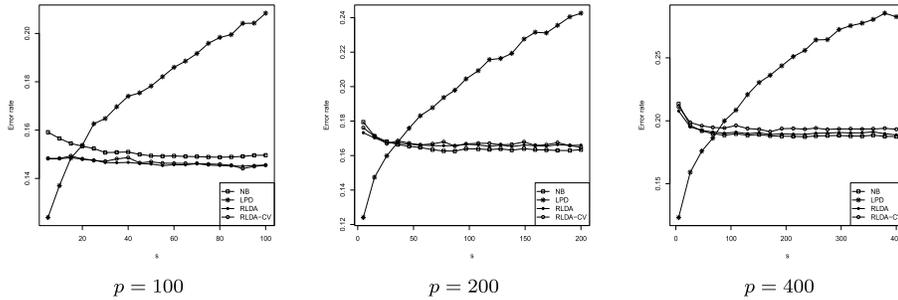
$p = 100$ $p = 200$ $p = 400$

FIG 6. *Misclassification rates for naive Bayes, LPD and RLDA with cross validation and estimated λ. The horizontal is the number of non-zeros of $\Sigma^{-1}(\mu_1 - \mu_2)$.*

with rnorm(s) represent s random variables generated independently from $N(0, 1)$, and in this study s is increased from 5 to p. Figure 6 presents the simulation results. We observe that when s is small, LPD is outstanding with the least error rates and the performance becomes poorer and poorer when s increases, while RLDA and NB are robust to different s. Furthermore, we find out that the version with estimated tuning parameter is comparable to the one with cross validation. Overall, we claim that LPD is applicable to the sparse cases while RLDA is favorable to the dense cases.

Acknowledgments

6. Appendix

Lemma 6.1. *Assuming $W_m \sim Wishart(I_p, m)$ where $m > p + 7$, for any non-random unit vector $e \in \mathbb{R}^p$, we have*

$$E(e^T W_m^{-1} e) = \frac{1}{m - p - 1}, \tag{6.1}$$

$$E(e^T W_m^{-2} e) = \frac{m - 1}{(m - p)(m - p - 1)(m - p - 3)}, \tag{6.2}$$

and

$$E\{(e^T W_m^{-1} e)^2\} = \frac{1}{(m - p - 1)(m - p - 3)}, \tag{6.3}$$

$$E\{(e^T W_m^{-2} e)^2\} = \frac{m^2}{(m - p)^6} \Big\{ 1 + O\Big(\frac{1}{m - p}\Big) \Big\}. \tag{6.4}$$

Proof of Lemma 6.1. For any non-random orthogonal matrix U, we have

$$U^{\mathrm{T}}W_m U \sim \mathrm{Wishart}(I_p, m),$$

and

$$e^{\mathrm{T}}W_m^{-1}e \overset{d}{=} (Ue)^{\mathrm{T}}W_m^{-1}Ue, \ \ e^{\mathrm{T}}W_m^{-2}e \overset{d}{=} (Ue)^{\mathrm{T}}W_m^{-2}Ue.$$

Setting $Ue = (1, 0, \cdots, 0)^{\mathrm{T}}$ yields

$$e^{\mathrm{T}}W_m^{-1}e \overset{d}{=} (W_m^{-1})_{11}, \ \ e^{\mathrm{T}}W_m^{-2}e \overset{d}{=} (W_m^{-2})_{11} = \sum_{i=1}^{p}(W_m^{-1})_{1i}^2.$$

Therefore, we can get the results (6.1)-(6.3) by von Rosen (1988) or the Proposition 2.1 of Cook and Forzani (2011). The order of (6.4) is derived based on Theorem 2 of Matsumoto (2012) and properties of the Weingarten function (Collins and Śniady, 2006) which involves lengthy and tedious calculations and is hence neglected. □

Lemma 6.2. *Assuming $Y \sim N(0, I_p)$ and Y is independent with S_n, we have*

$$\frac{1}{p}Y^{\mathrm{T}}B_n(\lambda)Y \overset{p}{\to} R_1(\lambda), \ \ \text{and} \ \ \frac{1}{p}Y^{\mathrm{T}}B_n^2(\lambda)Y \overset{p}{\to} R_2(\lambda),$$

where $B_n(\lambda)$ and $R_1(\lambda), R_2(\lambda)$ are defined as in (3.1) and Theorem 3.1 respectively.

Proof of Lemma 6.2. Let $\|\cdot\|$ be the matrix operator norm. By direct calculations can have,

$$E\Big\{\frac{1}{p}Y^{\mathrm{T}}B_n(\lambda)Y - \frac{1}{p}\mathrm{tr}B_n(\lambda)\Big\} = 0,$$

$$E\Big\{\frac{1}{p}Y^{\mathrm{T}}B_n(\lambda)Y - \frac{1}{p}\mathrm{tr}B_n(\lambda)\Big\}^2 = E\frac{2}{p^2}\mathrm{tr}B_n^2(\lambda) \leq \frac{2\|\Sigma\|^2}{\lambda^2 p} \to 0,$$

$$E\Big\{\frac{1}{p}Y^{\mathrm{T}}B_n^2(\lambda)Y - \frac{1}{p}\mathrm{tr}B_n^2(\lambda)\Big\} = 0,$$

$$E\Big\{\frac{1}{p}Y^{\mathrm{T}}B_n^2(\lambda)Y - \frac{1}{p}\mathrm{tr}B_n^2(\lambda)\Big\}^2 = E\frac{2}{p^2}\mathrm{tr}B_n^4(\lambda) \leq \frac{2\|\Sigma\|^4}{\lambda^4 p} \to 0.$$

Here, we used the fact that

$$B_n(\lambda) = \Sigma^{\frac{1}{2}}(S_n + \lambda I)^{-1}\Sigma^{\frac{1}{2}} \overset{d}{=} \big(\frac{1}{n-2}W_n + \lambda\Sigma^{-1}\big)^{-1},$$

where $W_n \sim \mathrm{Wishart}(I_p, n-2)$ is semidefinite and hence

$$\|B_n(\lambda)\| \leq \frac{1}{\lambda_{min}(\lambda\Sigma^{-1})} = \frac{\|\Sigma\|}{\lambda}.$$

Consequently we have,

$$\frac{1}{p}Y^\mathsf{T}B_n(\lambda)Y = \frac{1}{p}\text{tr}B_n(\lambda) + o_p(1), \ \frac{1}{p}Y^\mathsf{T}B_n^2(\lambda)Y = \frac{1}{p}\text{tr}B_n^2(\lambda) + o_p(1).$$

Together with Theorem 2 of Wang et al. (2015) we have,

$$\frac{1}{p}\text{tr}B_n(\lambda) \xrightarrow{p} R_1(\lambda), \ \frac{1}{p}\text{tr}B_n^2(\lambda) \xrightarrow{p} R_2(\lambda). \tag{6.5}$$

The proof is completed. □

Lemma 6.3. *Under the conditions (C1)-(C3),*

$$\mu^T B_n(\lambda)\mu \xrightarrow{p} \Delta^2 H_1(\lambda), \ \mu^T B_n^2(\lambda)\mu \xrightarrow{p} \Delta^2 H_2(\lambda).$$

Proof of Lemma 6.3. By El Karoui and Holger (2011) we have,

$$\mu^\mathsf{T}B_n(\lambda)\mu - \mu^\mathsf{T}\Sigma^{\frac{1}{2}}(\gamma_n\Sigma + \lambda I_p)^{-1}\Sigma^{\frac{1}{2}}\mu \xrightarrow{p} 0,$$
$$\mu^\mathsf{T}B_n^2(\lambda)\mu - (1+\epsilon_n)\mu^\mathsf{T}\Sigma^{\frac{1}{2}}(\gamma_n\Sigma + \lambda I_p)^{-1}\Sigma(\gamma_n\Sigma + \lambda I_p)^{-1}\Sigma^{\frac{1}{2}}\mu \xrightarrow{p} 0,$$

where

$$\gamma_n = \frac{1}{1 + \frac{1}{n-2}\text{tr}B_n(\lambda)}, \ \epsilon_n = \frac{\gamma_n^2}{n-2}\text{tr}B_n^2(\lambda).$$

From (6.5) we have

$$\gamma_n \xrightarrow{p} \frac{1}{1 + yR_1(\lambda)} = 1 - y(1 - \lambda m_0(-\lambda)) \overset{\text{def}}{=} \gamma,$$
$$\epsilon_n \xrightarrow{p} \frac{yR_2(\lambda)}{(1 + yR_1(\lambda))^2} = \frac{y(1 - \lambda m_0(-\lambda))}{1 - y(1 - \lambda m_0(-\lambda))} - \frac{y(\lambda m_0(-\lambda) - \lambda^2 m_0'(-\lambda))}{(1 - y(1 - \lambda m_0(-\lambda)))^2} \overset{\text{def}}{=} \epsilon.$$

Consequently it can be routinely shown that

$$\mu^\mathsf{T}B_n(\lambda)\mu - \mu^\mathsf{T}\Sigma^{\frac{1}{2}}(\gamma\Sigma + \lambda I_p)^{-1}\Sigma^{\frac{1}{2}}\mu \xrightarrow{p} 0,$$
$$\mu^\mathsf{T}B_n^2(\lambda)\mu - \frac{1+\epsilon}{\gamma^2}\Big\{\mu^\mathsf{T}\Big(I_p + \frac{\lambda}{\gamma}\Sigma^{-1}\Big)^{-2}\mu\Big\} \xrightarrow{p} 0,$$

By Condition (C3),

$$\mu^\mathsf{T}\Sigma^{\frac{1}{2}}(\gamma\Sigma + \lambda I_p)^{-1}\Sigma^{\frac{1}{2}}\mu = \gamma^{-1}\mu^\mathsf{T}\Big(I_p + \frac{\lambda}{\gamma}\Sigma^{-1}\Big)^{-1}\mu \to \gamma^{-1}h_1\Big(\frac{\lambda}{\gamma}\Big)\Delta^2,$$

and

$$\frac{1+\epsilon}{\gamma^2}\Big\{\mu^\mathsf{T}\Big(I_p + \frac{\lambda}{\gamma}\Sigma^{-1}\Big)^{-2}\mu\Big\} \to \frac{1+\epsilon}{\gamma^2}h_2\Big(\frac{\lambda}{\gamma}\Big)\Delta^2.$$

The proof is completed. □

6.1. Proof of Theorem 2.1

By the properties of Gaussian distributions,

$$\bar{X}_1 \overset{d}{=} \frac{1}{\sqrt{n_1}}\Sigma^{\frac{1}{2}}Y_1 + \mu_1, \ \bar{X}_2 \overset{d}{=} \frac{1}{\sqrt{n_2}}\Sigma^{\frac{1}{2}}Y_2 + \mu_2,$$

where $Y_1, Y_2 \sim N(0, I_p)$ and Y_1, Y_2 are independent. We then have,

$$2\Big(\mu_1 - \frac{\bar{X}_1 + \bar{X}_2}{2}\Big)^{\mathrm{T}}\Sigma^{-1}(\bar{X}_1 - \bar{X}_2) = \Delta^2 - \frac{2}{\sqrt{n_2}}\mu^{\mathrm{T}}Y_2 + \frac{1}{n_2}Y_2^{\mathrm{T}}Y_2 - \frac{1}{n_1}Y_1^{\mathrm{T}}Y_1$$

$$\overset{p}{\to} \Delta^2 + y_2 - y_1,$$

and

$$-2\Big(\mu_2 - \frac{\bar{X}_1 + \bar{X}_2}{2}\Big)^{\mathrm{T}}\Sigma^{-1}(\bar{X}_1 - \bar{X}_2) = \Delta^2 + \frac{2}{\sqrt{n_1}}\mu^{\mathrm{T}}Y_1 + \frac{1}{n_1}Y_1^{\mathrm{T}}Y_1 - \frac{1}{n_2}Y_2^{\mathrm{T}}Y_2$$

$$\overset{p}{\to} \Delta^2 + y_1 - y_2,$$

where μ is defined as in (3.1). For the denominator,

$$(\bar{X}_1 - \bar{X}_2)^{\mathrm{T}}\Sigma^{-1}(\bar{X}_1 - \bar{X}_2)$$
$$= \Delta^2 + \frac{1}{n_1}Y_1^{\mathrm{T}}Y_1 + \frac{1}{n_2}Y_2^{\mathrm{T}}Y_2 + 2\mu^{\mathrm{T}}\Big(\frac{1}{\sqrt{n_1}}Y_1 - \frac{1}{\sqrt{n_2}}Y_2\Big) - \frac{2}{\sqrt{n_1 n_2}}Y_1^{\mathrm{T}}Y_2$$
$$\overset{p}{\to} \Delta^2 + y_1 + y_2.$$

By the Slutsky's theorem and the continuous mapping theorem, the proof is completed. □

6.2. Proof of Theorem 2.2

Write

$$S_n \overset{d}{=} \frac{1}{n-2}\Sigma^{\frac{1}{2}}W_n\Sigma^{\frac{1}{2}},$$

where $W_n \sim \text{Wishart}(I_p, n-2)$. Then

$$\mathrm{R}_{\mathrm{LDA2}} \overset{d}{=} \Phi\Big(\frac{-\mu^{\mathrm{T}}W_n^{-1}\mu}{2\sqrt{\mu^{\mathrm{T}}W_n^{-2}\mu}}\Big).$$

By Lemma 6.1, when $p/n \to y \in (0, 1)$, we have

$$nE\mu^{\mathrm{T}}W_n^{-1}\mu = \frac{n\Delta^2}{n-p-3} \to \frac{\Delta^2}{1-y},$$
$$n^2E\mu^{\mathrm{T}}W_n^{-2}\mu = \frac{n^2(n-3)\Delta^2}{(n-p-2)(n-p-3)(n-p-5)} \to \frac{\Delta^2}{(1-y)^3},$$

and

$$Var(n\mu^{\mathrm{T}}W_n^{-1}\mu) = \Delta^4 \left\{ \frac{n^2}{(n-p-3)(n-p-5)} - \frac{n^2}{(n-p-3)^2} \right\} \to 0,$$

$$Var(n^2\mu^{\mathrm{T}}W_n^{-2}\mu)$$

$$= \Delta^4 \left\{ \frac{n^6}{(n-p)^6}(1+o(1)) - \frac{n^4(n-3)^2}{(n-p-2)^2(n-p-3)^2(n-p-5)^2} \right\} \to 0.$$

Therefore,

$$n\mu^{\mathrm{T}}W_n^{-1}\mu \xrightarrow{p} \frac{\Delta^2}{1-y}, \quad n^2\mu^{\mathrm{T}}W_n^{-2}\mu \xrightarrow{p} \frac{\Delta^2}{(1-y)^3}.$$

The proof is completed. $\qquad\qquad\square$

6.3. Proof of Theorem 2.3

We have

$$\bar{X}_1 \overset{d}{=} \frac{1}{\sqrt{n_1}}\Sigma^{\frac{1}{2}}Y_1 + \mu_1, \quad \bar{X}_2 \overset{d}{=} \frac{1}{\sqrt{n_2}}\Sigma^{\frac{1}{2}}Y_2 + \mu_2, S_n \overset{d}{=} \frac{1}{n-2}\Sigma^{\frac{1}{2}}W_n\Sigma^{\frac{1}{2}},$$

where $Y_1, Y_2 \sim N(0, I_p)$, $W_n \sim \text{Wishart}(I_p, n-2)$ and Y_1, Y_2, W_n are independent. Then

$$-\left(\mu_1 - \frac{\bar{X}_1 + \bar{X}_2}{2}\right)^{\mathrm{T}} S_n^{-1}(\bar{X}_1 - \bar{X}_2)$$

$$\overset{d}{=} -\frac{n-2}{2}\left(\mu - \frac{1}{\sqrt{n_1}}Y_1 - \frac{1}{\sqrt{n_2}}Y_2\right)^{\mathrm{T}} W_n^{-1}\left(\mu + \frac{1}{\sqrt{n_1}}Y_1 - \frac{1}{\sqrt{n_2}}Y_2\right)$$

$$= -\frac{n-2}{2}\left(\mu - \frac{1}{\sqrt{n_2}}Y_2\right)^{\mathrm{T}} W_n^{-1}\left(\mu - \frac{1}{\sqrt{n_2}}Y_2\right) + \frac{n-2}{2n_1}Y_1^{\mathrm{T}}W_n^{-1}Y_1.$$

By Lemma 6.1, we have,

$$EY_1^{\mathrm{T}}W_n^{-1}Y_1 = \frac{1}{n-p-3}EY_1^{\mathrm{T}}Y_1 = \frac{p}{n-p-3} \to \frac{y}{1-y},$$

and

$$Var(Y_1^{\mathrm{T}}W_n^{-1}Y_1) = E(Y_1^{\mathrm{T}}W_n^{-1}Y_1)^2 - (EY_1^{\mathrm{T}}W_n^{-1}Y_1)^2$$

$$= \frac{E(Y_1^{\mathrm{T}}Y_1)^2}{(n-p-3)(n-p-5)} - \frac{p^2}{(n-p-3)^2}$$

$$= \frac{p^2 + 2p}{(n-p-3)(n-p-5)} - \frac{p^2}{(n-p-3)^2} \to 0.$$

Similarly,

$$E(n-2)\left(\mu - \frac{1}{\sqrt{n_2}}Y_2\right)^{\mathrm{T}} W_n^{-1}\left(\mu - \frac{1}{\sqrt{n_2}}Y_2\right)$$

$$= \frac{n-2}{n-p-3} E\left(\mu - \frac{1}{\sqrt{n_2}} Y_2\right)^{\mathrm{T}} \left(\mu - \frac{1}{\sqrt{n_2}} Y_2\right)$$

$$= \frac{n-2}{n-p-3} \left(\Delta^2 + \frac{p}{n_2}\right) \to \frac{1}{1-y} (\Delta^2 + y_2),$$

and

$$Var\left\{ (n-2)\left(\mu - \frac{1}{\sqrt{n_2}} Y_2\right)^{\mathrm{T}} W_n^{-1} \left(\mu - \frac{1}{\sqrt{n_2}} Y_2\right) \right\}$$

$$= \frac{(n-2)^2 E\{(\mu - \frac{1}{\sqrt{n_2}} Y_2)^{\mathrm{T}} (\mu - \frac{1}{\sqrt{n_2}} Y_2)\}^2}{(n-p-3)(n-p-5)} - \frac{(n-2)^2 \left(\Delta^2 + \frac{p}{n_2}\right)^2}{(n-p-3)^2}$$

$$= \frac{(n-2)^2 \left(\Delta^4 + \frac{(2p+4)\Delta^2}{n_2} + \frac{p^2+2p}{n_2^2}\right)}{(n-p-3)(n-p-5)} - \frac{(n-2)^2 \left(\Delta^2 + \frac{p}{n_2}\right)^2}{(n-p-3)^2} \to 0.$$

Thus,

$$- \left(\mu_1 - \frac{\bar{X}_1 + \bar{X}_2}{2}\right)^{\mathrm{T}} S_n^{-1} (\bar{X}_1 - \bar{X}_2)$$

$$\overset{d}{=} - \frac{n-2}{2} \left(\mu - \frac{1}{\sqrt{n_2}} Y_2\right)^{\mathrm{T}} W_n^{-1} \left(\mu - \frac{1}{\sqrt{n_2}} Y_2\right) + \frac{n-2}{2n_1} Y_1^{\mathrm{T}} W_n^{-1} Y_1$$

$$\overset{p}{\to} - \frac{1}{2(1-y)} (\Delta^2 + y_2) + \frac{y_1}{2y} \frac{y}{1-y} = - \frac{\Delta^2 - y_1 + y_2}{2(1-y)}, \tag{6.6}$$

and by similar arguments, we have

$$\left(\mu_2 - \frac{\bar{X}_1 + \bar{X}_2}{2}\right)^{\mathrm{T}} S_n^{-1} (\bar{X}_1 - \bar{X}_2)$$

$$\overset{d}{=} - \frac{n-2}{2} \left(\mu + \frac{1}{\sqrt{n_1}} Y_1\right)^{\mathrm{T}} W_n^{-1} \left(\mu + \frac{1}{\sqrt{n_1}} Y_1\right) + \frac{n-2}{2n_2} Y_2^{\mathrm{T}} W_n^{-1} Y_2$$

$$\overset{p}{\to} - \frac{\Delta^2 + y_1 - y_2}{2(1-y)}. \tag{6.7}$$

For the denominator, we have

$$(\bar{X}_1 - \bar{X}_2)^{\mathrm{T}} S_n^{-1} \Sigma S_n^{-1} (\bar{X}_1 - \bar{X}_2)$$

$$\overset{d}{=} (n-2)^2 \left(\mu + \frac{1}{\sqrt{n_1}} Y_1 - \frac{1}{\sqrt{n_2}} Y_2\right)^{\mathrm{T}} W_n^{-2} \left(\mu + \frac{1}{\sqrt{n_1}} Y_1 - \frac{1}{\sqrt{n_2}} Y_2\right).$$

By Lemma 6.1, we have,

$$E(n-2)^2 \left(\mu + \frac{1}{\sqrt{n_1}} Y_1 - \frac{1}{\sqrt{n_2}} Y_2\right)^{\mathrm{T}} W_n^{-2} \left(\mu + \frac{1}{\sqrt{n_1}} Y_1 - \frac{1}{\sqrt{n_2}} Y_2\right)$$

$$= \frac{(n-2)^2 (n-3) E\left(\mu + \frac{1}{\sqrt{n_1}} Y_1 - \frac{1}{\sqrt{n_2}} Y_2\right)^{\mathrm{T}} \left(\mu + \frac{1}{\sqrt{n_1}} Y_1 - \frac{1}{\sqrt{n_2}} Y_2\right)}{(n-p-2)(n-p-3)(n-p-5)}$$

$$= \frac{(n-2)^2 (n-3) \left(\Delta^2 + \frac{p}{n_1} + \frac{p}{n_2}\right)}{(n-p-2)(n-p-3)(n-p-5)} \to \frac{\Delta^2 + y_1 + y_2}{(1-y)^3},$$

and

$$E\left\{(n-2)^2\left(\mu + \frac{1}{\sqrt{n_1}}Y_1 - \frac{1}{\sqrt{n_2}}Y_2\right)^{\mathrm{T}}W_n^{-2}\left(\mu + \frac{1}{\sqrt{n_1}}Y_1 - \frac{1}{\sqrt{n_2}}Y_2\right)\right\}^2$$

$$=\left\{\frac{n^6}{(n-p)^6} + o(1)\right\}E\left\{\left(\mu + \frac{1}{\sqrt{n_1}}Y_1 - \frac{1}{\sqrt{n_2}}Y_2\right)^{\mathrm{T}}\left(\mu + \frac{1}{\sqrt{n_1}}Y_1 - \frac{1}{\sqrt{n_2}}Y_2\right)\right\}^2$$

$$=\left\{\frac{n^6}{(n-p)^6} + o(1)\right\}\left\{\left(\Delta^2 + \frac{p}{n_1} + \frac{p}{n_2}\right)^2 + o(1)\right\} \to \frac{(\Delta^2 + y_1 + y_2)^2}{(1-y)^6}.$$

Thus,

$$(\bar{X}_1 - \bar{X}_2)^{\mathrm{T}}S_n^{-1}\Sigma S_n^{-1}(\bar{X}_1 - \bar{X}_2) \xrightarrow{p} \frac{\Delta^2 + y_1 + y_2}{(1-y)^3}. \tag{6.8}$$

Combing (6.6), (6.7) and (6.8), the proof is completed. $\qquad\square$

6.4. Proof of Proposition 2.1

With some simple calculation it can be shown that the misclassification rate of $\delta_{\mathrm{LDA}}^c(X)$ is

$$\mathrm{R}_{\mathrm{LDA}}^c = \frac{1}{2}\sum_{j=1}^2 \Phi\left(\frac{(-1)^j\left[(\mu_j - \frac{\bar{X}_1+\bar{X}_2}{2})^{\mathrm{T}}S_n^{-1}(\bar{X}_1 - \bar{X}_2) + \alpha\right]}{\sqrt{(\bar{X}_1 - \bar{X}_2)^{\mathrm{T}}S_n^{-1}\Sigma S_n^{-1}(\bar{X}_1 - \bar{X}_2)}}\right).$$

where $\alpha = \frac{n-2}{n-p-3}\left(\frac{p}{2n_1} - \frac{p}{2n_2}\right)$. Note that from (6.6) and the fact that $\alpha \to \frac{y_1-y_2}{2(1-y)}$, we have

$$-\left[\left(\mu_1 - \frac{\bar{X}_1 + \bar{X}_2}{2}\right)^{\mathrm{T}}S_n^{-1}(\bar{X}_1 - \bar{X}_2) + \alpha\right] \xrightarrow{p} -\frac{\Delta^2 - y_1 + y_2}{2(1-y)} - \frac{y_1 - y_2}{2(1-y)}$$

$$= -\frac{\Delta^2}{2(1-y)}.$$

Similarly, by (6.7) we have

$$\left[\left(\mu_2 - \frac{\bar{X}_1 + \bar{X}_2}{2}\right)^{\mathrm{T}}S_n^{-1}(\bar{X}_1 - \bar{X}_2) + \alpha\right] \xrightarrow{p} -\frac{\Delta^2}{2(1-y)}.$$

Combining with (6.8) we have

$$\mathrm{R}_{\mathrm{LDA}}^c \xrightarrow{p} \Phi\left(-\frac{\Delta^2}{2(1-y)}\frac{(1-y)^{3/2}}{\sqrt{\Delta^2 + y_1 + y_2}}\right)$$

$$= \Phi\left(-\frac{\Delta^2}{2\sqrt{\Delta^2 + y_1 + y_2}}\sqrt{1-y}\right).$$

6.5. Proof of Theorem 3.1

Write

$$T_{1n} = (2\mu_1 - \bar{X}_1 - \bar{X}_2)^{\mathsf{T}}(S_n + \lambda I_p)^{-1}(\bar{X}_1 - \bar{X}_2),$$
$$T_{2n} = -(2\mu_2 - \bar{X}_1 - \bar{X}_2)^{\mathsf{T}}(S_n + \lambda I_p)^{-1}(\bar{X}_1 - \bar{X}_2),$$
$$T_{3n} = (\bar{X}_1 - \bar{X}_2)^{\mathsf{T}}(S_n + \lambda I_p)^{-1}\Sigma(S_n + \lambda I_p)^{-1}(\bar{X}_1 - \bar{X}_2).$$

Since

$$\bar{X}_1 \overset{d}{=} \frac{1}{\sqrt{n_1}}\Sigma^{\frac{1}{2}}Y_1 + \mu_1, \ \bar{X}_2 \overset{d}{=} \frac{1}{\sqrt{n_2}}\Sigma^{\frac{1}{2}}Y_2 + \mu_2,$$

where $Y_1, Y_2 \sim N(0, I_p)$ and Y_1, Y_2, S_n are independent, we have

$$T_{1n} \overset{d}{=} \mu^{\mathsf{T}}B_n(\lambda)\mu - \frac{2}{\sqrt{n_2}}\mu^{\mathsf{T}}B_n(\lambda)Y_2 + \frac{1}{n_2}Y_2^{\mathsf{T}}B_n(\lambda)Y_2 - \frac{1}{n_1}Y_1^{\mathsf{T}}B_n(\lambda)Y_1,$$

$$T_{2n} \overset{d}{=} \mu^{\mathsf{T}}B_n(\lambda)\mu + \frac{2}{\sqrt{n_1}}\mu^{\mathsf{T}}B_n(\lambda)Y_1 + \frac{1}{n_1}Y_1^{\mathsf{T}}B_n(\lambda)Y_1 - \frac{1}{n_2}Y_2^{\mathsf{T}}B_n(\lambda)Y_2,$$

$$T_{3n} \overset{d}{=} \left(\mu + \frac{1}{\sqrt{n_1}}Y_1 - \frac{1}{\sqrt{n_2}}Y_2\right)^{\mathsf{T}}B_n^2(\lambda)\left(\mu + \frac{1}{\sqrt{n_1}}Y_1 - \frac{1}{\sqrt{n_2}}Y_2\right)$$
$$\overset{d}{=} \mu^{\mathsf{T}}B_n^2(\lambda)\mu + 2\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\mu^{\mathsf{T}}B_n^2(\lambda)Y_1 + \left(\frac{1}{n_1} + \frac{1}{n_2}\right)Y_1^{\mathsf{T}}B_n^2(\lambda)Y_1.$$

For the leading terms, by Lemmas 6.2 and 6.3,

$$\frac{1}{n_j}Y_j^{\mathsf{T}}B_n(\lambda)Y_j \overset{p}{\to} y_j R_1(\lambda), \ j = 1, 2$$

$$\left(\frac{1}{n_1} + \frac{1}{n_2}\right)Y_1^{\mathsf{T}}B_n^2(\lambda)Y_1 \overset{p}{\to} (y_1 + y_2)R_2(\lambda)$$

$$\mu^{\mathsf{T}}B_n(\lambda)\mu \overset{p}{\to} H_1(\lambda)\Delta^2, \ \mu^{\mathsf{T}}B_n^2(\lambda)\mu \overset{p}{\to} H_2(\lambda)\Delta^2.$$

For the cross sectional terms,

$$E\left\{\frac{1}{\sqrt{n_j}}\mu^{\mathsf{T}}B_n(\lambda)Y_j\right\}^2 = \frac{1}{n_j}E\mu^{\mathsf{T}}B_n^2(\lambda)\mu \le \frac{\Delta^2\|\Sigma\|^2}{n_j\lambda^2} \to 0,$$

and similarly,

$$E\left\{\frac{1}{\sqrt{n}}\mu^{\mathsf{T}}B_n^2(\lambda)Y_1\right\}^2 = \frac{1}{n}E\mu^{\mathsf{T}}B_n^4(\lambda)\mu \le \frac{\Delta^2\|\Sigma\|^4}{n_j\lambda^4} \to 0.$$

Thus,

$$\frac{1}{\sqrt{n_1}}\mu^{\mathsf{T}}B_n(\lambda)Y_1, \frac{1}{\sqrt{n_2}}\mu^{\mathsf{T}}B_n(\lambda)Y_2, \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\mu^{\mathsf{T}}B_n^2(\lambda)Y_1 \overset{p}{\to} 0.$$

Above all,

$$T_{1n} \overset{p}{\to} H_1(\lambda)\Delta^2 + (y_2 - y_1)R_1(\lambda), \ T_{2n} \overset{p}{\to} H_1(\lambda) + (y_1 - y_2)R_1(\lambda),$$
$$T_{3n} \overset{p}{\to} H_2(\lambda)\Delta^2 + (y_1 + y_2)R_2(\lambda).$$

By the continuous mapping theorem, the proof is completed. □

6.6. Proof of Proposition 3.1

When

$$h_1(t) = \int \frac{x}{x+t} dH(x), \ h_2(t) = \int \frac{x}{(x+t)^2} dH(x),$$

by the definitions of $H_1(\lambda)$ and $H_2(\lambda)$, we have

$$H_1(\lambda) = (1 + yR_1(\lambda))h_1(\lambda(1 + yR_1(\lambda))) = \int \frac{1}{\frac{\lambda}{x} + \frac{1}{1+yR_1(\lambda)}} dH(x),$$

and

$$H_2(\lambda) = \{(1 + yR_1(\lambda))^2 + yR_2(\lambda)\} \int \frac{x^2}{(x + \lambda(1 + yR_1(\lambda))^2)} dH(x)$$

$$= \int \frac{1 + \frac{yR_2(\lambda)}{(1+yR_1(\lambda))^2}}{(\frac{\lambda}{x} + \frac{1}{1+yR_1(\lambda)})^2} dH(x).$$

Compared with Lemma 2 of Wang et al. (2015), $H_1(\lambda) = R_1(\lambda)$ and $H_2(\lambda) = R_2(\lambda)$. The convergence of misclassification rate is a direct conclusion of Theorem 3.1. The proof is completed. □

References

ANDERSON, T. (2003). *An introduction to multivariate statistical analysis.* Wiley Series in Probability and Statistics. MR1990662

AOSHIMA, M. and YATA, K. (2014). A distance-based, misclassification rate adjusted classifier for multiclass, high-dimensional data. *Annals of the Institute of Statistical Mathematics* **66** 983–1010. MR3250825

BAI, Z., LIU, H. and WONG, W.-K. (2009). Enhancement of the applicability of Markowitz's portfolio optimization by utilizing random matrix theory. *Mathematical Finance* **19** 639–667. MR2583523

BAI, Z., LIU, H. and WONG, W. (2011). Asymptotic properties of eigenmatrices of a large sample covariance matrix. *The Annals of Applied Probability* **21** 1994–2015. MR2884057

BAI, Z., MIAO, B. and PAN, G. (2007). On asymptotics of eigenvectors of large sample covariance matrix. *Annals of Probability* **35** 1532–1572. MR2330979

BAI, Z. and SARANADASA, H. (1996). Effect of high dimension: by an example of a two sample problem. *Statistica Sinica* **2** 311–329. MR1399305

BAI, Z. and SILVERSTEIN, J. W. (2010). *Spectral analysis of large dimensional random matrices.* Springer. MR2567175

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. MR2108040

BÜHLMANN, P. (2013). Statistical significance in high-dimensional linear models. *Bernoulli* **19** 1212–1242. MR3102549

CAI, T. and LIU, W. (2011). A direct estimation approach to sparse linear discriminant analysis. *Journal of the American Statistical Association* **106** 1566–1577. MR2896857

CAI, T., LIU, W. and LUO, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association* **106** 594–607. MR2847973

CAI, T., LIU, W. and XIA, Y. (2014). Two-sample test of high dimensional means under dependence. *Journal of the Royal Statistical Society, Series B* **76** 349–372. MR3164870

CHAN, Y.-B. and HALL, P. (2009). Scale adjustments for classifiers in high-dimensional, low sample size settings. *Biometrika* **96** 469–478. MR2507156

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**. MR2896840

CHENG, Y. (2004). Asymptotic probabilities of misclassification of two discriminant functions in cases of high dimensional data. *Statistics & Probability Letters* **67** 9–17. MR2039928

COLLINS, B. and ŚNIADY, P. (2006). Integration with respect to the Haar measure on unitary, orthogonal and symplectic group. *Communications in Mathematical Physics* **264** 773–795. MR2217291

COOK, R. D. and FORZANI, L. (2011). On the mean and variance of the generalized inverse of a singular Wishart matrix. *Electronic Journal of Statistics* **5** 146–158. MR2786485

DOBRIBAN, E. and WAGER, S. (2018). High-dimensional asymptotics of prediction: Ridge regression and classification. *Annals of Statistics* **46** 247–279. MR3766952

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. MR1963389

EL KAROUI, N. (2008). Spectrum estimation for large dimensional covariance matrices using random matrix theory. *Annals of Statistics* **36** 2757–2790. MR2485012

EL KAROUI, N. (2010). High-dimensionality effects in the Markowitz problem and other quadratic programs with linear constraints: Risk underestimation. *Annals of Statistics* **38** 3487–3566. MR2766860

EL KAROUI, N. and HOLGER, K. (2011). Geometric sensitivity of random matrix results: consequences for shrinkage estimators of covariance and related statistical methods. *arXiv:1105.1404*.

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* **74** 745–771. MR2965958

FRIEDMAN, J. H. (1989). Regularized discriminant analysis. *Journal of the American Statistical Association* **84** 165–175. MR0999675

GUO, Y., HASTIE, T. and TIBSHIRANI, R. (2007). Regularized linear discriminant analysis and its application in microarrays. *Biostatistics* **8** 86–100.

HAND, D. (2006). Classifier technology and the illusion of progress. *Statistical Science* **21** 1–14. MR2275965

HUANG, S., TONG, T. and ZHAO, H. (2010). Bias-Corrected Diagonal Discriminant Rules for High-Dimensional Classification. *Biometrics* **66** 1096–1106. MR2758497

JIANG, T. and YANG, F. (2013). Central limit theorems for classical likelihood ratio tests for high-dimensional normal distributions. *Annals of Statistics* **41** 2029–2074. MR3127857

KUBOKAWA, T. and SRIVASTAVA, M. S. (2008). Estimation of the precision matrix of a singular Wishart distribution and its application in high-dimensional data. *Journal of Multivariate Analysis* **99** 1906–1928. MR2466543

LEDOIT, O. and PÉCHÉ, S. (2011). Eigenvectors of some large sample covariance matrix ensembles. *Probability Theory and Related Fields* **151** 233–264. MR2834718

LEDOIT, O. and WOLF, M. (2004). Honey, I shrunk the sample covariance matrix. *The Journal of Portfolio Management* **30** 110–119.

LI, Z. and YAO, J. (2016). On two simple and effective procedures for high dimensional classification of general populations. *Statistical Papers* **57** 381–405. MR3483195

MAI, Q., ZOU, H. and YUAN, M. (2012). A direct approach to sparse discriminant analysis in ultra-high dimensions. *Biometrika* **99** 29–42. MR2899661

MARČENKO, V. A. and PASTUR, L. A. (1967). Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik* **1** 457.

MATSUMOTO, S. (2012). General moments of the inverse real Wishart distribution and orthogonal Weingarten functions. *Journal of Theoretical Probability* **25** 798–822. MR2956213

MORAN, M. and MURPHY, B. (1979). A closer look at two alternative methods of statistical discrimination. *Applied Statistics* **3** 223–232. MR0552709

PAN, G. and ZHOU, W. (2011). Central limit theorem for Hotelling's T^2 statistic under large dimension. *The Annals of Applied Probability* 1860–1910. MR2884053

SARANADASA, H. (1993). Asymptotic expansion of the misclassification probabilities of D-and A-criteria for discrimination from two high dimensional populations using the theory of large dimensional random matrices. *Journal of Multivariate Analysis* **46** 154–174. MR1231251

SHAO, J., WANG, Y., DENG, X. and WANG, S. (2011). Sparse linear discriminant analysis by thresholding for high dimensional data. *Annals of Statistics* **39** 1241–1265. MR2816353

VON ROSEN, D. (1988). Moments for the inverted Wishart distribution. *Scandinavian Journal of Statistics* 97–109. MR0968156

WANG, X. and LENG, C. (2016). High dimensional ordinary least squares projection for screening variables. *Journal of the Royal Statistical Society, Series B* **78** 589–611. MR3506794

WANG, C., PAN, G., TONG, T. and ZHU, L. (2015). Shrinkage estimation of large dimensional precision matrix using random matrix theory. *Statistica Sinica* **25** 993–1008. MR3409734

ZOLLANVARI, A. and DOUGHERTY, E. R. (2013). Application of double asymptotics and random matrix theory in error estimation of regularized linear discriminant analysis. In *Global Conference on Signal and Information Processing (GlobalSIP), 2013 IEEE* 57–59. IEEE.

ZOLLANVARI, A. and DOUGHERTY, E. R. (2015). Generalized consistent error estimator of linear discriminant analysis. *IEEE Transactions on Signal Processing* **63** 2804–2814. MR3345588