The Annals of Statistics
2021, Vol. 49, No. 3, 1537–1568
https://doi.org/10.1214/20-AOS2012
© Institute of Mathematical Statistics, 2021

A CONVEX OPTIMIZATION APPROACH TO HIGH-DIMENSIONAL SPARSE QUADRATIC DISCRIMINANT ANALYSIS

BY T. TONY CAI[1] AND LINJUN ZHANG[2]

[1]*Department of Statistics, The Wharton School, University of Pennsylvania, tcai@wharton.upenn.edu*
[2]*Department of Statistics, Rutgers University, linjun.zhang@rutgers.edu*

In this paper, we study high-dimensional sparse Quadratic Discriminant Analysis (QDA) and aim to establish the optimal convergence rates for the classification error. Minimax lower bounds are established to demonstrate the necessity of structural assumptions such as sparsity conditions on the discriminating direction and differential graph for the possible construction of consistent high-dimensional QDA rules.

We then propose a classification algorithm called SDAR using constrained convex optimization under the sparsity assumptions. Both minimax upper and lower bounds are obtained and this classification rule is shown to be simultaneously rate optimal over a collection of parameter spaces, up to a logarithmic factor. Simulation studies demonstrate that SDAR performs well numerically. The algorithm is also illustrated through an analysis of prostate cancer data and colon tissue data. The methodology and theory developed for high-dimensional QDA for two groups in the Gaussian setting are also extended to multigroup classification and to classification under the Gaussian copula model.

1. Introduction. Gaussian distribution plays a fundamental role in statistics and machine learning. Discriminant analysis, which aims to classify two multivariate Gaussian distributions, has a wide range of contemporary applications including face recognition [27, 43], business forecasting [15, 24] and gene expression analysis [26, 29, 30]. It also yields fruitful theoretical results in the new era of machine learning, such as understanding the interpolation of neural network training [17, 28, 37] and investigating adversarial robustness [14, 16, 39]. In the ideal setting of two known normal distributions $N_p(\boldsymbol{\mu}_1, \Sigma_1)$ (class 1) and $N_p(\boldsymbol{\mu}_2, \Sigma_2)$ (class 2), the goal of the discriminant analysis is to classify a new observation z, which is drawn from one of the two distributions with prior probabilities π_1 and π_2, respectively, into one of the two classes. In the ideal setting where all the parameters $\boldsymbol{\theta} = (\pi_1, \pi_2, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma_1, \Sigma_2)$ are known, the optimal classifier is the quadratic discriminant rule is given by

$$G^*_{\boldsymbol{\theta}}(z) = \begin{cases} 1, & (z-\boldsymbol{\mu}_1)^\top D(z-\boldsymbol{\mu}_1) - 2\boldsymbol{\delta}^\top \Omega_2(z-\bar{\boldsymbol{\mu}}) - \log\left(\frac{|\Sigma_1|}{|\Sigma_2|}\right) \\ & \quad + 2\log\left(\frac{\pi_1}{\pi_2}\right) > 0, \\ 2, & (z-\boldsymbol{\mu}_1)^\top D(z-\boldsymbol{\mu}_1) - 2\boldsymbol{\delta}^\top \Omega_2(z-\bar{\boldsymbol{\mu}}) - \log\left(\frac{|\Sigma_1|}{|\Sigma_2|}\right) \\ & \quad + 2\log\left(\frac{\pi_1}{\pi_2}\right) \le 0, \end{cases} \tag{1.1}$$

where $\boldsymbol{\delta} = \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$, $\bar{\boldsymbol{\mu}} = \frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2}$, and $D = \Omega_2 - \Omega_1$ with $\Omega_i = \Sigma_i^{-1}$ for $i = 1, 2$; see, for example, Anderson [2]. When $\Sigma_1 = \Sigma_2$, the quadratic classification boundary in (1.1) becomes

Received September 2019; revised August 2020.
MSC2020 subject classifications. Primary 62H30; secondary 62C20, 62H12.
Key words and phrases. Classification, constrained ℓ_1 minimization, high-dimensional data, minimax lower bound, optimal rate of convergence, quadratic discriminant analysis.

linear, reducing the quadratic discriminant analysis (QDA) to the linear discriminant analysis (LDA).

QDA has been an important technique for classification and is more flexible than the LDA [22]. In practice, the parameters $\pi_1, \pi_2, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma_1$ and Σ_2 are usually unknown and instead one observes two independent random samples, $\boldsymbol{X}_1^{(1)}, \dots, \boldsymbol{X}_{n_1}^{(1)} \overset{\text{i.i.d.}}{\sim} N_p(\boldsymbol{\mu}_1, \Sigma_1)$ and $\boldsymbol{X}_1^{(2)}, \dots, \boldsymbol{X}_{n_2}^{(2)} \overset{\text{i.i.d.}}{\sim} N_p(\boldsymbol{\mu}_2, \Sigma_2)$. It is practically important to construct a data-driven classification rule based on the two samples. In the low-dimensional setting where the dimension p is small relative to the sample sizes, a natural approach is to simply plug the sample means and sample covariance matrices into the oracle QDA rule (1.1). This approach has been well studied; see, for example, Anderson [2]. Thanks to the explosive growth of big data, high-dimensional data, where the dimension p can be much larger than the sample sizes, are now routinely collected in scientific investigations in a wide range of fields. In such settings, the conventional LDA and QDA rules perform poorly.

For high-dimensional LDA, there already exist a number of proposals and theoretical studies. In particular, assuming sparsity on the discriminating direction, direct estimation methods have been introduced in Cai and Liu [7] and Mai, Zou and Yuan [35] and optimality theory is developed in Cai and Zhang [11]. In contrast, relatively few methods have been introduced for regularized QDA in the high-dimensional setting and developing an optimality theory is technically more challenging. Li and Shao [31] studied high-dimensional QDA by imposing sparsity assumptions on $\boldsymbol{\delta}$, Σ_1, Σ_2 and $\Sigma_1 - \Sigma_2$ separately, and then plugging the estimates of these quantities into the oracle QDA rule (1.1). Jiang, Wang and Leng [25] introduced a direct estimation approach by assuming that $\Omega_1 - \Omega_2$ and $(\Omega_1 + \Omega_2)\boldsymbol{\delta}$ are sparse, and proposed a consistent classification rule. However, it is still unclear whether any of these methods achieves the optimal convergence rate for the classification error.

In the present paper, by observing that the oracle rule (1.1) depends on $\boldsymbol{\theta}$ only through the discriminating direction $\boldsymbol{\beta} = \Omega_2\boldsymbol{\delta}$ and differential graph $D = \Omega_2 - \Omega_1$, we propose a sparse QDA rule by directly estimating D and $\boldsymbol{\beta}$ through convex optimization, and aim to establish the optimality of the proposed classifier in the high-dimensional settings. It is intuitively clear that QDA is a difficult problem in the high-dimensional setting. For example, it can be seen easily from (1.1) that knowledge of the log-determinant of the covariance matrices $\log(\frac{|\Sigma_1|}{|\Sigma_2|})$ is essential for the QDA. However, as shown in Cai, Liang and Zhou [8], there is no consistent estimator for the log-determinant of the covariance matrices in the high-dimensional setting even when they are known to be diagonal. We begin by establishing rigorously minimax lower bound results, which demonstrate that structural assumptions such as sparsity conditions on the discriminating direction $\boldsymbol{\beta}$ and differential graph D are necessary for the possible construction of consistent high-dimensional QDA rules. There are two key steps in obtaining the impossibility results: One is the reduction of the classification error to an alternative loss and another is a careful construction of a collection of least favorable multivariate normal distributions.

We then propose a classifier called SDAR (*S*parse *D*iscriminant *A*nalysis with *R*egularization) to solve the high-dimensional QDA problem under the sparsity assumptions. The SDAR algorithm proceeds by first estimating $\boldsymbol{\beta}$ and D through constrained convex optimization, and then using the estimators to construct a data-driven classification rule. The first estimation step is in a similar spirit to that in Jiang, Wang and Leng [25] by directly estimating the key quantities in the oracle QDA rule. The second classification step is based on a simple but important observation that $\log(|\Sigma_1|/|\Sigma_2|) = \log(|D\Sigma_1 + I_p|)$. As a result, we are able to derive an explicit convergence rate for the classification error of the proposed SDAR algorithm. In addition, we establish a matching minimax lower bound, up to a logarithm factor, that shows the near-optimality of the classifier. Both simulations and real data analysis are carried out to study the numerical performance of the proposed algorithm. The results show that the proposed SDAR algorithm outperforms existing methods in the literature. The methodology and theory developed for high-dimensional QDA for two groups in the Gaussian setting are also extended to multi-group classification and to classification under the Gaussian copula model.

The contributions of the present paper are three-fold. First, we address the necessity of structural assumptions on the parameters for the high-dimensional QDA problem by observing that consistent classification is impossible unless $p = o(n)$ without any such assumptions. Second, under the sparsity assumptions, we proposed the SDAR rule, and established an explicit convergence rate of classification error. To the best of our knowledge, this is the first explicit convergence rate for high-dimensional QDA. Lastly, we provide a minimax lower bound, which shows that the convergence rate obtained by the SDAR rule is optimal, up to a logarithmic factor.

The rest of the paper is organized as follows. In Section 2, minimax lower bounds are established to show the necessity of imposing structural assumptions for high-dimensional QDA. Section 3 presents in detail the data-driven classification procedure SDAR. Theoretical properties of SDAR are investigated in Section 4 under certain sparsity conditions. The upper and lower bounds together show that the SDAR rule achieves the optimal rate for the classification error up to a logarithmic factor. In Section 5, we consider the semiparametric copula model and introduced a new method called Copula SDAR (CSDAR) and developed corresponding theoretical results for this non-Gaussian model. Simulation studies are given in Section 6 where we compare the performance of the proposed algorithms to other existing classification methods in the literature. In addition, the merits of the SDAR and SDAR classifiers are illustrated through an analysis of a prostate cancer dataset and a colon tissue dataset. Section 7 discusses extensions to multigroup classification and to classification under the Gaussian copula model. The proofs of main results are given in Section 8, and proofs of other results are provided in the Supplementary Material [12].

Notation and definitions. We first introduce basic notation and definitions that will be used throughout the rest of the paper. For an event A, $\mathbb{1}\{A\}$ is the indicator function on A. For an integer $m \geq 1$, $[m]$ denotes the set $\{1, 2, \dots, m\}$. Throughout the paper, vectors are denoted by boldface letters. For a vector $\boldsymbol{u}$, $\|\boldsymbol{u}\|$, $\|\boldsymbol{u}\|_1$, $\|\boldsymbol{u}\|_\infty$ denotes the ℓ_2 norm, ℓ_1 norm, and ℓ_∞ norm, respectively. We use $\operatorname{supp}(\boldsymbol{u})$ to denote the support of the vector $\boldsymbol{u}$. $\mathbf{0}_p$ is a p-dimensional vector with elements being 0, and $\mathbf{1}_p$ is a p-dimensional vector with elements being 1. For $i \in [p]$, $\boldsymbol{e}_i$ is the ith standard basis. For a matrix $M \in \mathbb{R}^{p\times p}$, $\|M\|$, $\|M\|_F$, $\|M\|_1$ denote the spectral norm, Frobenius norm and matrix l_1 norm, respectively. In addition, $|M|_1 = \sum_{i,j} |M_{i,j}|$, $|M|_\infty = \max_{i,j} |M_{i,j}|$ and $|M|$ is the determinant of M. Let $\lambda_i(M)$ denote the ith eigenvalue of M with $\lambda_1(M) \geq \cdots \geq \lambda_p(M)$. Let $M \succ 0$ denote M to be a positive semidefinite matrix and I_p is the $p \times p$ identity matrix. In addition, $M_1 \otimes M_2$ denotes the Kronecker product and $\operatorname{vec}(M)$ is the $p^2 \times 1$ vector obtained by stacking the columns of M. $\operatorname{diag}(M)$ is the linear operator that sets all the off diagonal elements of M to 0. $E_{i,i}$ is a $p \times p$ matrix whose (i,i)th entry is 1 and 0 else. For a positive integer $s < p$, let $\Gamma(s; p) = \{\boldsymbol{u} \in \mathbb{R}^p : \|\boldsymbol{u}_{S^C}\|_1 \leq \|\boldsymbol{u}_S\|_1,\ \text{for some } S \subset [p] \text{ with } |S| = s\}$, where $\boldsymbol{u}_S$ denotes the subvector of $\boldsymbol{u}$ confined to S. For two sequences of positive numbers a_n and b_n, $a_n \lesssim b_n$ means that for some constant $c > 0$, $a_n \leq c \cdot b_n$ for all n, and $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$. $a_n \ll b_n$ means that $\lim_{n\to\infty} |a_n|/|b_n| = 0$. In our asymptotic framework, we let n be the driving asymptotic parameter, s and p approach infinity as n grows to infinity. We also use $c, c_1, c_2, \dots, C, C_1, C_2$ to denote constants that does not depend on n, p, and their values may vary from place to place.

2. The difficulties of high-dimensional QDA. As mentioned in the Introduction, high-dimensional QDA is a difficult problem. In this section, we establish explicit minimax lower bounds that show the necessity of structural assumptions on the discriminating direction $\boldsymbol{\beta} = \Omega_2\boldsymbol{\delta}$ and differential graph $D = \Omega_2 - \Omega_1$ for constructing consistent high-dimensional QDA rules.

2.1. *The setup.* Suppose one observes n_1 samples from class 1: $\boldsymbol{x}_1, \dots, \boldsymbol{x}_{n_1} \overset{\text{i.i.d.}}{\sim} N_p(\boldsymbol{\mu}_1, \Sigma_1)$, n_2 samples from class 2: $\boldsymbol{y}_1, \dots, \boldsymbol{y}_{n_2} \overset{\text{i.i.d.}}{\sim} N_p(\boldsymbol{\mu}_2, \Sigma_2)$, and the prior probabilities of class 1 and 2 are π_1 and π_2, respectively. The goal is to construct a classification rule $\hat{G}$, which is a function of $\boldsymbol{x}_i$'s and $\boldsymbol{y}_i$'s, to classify a future data point $\boldsymbol{z} \sim \pi_1 N_p(\boldsymbol{\mu}_1, \Sigma_1) + \pi_2 N_p(\boldsymbol{\mu}_2, \Sigma_2)$. This model is parametrized by $\boldsymbol{\theta} = (\pi_1, \pi_2, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma_1, \Sigma_2)$. Let $n = \min\{n_1, n_2\}$. For any classification rule $\hat{G} : \mathbb{R}^p \to \{1, 2\}$, the accuracy is measured by the classification error

$$R_{\boldsymbol{\theta}}(\hat{G}) = \mathbb{E}_{\boldsymbol{\theta}}\big[\mathbb{1}\{\hat{G}(\boldsymbol{z}) \neq L(\boldsymbol{z})\}\big], \tag{2.1}$$

where $L(\boldsymbol{z})$ denotes the true class label of $\boldsymbol{z}$, that is, $L(\boldsymbol{z}) = 1$ if $\boldsymbol{z} \sim N_p(\boldsymbol{\mu}_1, \Sigma_1)$, and 2 otherwise.

When $\boldsymbol{\theta} = (\pi_1, \pi_2, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma_1, \Sigma_2)$ is known in advance, the oracle classification rule in (1.1) is the Bayes rule and achieves the the minimal classification error; see Anderson [2]. For ease of presentation, let us define the discriminant function by

$$Q(\boldsymbol{z}; \boldsymbol{\theta}) = (\boldsymbol{z} - \boldsymbol{\mu}_1)^\top D(\boldsymbol{z} - \boldsymbol{\mu}_1) - 2\boldsymbol{\delta}^\top \Omega_2(\boldsymbol{z} - \bar{\boldsymbol{\mu}}) - \log\left(\frac{|\Sigma_1|}{|\Sigma_2|}\right) + 2\log\left(\frac{\pi_1}{\pi_2}\right). \tag{2.2}$$

Then $Q(\boldsymbol{z}; \boldsymbol{\theta}) = 0$ characterizes the classification boundary of the oracle QDA rule, and (1.1) can be rewritten as

$$G^*_{\boldsymbol{\theta}}(\boldsymbol{z}) = 1 + \mathbb{1}\{Q(\boldsymbol{z}; \boldsymbol{\theta}) \le 0\},$$

and $R_{\boldsymbol{\theta}}(G^*_{\boldsymbol{\theta}}) = \min_{G \in \mathcal{G}} R_{\boldsymbol{\theta}}(G)$, where $\mathcal{G}$ is the set of all classification rules.

In the following, the Bayes classification risk $R_{\boldsymbol{\theta}}(G^*_{\boldsymbol{\theta}})$ is used as the benchmark and the excess risk $R_{\boldsymbol{\theta}}(\hat{G}) - R_{\boldsymbol{\theta}}(G^*_{\boldsymbol{\theta}})$ is used to evaluate the performance of a data-driven classification rule $\hat{G}$. We say $\hat{G}$ is consistent, or $G^*_{\boldsymbol{\theta}}$ can be mimicked by $\hat{G}$, if the excess risk $R_{\boldsymbol{\theta}}(\hat{G}) - R_{\boldsymbol{\theta}}(G^*_{\boldsymbol{\theta}}) \to 0$ as the sample size $n \to \infty$.

2.2. *Impossibility of QDA in high dimensions.* We now characterize the fundamental limits of QDA by showing that, without structural assumptions, $G^*_{\boldsymbol{\theta}}$ cannot be mimicked unless $p \ll n$, which precludes the framework in the high-dimensional settings that motivates our study.

We first consider the simple case where $\Sigma_1 = \Sigma_2 = \Sigma$, and in which case the QDA is reduced to the LDA problem. Under the LDA model in the high-dimensional regime, Bickel and Levina [6] and Cai, Zhang and Ma [9] proposed consistent classification rules under stringent structural conditions on $(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma)$. In this paper, we demonstrate the the necessity of these structural assumptions by showing that without structural assumptions, a consistent classification rule is impossible in the high-dimensional LDA problem.

We first consider the parameter space

$$\Theta_p^{(1)} = \{\boldsymbol{\theta} = (1/2, 1/2, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, I_p, I_p) : \boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in \mathbb{R}^p, c_1 \le \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\| \le c_2\},$$

for some constant $c_1, c_2 > 0$.

THEOREM 2.1. *Suppose that $\hat{G}$ is any classification rule constructed based on the observations* $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n \stackrel{\text{i.i.d.}}{\sim} N_p(\boldsymbol{\mu}_1, I_p)$, $\boldsymbol{y}_1, \dots, \boldsymbol{y}_n \stackrel{\text{i.i.d.}}{\sim} N_p(\boldsymbol{\mu}_2, I_p)$ *with* $\boldsymbol{\theta} = (1/2, 1/2, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, I_p, I_p) \in \Theta_p^{(1)}$, *then when n is sufficiently large*,

$$\inf_{\hat{G}} \sup_{\boldsymbol{\theta} \in \Theta_p^{(1)}} \mathbb{E}\big[R_{\boldsymbol{\theta}}(\hat{G}) - R_{\boldsymbol{\theta}}\big(G_{\boldsymbol{\theta}}^*\big)\big] \gtrsim \frac{p}{n} \wedge 1.$$

This theorem implies that even when the covariance matrices are equal and known to be identity matrices, as long as the mean vectors $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ are unknown, no data-driven method is able to mimic $G_{\boldsymbol{\theta}}^*$ in the high-dimensional setting where $p \gtrsim n$. Structural assumptions are $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are necessary for a consistent classification rule.

However, for high-dimensional QDA, structural assumptions on $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are not enough and more assumptions are needed. To this end, we consider another scenario where $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are known exactly. Let $\boldsymbol{\mu}_1^*, \boldsymbol{\mu}_2^* \in \mathbb{R}^p$ be two given vectors and define the parameter space

$$\Theta_p^{(2)}(\boldsymbol{\mu}_1^*, \boldsymbol{\mu}_2^*) = \{\boldsymbol{\theta} = \big(1/2, 1/2, \boldsymbol{\mu}_1^*, \boldsymbol{\mu}_2^*, \Sigma_1, \Sigma_2\big) : \Sigma_1, \Sigma_2 \text{ are diagonal matrices}\}.$$

THEOREM 2.2. *Suppose $\hat{G}$ is constructed based on the observations* $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n \stackrel{\text{i.i.d.}}{\sim} N_p(\boldsymbol{\mu}_1, \Sigma_1)$, $\boldsymbol{y}_1, \dots, \boldsymbol{y}_n \stackrel{\text{i.i.d.}}{\sim} N_p(\boldsymbol{\mu}_2, \Sigma_2)$. *For any given* $\boldsymbol{\mu}_1^*, \boldsymbol{\mu}_2^* \in \mathbb{R}^p$ *with* $\|\boldsymbol{\mu}_1^* - \boldsymbol{\mu}_2^*\|_2 \le C$ *where $C > 0$ is some constant, when* $\boldsymbol{\theta} = (1/2, 1/2, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma_1, \Sigma_2) \in \Theta_p^{(2)}(\boldsymbol{\mu}_1^*, \boldsymbol{\mu}_2^*)$, *we have for sufficiently large n*,

$$\inf_{\hat{G}} \sup_{\boldsymbol{\theta} \in \Theta_p^{(2)}(\boldsymbol{\mu}_1^*, \boldsymbol{\mu}_2^*)} \mathbb{E}\big[R_{\boldsymbol{\theta}}(\hat{G}) - R_{\boldsymbol{\theta}}\big(G_{\boldsymbol{\theta}}^*\big)\big] \gtrsim \frac{p}{n} \wedge 1.$$

This theorem implies that even if we have the prior information that $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ are known and Σ_1, Σ_2 are both diagonal, the quadratic discriminant rule $G_{\boldsymbol{\theta}}^*$ cannot be mimicked consistently if $p \gtrsim n$. The construction of consistent classification rules requires stronger assumptions.

The main strategy of these proofs are discussed in Section 4.2, and the detailed proofs of these lower bound results are provided in Section 8.1. In addition, the lower bounds are tight, up to a logarithmic factor. Specifically, by using the techniques similar to that in Theorem 4.2, the plug-in classification rule $\hat{G}$, which is obtained by plugging in sample means and sample covariance matrices in (1.1), satisfies that $R_{\boldsymbol{\theta}}(\hat{G}) - R_{\boldsymbol{\theta}}(G_{\boldsymbol{\theta}}^*) \lesssim \frac{p \log^2 n}{n} \wedge 1$. This result is further discussed in the Supplementary Material.

3. Sparse quadratic discriminant analysis. The inconsistency results in Theorems 2.1 and 2.2 imply the necessity of imposing structural assumptions on both the mean vectors and covariance matrices. In this section, we consider the QDA problem under the assumptions that the discriminating direction $\boldsymbol{\beta} = \Omega_2 \boldsymbol{\delta}$ and the differential graph D are both sparse. This sparsity assumption, according to (2.2), implies that the classification boundary of the oracle rule depends only on a small number of features in $\boldsymbol{z}$. It is also worth noting that the differential graph D corresponds to the change of interactions in two different graphs Ω_1 and Ω_2. The problem of interaction selection is important in its own right and has been studied extensively recently in dynamic network analysis under various environmental and experimental conditions; see Bandyopadhyay et al. [4], Zhao, Cai and Li [47], Xia, Cai and Cai [44], Hill et al. [23].

To see that these two sparsity assumptions are sufficient to obtain a consistent estimator for the optimal classification rule $G_{\boldsymbol{\theta}}^*$, we begin by rewriting $Q(\boldsymbol{z}; \boldsymbol{\theta})$, defined in (2.2). Recall that $\boldsymbol{\delta} = \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$, $\bar{\boldsymbol{\mu}} = \frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2}$, $D = \Omega_2 - \Omega_1$ and $\boldsymbol{\beta} = \Omega_2\boldsymbol{\delta}$, then

$$\begin{aligned} Q(z;\boldsymbol{\theta}) &= (z-\boldsymbol{\mu}_1)^\top D(z-\boldsymbol{\mu}_1) - 2\boldsymbol{\beta}^\top(z-\bar{\boldsymbol{\mu}}) - \log\left(\frac{|\Sigma_1|}{|\Sigma_2|}\right) + 2\log\left(\frac{\pi_1}{\pi_2}\right) \\ &= (z-\boldsymbol{\mu}_1)^\top D(z-\boldsymbol{\mu}_1) - 2\boldsymbol{\beta}^\top(z-\bar{\boldsymbol{\mu}}) - \log(|D\Sigma_1 + I_p|) + 2\log\left(\frac{\pi_1}{\pi_2}\right). \end{aligned} \tag{3.1}$$

A simple but essential observation of (3.1) is that the first three quantities in the above oracle QDA rule $G^*_{\boldsymbol{\theta}}$ depends on either D or $\boldsymbol{\beta}$, and the forth term $\log(\pi_1/\pi_2)$ is easy to estimate. In the present paper, we shall show that under the sparsity assumptions on these two quantities, D and $\boldsymbol{\beta}$ can be estimated directly and efficiently, and the classification rule based on these two estimates enjoys desirable theoretical guarantees.

REMARK 1. By symmetry, $Q(z;\boldsymbol{\theta})$ can also be rewritten in a form that depends on $(\Omega_1+\Omega_2)\boldsymbol{\delta}$ and D. The reason that we consider $(\Omega_2\boldsymbol{\delta}, D)$ as the key quantity is that this could be easily extended to the case with K multiple groups. In this generalized setting, we consider using the first group as a benchmark, and computing the likelihood ratio of other groups versus the first one. As a result, the key quantity in the multiple classification case is $\{(\Omega_k(\boldsymbol{\mu}_k - \boldsymbol{\mu}_1), \Omega_k - \Omega_1)]\}_{k=2}^K$. See more discussion in Section 7.

In the following, we proceed to estimate D and $\boldsymbol{\beta}$ through constrained convex optimization. Let the first sample covariance matrix be $\hat{\Sigma}_1 = n_1^{-1}\sum_{i=1}^{n_1}(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_1)(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_1)^\top$, where $\hat{\boldsymbol{\mu}}_1 = n_1^{-1}\sum_{i=1}^{n_1}\boldsymbol{x}_i$ and define $\hat{\Sigma}_2$ and $\hat{\boldsymbol{\mu}}_2$ similarly. Since D satisfies the equation $\Sigma_1 D\Sigma_2 = \Sigma_1 - \Sigma_2$ and $\Sigma_2 D\Sigma_1 = \Sigma_1 - \Sigma_2$, a sensible estimation procedure is to solve $\hat{\Sigma}_1 D\hat{\Sigma}_2/2 + \hat{\Sigma}_2 D\hat{\Sigma}_1/2 - \hat{\Sigma}_1 + \hat{\Sigma}_2 = 0$ for D. We estimate D through the following constrained ℓ_1 minimization approach:

$$\hat{D} = \arg\min_{D\in\mathbb{R}^{p\times p}}\left\{|D|_1 : \left|\frac{1}{2}\hat{\Sigma}_1 D\hat{\Sigma}_2 + \frac{1}{2}\hat{\Sigma}_2 D\hat{\Sigma}_1 - \hat{\Sigma}_1 + \hat{\Sigma}_2\right|_\infty \le \lambda_{1,n}\right\}, \tag{3.2}$$

where $\lambda_{1,n} = c_1\sqrt{\frac{\log p}{n}}$ is a tuning parameter with some constant $c_1 > 0$ that will be specified later.

REMARK 2. The estimator $\hat{D}$ defined in (3.2) is similar to that in Zhao, Cai and Li [47], but has better numerical performance due to symmetrization. In addition, we are able to solve (3.2) in a more computationally efficient way. Zhao, Cai and Li [47] vectorized D and transformed the optimization problem (3.2) to a linear programming with a $p^2 \times p^2$ constraint matrix $\hat{\Sigma}_1 \otimes \hat{\Sigma}_2$, which is computationally demanding for large p. In contrast, we solve (3.2) by using the primal-dual interior point method [13], and keep the matrix form of D in each step of conjugate gradient descent, by using the matrix multiplications $\frac{1}{2}\hat{\Sigma}_1 D\hat{\Sigma}_2 + \frac{1}{2}\hat{\Sigma}_2 D\hat{\Sigma}_1$ instead of computing $(\frac{1}{2}\hat{\Sigma}_1 \otimes \hat{\Sigma}_2 + \frac{1}{2}\hat{\Sigma}_2 \otimes \hat{\Sigma}_1)\operatorname{vec}(D)$ repeatedly. As a result, the computational complexity is reduced to $O(p^3)$ from $O(p^4)$, and our method is able to handle the problem with larger dimension p. The code is available at https://github.com/linjunz/SDAR.

We then proceed to estimating $\boldsymbol{\beta}$. Similarly, since the true $\boldsymbol{\beta}$ satisfies that $\Sigma_2\boldsymbol{\beta} = \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$, following Cai and Liu [7], $\boldsymbol{\beta}$ can be estimated by the following procedure:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p}\{\|\boldsymbol{\beta}\|_1 : \|\hat{\Sigma}_2\boldsymbol{\beta} - \hat{\boldsymbol{\mu}}_2 + \hat{\boldsymbol{\mu}}_1\|_\infty \le \lambda_{2,n}\}, \tag{3.3}$$

where $\lambda_{2,n} = c_2\sqrt{\frac{\log p}{n}}$ is a tuning parameter with some constant $c_2 > 0$.

We estimate π_1 and π_2 by $\hat{\pi}_1 = \frac{n_1}{n_1+n_2}$ and $\hat{\pi}_2 = \frac{n_2}{n_1+n_2}$, respectively. Given the solutions $\hat{D}$ and $\hat{\boldsymbol{\beta}}$ to (3.2) and (3.3) and the estimates $\hat{\pi}_1$ and $\hat{\pi}_2$, we then propose the following classification rule: classify z to class 1 if and and only if

$$(z - \hat{\boldsymbol{\mu}}_1)^\top \hat{D}(z - \hat{\boldsymbol{\mu}}_1) - 2\hat{\boldsymbol{\beta}}^\top \left(z - \frac{\hat{\boldsymbol{\mu}}_1 + \hat{\boldsymbol{\mu}}_2}{2}\right) - \log(|\hat{D}\hat{\Sigma}_1 + I_p|) + \log\left(\frac{\hat{\pi}_1}{\hat{\pi}_2}\right) > 0.$$

We shall call this rule the Sparse quadratic Discriminant Analysis rule with Regularization (SDAR), and denote it by $\hat{G}_{\mathrm{SDAR}}$. Analytically, it is written as

$$\begin{aligned} &\hat{G}_{\mathrm{SDAR}}(z) \\ &= 1 + \mathbb{1}\left\{(z - \hat{\boldsymbol{\mu}}_1)^\top \hat{D}(z - \hat{\boldsymbol{\mu}}_1) - 2\hat{\boldsymbol{\beta}}^\top \left(z - \frac{\hat{\boldsymbol{\mu}}_1 + \hat{\boldsymbol{\mu}}_2}{2}\right)\right. \\ &\qquad \left. - \log(|\hat{D}\hat{\Sigma}_1 + I_p|) + \log\left(\frac{\hat{\pi}_1}{\hat{\pi}_2}\right) \le 0\right\}. \end{aligned} \tag{3.4}$$

The SDAR rule is easy to implement as both (3.2) and (3.3) can be solved by linear programming. We shall show in the next sections that the SDAR rule has desirable properties both theoretically and numerically.

4. Theoretical guarantees. We now study the accuracy of the estimators $\hat{D}$ and $\hat{\boldsymbol{\beta}}$ in (3.2) and (3.3), and the performance of the resulting classifier $\hat{G}_{\mathrm{SDAR}}$ in (3.4). We first establish the rates of convergence for the estimation and classification error and then provide matching minimax lower bounds, up to logarithm factors. These results together show the near-optimality of the SDAR rule.

4.1. *Upper bounds.* To overcome the limitations illustrated in Section 2, we consider the following parameter space of $\boldsymbol{\theta} = (\pi_1, \pi_2, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma_1, \Sigma_2)$. Especially, we assume here that both the discriminating direction $\boldsymbol{\beta}$ and the differential graph D are sparse. Let $f_{Q,\boldsymbol{\theta}}$ be the probability density of $Q(z; \boldsymbol{\theta})$ defined in (2.2), we consider the following parameter space:

$$\begin{aligned} \Theta_p(s_1, s_2) = \Big\{ & \boldsymbol{\theta} = (\pi_1, \pi_2, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma_1, \Sigma_2) : \boldsymbol{\mu}_1, \\ & \boldsymbol{\mu}_2 \in \mathbb{R}^p, \Sigma_1, \Sigma_2 \succ 0, |D|_0 \le s_1, \|\boldsymbol{\beta}\|_0 \le s_2 \\ & \|D\|_F, \|\boldsymbol{\beta}\|_2 \le M_0, M_1^{-1} \le \lambda_{\min}(\Sigma_k) \le \lambda_{\max}(\Sigma_k) \le M_1, k = 1, 2, \\ & \sup_{|x|<\delta} f_{Q,\boldsymbol{\theta}}(x) < M_2, c \le \pi_1, \pi_2 \le 1 - c \Big\}, \end{aligned} \tag{4.1}$$

for some constants $M_0 > 0$, $M_1 > 1$, $\delta, M_2 > 0$ and $c \in (0, 1/2)$.

REMARK 3. Note that we assume sparsity on both the discriminant direction $\boldsymbol{\beta}$ and the differential graph D, whose necessities are shown by Theorem 2.1 and 2.2. The upper bound on $\|\boldsymbol{\beta}\|_2$ is a general assumption in LDA (see Cai and Liu [7], Neykov et al. [38]; and Cai, Zhang and Ma [9]), and we assume the same on $\|\mathrm{vec}(D)\|_2 = \|D\|_F$ in the QDA setting. Moreover, the condition on the bounded density is commonly assumed in discriminant analysis; see the margin assumption in Mammen and Tsybakov [36], condition (C1) in Cai and Liu [7] and discussions in Li and Shao [31] and Jiang, Wang and Leng [25]. In the following, we present a condition on $\boldsymbol{\theta}$ such that this bounded density assumption holds. Note that the term $z^\top Dz + \boldsymbol{\beta}^\top z$ is equal in distribution to a weighted noncentral chi-square distribution, by using the similar proof as that of Lemma 7.2 in Xu, Zhang and Wu

[45], the condition $\sup_{|x|<\delta} f_{Q,\boldsymbol{\theta}}(x) < M_2$ holds when either the two largest positive eigenvalues of D $\lambda_1(D), \lambda_2(D)$ or the two largest negative eigenvalues of D $\tilde{\lambda}_1(D), \tilde{\lambda}_2(D)$ are of the same order, that is, $0 < \liminf_{n\to\infty} \frac{\lambda_1(D)}{\lambda_1(D)+\lambda_2(D)} < \limsup_{n\to\infty} \frac{\lambda_1(D)}{\lambda_1(D)+\lambda_2(D)} < 1$ or $0 < \liminf_{n\to\infty} \frac{\tilde{\lambda}_1(D)}{\tilde{\lambda}_1(D)+\tilde{\lambda}_2(D)} < \limsup_{n\to\infty} \frac{\tilde{\lambda}_1(D)}{\tilde{\lambda}_1(D)+\tilde{\lambda}_2(D)} < 1$.

At first, we show that over the parameter space $\Theta_p(s_1, s_2)$, the estimators $\hat{D}$, $\hat{\boldsymbol{\beta}}$ obtained in (3.2) and (3.3) converge to the true parameters D and $\boldsymbol{\beta}$. This theorem will then be used to establish the consistency of the proposed classification rule.

THEOREM 4.1. *Consider the parameter space $\Theta_p(s_1, s_2)$, and assume that $n_1 \asymp n_2$, $s_1 + s_2 \lesssim \frac{n}{\log p}$, where $n = \min\{n_1, n_2\}$. In optimization problems* (3.2) *and* (3.3), *let $\lambda_{i,n} = c_i\sqrt{\log p/n}$ with $c_i > 0$, $i = 1, 2$ being sufficiently large constants. Then the estimators obtained in* (3.2) *and* (3.3) *satisfies that, with probability at least $1 - p^{-1}$,*

$$\|\hat{D} - D\|_F \lesssim \sqrt{\frac{s_1 \log p}{n}}; \qquad \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 \lesssim \sqrt{\frac{s_2 \log p}{n}}.$$

The above theorem shows that although our estimating procedure (3.3) is different from Zhao, Cai and Li [47], the same convergence rate can be obtained and requires milder theoretical conditions. In fact, Zhao, Cai and Li [47] assumes that $\|\Omega_1\|_1$ and $\|\Omega_2\|_1$ are both bounded, and additionally requires that the off-diagonal elements of Σ_1 and Σ_2 are vanishing as $n \to \infty$, which is much stronger than conditions in (4.1). In addition, the above bound implies that when $\Sigma_1 = \Sigma_2$, that is, $s_1 = 0$, we have $\hat{D} = D = 0$ when $\lambda_{1,n}$ is suitably chosen. This implies that when the two covariance matrices are equal, SDAR rule (3.4) would adaptively be reduced to the LPD rule in Cai and Liu [7] designed for high-dimensional LDA.

We now turn to the performance of the classification rule $\hat{G}_{\mathrm{SDAR}}$. The behavior of $\hat{G}_{\mathrm{SDAR}}$ is measured by the excess risk $R_{\boldsymbol{\theta}}(\hat{G}_{\mathrm{SDAR}}) - R_{\boldsymbol{\theta}}(G^*_{\boldsymbol{\theta}})$, defined in (2.1). The following theorem provides the upper bound for the excess classification error.

THEOREM 4.2. *Consider the parameter space $\Theta_p(s_1, s_2)$, and assume that $n_1 \asymp n_2$, $s_1 + s_2 \lesssim \frac{n}{\log p \cdot \log^2 n}$. Then the proposed SDAR classification rule in* (3.4) *satisfies that*

$$\sup_{\boldsymbol{\theta}\in\Theta_p(s_1,s_2)} \mathbb{E}\big[R_{\boldsymbol{\theta}}(\hat{G}_{\mathrm{SDAR}}) - R_{\boldsymbol{\theta}}(G^*_{\boldsymbol{\theta}})\big] \lesssim (s_1 + s_2) \cdot \frac{\log p}{n} \cdot \log^2 n.$$

The result in Theorem 4.2 shows that $\hat{G}_{\mathrm{SDAR}}$ is able to mimic $G^*_{\boldsymbol{\theta}}$ consistently over the parameter space $\Theta_p(s_1, s_2)$, and to the best of our knowledge, gives the first explicit convergence rate of classification error for the high-dimensional QDA problem.

REMARK 4. Related work studying the convergence of classification error includes Li and Shao [31] and Jiang, Wang and Leng [25], but both Theorem 3 in Li and Shao [31] and Theorem 4 in Jiang, Wang and Leng [25] only show the consistency of their proposed classification rules instead of explicit convergence rates. Although in Corollary 3 of Jiang, Wang and Leng [25], the authors showed a convergence rate for the classification error of order $s_1 s_2^2 \sqrt{\log p/n}$ under some regularity conditions, this result is based on the assumption that an intercept term η, defined in their paper, is known. Jiang, Wang and Leng [25] proposed to estimate η based on the idea of cross-validation and in their Theorem 3 they showed the consistency of this estimation without explicit convergence rate. In contrast, our paper shows that the convergence rate $O((s_1 + s_2)\log p \cdot \log^2 n/n)$ is achievable, which is much faster than their results. In addition, the assumptions here are weaker.

The major technical challenge of this improvement is the characterization of the distribution of $Q(z;\boldsymbol{\theta})$, which involves the sum of weighted noncentral chi-square random variables. In the next section, we will show that this convergence rate is indeed optimal up to logarithm factors.

4.2. *Minimax lower bound for sparse QDA*. In this section, we establish the minimax lower bound for the convergence rate of $R_{\boldsymbol{\theta}}(\hat{G}) - R_{\boldsymbol{\theta}}(G^*_{\boldsymbol{\theta}})$, and thus show the optimality of $\hat{G}_{\mathrm{SDAR}}$ up to logarithm factors.

THEOREM 4.3. *Consider the parameter space $\Theta_p(s_1, s_2)$ defined in* (4.1). *Suppose $n_1 \asymp n_2$, $1 \le s_1, s_2 \le o(\frac{n}{\log p})$, and $\hat{G}$ is constructed based on the observations $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n \overset{\text{i.i.d.}}{\sim} N_p(\boldsymbol{\mu}_1, \Sigma_1)$, $\boldsymbol{y}_1, \dots, \boldsymbol{y}_n \overset{\text{i.i.d.}}{\sim} N_p(\boldsymbol{\mu}_2, \Sigma_2)$. Then the minimax risk of the classification error over $\Theta_p(s_1, s_2)$ satisfies*

$$\inf_{\hat{G}} \sup_{\boldsymbol{\theta} \in \Theta_p(s_1, s_2)} \mathbb{E}\big[R_{\boldsymbol{\theta}}(\hat{G}) - R_{\boldsymbol{\theta}}\big(G^*_{\boldsymbol{\theta}}\big)\big] \gtrsim (s_1 + s_2) \cdot \frac{\log p}{n}.$$

REMARK 5. Theorems 4.2 and 4.3 together show that the proposed SQDA rule is optimal for classifying Gaussian data under mild regularity conditions. No other method can have a faster convergence rate of misclassification error in this region. The method and results can be further extended beyond the Gaussian setting. See Section 5 for a detailed discussion on the extension.

The challenge of proving Theorem 4.3 is that the excess risk $R_{\boldsymbol{\theta}}(\hat{G}) - R_{\boldsymbol{\theta}}(G^*_{\boldsymbol{\theta}})$ does not satisfy the triangle inequality (or subadditivity), which is essential to the standard minimax lower bound techniques. To overcome this challenge, we define an alternative risk function $L_{\boldsymbol{\theta}}(\hat{G})$ as follows:

$$L_{\boldsymbol{\theta}}(\hat{G}) := \mathbb{P}_{\boldsymbol{\theta}}\big(\hat{G}(z) \neq G^*_{\boldsymbol{\theta}}(z)\big). \tag{4.2}$$

This loss function $L_{\boldsymbol{\theta}}(\hat{G})$ is essentially the probability that $\hat{G}$ produces a different label than $G^*_{\boldsymbol{\theta}}$, and satisfies the triangle inequality, as shown in Lemma 8.1. The connection between $R_{\boldsymbol{\theta}}(\hat{G}) - R_{\boldsymbol{\theta}}(G^*_{\boldsymbol{\theta}})$ and $L_{\boldsymbol{\theta}}(\hat{G})$ is presented by the following lemma, which shows that it is sufficient to provide a lower bound for $L_{\boldsymbol{\theta}}(\hat{G})$ to prove Theorem 4.3.

LEMMA 4.1. *Suppose $\boldsymbol{\theta} \in \Theta_p(s_1, s_2)$. There exists a constant $c > 0$, does not depend on n, p, such that for some classification rule G, if $L_{\boldsymbol{\theta}}(G) < c$, then*

$$L^2_{\boldsymbol{\theta}}(G) \lesssim \mathbb{P}_{\boldsymbol{\theta}}\big(G(z) \neq L(z)\big) - \mathbb{P}_{\boldsymbol{\theta}}\big(G_{\boldsymbol{\theta}}(z) \neq L(z)\big).$$

Based on Lemma 4.1, we use Fano's inequality on a carefully designed least favorable multivariate normal distributions to complete the proof of Theorems 2.2 and 4.3. The details are shown in Section 8.

5. Extension to the non-Gaussian distributions. The Gaussianity assumption can be relaxed by incorporating semiparametric Gaussian copula model into the QDA framework. This larger semiparametric Gaussian copula model enables robust estimation and classification, and has been studied widely in statistics and machine learning, including linear discriminant analysis (LDA) [21, 34], correlation matrix estimation [20], graphical models [5, 33, 46] and linear regression [10].

The Semiparametric Discriminant Analysis (SeDA) model, introduced by Lin and Jeon [32], assumes that there are two groups of p-dimensional observations $\boldsymbol{x}_1^{(1)}, \dots, \boldsymbol{x}_{n_1}^{(1)} \sim \boldsymbol{X}^{(1)}$, $\boldsymbol{x}_1^{(2)}, \dots, \boldsymbol{x}_{n_2}^{(2)} \sim \boldsymbol{X}^{(2)}$, and there are some unknown strictly increasing functions $f_1, \dots, f_p$, such that

$$\boldsymbol{f}(\boldsymbol{X}^{(k)}) \stackrel{\text{def}}{=} \big(f_1(X_1^{(k)}), \dots, f_p(X_p^{(k)})\big) \sim N_p(\boldsymbol{\mu}_k, \Sigma_k) \quad \text{for } k = 1, 2. \tag{5.1}$$

By properties of the Gaussian distribution, f_j's are only unique up to location and scale shifts. Therefore, for identifiability, same as Mai and Zou [34], we assume, for $j = 1, \dots, p$,

$$\mathbb{E}[f_j(X_j^{(1)})] = \mu_j^{(1)} = 0; \qquad \operatorname{Var}\big(f_j(X_j^{(1)})\big) = \sigma_{jj}^{(1)} = 1. \tag{5.2}$$

The SeDA model in the high-dimensional LDA setting was recently studied by Mai and Zou [34] and Han, Zhao and Liu [21] under the assumption that Σ_k's are all equal. By applying the LPD idea in Cai and Liu [7], consistent classification rules were proposed under this semiparametric linear discriminant analysis model.

The current paper presents a framework to extend the high-dimensional semiparametric LDA to high-dimensional semiparametric QDA. Estimating the mean vectors and covariance matrices similarly as in Mai and Zou [34], Han, Zhao and Liu [21] and then plugging these estimators in (3.2) and (3.3) would lead to a generalized classification rule under the semiparametric quadratic discriminant analysis model. Specifically, for $j \in \{0, 1, 2, \dots, p\}$, let $\hat{F}_j^{(k)}(t)$ be the empirical cumulative distribution function of $\{x_{ij}^{(k)}\}_{i=1}^{n_k}$ Winsorized at $(1/n_k^2, 1 - 1/n_k^2)$ [34], with $j \in \{1, 2, \dots, p\}$. In addition, we estimate the mean $\mu_j^{(2)}$ and variance $\sigma_{jj}^{(2)}$, respectively, by

$$\hat{\mu}_j^{(2)} = \frac{1}{n_2} \sum_{i=1}^{n_2} \Phi^{-1} \circ \hat{F}_j^{(1)}(x_{ij}^{(2)}), \qquad \hat{\sigma}_{jj}^{(2)} = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} \big(\Phi^{-1} \circ \hat{F}_1^{(1)}(x_i^{(2)}) - \hat{\mu}_j^{(2)}\big)^2. \tag{5.3}$$

Here, we note that by the identifiability assumption (5.2), we have $\hat{\mu}_j^{(1)} = 0$, $\hat{\sigma}_{jj}^{(1)} = 1$ for $j = 1, 2, \dots, p$.

Then we estimate the correlation matrices of $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$ the same way as Han, Zhao and Liu [21]. For $j_1 \neq j_2 \in [p]$, we first let the Kendall'ss tau be

$$\hat{\tau}_{j_1, j_2}^{(k)} = \frac{2}{n(n-1)} \sum_{i, i' \in [n]} \operatorname{sign}\big\{(X_{ij_1}^{(k)} - X_{i'j_1}^{(k)})(X_{ij_2}^{(k)} - X_{i'j_2}^{(k)})\big\},$$

and then estimate the correlation matrices $R^{(k)} = (R_{j_1, j_2}^{(k)})_{j_1, j_2 \in [p]}$ by

$$\hat{R}_{j_1, j_2}^{(k)} = \sin\left(\frac{\pi}{2} \hat{\tau}_{j_1, j_2}^{(k)}\right) \cdot \mathbb{1}\{j_1 \neq j_2\} + 1 \cdot \mathbb{1}\{j_1 = j_2\} \quad \text{for } k = 1, 2.$$

At last, we let $\hat{D}_V^{(k)} = \operatorname{diag}((\tilde{\sigma}_{11}^{(k)})^{1/2}, \dots, (\tilde{\sigma}_{pp}^{(k)})^{1/2})$, and estimate Σ_k by

$$\tilde{\Sigma}_k = \hat{D}_V^{(k)} \hat{R} \hat{D}_V^{(k)} \quad \text{for } k = 1, 2.$$

Moreover, we estimate the monotone transformation in a pooled way as

$$\begin{aligned} \hat{f}_j(t) = \frac{1}{n_1 + n_2} & \big(n_1\big(\hat{\mu}_j^{(1)} + (\hat{\sigma}_{jj}^{(1)})^{1/2} \cdot \Phi^{-1}\big(\hat{F}_j^{(1)}(t)\big)\big) \\ & + n_2\big(\hat{\mu}_j^{(2)} + (\hat{\sigma}_{jj}^{(2)})^{1/2} \cdot \Phi^{-1}\big(\hat{F}_j^{(2)}(t)\big)\big)\big), \end{aligned}$$

where $\tilde{\mu}_j^{(k)} = \frac{1}{n_k} \sum_{i=1}^{n_k} x_{ij}^{(k)}$, and $\tilde{\sigma}_{jj}^{(k)} = \frac{1}{n_k - 1} \sum_{i=1}^{n_k} (x_{ij}^{(k)} - \tilde{\mu}_j^{(k)})^2$.

After we obtain the estimators $\tilde{\boldsymbol{\mu}}_1, \tilde{\boldsymbol{\mu}}_2, \tilde{\Sigma}_1, \tilde{\Sigma}_2, \hat{\boldsymbol{f}}$, we can then apply the framework we developed in previous sections for the copula QDA model as follows.

First, we estimate the $\hat{D}$ and $\hat{\boldsymbol{\beta}}$ by plugging them into (3.2) and (3.3) to get $\tilde{D}$ and $\tilde{\boldsymbol{\beta}}$, respectively.

Under the SeDA model, the oracle classification rule is given by

$$G_{\boldsymbol{\theta}}^{\text{copula}}(z) = \begin{cases} 1, & (f(z)-\boldsymbol{\mu}_1)^\top D(f(z)-\boldsymbol{\mu}_1) - 2\boldsymbol{\delta}^\top \Omega_2(f(z)-\bar{\boldsymbol{\mu}}) \\ & \quad -\log\left(\frac{|\Sigma_1|}{|\Sigma_2|}\right) + 2\log\left(\frac{\pi_1}{\pi_2}\right) > 0, \\ 2, & (f(z)-\boldsymbol{\mu}_1)^\top D(f(z)-\boldsymbol{\mu}_1) - 2\boldsymbol{\delta}^\top \Omega_2(f(z)-\bar{\boldsymbol{\mu}}) \\ & \quad -\log\left(\frac{|\Sigma_1|}{|\Sigma_2|}\right) + 2\log\left(\frac{\pi_1}{\pi_2}\right) \le 0. \end{cases} \tag{5.4}$$

Therefore, for a new observation $\boldsymbol{z}$, we propose the following extended classification rule Copula SDAR (CSDAR) for the QDA under the copula model.

$$\begin{aligned} &\hat{G}_{\text{CSDAR}}(z) \\ &\quad = 1 + \mathbb{1}\left\{(\hat{\boldsymbol{f}}(z)-\tilde{\boldsymbol{\mu}}_1)^\top \hat{D}(\hat{\boldsymbol{f}}(z)-\tilde{\boldsymbol{\mu}}_1) - 2\tilde{\boldsymbol{\beta}}^\top\left(\hat{\boldsymbol{f}}(z) - \frac{\tilde{\boldsymbol{\mu}}_1+\tilde{\boldsymbol{\mu}}_2}{2}\right)\right. \\ &\qquad \left. -\log(|\tilde{D}\tilde{\Sigma}_1 + I_p|) + \log\left(\frac{\hat{\pi}_1}{\hat{\pi}_2}\right) \le 0\right\}. \end{aligned} \tag{5.5}$$

We then derive the theoretical properties for this extended SDA rule. At first, we have the following bounds on estimating $\boldsymbol{\beta}$ and D in this non-Gaussian setting.

THEOREM 5.1. *Consider the parameter space $\Theta_p(s_1, s_2)$, and assume that $n_1 \asymp n_2$, $(s_1+s_2)\cdot n^{-1} \to 0$ depending on $\{F_i(z_i)\}_{i=1}^p$. Then the proposed SDAR classification rule in* (3.4) *satisfies that, with probability at least $1 - O(p^{-1})$,*

$$\|\tilde{D} - D\|_F \lesssim \sqrt{\frac{s_1 \log p}{n}}; \qquad \|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 \lesssim \sqrt{\frac{s_2 \log p}{n}}.$$

We then analyze the misclassification error of this extended SDAR rule. In addition to parameter estimation, under the SeDA model, it's required to estimate $\{f_j(t)\}_{j=1}^p$. Since the estimation of f_j is only accurate when there are sufficient samples around t, we define the following region: let $S = j_1, \dots, j_s$ be the joint set of the row support of D and support of $\boldsymbol{\beta}$, and $\gamma \in (0,1)$, define $M_n \subset \mathbb{R}^p$:

$$\begin{aligned} M_n = \{&\boldsymbol{x} \in \mathbb{R}^p : \boldsymbol{x}_S \in [f_{j_1}^{-1}(-\sqrt{2\gamma \log n}), f_{j_1}^{-1}(\sqrt{2\gamma \log n})] \times \dots \\ &\times [f_{j_s}^{-1}(-\sqrt{2\gamma \log n}), f_{j_s}^{-1}(\sqrt{2\gamma \log n})]\}, \end{aligned}$$

which is a high-probability event with $\mathbb{P}(M_n) \ge 1 - C \cdot s \cdot n^{-\gamma}$.

We then define the misclassification error for the copula model:

$$\tilde{R}_{\boldsymbol{\theta}}(\hat{G}) = \mathbb{E}_{\boldsymbol{\theta}}[\mathbb{1}\{\hat{G}(z) \ne L(z) \mid z \in M_n\}]. \tag{5.6}$$

Similar construction of M_n has been considered in all previous papers considering the SeDA model [21, 34, 48]. We then have the following result for the misclassification error.

THEOREM 5.2. *Consider the parameter space $\Theta_p(s_1, s_2)$. Under the same condition as in Theorem* 5.1, *and $\gamma \in (0,1)$ satisfies $s \cdot n^{-\gamma} \to 0$, then the proposed SDAR classification rule in* (3.4) *satisfies that, for sufficiently large n,*

$$\sup_{\boldsymbol{\theta}\in\Theta_p(s_1,s_2)} \mathbb{E}[\tilde{R}_{\boldsymbol{\theta}}(\hat{G}_{\text{CSDAR}}) - \tilde{R}_{\boldsymbol{\theta}}(G_{\boldsymbol{\theta}}^{\text{copula}})] \lesssim (s_1+s_2)\cdot \frac{\log p}{n^{1-\gamma}} \cdot \log^2 n.$$

TABLE 1
Average classification errors (s.e.) based on $n = 100$ test samples from 100 replications under the setting where covariance matrices are known to be identity

	n	$R_{\boldsymbol{\theta}}(\hat{G})$	$R_{\boldsymbol{\theta}}(G_{\text{opt}})$
$p = 200$	100	0.242 (0.054)	0.155 (0.035)
	150	0.232 (0.051)	0.155 (0.035)
	200	0.219 (0.039)	0.155 (0.035)
$p = 300$	100	0.265 (0.048)	0.149 (0.032)
	200	0.223 (0.047)	0.149 (0.032)
	300	0.208 (0.038)	0.149 (0.032)
$p = 600$	200	0.269 (0.045)	0.158 (0.035)
	400	0.230 (0.035)	0.158 (0.035)
	600	0.201 (0.035)	0.158 (0.035)

REMARK 6. The additional term n^{γ} commonly appeared in recent literature studying Gaussian copula models, especially for the classification setting (see Han, Zhao and Liu [21], Zhao and Wegkamp [48] and Mai and Zou [34]), and this term occurs due to the necessity of estimating f_j's. We improve the the convergence rate of $\frac{\sqrt{s}}{n^{(1-\gamma)/2}}$ from prior works to $\frac{s}{n^{1-\gamma}}$ in Theorem 5.2.

6. Numerical studies. In this section, we first conduct simulation studies to investigate the impossibility results shown in Section 2.2, and then study numerical properties of the proposed SDAR and CSDAR methods under various settings.

6.1. *Impossibility results.* We would like to illustrate the impossibility results Theorem 2.1 and Theorem 2.2 in a numerical fashion in this subsection.

Let us start with Theorem 2.1, which shows the sparsity condition on $\boldsymbol{\beta}$ is necessary. In the simulation, we consider the simple case where both covariance matrices are known to be identity but the means are unknown: $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \sim N_p(\boldsymbol{\mu}_1, I_p)$ and $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n \sim N_p(\boldsymbol{\mu}_2, I_p)$ and let $\boldsymbol{\mu}_1 = -\boldsymbol{\mu}_2 = \boldsymbol{\mu} = \frac{1}{\sqrt{p}} \cdot \mathbf{1}_p$, satisfying $\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_2 = 2$.

We consider nine cases where $(n, p) = (100, 200), (150, 200), (200, 200), (100, 300), (200, 300), (300, 300), (200, 600), (400, 600), (600, 600)$. In each setting, we compare the oracle classification rule $G^*_{\boldsymbol{\theta}}$ in (1.1) with the plug-in classification rule $\hat{G}$ where we estimate $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ by the sample means. The testing sample size is set to 100 and the simulation is repeated 100 times in each setting. The simulations results is summarized in Table 1.

To illustrate Theorem 2.2, we consider a simple case where $\boldsymbol{\mu}_1 = -\boldsymbol{\mu}_2 = (1, 0, 0, \ldots, 0)^{\top}$ and the covariance matrices are known to be diagonal. Two classes are $N_p(\boldsymbol{\mu}_1, I_p)$ and $N_p(\boldsymbol{\mu}_2, \Sigma_2)$, where $\Sigma_2 = (I_p + \sum_{i=1}^{p/2} \frac{2}{\sqrt{p}} E_{i,i})^{-1}$ and $E_{i,i}$ is a $p \times p$ matrix whose (i, i)th entry is 1 and 0 else.

We consider nine cases where $(n, p) = (100, 200), (150, 200), (200, 200), (100, 300), (200, 300), (300, 300), (200, 600), (400, 600), (600, 600)$. In each setting, we compare the oracle classification rule G_{opt}, that is (1.1), with the plug-in classification rule $\hat{G}$ where we estimate Σ_1, Σ_2 by the diagonals of sample covariance matrices. Table 2 summarizes the simulation results where the testing sample size is set to 100 and the simulation is repeated 100 times.

6.2. *SDAR on synthetic data.* In this section, we provide extensive numerical evidence to show the empirical performance of SDAR by comparing it to its competitors, including

TABLE 2
Average classification errors (s.e.) based on $n = 100$ test samples from 100 replications under the setting where means are known to be $\mathbf{0}_p$ and covariance matrices are known to be diagonal

	n	$R_{\boldsymbol{\theta}}(\hat{G})$	$R_{\boldsymbol{\theta}}(G_{\text{opt}})$
$p = 200$	100	0.274 (0.049)	0.193 (0.038)
	150	0.260 (0.036)	0.193 (0.038)
	200	0.252 (0.033)	0.193 (0.038)
$p = 300$	100	0.271 (0.043)	0.151 (0.034)
	200	0.238 (0.048)	0.151 (0.034)
	300	0.224 (0.039)	0.151 (0.034)
$p = 600$	200	0.296 (0.032)	0.183 (0.046)
	400	0.255 (0.055)	0.183 (0.046)
	600	0.245 (0.037)	0.183 (0.046)

the sparse QDA (SQDA, Li and Shao [31]), the direct approach for sparse LDA (LPD, Cai and Liu [7]), the conventional LDA (LDA), the conventional QDA (QDA) and the oracle procedure (Oracle). The oracle procedure uses the true underlying model and serves as the optimal risk bound for comparison. We also compare SDAR with other popular classifiers, including sparse logistic regression (SLR), random forest (RF), AdaBoost (AB) and Kernel SVM (KSVM). We evaluate all methods via three synthetic datasets.

In all simulations, the sample size is $n_1 = n_2 = 200$ while the number of variables p varies from $100, 200, 400$ to 600. The sparsity levels are set to be $s_1 = 10, s_2 = 20$. The discriminating direction $\boldsymbol{\beta} = (1, \dots, 1, 0, \dots, 0)^\top$ is sparse such that only the first $s_1 = 10$ entries are nonzero. Given the inverse covariance matrix of the second sample Ω_2, the mean for class 1 is $\boldsymbol{\mu}_1 = (0, \dots, 0)^\top$ and the mean for class 2 is set to be $\boldsymbol{\mu}_2 = \boldsymbol{\mu}_1 - \Sigma_2\boldsymbol{\beta}$. In addition, the differential graph D is a random sparse symmetric matrix with its nonzero positions generated by uniform sample. Each nonzero entry on D is i.i.d. and from a standard normal distribution $N(0, 1)$. Lastly, we let $\Omega_1 = D + \Omega_2$, and $\Omega_1 = \Sigma_1^{-1}$, $\Omega_2 = \Sigma_2^{-1}$. We use the following three models to generate Ω_2.

Model 1: *AR(1) model*: $\Omega_2 = (\Omega_{ij})_{p\times p}$ with $\Omega_{ij} = \rho^{|i-j|}$. In the simulation, the tuning parameters for the SDAR method are chosen by cross-validation over a grid $\{\frac{k}{4}\sqrt{\frac{\log p}{n}}\}_{k=1:15}$. The simulation results from 100 replications are summerized as follows, with $\rho = 0.5$.

Model 2: *Block sparse model*: We generate $\Omega_2 = U^T \Lambda U$, where $\Lambda \in \mathbb{R}^{p\times p}$ is a diagonal matrix and its entries are i.i.d. and uniform on $[1, 2]$, and $U \in \mathbb{R}^{p\times p}$ is a random matrix with i.i.d. entries from $N(0, 1)$. In the simulation, the tuning parameters for SDAR method are chosen over a grid $\{\frac{k}{2}\sqrt{\frac{\log p}{n}}\}_{k=1:15}$.

Model 3: *Erdős–Rényi random graph*: Let $\tilde{\Omega}_2 = (\tilde{\omega}_{ij})$ where $\tilde{\omega}_{ij} = u_{ij}\delta_{ij}$, $\delta_{ij} \sim \text{Ber}(1, \rho)$ being the Bernoulli random variable with success probability 0.05 and $u_{ij} \sim \text{Unif}[0.5, 1] \cup [-1, -0.5]$. After symmetrizing $\tilde{\Omega}_2$, set $\Omega_2 = \tilde{\Omega}_2 + \{\max(-\phi_{\min}(\tilde{\Omega}_2), 0) + 0.05\}\mathbf{I}_p$ to ensure the positive definiteness. In the simulation, the tuning parameters for SDAR method are chosen over a grid $\{\frac{k}{2}\sqrt{\frac{\log p}{n}}\}_{k=1:15}$.

In each model, the number of repetition is set to be 100, and the classification errors are evaluated based on the test data with size 200 that is generated from a Gaussian mixture model $\frac{1}{2}N_p(\boldsymbol{\mu}_1, \Sigma_1) + \frac{1}{2}N_p(\boldsymbol{\mu}_2, \Sigma_2)$. We compare the proposed SDAR method with the oracle QDA rule (1.1). The simulation results are summarized in Table 3 and visualized in Figure 1.

This simulation results show that the proposed SDAR algorithm outperforms the LPD algorithm when there are strong interactions among features ($D \neq 0$). As expected, the conven-

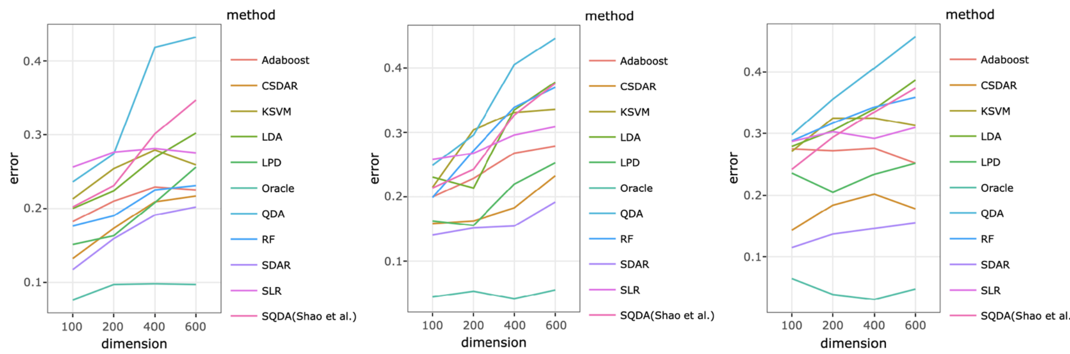

FIG. 1. *Visualization of the results in Table 3.*

tional LDA and QDA works poorly in the high-dimensional setting, and the performance of conventional QDA is even worse due to overfitting. Comparing to the model-free classifiers, we found that they perform better than LDA/QDA, but still have higher misclassification error than the proposed SDAR algorithm since the latter incorporates more model information such as Gaussianity and sparsity. In the setting where $D=0$, the estimated $\hat{D}$ would equal to $D=0$ for properly chosen λ_1, according to Theorem 4.1. As we estimate $\boldsymbol{\beta}$ and D separately, the proposed SDAR rule in this case would adaptively reduced to LPD. For reasons of space, we do not present the detailed numerical results for this case.

6.3. *CSDAR on synthetic data under Gaussian copula model.* Same as the previous discussion, in this section, we compare the performance of CSDAR with its competitors, including LDA, QDA, SQDA, LPD, SLR, RF, AB and KSVM. For the synthetic data generation, we use the same parameter settings as Model 1–Model 3 to generate mean vectors and covariance matrices, and call them Model 4–6. Additionally, after the generation of Gaussian distributed data, for each model, we apply the following monotone transformations: $f_1(x)=x^3$, $f_2(x)=\arctan(x)$, $f_3(x)=\arctan^3(x)$, $f_4(x)=x^5$ to the 1st–5th, 11th–15th, 2st–50th and 51st–85th entries, respectively. The simulation results are summarized in Table 4 and visualized in Figure 2.

This simulation results show that all Gaussian-model based algorithms fail in this setting, while the proposed CSDAR classifier and model free algorithms still maintain their good performances. Further, due to the incorporation of the model information such as Gaussian copula and sparsity, the CSDAR algorithm has smaller misclassification errors than RF, AB, SLR and KSVM in most cases.

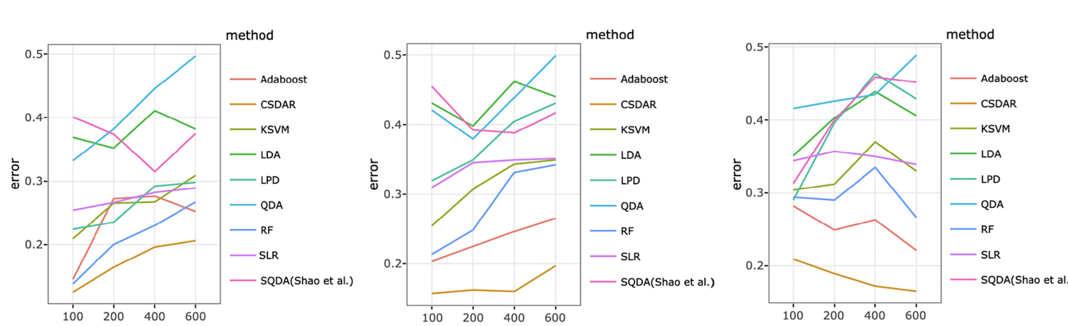

FIG. 2. *Visualization of the results in Table 4.*

TABLE 3
Average classification errors (s.d.) based on $n = 200$ test samples from 100 replications under three different models (Gaussian setting)

	p	100	200	400	600
Model 1	LDA	0.200 (0.019)	0.224 (0.028)	0.269 (0.022)	0.302 (0.024)
	QDA	0.236 (0.026)	0.274 (0.023)	0.418 (0.025)	0.432 (0.027)
	SQDA (Shao et al.)	0.202 (0.022)	0.231 (0.027)	0.301 (0.023)	0.347 (0.025)
	LPD	0.151 (0.020)	0.163 (0.021)	0.208 (0.028)	0.256 (0.025)
	SLR	0.256 (0.024)	0.276 (0.007)	0.281 (0.013)	0.275 (0.020)
	RF	0.176 (0.021)	0.190 (0.022)	0.225 (0.018)	0.231 (0.013)
	Adaboost	0.182 (0.018)	0.210 (0.029)	0.229 (0.017)	0.225 (0.026)
	KSVM	0.213 (0.027)	0.254 (0.016)	0.279 (0.022)	0.259 (0.029)
	SDAR	**0.117 (0.019)**	**0.159 (0.022)**	**0.191 (0.029)**	**0.202 (0.027)**
	CSDAR	0.132 (0.017)	0.173 (0.025)	0.209 (0.024)	0.217 (0.022)
	Oracle	0.076 (0.010)	0.097 (0.007)	0.098 (0.010)	0.097 (0.009)
Model 2	LDA	0.231 (0.022)	0.214 (0.021)	0.335 (0.025)	0.378 (0.027)
	QDA	0.249 (0.025)	0.296 (0.029)	0.405 (0.026)	0.446 (0.028)
	SQDA (Shao et al.)	0.214 (0.023)	0.243 (0.024)	0.327 (0.023)	0.376 (0.025)
	LPD	0.163 (0.018)	0.156 (0.019)	0.220 (0.027)	0.253 (0.024)
	SLR	0.258 (0.015)	0.268 (0.015)	0.296 (0.012)	0.309 (0.025)
	RF	0.199 (0.027)	0.272 (0.020)	0.339 (0.038)	0.370 (0.029)
	Adaboost	0.200 (0.018)	0.229 (0.017)	0.268 (0.027)	0.279 (0.031)
	KSVM	0.215 (0.031)	0.304 (0.021)	0.331 (0.022)	0.336 (0.018)
	SDAR	**0.141 (0.015)**	**0.152 (0.019)**	**0.155 (0.020)**	**0.192 (0.019)**
	CSDAR	0.159 (0.021)	0.163 (0.019)	0.183 (0.026)	0.233 (0.027)
	Oracle	0.045 (0.010)	0.054 (0.007)	0.042 (0.008)	0.056 (0.008)
Model 3	LDA	0.279 (0.028)	0.305 (0.032)	0.340 (0.031)	0.387 (0.029)
	QDA	0.298 (0.024)	0.356 (0.025)	0.406 (0.026)	0.457 (0.025)
	SQDA (Shao et al.)	0.242 (0.024)	0.294 (0.029)	0.335 (0.026)	0.374 (0.026)
	LPD	0.236 (0.023)	0.205 (0.020)	0.234 (0.031)	0.252 (0.027)
	SLR	0.287 (0.015)	0.303 (0.015)	0.292 (0.011)	0.310 (0.026)
	RF	0.288 (0.014)	0.317 (0.022)	0.343 (0.024)	0.359 (0.027)
	Adaboost	0.275 (0.028)	0.272 (0.016)	0.276 (0.018)	0.252 (0.019)
	KSVM	0.271 (0.034)	0.325 (0.025)	0.325 (0.036)	0.313 (0.027)
	SDAR	**0.115 (0.022)**	**0.137 (0.026)**	**0.146 (0.028)**	**0.155 (0.026)**
	CSDAR	0.143 (0.019)	0.184 (0.024)	0.202 (0.032)	0.178 (0.023)
	Oracle	0.065 (0.013)	0.039 (0.009)	0.031 (0.008)	0.048 (0.010)

6.4. *Real data.* In addition to the simulation studies, we also illustrate the merits of the SDAR classifier in the analysis of two real datasets to further investigate the numerical performance of the proposed method. One is the prostate cancer data in Singh et al. [40], which is available at ftp://stat.ethz.ch/Manuscripts/dettling/prostate.rda, and another dataset is the colon tissues data analyzed in Alon et al. [1] by using the Oligonucleotide microarray technique, available at http://microarray.princeton.edu/oncology/affydata/index.html. These two datasets were frequently used for illustrating the empirical performance of the classifier for high-dimensional data in recent literature; see Dettling [18] and Efron [19]. We will compare SDAR with the existing methods, including the sparse QDA (SQDA, Li and Shao [31]), the direct approach for sparse LDA (LPD, Cai and Liu [7]), the conventional LDA (LDA), the conventional QDA (QDA).

6.4.1. *Prostate cancer data.* The prostate cancer data consists of genetic expression levels for $p = 6033$ genes from 102 individuals (50 normal control subjects and 52 prostate cancer patients). The SDAR classifier allows us to model the interactions among genes, and

TABLE 4
Average classification errors (s.d.) based on $n = 200$ test samples from 100 replications under three different models (Gaussian copula setting)

	p	100	200	400	600
Model 4	LDA	0.369 (0.011)	0.362 (0.013)	0.411 (0.023)	0.382 (0.014)
	QDA	0.332 (0.012)	0.382 (0.008)	0.446 (0.009)	0.497 (0.002)
	SQDA (Shao et al.)	0.401 (0.019)	0.374 (0.034)	0.315 (0.032)	0.345 (0.027)
	LPD	0.224 (0.026)	0.235 (0.045)	0.292 (0.041)	0.298 (0.008)
	SLR	0.254 (0.021)	0.266 (0.017)	0.282 (0.016)	0.289 (0.019)
	RF	0.138 (0.005)	0.200 (0.008)	0.230 (0.011)	0.267 (0.023)
	Adaboost	0.145 (0.004)	0.201 (0.005)	0.219 (0.012)	0.232 (0.006)
	KSVM	0.209 (0.008)	0.265 (0.010)	0.267 (0.007)	0.309 (0.008)
	CSDAR	**0.125 (0.005)**	**0.164 (0.005)**	**0.196 (0.002)**	**0.206 (0.005)**
Model 5	LDA	0.431 (0.008)	0.398 (0.007)	0.462 (0.008)	0.440 (0.011)
	QDA	0.421 (0.008)	0.379 (0.008)	0.439 (0.009)	0.499 (0.001)
	SQDA (Shao et al.)	0.455 (0.011)	0.392 (0.013)	0.388 (0.017)	0.417 (0.020)
	LPD	0.319 (0.017)	0.349 (0.022)	0.405 (0.020)	0.431 (0.024)
	SLR	0.309 (0.010)	0.345 (0.027)	0.349 (0.028)	0.351 (0.022)
	RF	0.213 (0.007)	0.248 (0.007)	0.331 (0.011)	0.342 (0.006)
	Adaboost	0.203 (0.007)	0.225 (0.005)	0.246 (0.006)	0.265 (0.009)
	KSVM	0.254 (0.006)	0.307 (0.007)	0.343 (0.008)	0.349 (0.007)
	CSDAR	**0.157 (0.018)**	**0.162 (0.005)**	**0.160 (0.006)**	**0.197 (0.009)**
Model 6	LDA	0.351 (0.016)	0.403 (0.009)	0.439 (0.010)	0.406 (0.007)
	QDA	0.416 (0.006)	0.426 (0.011)	0.435 (0.007)	0.489 (0.003)
	SQDA (Shao et al.)	0.313 (0.016)	0.400 (0.022)	0.459 (0.022)	0.452 (0.012)
	LPD	0.290 (0.013)	0.396 (0.018)	0.464 (0.014)	0.429 (0.013)
	SLR	0.344 (0.024)	0.357 (0.022)	0.350 (0.032)	0.339 (0.028)
	RF	0.294 (0.007)	0.290 (0.008)	0.335 (0.006)	0.266 (0.010)
	Adaboost	0.282 (0.008)	0.249 (0.007)	0.263 (0.009)	0.221 (0.015)
	KSVM	0.304 (0.007)	0.312 (0.008)	0.370 (0.012)	0.330 (0.014)
	CSDAR	**0.209 (0.008)**	**0.189 (0.004)**	**0.172 (0.007)**	**0.165 (0.005)**

thus improve the classification accuracy. For this data, we follow the same data cleaning routine in Cai and Liu [7], retaining only the top 200 genes with the largest absolute values of the two sample t-statistics. The average classification errors using 5-fold cross-validation for various methods with 50 repetitions are reported in Table 5. The proposed SDAR method outperforms all the other methods

6.4.2. *Colon tissues data*. The colon tissues data analyzed gene expression difference between tumor and normal colon tissues using the Oligonucleotide microarray technique, consisting 20 observations from normal tissues and 42 observations from tumor tissues, measured in $p = 2000$ genes.

TABLE 5
Classification error (%) with s.d. of prostate cancer data by various methods

	SDAR	SQDA (Shao et al.)	LPD	LDA	QDA
Testing error	2.20 (1.11)	3.10 (1.26)	11.20 (1.87)	32.20 (3.67)	35.30 (4.18)
	CSDAR	RF	AB	SLR	KSVM
Testing error	4.27 (0.15)	3.10 (4.26)	5.36 (4.89)	4.13 (1.02)	5.36 (2.39)

TABLE 6
Classification error (%) with s.d. of colon tissues data by various methods

	SDAR	SQDA (Shao et al.)	LPD	LDA	QDA
Testing error	19.05 (2.40)	23.20 (2.36)	26.67 (2.75)	38.20 (3.14)	39.30 (4.71)
	CSDAR	RF	AB	SLR	KSVM
Testing error	22.27 (2.41)	25.33 (4.24)	23.81 (3.72)	20.20 (2.31)	25.71 (4.24)

Similar to the analysis of the prostate cancer data, to control the computational costs, we use 200 genes with the largest absolute values of the two sample t-statistics. Classification results by using 5-fold cross-validation with 50 repetitions are summarized in Table 6. In this example, the SDAR is still the best among all classifiers.

7. Extension to the multigroup classification. We have so far focused on high-dimensional QDA for two groups in the Gaussian setting. The methodology and theory developed in the earlier sections can be extended to multigroup classification and to classification under the Gaussian copula model.

7.1. *Multigroup classification.* We first turn to multigroup classification. Suppose there are K classes $N_p(\boldsymbol{\mu}_k, \Sigma_k)$ with prior probability π_k for $1 \le k \le K$, respectively, and an observation z is drawn from the same distribution. In the ideal setting where all the parameters are known, the oracle rule classifies z to class k if and only if

$$k = \arg\min_{k\in[K]} \{Q_k(z)\},$$

where the discriminating function $Q_k(z)$ is

$$Q_k(z) = \begin{cases} 1, & k = 1, \\ \frac{1}{2}(z - \boldsymbol{\mu}_k)^\top D_k(z - \boldsymbol{\mu}_k) - \boldsymbol{\beta}_k^\top (z - \bar{\boldsymbol{\mu}}_k) - \frac{1}{2}\log|D_k\Sigma_1 + I_p| + \log \pi_k, & k \ge 2, \end{cases}$$

with $\bar{\boldsymbol{\mu}}_k = \frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_k}{2}$, $D_k = \Omega_1 - \Omega_k$, $\boldsymbol{\beta}_k = \Omega_1(\boldsymbol{\mu}_k - \boldsymbol{\mu}_1)$, and $\Omega_k = \Sigma_k^{-1}$. When the parameters are unknown and random samples from K classes (with prior probabilities $\{\pi_k\}_{k=1}^K$) are available: $\boldsymbol{x}_1^{(k)}, \dots, \boldsymbol{x}_{n_k}^{(k)} \overset{\text{i.i.d.}}{\sim} N_p(\boldsymbol{\mu}_k, \Sigma_k)$, $k = 1, \dots, K$, by assuming the sparsity on D_k's and $\boldsymbol{\beta}_k$'s, they can then be estimated by solving a similar linear programming as in (3.2) and (3.3). For $k = 2, 3, \dots, K$, D_k and $\boldsymbol{\beta}_k$ are estimated by

$$\hat{D}_k = \arg\min_{D\in\mathbb{R}^{p\times p}} \left\{|D|_1 : \left|\frac{1}{2}\hat{\Sigma}_1 D\hat{\Sigma}_k + \frac{1}{2}\hat{\Sigma}_2 D\hat{\Sigma}_1 - \hat{\Sigma}_1 + \hat{\Sigma}_k\right|_\infty \le \lambda_{1,n}\right\}, \tag{7.1}$$

where $\lambda_{1,n}$ is a tuning parameter with constant $c_1 > 0$.

$$\hat{\boldsymbol{\beta}}_k = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \{\|\boldsymbol{\beta}\|_1 : \|\hat{\Sigma}_1\boldsymbol{\beta} - \hat{\boldsymbol{\mu}}_k + \hat{\boldsymbol{\mu}}_1\|_\infty \le \lambda_{2,n}\}, \tag{7.2}$$

where $\lambda_{2,n}$ is a tuning parameter with constant $c_2 > 0$.

Given these estimators and $\hat{\pi}_k = n_k/(\sum_{k=1}^K n_k)$, the discriminating function is then estimated by

$$\hat{Q}_k(z) = \begin{cases} 1, & k = 1, \\ \frac{1}{2}(z - \hat{\boldsymbol{\mu}}_k)^\top \hat{D}_k(z - \hat{\boldsymbol{\mu}}_k) - \hat{\boldsymbol{\beta}}_k^\top (z - \hat{\bar{\boldsymbol{\mu}}}_k) - \frac{1}{2}\log|\hat{D}_k\hat{\Sigma}_1 + I_p| + \log \hat{\pi}_k, & k \ge 2. \end{cases}$$

Then the SDAR classification rule for multigroup classification is constructed as

$$\hat{G}(z) = \underset{k\in[K]}{\arg\min}\{\hat{Q}_k(z)\}.$$

By applying the same techniques we developed for Theorems 4.1 and 4.2, similar convergence rates can be obtained for both estimation and classification errors.

8. Proofs. We present the proofs of Theorems 2.1, 2.2, 4.1, 4.2 in this section. The proof of Theorem 4.3 is similar to Theorems 2.1, 2.2, so we present its proof in the Supplementary Material.

8.1. *Proof of Theorem* 2.1 *and* 2.2. We prove Theorem 2.1 and 2.2 for the case where $p \lesssim n$. In the case where $\limsup_{n\to\infty} p/n = \infty$, the right hand side of Theorem 2.1 (and 2.2) is of constant order and we can consider only the first n-dimension of p-dimensional vector, and assume the rest is known.

We begin by collecting a few important technical lemmas that will be used in the proofs of the minimax lower bounds.

8.1.1. *Technical lemmas.*

LEMMA 8.1 ([3]). *For any $\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}} \in \Theta_p(s_1, s_2)$ and any classification rule $\hat{G}$, recall that $G^*_{\tilde{\boldsymbol{\theta}}}$ is the optimal rule w.r.t. $\tilde{\boldsymbol{\theta}}$. If*

$$L_{\boldsymbol{\theta}}(G^*_{\tilde{\boldsymbol{\theta}}}) + L_{\boldsymbol{\theta}}(\hat{G}) + \sqrt{\frac{\mathrm{KL}(\mathbb{P}_{\boldsymbol{\theta}}, \mathbb{P}_{\tilde{\boldsymbol{\theta}}})}{2}} \le 1/2,$$

then

$$L_{\boldsymbol{\theta}}(G^*_{\tilde{\boldsymbol{\theta}}}) - L_{\boldsymbol{\theta}}(\hat{G}) - \sqrt{\frac{\mathrm{KL}(\mathbb{P}_{\boldsymbol{\theta}}, \mathbb{P}_{\tilde{\boldsymbol{\theta}}})}{2}} \le L_{\tilde{\boldsymbol{\theta}}}(\hat{G}) \le L_{\boldsymbol{\theta}}(G^*_{\tilde{\boldsymbol{\theta}}}) + L_{\boldsymbol{\theta}}(\hat{G}) + \sqrt{\frac{\mathrm{KL}(\mathbb{P}_{\boldsymbol{\theta}}, \mathbb{P}_{\tilde{\boldsymbol{\theta}}})}{2}},$$

where the KL divergence of two probability density functions $\mathbb{P}_{\boldsymbol{\theta}_1}$ and $\mathbb{P}_{\boldsymbol{\theta}_2}$ is defined by

$$\mathrm{KL}(\mathbb{P}_{\boldsymbol{\theta}_1}, \mathbb{P}_{\boldsymbol{\theta}_2}) = \int \mathbb{P}_{\boldsymbol{\theta}_1}(x) \log \frac{\mathbb{P}_{\boldsymbol{\theta}_1}(x)}{\mathbb{P}_{\boldsymbol{\theta}_2}(x)}\, dz.$$

LEMMA 8.2 ([41]). *Let $M \ge 0$ and $\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_M \in \Theta_p(s_1, s_2)$. For some constants $\alpha \in (0, 1/8), \gamma > 0$, and any classification rule $\hat{G}$, if $\mathrm{KL}(\mathbb{P}_{\boldsymbol{\theta}_i}, \mathbb{P}_{\boldsymbol{\theta}_0}) \le \alpha \log M/n$ for all $1 \le i \le M$, and $L_{\boldsymbol{\theta}_i}(\hat{G}) < \gamma$ implies $L_{\boldsymbol{\theta}_j}(\hat{G}) \ge \gamma$ for all $0 \le i \ne j \le M$, then*

$$\inf_{\hat{G}} \sup_{i\in[M]} \mathbb{E}_{\boldsymbol{\theta}_i}[L_{\boldsymbol{\theta}_i}(\hat{G})] \gtrsim \gamma.$$

To use Fano's type minimax lower bound, we need a covering number argument, provided by the following Lemma 8.3.

LEMMA 8.3 ([41]). *Define $\mathcal{A}_{p,s} = \{\boldsymbol{u} : \boldsymbol{u} \in \{0, 1\}^p, \|\boldsymbol{u}\|_0 = s\}$. If $p \ge 4s$, then there exists a subset $\{\boldsymbol{u}_0, \boldsymbol{u}_1, \dots, \boldsymbol{u}_M\} \subset \mathcal{A}_{p,s}$ such that $\boldsymbol{u}_0 = \{0, \dots, 0\}^\top$, $\rho_H(\boldsymbol{u}_i, \boldsymbol{u}_j) \ge s/2$ and $\log(M + 1) \ge \frac{s}{5}\log(\frac{p}{s})$, where ρ_H denotes the Hamming distance.*

8.1.2. *Main proof of Theorem* 2.1. At first we construct the following least favorable subset, which characterizes the difficulty of the general QDA problem. Let us consider the parameter space

$$\Theta_1 = \Biggl\{\boldsymbol{\theta}_{\boldsymbol{u}} = (1/2, 1/2, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, I_p, I_p) : \\ \boldsymbol{\mu}_1 = \lambda_1 \boldsymbol{e}_1 + \sum_{i=2}^{p} \frac{\lambda_2}{\sqrt{n}} \cdot u_i \cdot \boldsymbol{e}_i, \boldsymbol{u} \in \mathcal{A}_{p,p/4}, \boldsymbol{\mu}_2 = \mathbf{0}_p\Biggr\},$$

where $\mathcal{A}_{p,p/4}$ is defined in Lemma 8.3, and λ_1, λ_2 are of constant order and chosen later.

According to Lemma 8.3, there is a subset of Θ_1 with logarithm cardinality being of order p, such that for any $\boldsymbol{\theta}_{\boldsymbol{u}}, \boldsymbol{\theta}_{\boldsymbol{u}'}$ in this subset, we have $\rho_H(\boldsymbol{u}, \boldsymbol{u}') \geq p/8$. We are going to apply Lemma 8.2 to this subset to complete the proof of Theorem 2.1.

For $\boldsymbol{u} \in \mathcal{A}_{p,p/4}$, let $\boldsymbol{\mu}_{\boldsymbol{u}} = \lambda_1 \boldsymbol{e}_1 + \sum_{i=2}^{p} \frac{\lambda_2}{\sqrt{n}} \cdot u_i \cdot \boldsymbol{e}_i$. Note that for two multivariate normal distributions $\mathbb{P}_{\boldsymbol{\theta}_{\boldsymbol{u}}} = N_p(\boldsymbol{\mu}_{\boldsymbol{u}}, I_p)$ and $\mathbb{P}_{\boldsymbol{\theta}_{\boldsymbol{u}'}} = N_p(\boldsymbol{\mu}_{\boldsymbol{u}'}, I_p)$, the KL divergence between them are upper bounded by

$$\mathrm{KL}(\mathbb{P}_{\boldsymbol{\theta}_{\boldsymbol{u}}}, \mathbb{P}_{\boldsymbol{\theta}_{\boldsymbol{u}'}}) = \frac{1}{2}\|\boldsymbol{\mu}_{\boldsymbol{u}} - \boldsymbol{\mu}_{\boldsymbol{u}'}\|_2^2 \leq \frac{\lambda_2^2 \cdot p}{4n}.$$

To use Lemma 8.2 to prove Theorem 2.1, we further need to show that for any $\boldsymbol{\theta}_{\boldsymbol{u}}, \boldsymbol{\theta}_{\boldsymbol{u}'}$,

$$\bigl[R_{\boldsymbol{\theta}}(G) - R_{\boldsymbol{\theta}}(G^*_{\boldsymbol{\theta}_{\boldsymbol{u}}})\bigr] + \bigl[R_{\boldsymbol{\theta}}(G) - R_{\boldsymbol{\theta}}(G^*_{\boldsymbol{\theta}_{\boldsymbol{u}'}})\bigr] \gtrsim \frac{p}{n}.$$

By Lemma 4.1 and 8.1,

$$\bigl[R_{\boldsymbol{\theta}}(G) - R_{\boldsymbol{\theta}}(G^*_{\boldsymbol{\theta}_{\boldsymbol{u}}})\bigr] + \bigl[R_{\boldsymbol{\theta}}(G) - R_{\boldsymbol{\theta}}(G^*_{\boldsymbol{\theta}_{\boldsymbol{u}'}})\bigr] \\ \gtrsim L^2_{\boldsymbol{\theta}_{\boldsymbol{u}}}(G) + L^2_{\boldsymbol{\theta}_{\boldsymbol{u}'}}(G) \geq \frac{1}{2}\bigl(L_{\boldsymbol{\theta}_{\boldsymbol{u}}}(G) + L_{\boldsymbol{\theta}_{\boldsymbol{u}'}}(G)\bigr)^2 \geq \frac{1}{2}\Biggl(L_{\boldsymbol{\theta}_{\boldsymbol{u}}}(G^*_{\boldsymbol{\theta}_{\boldsymbol{u}'}}) - \sqrt{\frac{\mathrm{KL}(\mathbb{P}_{\boldsymbol{\theta}_{\boldsymbol{u}}}, \mathbb{P}_{\boldsymbol{\theta}_{\boldsymbol{u}'}})}{2}}\Biggr)^2.$$

Since now that $\mathrm{KL}(\mathbb{P}_{\boldsymbol{\theta}_{\boldsymbol{u}}}, \mathbb{P}_{\boldsymbol{\theta}_{\boldsymbol{u}'}}) \leq \frac{\lambda_2^2 \cdot p}{4n}$, it is then sufficient to show $L_{\boldsymbol{\theta}_{\boldsymbol{u}}}(G^*_{\boldsymbol{\theta}_{\boldsymbol{u}'}}) \geq c\sqrt{\frac{p}{n}}$ for some $c > \frac{\lambda_2}{2\sqrt{2}}$.

Without loss of generality, we assume that the coordinates of $\boldsymbol{u}$ and $\boldsymbol{u}'$ are ordered such that $u_i = u'_i = 1$ for $i = 2, \dots, m_1$, $u_i = 1 - u'_i = 1$ for $i = m_1 + 1, \dots, m_2$, $u_i = 1 - u'_i = 0$ for $i = m_2 + 1, \dots, m_3$ and $u_i = u'_i = 0$ for $i = m_3 + 1, \dots, p$. We then have $\rho_H(\boldsymbol{u}, \boldsymbol{u}') = m_3 - m_1 \geq \frac{p}{8}$.

Recall that when $\Sigma_1 = \Sigma_2 = I_p$ and $\boldsymbol{\mu}_2 = \mathbf{0}_p$, the oracle rule is given by

$$G^*_{\boldsymbol{\theta}}(\boldsymbol{z}) = 1 + \mathbb{1}\biggl\{-\boldsymbol{\mu}_1^\top\biggl(\boldsymbol{z} - \frac{\boldsymbol{\mu}_1}{2}\biggr) > 0\biggr\}.$$

Then

$$G^*_{\boldsymbol{\theta}_{\boldsymbol{u}}}(\boldsymbol{z}) = 1 + \mathbb{1}\Biggl\{-\frac{\lambda_2}{\sqrt{n}}\Biggl(\sum_{i=2}^{m_1} z_i + \sum_{i=m_1+1}^{m_2} z_i\Biggr) - \lambda_1 z_1 + \frac{1}{2}\lambda_1^2 + \frac{\lambda_2^2(p-1)}{8n} > 0\Biggr\},$$

and

$$G^*_{\boldsymbol{\theta}_{\boldsymbol{u}'}}(\boldsymbol{z}) = 1 + \mathbb{1}\Biggl\{-\frac{\lambda_2}{\sqrt{n}}\Biggl(\sum_{i=2}^{m_1} z_i + \sum_{i=m_2+1}^{m_3} z_i\Biggr) - \lambda_1 z_1 + \frac{1}{2}\lambda_1^2 + \frac{\lambda_2^2(p-1)}{8n} > 0\Biggr\}.$$

Let $Z_1 = -\lambda_1 z_1 - \frac{\lambda_2}{\sqrt{n}}\sum_{i=2}^{m_1} z_i + \frac{1}{2}\lambda_1^2 + \frac{\lambda_2^2(p-1)}{8n}$, $Z_2 = \frac{\lambda_2}{\sqrt{n}}\sum_{i=m_1+1}^{m_2} z_i$ and $Z_3 = \frac{\lambda_2}{\sqrt{n}}\sum_{i=m_2+1}^{m_3} z_i$, then

$$G^*_{\boldsymbol{\theta}_{\boldsymbol{u}}}(z) = 1 + \mathbb{1}\{Z_1 - Z_2 > 0\} \quad \text{and} \quad G^*_{\boldsymbol{\theta}_{\boldsymbol{u}'}}(z) = 1 + \mathbb{1}\{Z_1 - Z_3 > 0\}$$

and, therefore,

$$\begin{aligned} L_{\boldsymbol{\theta}_{\boldsymbol{u}}}\big(G^*_{\boldsymbol{\theta}_{\boldsymbol{u}'}}\big) &= \mathbb{P}_{\boldsymbol{\theta}_{\boldsymbol{u}}}\big(G^*_{\boldsymbol{\theta}_{\boldsymbol{u}'}}(z) \neq G^*_{\boldsymbol{\theta}_{\boldsymbol{u}}}(z)\big) \\ &= \mathbb{P}_{\boldsymbol{\theta}_{\boldsymbol{u}}}(Z_2 \le Z_1 \le Z_3) + \mathbb{P}_{\boldsymbol{\theta}_{\boldsymbol{u}}}(Z_3 \le Z_1 \le Z_2) \\ &\ge \mathbb{P}_{\boldsymbol{\theta}_{\boldsymbol{u}}}(Z_2 \le Z_1 \le Z_3) \\ &= \frac{1}{2}\mathbb{P}_{z\sim N_p(\boldsymbol{\mu}_{\boldsymbol{u}}, I_p)}(Z_2 \le Z_1 \le Z_3) + \frac{1}{2}\mathbb{P}_{z\sim N_p(\mathbf{0}_p, I_p)}(Z_2 \le Z_1 \le Z_3) \\ &\ge \frac{1}{2}\mathbb{P}_{z\sim N_p(\mathbf{0}_p, I_p)}(Z_2 \le Z_1 \le Z_3). \end{aligned}$$

Then, since $Z_1 \sim N(\frac{1}{2}\lambda_1^2 + \frac{\lambda_2^2(p-1)}{8n}, \lambda_1^2 + \lambda_2^2 p/(4n))$, the density of Z_1, $f(z)$ satisfies

$$f(z) \ge \frac{1}{\sqrt{2\pi(\lambda_1^2 + \lambda_2^2 p/(4n))}} \exp\left(-\frac{(z - \lambda_1^2/2 - \lambda_2^2(p-1)/(8n))^2}{2(\lambda_1^2 + \lambda_2^2 p/(4n))^2}\right),$$

leading to

$$f(z) \ge c_1(\lambda_1, \lambda_2), \quad \text{for } z \in [-\lambda_2\sqrt{p/n}, \lambda_2\sqrt{p/n}]$$

for some constant $c_1(\lambda_1, \lambda_2) = \frac{1}{\sqrt{2\pi(\lambda_1^2+\lambda_2^2 p/(4n))}} \exp(-\frac{(\lambda_2\sqrt{p/n}+\lambda_1^2/2+\lambda_2^2(p-1)/(8n))^2}{2(\lambda_1^2+\lambda_2^2 p/(4n))^2})$.

In addition, since $m_3 - m_1 \in (\frac{p}{8}, \frac{p}{2})$, $Z_3 - Z_2$ is normally distributed with mean 0 and variance of order $\frac{p}{n}$ and, therefore, we claim that for some constant c_2,

$$\mathbb{E}\left[(Z_3 - Z_2) \cdot \mathbb{1}\left\{-\lambda_2\sqrt{\frac{p}{n}} < Z_2 < Z_3 < \lambda_2\sqrt{\frac{p}{n}}\right\}\right] \ge c_2\lambda_2\sqrt{\frac{p}{n}}.$$

In fact,

$$\begin{aligned} &\mathbb{E}\left[(Z_3 - Z_2) \cdot \mathbb{1}\left\{-\lambda_2\sqrt{\frac{p}{n}} < Z_2 < Z_3 < \lambda_2\sqrt{\frac{p}{n}}\right\}\right] \\ &\quad \ge \mathbb{E}\left[(Z_3 - Z_2) \cdot \mathbb{1}\left\{-\lambda_2\sqrt{\frac{p}{n}} < Z_2 < -\frac{\lambda_2}{2}\sqrt{\frac{m_2 - m_1}{n}}, \frac{\lambda_2}{2}\sqrt{\frac{m_3 - m_2}{n}} < Z_3 < \lambda_2\sqrt{\frac{p}{n}}\right\}\right] \\ &\quad \ge \lambda_2\sqrt{\frac{p}{n}} \cdot \mathbb{P}\left(-\lambda_2\sqrt{\frac{p}{n}} < Z_2 < -\frac{\lambda_2}{2}\sqrt{\frac{m_2 - m_1}{n}}\right) \cdot \mathbb{P}\left(\frac{\lambda_2}{2}\sqrt{\frac{m_3 - m_2}{n}} < Z_3 < \lambda_2\sqrt{\frac{p}{n}}\right) \\ &\quad \ge \lambda_2\sqrt{\frac{p}{8n}} \cdot \mathbb{P}_{Z\sim N(0,1)}\left(-\sqrt{\frac{p}{m_2 - m_1}} < Z < -\frac{1}{2}\right) \cdot \mathbb{P}_{Z\sim N(0,1)}\left(\frac{1}{2} < Z < \sqrt{\frac{p}{m_3 - m_2}}\right) \\ &\quad \ge \lambda_2\sqrt{\frac{p}{8n}} \cdot \mathbb{P}_{Z\sim N(0,1)}\left(-\sqrt{2} < Z < -\frac{1}{2}\right) \cdot \mathbb{P}_{Z\sim N(0,1)}\left(\frac{1}{2} < Z < \sqrt{2}\right) := c_2\lambda_2\sqrt{\frac{p}{n}}, \end{aligned}$$

where $c_2 = \sqrt{\frac{1}{8}}\mathbb{P}_{Z\sim N(0,1)}(-\sqrt{2} < Z < -\frac{1}{2}) \cdot \mathbb{P}_{Z\sim N(0,1)}(\frac{1}{2} < Z < \sqrt{2})$ is of constant order and the inequality above uses $\sqrt{m_2 - m_1} + \sqrt{m_3 - m_2} \ge \sqrt{m_3 - m_1} \ge \sqrt{p/8}$, $m_2 - m_1, m_3 - m_2 \le m_3 - m_1 \le p/2$.

Then we have

$$
\begin{aligned}
&\mathbb{P}_{z\sim N_p(\mathbf{0}_p, I_p)}(Z_2 \le Z_1 \le Z_3)\\
&\quad \ge \mathbb{P}_{z\sim N_p(\mathbf{0}_p, I_p)}\left(Z_2 \le Z_1 \le Z_3, -\lambda_2\sqrt{\frac{p}{n}} < Z_2 < Z_3 < \lambda_2\sqrt{\frac{p}{n}}\right)\\
&\quad = \mathbb{E}_{Z_2}\left[\int_{Z_2}^{Z_3} f(z_1)\,dz_1 \cdot \mathbb{1}\left\{-\lambda_2\sqrt{\frac{p}{n}} < Z_2 < Z_3 < \lambda_2\sqrt{\frac{p}{n}}\right\}\right]\\
&\quad \ge c_1(\lambda_1,\lambda_2)\cdot \cdot \mathbb{E}_{Z_2}\left[(Z_3 - Z_2)\cdot \mathbb{1}\left\{-\lambda_2\sqrt{\frac{p}{n}} < Z_2 < Z_3 < \lambda_2\sqrt{\frac{p}{n}}\right\}\right]\\
&\quad \ge c_1(\lambda_1,\lambda_2) c_2 \lambda_2 \cdot \sqrt{\frac{p}{n}}.
\end{aligned}
$$

Since $p \lesssim n$, we have $c_1(\lambda_1,\lambda_2) \to \infty$ when $\lambda_1, \lambda_2 \to 0$. Therefore, we can choose λ_1, λ_2 to be sufficiently small such that $c_1(\lambda_1,\lambda_2)c_2\lambda_2\sqrt{\frac{p}{n}} \ge \frac{\lambda_2}{2\sqrt{2}}\sqrt{\frac{p}{n}}$. This completes the proof.

8.1.3. *Proof of Theorem* 2.2. At first, we construct the following least favorable subset, which characterizes the difficulty of the general QDA problem. For simplicity of notation, we use the letters λ_1, λ_2 in this section, whose values are different from those in Section 8.1.2.

Since the KL-divergence and ℓ_2 norm are invariant to translations and orthogonal transformations, without loss of generality, we assume that $\boldsymbol{\mu}_1^* = -\boldsymbol{\mu}_2^* = \lambda_1 \boldsymbol{e}_1 + \tilde{\lambda}_1 \boldsymbol{e}_2$ for some constants $\lambda_1, \tilde{\lambda}_1 > 0$ whose values are determined later, with $2\sqrt{\lambda_1^2 + \tilde{\lambda}_1^2} = \|\boldsymbol{\mu}_1^* - \boldsymbol{\mu}_2^*\|_2$. In addition, we assume that $p/4$ is an integer.

Now let us consider

$$
\Theta_2 = \left\{\boldsymbol{\theta}_{\boldsymbol{u}} = \big(1/2, 1/2, \lambda_1\boldsymbol{e}_1 + \tilde{\lambda}_1\boldsymbol{e}_2, -\lambda_1\boldsymbol{e}_1 - \tilde{\lambda}_1\boldsymbol{e}_2, \Sigma_1^{\boldsymbol{u}}, \Sigma_2\big):\right.
$$
$$
\left.\Sigma_1^{\boldsymbol{u}} = \left(I_p + \tilde{\lambda}_2 E_{2,2} + \frac{\lambda_2}{\sqrt{n}}\sum_{i=3}^{p/2} u_i E_{i,i}\right)^{-1}, \boldsymbol{u} \in \mathcal{A}_{p,p/4}, \Sigma_2 = I_p + \tilde{\lambda}_2 E_{2,2}\right\},
$$

where $\mathcal{A}_{p,p/4}$ is defined in Lemma 8.3.

According to Lemma 8.3, there is a subset of Θ_1 with logarithm cardinality being of order p, such that for any $\boldsymbol{\theta}_{\boldsymbol{u}}, \boldsymbol{\theta}_{\boldsymbol{u}'}$ in this subset, we have $\rho_H(\boldsymbol{u}, \boldsymbol{u}') \ge p/8$. We are going to apply Lemma 8.2 to this subset to complete the proof of Theorem 2.2.

At first, we note that for two multivariate normal distribution $N_p(\boldsymbol{\mu}_1^*, \Sigma_1^{\boldsymbol{u}})$ and $N_p(\boldsymbol{\mu}_1^*, \Sigma_1^{\boldsymbol{u}'})$, using the fact that $\log(1+x) \asymp x - x^2/2 + o(x^2)$ for $x = o(1)$, the KL divergence between them are upper bounded by

$$
\begin{aligned}
\mathrm{KL} &= \frac{1}{2}\left[\log\frac{|\Sigma_1^{\boldsymbol{u}'}|}{|\Sigma_1^{\boldsymbol{u}}|} - p + \mathrm{tr}\big((\Sigma_1^{\boldsymbol{u}'})^{-1}\Sigma_1^{\boldsymbol{u}}\big)\right]\\
&= \frac{1}{2}\left[\sum_{i=3}^{p}\log\frac{1 + \frac{\lambda_2}{\sqrt{n}}u_i'}{1 + \frac{\lambda_2}{\sqrt{n}}u_i} - \rho_H(\boldsymbol{u}, \boldsymbol{u}') + \sum_{i=3}^{p}\frac{1 + \frac{\lambda_2}{\sqrt{n}}u_i}{1 + \frac{\lambda_2}{\sqrt{n}}u_i'}\right]\\
&= \frac{1}{2}\left[-\sum_{i=3}^{p}\log\left(1 + \frac{\frac{\lambda_2}{\sqrt{n}}(u_i - u_i')}{1 + \frac{\lambda_2}{\sqrt{n}}u_i'}\right) + \sum_{i=3}^{p}\frac{\frac{\lambda_2}{\sqrt{n}}(u_i - u_i')}{1 + \frac{\lambda_2}{\sqrt{n}}u_i'}\right]\\
&= \frac{1}{4}\sum_{i=3}^{p}\frac{1}{n}(u_i - u_i')^2 + o\left(\frac{p}{n}\right) \le \frac{\lambda_2^2 p}{16n} + o\left(\frac{p}{n}\right) \le \frac{\lambda_2^2 p}{8n}.
\end{aligned}
$$

Therefore, we have $\mathrm{KL}(\mathbb{P}_{\boldsymbol{\theta_u}}, \mathbb{P}_{\boldsymbol{\theta_{u'}}}) \le \lambda_2^2 p/(8n)$. To use Lemma 8.2 to prove Theorem 2.2, we further need to show that for any $\boldsymbol{\theta_u}, \boldsymbol{\theta_{u'}}$,

$$\big[R_{\boldsymbol\theta}(G) - R_{\boldsymbol\theta}\big(G^*_{\boldsymbol{\theta_u}}\big)\big] + \big[R_{\boldsymbol\theta}(G) - R_{\boldsymbol\theta}\big(G^*_{\boldsymbol{\theta_{u'}}}\big)\big] \gtrsim \frac{p}{n}.$$

By Lemma 4.1 and 8.1,

$$\begin{aligned}&\big[R_{\boldsymbol\theta}(G) - R_{\boldsymbol\theta}\big(G^*_{\boldsymbol{\theta_u}}\big)\big] + \big[R_{\boldsymbol\theta}(G) - R_{\boldsymbol\theta}\big(G^*_{\boldsymbol{\theta_{u'}}}\big)\big]\\ &\quad \ge L^2_{\boldsymbol{\theta_u}}(\hat G) + L^2_{\boldsymbol{\theta_{u'}}}(\hat G) \ge \frac{1}{2}\big(L_{\boldsymbol{\theta_u}}(\hat G) + L_{\boldsymbol{\theta_{u'}}}(\hat G)\big)^2 \ge \frac{1}{2}\bigg(L_{\boldsymbol{\theta_u}}\big(G^*_{\boldsymbol{\theta_{u'}}}\big) - \sqrt{\frac{\mathrm{KL}(\mathbb{P}_{\boldsymbol{\theta_u}}, \mathbb{P}_{\boldsymbol{\theta_{u'}}})}{2}}\bigg)^2.\end{aligned}$$

Since now that $\mathrm{KL}(\mathbb{P}_{\boldsymbol{\theta_u}}, \mathbb{P}_{\boldsymbol{\theta_{u'}}}) \le \lambda_2^2 \frac{p}{8n}$, it is then sufficient to show $L_{\boldsymbol{\theta_u}}(G^*_{\boldsymbol{\theta_{u'}}}) \ge c\sqrt{\frac{p}{n}}$ for some $c > \lambda_2/4$.

Recall that

$$G^*_{\boldsymbol\theta}(\boldsymbol z) = \mathbb{1}\bigg\{(\boldsymbol z - \boldsymbol\mu_1)^\top D(\boldsymbol z - \boldsymbol\mu_1) - 2\boldsymbol\delta^\top \Omega_2(\boldsymbol z - \boldsymbol\mu_1) + \boldsymbol\delta^\top\Omega_2\boldsymbol\delta - \log\bigg(\frac{|\Sigma_1|}{|\Sigma_2|}\bigg) > 0\bigg\},$$

where $\boldsymbol\delta = \boldsymbol\mu_2 - \boldsymbol\mu_1$, $D = \Omega_2 - \Omega_1$.

Without loss of generality, we assume that $u_i = u'_i = 1$ when $i = 3, \dots, m_1$, $u_i = 1 - u'_i = 1$ when $i = m_1 + 1, \dots, m_2$, $u_i = 1 - u'_i = 0$ when $i = m_2 + 1, \dots, m_3$ and $u_i = u'_i = 0$ when $i = m_3 + 1, \dots, p$.

Then with a little abuse of notation, we have $\boldsymbol z \sim \frac{1}{2}N_p(\boldsymbol\mu_1, \Sigma_1^{\boldsymbol u}) + \frac{1}{2}N_p(\boldsymbol\mu_2, \Sigma_2)$ with $\boldsymbol\mu_!1 - \boldsymbol\mu_2 = \lambda_1\boldsymbol e_1 + \tilde\lambda_1\boldsymbol e_2$. Using the fact that $\log(1 + \frac{\lambda_2}{\sqrt n}) = \frac{\lambda_2}{\sqrt n} - \frac{\lambda_2^2}{2n} + o(\frac{1}{n})$, we have

$$\begin{aligned}G^*_{\boldsymbol{\theta_u}}(\boldsymbol z) = 1 + \mathbb{1}\bigg\{&\frac{\lambda_2}{\sqrt n}\bigg(\sum_{i=3}^{m_1}(z_i^2 - 1) + \sum_{i=m_1+1}^{m_2}(z_i^2 - 1)\bigg)\\ &+ 4\lambda_1 z_1 + 4\frac{\tilde\lambda_1}{1 + \tilde\lambda_2}z_2 + \frac{p}{8n} + o\bigg(\frac{p}{n}\bigg) > 0\bigg\}\end{aligned}$$

and

$$\begin{aligned}G^*_{\boldsymbol{\theta_{u'}}}(\boldsymbol z) = 1 + \mathbb{1}\bigg\{&\frac{\lambda_2}{\sqrt n}\bigg(\sum_{i=3}^{m_1}(z_i^2 - 1) + \sum_{i=m_2+1}^{m_3}(z_i^2 - 1)\bigg)\\ &+ 4\lambda_1 z_1 + 4\frac{\tilde\lambda_1}{1 + \tilde\lambda_2}z_2 + \frac{p}{8n} + o\bigg(\frac{p}{n}\bigg) > 0\bigg\}.\end{aligned}$$

Let $Z_1 = -(4\lambda_1 z_1 + 4\frac{\tilde\lambda_1}{1+\tilde\lambda_2}z_2 + \frac{\lambda_2}{\sqrt n}\sum_{i=3}^{m_1}(z_i^2 - 1) + \frac{p}{8n})$, $Z_2 = \frac{\lambda_2}{\sqrt n}\sum_{i=m_1+1}^{m_2}(z_i^2 - 1)$, $Z_3 = \frac{\lambda_2}{\sqrt n}\sum_{i=m_2+1}^{m_3}(z_i^2 - 1)$, then

$$G^*_{\boldsymbol{\theta_u}}(\boldsymbol z) = \mathbb{1}\bigg\{-Z_1 + Z_2 + o\bigg(\frac{p}{n}\bigg) > 0\bigg\} \quad \text{and} \quad G^*_{\boldsymbol{\theta_{u'}}}(\boldsymbol z) = \mathbb{1}\bigg\{-Z_1 + Z_3 + o\bigg(\frac{p}{n}\bigg) > 0\bigg\},$$

and

$$\begin{aligned}L_{\boldsymbol{\theta_u}}\big(G^*_{\boldsymbol{\theta_{u'}}}\big) &= \mathbb{P}_{\boldsymbol{\theta_u}}\big(G^*_{\boldsymbol{\theta_{u'}}}(\boldsymbol z) \ne G^*_{\boldsymbol{\theta_u}}(\boldsymbol z)\big)\\ &\ge \frac{1}{2}\mathbb{P}_{\boldsymbol z \sim N_p(\boldsymbol\mu_1, \Sigma_1^{\boldsymbol u})}\bigg(Z_2 + o\bigg(\frac{p}{n}\bigg) \le Z_1 \le Z_3 + o\bigg(\frac{p}{n}\bigg)\bigg)\\ &\quad + \frac{1}{2}\mathbb{P}_{\boldsymbol z \sim N_p(\boldsymbol\mu_2, \Sigma_2)}\bigg(Z_3 + o\bigg(\frac{p}{n}\bigg) \le Z_1 \le Z_2 + o\bigg(\frac{p}{n}\bigg)\bigg)\\ &\ge \frac{1}{2}\mathbb{P}_{\boldsymbol z \sim N_p(\boldsymbol\mu_1, \Sigma_2)}(Z_2 \le Z_1 \le Z_3) + o\bigg(\frac{p}{n}\bigg).\end{aligned}$$

By the central limit theorem, $\frac{\sqrt{n}}{\lambda_2\sqrt{m_2-m_1}}Z_2$, $\frac{\sqrt{n}}{\lambda_2\sqrt{m_3-m_2}}Z_3$ converges to the standard normal distribution $N(0,1)$. Since $m_3-m_2=\rho_H(\boldsymbol{u},\boldsymbol{u}')\geq p/8$, and $\limsup_{n,p\to\infty}\frac{p}{n}\leq C_1$, similar as the derivation in Section 8.1.2, there exists a constant c_2, such that n,p are sufficiently large,

$$\begin{aligned}
&\mathbb{E}\left[(Z_3-Z_2)\cdot\mathbb{1}\left\{-\lambda_2\sqrt{\frac{p}{n}}<Z_2<Z_3<\lambda_2\sqrt{\frac{p}{n}}\right\}\right]\\
&\quad\geq\mathbb{E}\left[(Z_3-Z_2)\cdot\mathbb{1}\left\{-\lambda_2\sqrt{\frac{p}{n}}<Z_2<-\frac{\lambda_2}{2}\sqrt{\frac{p}{n}},\frac{\lambda_2}{2}\sqrt{\frac{p}{n}}<Z_3<\lambda_2\sqrt{\frac{p}{n}}\right\}\right]\\
&\quad\geq\lambda_2\sqrt{\frac{p}{n}}\cdot\mathbb{P}\left(-\lambda_2\sqrt{\frac{p}{n}}<Z_2<-\frac{\lambda_2}{2}\sqrt{\frac{m_2-m_1}{n}}\right)\cdot\mathbb{P}\left(\frac{\lambda_2}{2}\sqrt{\frac{m_3-m_1}{n}}<Z_3<\lambda_2\sqrt{\frac{p}{n}}\right)\\
&\quad\geq\lambda_2\sqrt{\frac{p}{8n}}\cdot\mathbb{P}_{Z\sim N(0,1)}\left(-\sqrt{\frac{p}{m_2-m_1}}<Z<-\frac{1}{2}\right)\cdot\mathbb{P}_{Z\sim N(0,1)}\left(\frac{1}{2}<Z<\sqrt{\frac{p}{m_3-m_2}}\right)\\
&\quad\geq\lambda_2\sqrt{\frac{p}{8n}}\cdot\mathbb{P}_{Z\sim N(0,1)}\left(-\sqrt{2}<Z<-\frac{1}{2}\right)\cdot\mathbb{P}_{Z\sim N(0,1)}\left(\frac{1}{2}<Z<\sqrt{2}\right)\geq c_2\lambda_2\sqrt{\frac{p}{n}}.
\end{aligned}$$

Similar to that in Section 8.1.2, let us denote the probability density function of Z_1 by f. Use the central limit theorem again, when $\boldsymbol{z}\sim N_p(\boldsymbol{\mu}_1,\Sigma_2)$, $p\lesssim n$, and n,p are sufficiently large, $Z_1\approx N(-4\lambda_1^2-\frac{4\tilde{\lambda}_1^2}{1+\tilde{\lambda}_2}+\frac{p}{8n},\lambda_1^2+\frac{\tilde{\lambda}_1^2}{1+\tilde{\lambda}_2}+\frac{2(m_1-2)\lambda_2^2}{n})$ if $m_1\to\infty$. Therefore, there exists constant $c_1(\lambda_1,\tilde{\lambda}_1,\lambda_2,\tilde{\lambda}_2)$, such that $\inf_{|x|<\lambda_2\sqrt{p/n}}f(x)>c_1(\lambda_1,\tilde{\lambda}_1,\lambda_2,\tilde{\lambda}_2)$, and $c_1(\lambda_1,\tilde{\lambda}_1,\lambda_2,\tilde{\lambda}_2)$ goes to infinity when $\lambda_1,\lambda_2\to 0$, $\tilde{\lambda}_2\to\infty$, and $\tilde{\lambda}_1$ is chosen such that $\sqrt{\lambda_1^2+\tilde{\lambda}_1^2}=\|\boldsymbol{\mu}_1^*-\boldsymbol{\mu}_2^*\|_2/2$,

$$\begin{aligned}
&\mathbb{P}_{\boldsymbol{z}\sim N_p(\boldsymbol{\mu}_1,\Sigma_2)}(Z_2\leq Z_1\leq Z_3)\\
&\quad\geq\mathbb{P}_{\boldsymbol{z}\sim N_p(\boldsymbol{\mu}_1,\Sigma_2)}\left(Z_2\leq Z_1\leq Z_3,-\lambda_2\sqrt{\frac{p}{n}}<Z_2<Z_3<\lambda_2\sqrt{\frac{p}{n}}\right)\\
&\quad=\mathbb{E}_{Z_2}\left[\int_{Z_2}^{Z_3}f(z_1)\,dz_1\cdot\mathbb{1}\left\{-\lambda_2\sqrt{\frac{p}{n}}<Z_2<Z_3<\lambda_2\sqrt{\frac{p}{n}}\right\}\right]\\
&\quad\geq c_1(\lambda_1,\tilde{\lambda}_1,\lambda_2,\tilde{\lambda}_2)\cdot\cdot\mathbb{E}_{Z_2}\left[(Z_3-Z_2)\cdot\mathbb{1}\left\{-\lambda_2\sqrt{\frac{p}{n}}<Z_2<Z_3<\lambda_2\sqrt{\frac{p}{n}}\right\}\right]\\
&\quad\geq c_1(\lambda_1,\tilde{\lambda}_1,\lambda_2,\tilde{\lambda}_2)c_2\lambda_2\cdot\sqrt{\frac{p}{n}}.
\end{aligned}$$

Therefore, by choosing sufficiently small λ_1,λ_2 and large $\tilde{\lambda}_2$ (does not depend on n,p), we have $c_2c_1(\lambda_1,\tilde{\lambda}_1,\lambda_2,\tilde{\lambda}_2)\cdot\lambda_2\sqrt{\frac{p}{n}}\geq\frac{\lambda_2}{4}\sqrt{\frac{p}{n}}$. □

8.2. *Proof of the Theorem* 4.1. To prove Theorem 4.1, we begin by collecting a few important technical lemmas that will be used in the main proofs.

8.2.1. *Auxiliary lemmas.*

LEMMA 8.4. *Suppose $\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n$ i.i.d. $\sim N_p(\boldsymbol{\mu},\Sigma)$, and assume that $\hat{\boldsymbol{\mu}}$, $\hat{\Sigma}$ are the sample mean and sample covariance matrix, respectively. Let $\Gamma(s;p)=\{\boldsymbol{u}\in\mathbb{R}^p:\|\boldsymbol{u}\|_2=1,\|\boldsymbol{u}_{S^C}\|_1\leq\|\boldsymbol{u}_S\|_1$, for some $S\subset[p]$ with $|S|=s\}$, then with probability at least $1-p^{-1}$,*

$$\sup_{\boldsymbol{u}\in\Gamma(s;p)}\boldsymbol{u}^\top(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu})\lesssim\sqrt{\frac{s\log p}{n}};$$

$$\sup_{\boldsymbol{u},\boldsymbol{v}\in\Gamma(s;p)} \boldsymbol{u}^\top(\hat{\Sigma}-\Sigma)\boldsymbol{v} \lesssim \sqrt{\frac{s\log p}{n}}; \qquad \sup_{\boldsymbol{a}\in\Gamma(s;p^2)} \boldsymbol{a}^\top \operatorname{vec}(\hat{\Sigma}-\Sigma) \lesssim \sqrt{\frac{s\log p}{n}}.$$

LEMMA 8.5. *Suppose $\boldsymbol{X}_1,\dots,\boldsymbol{X}_{n_1}$ i.i.d. $\sim N_p(\boldsymbol{\mu}_1,\Sigma_1)$, $\boldsymbol{Y}_1,\dots,\boldsymbol{Y}_{n_2}$ i.i.d. $\sim N_p(\boldsymbol{\mu}_2,\Sigma_2)$, $n=\min(n_1,n_2)$ and assume that $\hat{\boldsymbol{\mu}}_1,\hat{\boldsymbol{\mu}}_2$, $\hat{\Sigma}_1,\hat{\Sigma}_2$ are the sample means and sample covariance matrices. Denote $V=\frac{1}{2}\Sigma_1\otimes\Sigma_2+\frac{1}{2}\Sigma_2\otimes\Sigma_1$ and $\hat{V}=\frac{1}{2}\hat{\Sigma}_1\otimes\hat{\Sigma}_2+\frac{1}{2}\hat{\Sigma}_2\otimes\hat{\Sigma}_1$. Assume that $\boldsymbol{\beta}=\Omega_2(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)$ and $\operatorname{vec}(D)$ has bounded ℓ_2 norm, then with probability at least $1-p^{-1}$,*

$$\|\hat{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_k\|_\infty \lesssim \sqrt{\frac{\log p}{n}}, \qquad \|(\hat{\Sigma}_k-\Sigma_k)\boldsymbol{\beta}\|_\infty \lesssim \sqrt{\frac{\log p}{n}}, \quad k=1,2;$$

$$\|\operatorname{vec}(\hat{\Sigma}-\Sigma)\|_\infty \lesssim \sqrt{\frac{\log p}{n}}; \qquad \|(\hat{V}-V)\operatorname{vec}(D)\|_\infty \lesssim \sqrt{\frac{\log p}{n}}.$$

LEMMA 8.6. *Suppose $\boldsymbol{x},\boldsymbol{y}\in\mathbb{R}^p$. Let $\boldsymbol{h}=\boldsymbol{x}-\boldsymbol{y}$. Denote $\mathcal{S}=\operatorname{supp}(\boldsymbol{y})$ and $s=|S|$. If $\|\boldsymbol{x}\|_1\le\|\boldsymbol{y}\|_1$, then $h\in\Gamma(s;p)$, that is,*

$$\|\boldsymbol{h}_{\mathcal{S}^c}\|_1\le\|\boldsymbol{h}_S\|_1.$$

LEMMA 8.7. *For any two matrices $A,B\in\mathbb{R}^{p\times p}$ with nonnegative eigenvalues,*

$$\bigl|\log|A|-\log|B|\bigr| \le \max\bigl\{\bigl|\operatorname{tr}\bigl(B^{-1}(A-B)\bigr)\bigr|, \bigl|\operatorname{tr}\bigl(A^{-1}(B-A)\bigr)\bigr|\bigr\}.$$

8.2.2. *Main proofs.* We prove the consistency of estimation of D first. The consistency of estimating $\boldsymbol{\beta}$ can be derived similarly.

Recall that

$$\begin{aligned}\hat{D}=\arg\min_{D\in\mathbb{R}^{p\times p}}\biggl\{|D|_1: \biggl\|\biggl(\frac{1}{2}\hat{\Sigma}_1\otimes\hat{\Sigma}_2+\frac{1}{2}\hat{\Sigma}_2\otimes\hat{\Sigma}_1\biggr)\operatorname{vec}(D)-\operatorname{vec}(\hat{\Sigma}_1)+\operatorname{vec}(\hat{\Sigma}_2)\biggr\|_\infty \\ \le \lambda_{1,n}\biggr\}.\end{aligned} \tag{8.1}$$

By Lemma 8.5, D is a feasible solution to (8.1) with $\lambda_{1,n}=c_1\sqrt{\frac{\log p}{n}}$ when c_1 is a sufficiently large constant. Then using Lemma 8.6, we have $\operatorname{vec}(D-\hat{D})\in\Gamma(s_1;p^2)$.

Denote $V=\frac{1}{2}\Sigma_1\otimes\Sigma_2+\frac{1}{2}\Sigma_2\otimes\Sigma_1$, $\boldsymbol{v}_\Sigma=\operatorname{vec}(\Sigma_1)-\operatorname{vec}(\Sigma_2)$ and $\hat{V}=\frac{1}{2}\hat{\Sigma}_1\otimes\hat{\Sigma}_2+\frac{1}{2}\hat{\Sigma}_2\otimes\hat{\Sigma}_1$, $\widehat{\boldsymbol{v}_\Sigma}=\operatorname{vec}(\hat{\Sigma}_1)-\operatorname{vec}(\hat{\Sigma}_2)$.

We have

$$\begin{aligned}V\operatorname{vec}(D)&=\biggl(\frac{1}{2}\Sigma_1\otimes\Sigma_2+\frac{1}{2}\Sigma_2\otimes\Sigma_1\biggr)\operatorname{vec}(D)=\operatorname{vec}\biggl(\frac{1}{2}\Sigma_1 D\Sigma_2+\frac{1}{2}\Sigma_2 D\Sigma_1\biggr)\\ &=\operatorname{vec}(\Sigma_1-\Sigma_2)=\boldsymbol{v}_\Sigma.\end{aligned}$$

In addition, over the parameter space $\Theta_p(s_1,s_2)$,

$$\|V^{-1}\|_2=\|\Omega_1\otimes\Omega_2\|_2=\|\Omega_1\|_2\cdot\|\Omega_2\|_2\le M_1^2,$$

which is followed by $\lambda_{\min}(V)\ge M_1^{-2}$.

As a consequence, by Lemma 8.4, with probability at least $1-3p^{-1}$,

$$
\begin{aligned}
&\big|(\mathrm{vec}(\hat{D})-\mathrm{vec}(D))^\top V(\mathrm{vec}(\hat{D})-\mathrm{vec}(D))\big| \\
&\quad\le \big|(\mathrm{vec}(\hat{D})-\mathrm{vec}(D))^\top(\hat{V}\,\mathrm{vec}(\hat{D})-\widehat{\boldsymbol{v}_\Sigma})\big| \\
&\qquad+\big|(\mathrm{vec}(\hat{D})-\mathrm{vec}(D))^\top(\hat{V}-V)\,\mathrm{vec}(\hat{D}))\big| \\
&\qquad+\big|(\mathrm{vec}(\hat{D})-\mathrm{vec}(D))^\top(\boldsymbol{v}_\Sigma-\widehat{\boldsymbol{v}_\Sigma})\big| \\
&\quad\lesssim \sqrt{s_1}\|\mathrm{vec}(\hat{D})-\mathrm{vec}(D)\|_2\cdot\|\hat{V}\,\mathrm{vec}(\hat{D})-\widehat{\boldsymbol{v}_\Sigma}\|_\infty \\
&\qquad+\|\mathrm{vec}(\hat{D})-\mathrm{vec}(D)\|_2\cdot\sqrt{\frac{s_1\log p}{n}}\cdot\|\mathrm{vec}(D)-\mathrm{vec}(\hat{D})\|_2 \\
&\qquad+\|\mathrm{vec}(\hat{D})-\mathrm{vec}(\hat{D})\|_2\sqrt{\frac{s_1\log p}{n}}\cdot\|\mathrm{vec}(D)\|_2 \\
&\qquad+\|\mathrm{vec}(D)-\mathrm{vec}(\hat{D})\|_2\sqrt{\frac{s_1\log p}{n}}.
\end{aligned}
\tag{8.2}
$$

In addition, since $|(\mathrm{vec}(\hat{D})-\mathrm{vec}(D))^\top V(\mathrm{vec}(\hat{D})-\mathrm{vec}(D))|\ge\lambda_{\min}(V)\|\mathrm{vec}(\hat{D})-\mathrm{vec}(D)\|_2^2\ge M_1^{-2}\|\mathrm{vec}(\hat{D})-\mathrm{vec}(D)\|_2^2$, we then have

$$
\|D-\hat{D}\|_F=\|\mathrm{vec}(\hat{D})-\mathrm{vec}(D)\|_2\lesssim\sqrt{\frac{s_1\log p}{n}}.
$$

The estimation error of $\boldsymbol{\beta}$ can be derived similarly. By Lemma 8.5, $\boldsymbol{\beta}$ is a feasible solution to (3.3) with $\lambda_{2,n}=c_2\sqrt{\frac{\log p}{n}}$ when c_2 is sufficiently large. Then using Lemma 8.6, we have $\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\in\Gamma(s_2;p)$.

Then with probability at least $1-3p^{-1}$,

$$
\begin{aligned}
&\big|(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^\top\Sigma_2(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\big| \\
&\quad\le\big|(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^\top(\hat{\Sigma}_2\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\delta}})\big|+\big|(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^\top(\hat{\Sigma}_2-\Sigma_2)\hat{\boldsymbol{\beta}})\big|+\big|(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^\top(\boldsymbol{\delta}-\hat{\boldsymbol{\delta}})\big| \\
&\quad\lesssim\sqrt{s_2}\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_2\cdot\|\hat{\Sigma}\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\delta}}\|_\infty+\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_2\cdot\sqrt{\frac{s_2\log p}{n}}\cdot\|\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\|_2 \\
&\qquad+\|\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\|_2\sqrt{\frac{s_2\log p}{n}}\cdot\|\boldsymbol{\beta}\|_2+\|\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\|_2\sqrt{\frac{s_2\log p}{n}}.
\end{aligned}
\tag{8.3}
$$

Similarly, since $\lambda_{\min}(\Sigma_2)\ge M_1^{-1}$, we have with probability at least $1-p^{-1}$,

$$
\|\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\|_2\lesssim\sqrt{\frac{s_2\log p}{n}}.
$$

8.3. *Proof of Theorem* 4.2. We note here that the notation c, C denote generic constants and their values might vary line by line. Recall that the QDA rule is

$$
1+\mathbb{1}\left\{(\boldsymbol{z}-\boldsymbol{\mu}_1)^\top D(\boldsymbol{z}-\boldsymbol{\mu}_1)-2\boldsymbol{\beta}^\top(\boldsymbol{z}-\bar{\boldsymbol{\mu}})-\log(|D\Sigma_1+I_p|)+2\log\left(\frac{\pi_1}{\pi_2}\right)>0\right\}.
$$

Let $\bar{\boldsymbol{\mu}}=(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)/2$, $Q(\boldsymbol{z})=(\boldsymbol{z}-\boldsymbol{\mu}_1)^\top D(\boldsymbol{z}-\boldsymbol{\mu}_1)-2\boldsymbol{\beta}^\top(\boldsymbol{z}-\bar{\boldsymbol{\mu}})-\log(|D\Sigma_1+I_p|)+2\log(\frac{\pi_1}{\pi_2})$, $\hat{Q}(\boldsymbol{z})=(\boldsymbol{z}-\hat{\boldsymbol{\mu}}_1)^\top\hat{D}(\boldsymbol{z}-\hat{\boldsymbol{\mu}}_1)-2\hat{\boldsymbol{\beta}}^\top(\boldsymbol{z}-\frac{\hat{\boldsymbol{\mu}}_1+\hat{\boldsymbol{\mu}}_2}{2})-\log(|\hat{D}\hat{\Sigma}_1+I_p|)+\log(\frac{\hat{\pi}_1}{\hat{\pi}_2})$, and $M(z) = Q(z) - \hat{Q}(z)$, we are going to show that there exist some constants $c, C > 0$, such that for any $M > 0$,

$$\mathbb{P}_{z\sim N_p(\boldsymbol{\mu}_1,\Sigma_1)}\left(|M(z)| > M\sqrt{\frac{(s_1+s_2)\log p}{n}}\right) \le e^{-cM} + Cp^{-1},$$

note that the above probability is taken with respect to the random samples $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{n_1}$ i.i.d. $\sim N_p(\boldsymbol{\mu}_1, \Sigma_1)$, $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_{n_2}$ i.i.d. $\sim N_p(\boldsymbol{\mu}_2, \Sigma_2)$, and $\boldsymbol{z} \sim N_p(\boldsymbol{\mu}_1, \Sigma_1)$. We will later see how we reduce the mixed distribution of the test sample to the single distribution when we calculate the classification error.

Rewrite the QDA rule as

$$\mathbb{1}\left\{(z-\boldsymbol{\mu}_1)^\top D(z-\boldsymbol{\mu}_1) - 2\boldsymbol{\beta}^\top(z-\boldsymbol{\mu}_1) + \boldsymbol{\beta}^\top(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1) - \log(|D\Sigma_1 + I_p|) + 2\log\left(\frac{\pi_1}{\pi_2}\right) > 0\right\}.$$

We first bound the estimation error of the constant term $\boldsymbol{\beta}^\top(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$. We have with probability at least $1 - p^{-1}$,

$$\begin{aligned}
&|\boldsymbol{\beta}^\top(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1) - \hat{\boldsymbol{\beta}}^\top(\hat{\boldsymbol{\mu}}_2-\hat{\boldsymbol{\mu}}_1)| \\
&\quad \le |\hat{\boldsymbol{\beta}}^\top(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1-\hat{\boldsymbol{\mu}}_2+\hat{\boldsymbol{\mu}}_1)| + \|(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^\top(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)\|_2 \\
&\quad \le \|\hat{\boldsymbol{\beta}}\|_1 \cdot \|\boldsymbol{\mu}_2-\boldsymbol{\mu}_1-\hat{\boldsymbol{\mu}}_2+\hat{\boldsymbol{\mu}}_1\|_\infty + \|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_2\|\boldsymbol{\mu}_2-\boldsymbol{\mu}_1\|_2 \\
&\quad \le \|\boldsymbol{\beta}\|_1 \cdot \|\boldsymbol{\mu}_2-\boldsymbol{\mu}_1-\hat{\boldsymbol{\mu}}_2+\hat{\boldsymbol{\mu}}_1\|_\infty + \|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_2\|\boldsymbol{\mu}_2-\boldsymbol{\mu}_1\|_2 \\
&\quad \le \sqrt{s}_2\|\boldsymbol{\beta}\|_2 \cdot \|\boldsymbol{\mu}_2-\boldsymbol{\mu}_1-\hat{\boldsymbol{\mu}}_2+\hat{\boldsymbol{\mu}}_1\|_\infty + \|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_2\|\boldsymbol{\mu}_2-\boldsymbol{\mu}_1\|_2 \lesssim \sqrt{\frac{s_2\log p}{n}}.
\end{aligned}$$

For $\log|D\Sigma_1 + I_p|$, notice that $D\Sigma_1 + I_p = \Omega_2\Sigma_1$ and the product of two positive semidefinite and symmetric matrices has nonnegative eigenvalues, followed by $(D\Sigma_1 + I_p)^{-1} = \Omega_1\Sigma_2 = (\Omega_2 - D)\Sigma_2 = I_p - D\Sigma_2$, then

$$\begin{aligned}
&\log|D\Sigma_1 + I_p| - \log|\hat{D}\hat{\Sigma}_1 + I_p| \\
&\quad \le \operatorname{tr}((D\Sigma_1+I_p)^{-1}(D\Sigma_1 - \hat{D}\hat{\Sigma}_1)) \\
&\quad = \operatorname{tr}((-D\Sigma_2+I_p)(D\Sigma_1 - \hat{D}\hat{\Sigma}_1)) \\
&\quad = \operatorname{tr}((-D\Sigma_2)(D\Sigma_1 - \hat{D}\hat{\Sigma}_1)) + \operatorname{tr}(D\Sigma_1 - \hat{D}\hat{\Sigma}_1) \qquad (8.4)\\
&\quad \le \|D\Sigma_2\|_F \cdot \|D\Sigma_1 - \hat{D}\hat{\Sigma}_1\|_F + \operatorname{tr}(D\Sigma_1 - \hat{D}\hat{\Sigma}_1) \\
&\quad \le \|D\|_F\|\Sigma_2\|_2 \cdot \|D\Sigma_1 - \hat{D}\hat{\Sigma}_1\|_F + \operatorname{tr}(D\Sigma_1 - \hat{D}\hat{\Sigma}_1) \\
&\quad \le \|D\|_F\|\Sigma_2\|_2 \cdot \|D\Sigma_1 - \hat{D}\hat{\Sigma}_1\|_F + |\operatorname{tr}(\hat{D}\Sigma_1 - \hat{D}\hat{\Sigma}_1)| + \operatorname{tr}(D\Sigma_1 - \hat{D}\Sigma_1).
\end{aligned}$$

In addition, with probability at least $1 - p^{-1}$,

$$\begin{aligned}
&\|D\Sigma_1 - \hat{D}\hat{\Sigma}_1\|_F \\
&\quad \le \|D\Sigma_1 - \hat{D}\Sigma_1\|_F + \|\hat{D}(\Sigma_1 - \hat{\Sigma}_1)\|_F \\
&\quad \le \|\Sigma_1\|_2\|D - \hat{D}\|_F + \|\Sigma_1 - \hat{\Sigma}_1\|_{2,s_1}\|\hat{D}\|_F \\
&\quad \lesssim \sqrt{\frac{s_1\log p}{n}} + \|\Sigma_1 - \hat{\Sigma}_1\|_{2,s_1}\left(\|D\|_F + \sqrt{\frac{s_1\log p}{n}}\right)
\end{aligned}$$

$$\leq \sqrt{\frac{s_1 \log p}{n}} + \sqrt{\frac{s_1 \log p}{n}}\left(\|D\|_F + \sqrt{\frac{s_1 \log p}{n}}\right) \lesssim \sqrt{\frac{s_1 \log p}{n}},$$

where $\|\Sigma_1 - \hat{\Sigma}_1\|_{2,s_1}$ is defined as

$$\|\Sigma_1 - \hat{\Sigma}_1\|_{2,s_1} := \sup_{\|\boldsymbol{u}\|_0 \leq s_1, \|\boldsymbol{u}\|_2 = 1} \|(\Sigma_1 - \hat{\Sigma}_1)\boldsymbol{u}\|_2 \lesssim \sqrt{\frac{s_1 \log p}{n}},$$

where the last inequality is similarly proved as Lemma 8.4, by using the packing number argument.

In addition, with probability at least $1 - p^{-1}$,

$$\left|\mathrm{tr}(\hat{D}\Sigma_1 - \hat{D}\hat{\Sigma}_1)\right| \leq \sqrt{s}_1 |\Sigma_1 - \hat{\Sigma}_1|_\infty \|\hat{D}\|_F \lesssim \sqrt{\frac{s_1 \log p}{n}}.$$

There is still a remaining term $\mathrm{tr}(D\Sigma_1 - \hat{D}\Sigma_1)$ in (8.4), we will leave it there and use it when we derive the distribution of the term involving z. The other direction, the upper bound of $\mathrm{tr}(D\Sigma_1 - \hat{D}\Sigma_1) - (\log|D\Sigma_1 + I_p| - \log|\hat{D}\hat{\Sigma}_1 + I_p|)$, can be derived similarly. Therefore, by symmetry, we have with probability at least $1 - p^{-1}$

$$\left|(\log|D\Sigma_1 + I_p| - \log|\hat{D}\hat{\Sigma}_1 + I_p|) - (\mathrm{tr}(D\Sigma_1 - \hat{D}\Sigma_1))\right| \lesssim \sqrt{\frac{s_1 \log p}{n}}.$$

For the term involving z, when $z \sim N_p(\boldsymbol{\mu}_1, \Sigma_1)$, we have

$$\begin{aligned}
&(z - \boldsymbol{\mu}_1)^\top D(z - \boldsymbol{\mu}_1) - (z - \boldsymbol{\mu}_1)^\top \hat{D}(z - \boldsymbol{\mu}_1) - (\mathrm{tr}(D\Sigma_1 - \hat{D}\Sigma_1)) \\
&\quad = (z - \boldsymbol{\mu}_1)^\top (\hat{D} - D)(z - \boldsymbol{\mu}_1) - (\mathrm{tr}(D\Sigma_1 - \hat{D}\Sigma_1)) \\
&\quad \overset{d}{=} z_0^\top \Sigma_1^{1/2}(\hat{D} - D)\Sigma_1^{1/2} z_0 - \mathrm{tr}(\Sigma_1^{1/2}(\hat{D} - D)\Sigma_1^{1/2}) \overset{\text{def}}{=} \sum_{i=1}^p \lambda_i (z_{0i}^2 - 1),
\end{aligned}$$

where λ_i's are the eigenvalues of $\Sigma_1^{1/2}(\hat{D} - D)\Sigma_1^{1/2}$.

Since with probability at least $1 - p^{-1}$,

$$\sqrt{\sum_{i=1}^p \lambda_i^2} = \|\Sigma_1^{1/2}(\hat{D} - D)\Sigma_1^{1/2}\|_F \leq \|\Sigma_1\|_2 \|\hat{D} - D\|_F \lesssim \sqrt{\frac{s_1 \log p}{n}},$$

and with probability at least $1 - p^{-1}$,

$$\max_i |\lambda_i| \leq \|\Sigma_1^{1/2}(\hat{D} - D)\Sigma_1^{1/2}\|_2 \leq \|\Sigma_1\|_2 \|\hat{D} - D\|_2 \lesssim \sqrt{\frac{s_1 \log p}{n}},$$

by Bernstein type inequality for subexponential random variables (see Vershynin [42]), we have for some $\tilde{c}_1 > 0$,

$$\mathbb{P}\left(\left|\sum_{i=1}^p \lambda_i (z_{0i}^2 - 1)\right| \geq t\right) \leq 2\exp\left\{-\tilde{c}_1 \min\left\{\frac{t^2}{s_1 \log p/n}, \frac{t}{\sqrt{s_1 \log p/n}}\right\}\right\},$$

which implies that for some $c_1 > 0$,

$$\mathbb{P}\left(\left|\sum_{i=1}^p \lambda_i (z_{0i}^2 - 1)\right| \geq M\sqrt{\frac{s_1 \log p}{n}}\right) \leq e^{-c_1 M} + Cp^{-1}.$$

For $(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^\top z$, when $z \sim N_p(\boldsymbol{\mu}_1, \Sigma_1)$, we have

$$(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^\top z \sim N\big((\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^\top \boldsymbol{\mu}_1, (\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^\top \Sigma_1(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\big).$$

Since with probability at least $1-p^{-1}$,

$$\big|(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^\top \boldsymbol{\mu}_1\big| \le \|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_2 \cdot \|\boldsymbol{\mu}_1\|_2 \lesssim \sqrt{\frac{s_2 \log p}{n}},$$

and with probability at least $1-p^{-1}$,

$$\big|(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^\top \Sigma_1(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\big| \le \|\Sigma_1\|_2 \cdot \|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_2^2 \le \frac{s_2 \log p}{n},$$

we have for some $c_2 > 0$,

$$\mathbb{P}\bigg(\big|(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^\top z\big| > M\sqrt{\frac{s_2 \log p}{n}}\bigg) \le e^{-c_2 M^2} + Cp^{-1}.$$

Lastly,

$$\bigg|2\log\bigg(\frac{\pi_1}{\pi_2}\bigg) - \log\bigg(\frac{\hat{\pi}_1}{\hat{\pi}_2}\bigg)\bigg| \lesssim |\hat{\pi}_1 - \pi_1| + |\hat{\pi}_2 - \pi_2|.$$

and by Hoeffding inequality, for $k \in [2]$, there are some constant $c_H > 0$, such that

$$\mathbb{P}\big(|\hat{\pi}_k - \pi_k| > t\big) \le \exp\big(-c_H \cdot nt^2\big).$$

We have for some constant $c, M_H > 0$,

$$\mathbb{P}\bigg(\bigg|2\log\bigg(\frac{\pi_1}{\pi_2}\bigg) - \log\bigg(\frac{\hat{\pi}_1}{\hat{\pi}_2}\bigg)\bigg| > M_H\sqrt{\frac{1}{n}}\bigg) \le e^{-cM_H}.$$

Therefore, there exists some $c > 0$, such that for any $M > 0$,

$$\mathbb{P}_{z\sim N_p(\boldsymbol{\mu}_1,\Sigma_1)}\bigg(M(z) > M\sqrt{\frac{(s_1+s_2)\log p}{n}}\bigg) \le e^{-cM} + Cp^{-1}.$$

Then it follows that

$$\begin{aligned}
&R(\hat{G}_{\mathrm{SDAR}}) - R_{\boldsymbol{\theta}}(G_{\boldsymbol{\theta}}^*) \\
&= \frac{1}{2}\int_{Q(z)>0} \frac{\pi_1}{(2\pi)^{p/2}|\Sigma_1|^{1/2}} e^{-1/2\cdot(z-\boldsymbol{\mu}_1)^\top \Omega_1(z-\boldsymbol{\mu}_1)}\,dz \\
&\quad + \frac{1}{2}\int_{Q(z)\le 0} \frac{\pi_2}{(2\pi)^{p/2}|\Sigma_2|^{1/2}} e^{-1/2\cdot(z-\boldsymbol{\mu}_2)^\top \Omega_2(z-\boldsymbol{\mu}_2)}\,dz \\
&\quad - \frac{1}{2}\int_{\hat{Q}(z)>0} \frac{\pi_1}{(2\pi)^{p/2}|\Sigma_1|^{1/2}} e^{-1/2\cdot(z-\boldsymbol{\mu}_1)^\top \Omega_1(z-\boldsymbol{\mu}_1)}\,dz \\
&\quad - \frac{1}{2}\int_{\hat{Q}(z)\le 0} \frac{\pi_2}{(2\pi)^{p/2}|\Sigma_2|^{1/2}} e^{-1/2\cdot(z-\boldsymbol{\mu}_2)^\top \Omega_2(z-\boldsymbol{\mu}_2)}\,dz.
\end{aligned}$$

$$\begin{aligned}
&R(\hat{G}_{\mathrm{SDAR}}) - R_{\boldsymbol{\theta}}(G_{\boldsymbol{\theta}}^*) \\
&= \frac{1}{2}\int_{Q(z)>0} \frac{1}{(2\pi)^{p/2}} e^{-1/2\cdot(z-\boldsymbol{\mu}_1)^\top \Omega_1(z-\boldsymbol{\mu}_1) - \log|\Sigma_1|/2 + \log\pi_1} \\
&\quad - \frac{1}{(2\pi)^{p/2}} e^{-1/2\cdot(z-\boldsymbol{\mu}_2)^\top \Omega_2(z-\boldsymbol{\mu}_2) - \log|\Sigma_2|/2 + \log\pi_2}\,dz
\end{aligned}$$

$$
\begin{aligned}
&- \frac{1}{2} \int_{\hat{Q}(z)>0} \frac{1}{(2\pi)^{p/2}} e^{-1/2\cdot(z-\boldsymbol{\mu}_1)^\top \Omega_1 (z-\boldsymbol{\mu}_1) - \log|\Sigma_1|/2 + \log \pi_1} \\
&- \frac{1}{(2\pi)^{p/2}} e^{-1/2\cdot(z-\boldsymbol{\mu}_2)^\top \Omega_2 (z-\boldsymbol{\mu}_2) - \log|\Sigma_2|/2 + \log \pi_2} \, dz \\
= {} & \frac{1}{2} \int_{Q(z)>0} \frac{1}{(2\pi)^{p/2}} e^{-1/2\cdot(z-\boldsymbol{\mu}_1)^\top \Omega_1 (z-\boldsymbol{\mu}_1) - \log|\Sigma_1|/2} \big(1 - e^{-Q(z)}\big) \, dz \\
&- \frac{1}{2} \int_{\hat{Q}(z)>0} \frac{1}{(2\pi)^{p/2}} e^{-1/2\cdot(z-\boldsymbol{\mu}_1)^\top \Omega_1 (z-\boldsymbol{\mu}_1) - \log|\Sigma_1|/2} \big(1 - e^{-Q(z)}\big) \, dz
\end{aligned}
$$

Then it follows

$$
\begin{aligned}
& R(\hat{G}_{\text{SDAR}}) - R_{\boldsymbol{\theta}}(G^*_{\boldsymbol{\theta}}) \\
& \quad \le \frac{1}{2} \int_{Q(z)>0, \hat{Q}(z)\le 0} \frac{1}{(2\pi)^{p/2}} e^{-1/2\cdot(z-\boldsymbol{\mu}_1)^\top \Omega_1 (z-\boldsymbol{\mu}_1) - \log|\Sigma_1|/2} \big(1 - e^{-Q(z)}\big) \, dz \\
& \quad = \frac{1}{2} \int_{Q(z)>0, Q(z)\le Q(z)-\hat{Q}(z)} \frac{1}{(2\pi)^{p/2}} e^{-1/2\cdot(z-\boldsymbol{\mu}_1)^\top \Omega_1 (z-\boldsymbol{\mu}_1) - \log|\Sigma_1|/2} \big(1 - e^{-Q(z)}\big) \, dz \\
& \quad = \frac{1}{2} \mathbb{E}_{z\sim N_p(\boldsymbol{\mu}_1, \Sigma_1)} \big[\big(1 - e^{-Q(z)}\big) \mathbb{1}\{0 < Q(z) \le M(z)\}\big] \\
& \quad = \frac{1}{2} \mathbb{E}_{z\sim N_p(\boldsymbol{\mu}_1, \Sigma_1)} \bigg[\big(1 - e^{-Q(z)}\big) \mathbb{1}\{0 < Q(z) \le M(z)\} \\
& \qquad \times \mathbb{1}\bigg\{M(z) < M \log n \sqrt{\frac{(s_1+s_2)\log p}{n}}\bigg\}\bigg] \\
& \qquad + \frac{1}{2} \mathbb{E}_{z\sim N_p(\boldsymbol{\mu}_1, \Sigma_1)} \bigg[\big(1 - e^{-Q(z)}\big) \mathbb{1}\{0 < Q(z) \le M(z)\} \\
& \qquad \times \mathbb{1}\bigg\{M(z) \ge M \log n \sqrt{\frac{(s_1+s_2)\log p}{n}}\bigg\}\bigg] \\
& \quad \le \frac{1}{2} \mathbb{E}_{z\sim N_p(\boldsymbol{\mu}_1, \Sigma_1)} \bigg[\big(1 - e^{-Q(z)}\big) \mathbb{1}\{0 < Q(z) \le M(z)\} \\
& \qquad \times \mathbb{1}\bigg\{M(z) < M \log n \sqrt{\frac{(s_1+s_2)\log p}{n}}\bigg\}\bigg] \\
& \qquad + \mathbb{P}_{z\sim N_p(\boldsymbol{\mu}_1, \Sigma_1)} \bigg(M(z) \ge M \log n \sqrt{\frac{(s_1+s_2)\log p}{n}}\bigg) \\
& \quad \lesssim \mathbb{E}_{z\sim N_p(\boldsymbol{\mu}_1, \Sigma_1)} \bigg[\big(1 - e^{-Q(z)}\big) \mathbb{1}\{0 < Q(z) \le M(z)\} \\
& \qquad \times \mathbb{1}\bigg\{M(z) < M \log n \sqrt{\frac{(s_1+s_2)\log p}{n}}\bigg\}\bigg] \\
& \qquad + n^{-1} + p^{-1} \\
& \quad \lesssim \log n \cdot \sqrt{\frac{(s_1+s_2)\log p}{n}}
\end{aligned}
$$

$$\times \mathbb{E}_{z\sim N_p(\boldsymbol{\mu}_1,\Sigma_1)}\left[\mathbb{1}\left\{0 < Q(z) \le M\log n\sqrt{\frac{(s_1+s_2)\log p}{n}}\right\}\right] + n^{-1} + p^{-1}$$
$$\lesssim \log^2 n \cdot \frac{(s_1+s_2)\log p}{n},$$

where the last inequality uses the assumption that $\sup_{|x|<\delta} f_{Q,\boldsymbol{\theta}}(x) < M_2$.

Acknowledgments. The authors would like to thank the anonymous referees, an Associate Editor and the editor for their constructive comments that improved the quality of this paper.

Funding. The first author was supported by NSF Grant DMS-1712735 and NIH Grant R01 GM-123056.

The second author was supported by NSF Grant DMS-2015378.

SUPPLEMENTARY MATERIAL

Supplement to "A convex optimization approach to high-dimensional sparse quadratic discriminant analysis" (DOI: 10.1214/20-AOS2012SUPP; .pdf). The supplement provides a detailed proof of Theorem 4.3, which is the lower bound of the misclassification error for the high-dimensional QDA problem with sparsity assumptions, and proofs of Theorem 5.1 and 5.2, the convergence rate of CSQDA under the Gaussian Copula Model. In addition, proofs of the technical lemmas used in the proofs of the main results are given. Some additional simulation results are also presented in the supplement.

REFERENCES

[1] ALON, Y., BARKAI, N., NOTTERMAN, DA., GISH, K., YBARRA, S., MACK, D. and LEVINE, AJ. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *PNAS* **96(12)** 6745–6750. https://doi.org/10.1073/pnas.96.12.6745

[2] ANDERSON, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*, 3rd ed. *Wiley Series in Probability and Statistics*. Wiley Interscience, Hoboken, NJ. MR1990662

[3] AZIZYAN, M., SINGH, A. and WASSERMAN, L. (2013). Minimax theory for high-dimensional Gaussian mixtures with sparse mean separation. In *Advances in Neural Information Processing Systems* 2139–2147.

[4] BANDYOPADHYAY, S., MEHTA, M., KUO, D., SUNG, M.-K., CHUANG, R., JAEHNIG, E. J., BODENMILLER, B., LICON, K., COPELAND, W. et al. (2010). Rewiring of genetic networks in response to DNA damage. *Science* **330** 1385–1389.

[5] BARBER, R. F. and KOLAR, M. (2018). ROCKET: robust confidence intervals via Kendall's tau for transelliptical graphical models. *Ann. Statist.* **46** 3422–3450. MR3852657 https://doi.org/10.1214/17-AOS1663

[6] BICKEL, P. J. and LEVINA, E. (2004). Some theory of Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations. *Bernoulli* **10** 989–1010. MR2108040 https://doi.org/10.3150/bj/1106314847

[7] CAI, T. and LIU, W. (2011). A direct estimation approach to sparse linear discriminant analysis. *J. Amer. Statist. Assoc.* **106** 1566–1577. MR2896857 https://doi.org/10.1198/jasa.2011.tm11199

[8] CAI, T. T., LIANG, T. and ZHOU, H. H. (2015). Law of log determinant of sample covariance matrix and optimal estimation of differential entropy for high-dimensional Gaussian distributions. *J. Multivariate Anal.* **137** 161–172. MR3332804 https://doi.org/10.1016/j.jmva.2015.02.003

[9] CAI, T. T., MA, J. and ZHANG, L. (2019). CHIME: Clustering of high-dimensional Gaussian mixtures with EM algorithm and its optimality. *Ann. Statist.* **47** 1234–1267. MR3911111 https://doi.org/10.1214/18-AOS1711

[10] CAI, T. T. and ZHANG, L. (2018). High-dimensional Gaussian copula regression: Adaptive estimation and statistical inference. *Statist. Sinica* **28** 963–993. MR3791096

[11] CAI, T. T. and ZHANG, L. (2019). High dimensional linear discriminant analysis: Optimality, adaptive algorithm and missing data. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **81** 675–705. MR3997097

[12] CAI, T. T and ZHANG, L. (2021). Supplement to "A convex optimization approach to high-dimensional sparse quadratic discriminant analysis." https://doi.org/10.1214/20-AOS2012SUPP

[13] CANDES, E. and ROMBERG, J. (2005). l1-magic: Recovery of sparse signals via convex programming.

[14] CARMON, Y., RAGHUNATHAN, A., SCHMIDT, L., DUCHI, J. C. and LIANG, P. S. (2019). Unlabeled data improves adversarial robustness. In *Advances in Neural Information Processing Systems* 11192–11203.

[15] CHURCHILL, G. A. and IACOBUCCI, D. (2006). *Marketing Research*: *Methodological Foundations*. Dryden Press, New York.

[16] DAN, C., WEI, Y. and RAVIKUMAR, P. (2020). Sharp statistical guarantees for adversarially robust Gaussian classification. Preprint. Available at arXiv:2006.16384.

[17] DENG, Z., KAMMOUN, A. and THRAMPOULIDIS, C. (2019). A model of double descent for high-dimensional binary linear classification. Preprint. Available at arXiv:1911.05822.

[18] DETTLING, M. (2004). BagBoosting for tumor classification with gene expression data. *Bioinformatics* **20(18)** 3853–3593. https://doi.org/10.1093/bioinformatics/bth447

[19] EFRON, B. (2010). *Large-Scale Inference*: *Empirical Bayes Methods for Estimation*, *Testing*, *and Prediction*, Cambridge University Press, New York.

[20] HAN, F. and LIU, H. (2017). Statistical analysis of latent generalized correlation matrix estimation in transelliptical distribution. *Bernoulli* **23** 23–57. MR3556765 https://doi.org/10.3150/15-BEJ702

[21] HAN, F., ZHAO, T. and LIU, H. (2013). CODA: High dimensional copula discriminant analysis. *J. Mach. Learn. Res.* **14** 629–671. MR3033343

[22] HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2009). *The Elements of Statistical Learning*: *Data Mining*, *Inference*, *and Prediction*, 2nd ed. *Springer Series in Statistics*. Springer, New York. MR2722294 https://doi.org/10.1007/978-0-387-84858-7

[23] HILL, S. M., HEISER, L. M., COKELAER, T., UNGER, M., NESSER, N. K., CARLIN, D. E., ZHANG, Y., SOKOLOV, A., PAULL, E. O. et al. (2016). Inferring causal molecular networks: Empirical assessment through a community-based effort. *Nat. Methods* **13** 310–318.

[24] INAM, F., INAM, A., MIAN, M. A., SHEIKH, A. A. and AWAN, H. M. (2018). Forecasting bankruptcy for organizational sustainability in Pakistan: Using artificial neural networks, logit regression, and discriminant analysis. *J. Econ. Adm. Sci.*

[25] JIANG, B., WANG, X. and LENG, C. (2018). A direct approach for sparse quadratic discriminant analysis. *J. Mach. Learn. Res.* **19** Paper No. 31, 37. MR3862438

[26] JOMBART, T., DEVILLARD, S. and BALLOUX, F. (2010). Discriminant analysis of principal components: A new method for the analysis of genetically structured populations. *BMC Genet.* **11** 94.

[27] JU, F., SUN, Y., GAO, J., HU, Y. and YIN, B. (2019). Probabilistic linear discriminant analysis with vectorial representation for tensor data. *IEEE Trans. Neural Netw. Learn. Syst.* **30** 2938–2950. MR4015091

[28] KINI, G. and THRAMPOULIDIS, C. (2020). Analytic study of double descent in binary classification: The impact of loss. Preprint. Available at arXiv:2001.11572.

[29] KOÇHAN, N., TÜTÜNCÜ, G. Y., SMYTH, G. K., GANDOLFO, L. C. and GINER, G. (2019). qtQDA: Quantile transformed quadratic discriminant analysis for high-dimensional RNA-seq data. *BioRxiv* 751370.

[30] LI, H.-D., XU, Q.-S. and LIANG, Y.-Z. (2018). libPLS: An integrated library for partial least squares regression and linear discriminant analysis. *Chemom. Intell. Lab. Syst.* **176** 34–43.

[31] LI, Q. and SHAO, J. (2015). Sparse quadratic discriminant analysis for high dimensional data. *Statist. Sinica* **25** 457–473. MR3379082

[32] LIN, Y. and JEON, Y. (2003). Discriminant analysis through a semiparametric model. *Biometrika* **90** 379–392. MR1986654 https://doi.org/10.1093/biomet/90.2.379

[33] LIU, H., HAN, F., YUAN, M., LAFFERTY, J. and WASSERMAN, L. (2012). High-dimensional semiparametric Gaussian copula graphical models. *Ann. Statist.* **40** 2293–2326. MR3059084 https://doi.org/10.1214/12-AOS1037

[34] MAI, Q. and ZOU, H. (2015). Sparse semiparametric discriminant analysis. *J. Multivariate Anal.* **135** 175–188. MR3306434 https://doi.org/10.1016/j.jmva.2014.12.009

[35] MAI, Q., ZOU, H. and YUAN, M. (2012). A direct approach to sparse discriminant analysis in ultra-high dimensions. *Biometrika* **99** 29–42. MR2899661 https://doi.org/10.1093/biomet/asr066

[36] MAMMEN, E. and TSYBAKOV, A. B. (1999). Smooth discrimination analysis. *Ann. Statist.* **27** 1808–1829. MR1765618 https://doi.org/10.1214/aos/1017939240

[37] MIGNACCO, F., KRZAKALA, F., LU, Y. M. and ZDEBOROVÁ, L. (2020). The role of regularization in classification of high-dimensional noisy Gaussian mixture. *ICML* 2020.

[38] NEYKOV, M., NING, Y., LIU, J. S. and LIU, H. (2018). A unified theory of confidence regions and testing for high-dimensional estimating equations. *Statist. Sci.* **33** 427–443. MR3843384 https://doi.org/10.1214/18-STS661

[39] SCHMIDT, L., SANTURKAR, S., TSIPRAS, D., TALWAR, K. and MADRY, A. (2018). Adversarially robust generalization requires more data. In *Advances in Neural Information Processing Systems* 5014–5026.

[40] SINGH, D., FEBBO, PG., ROSS, K., JACKSON, D., MANOLA, J., LADD, C., TAMAYO, P., RENSHAW, A., D'AMICO, A., RICHIE, JP., LANDER, ES., LODA, M., KANTOFF, P., GOLUB, TR. and SELLERS, WR. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell* **1(2)** 203–209. https://doi.org/10.1016/s1535-6108(02)00030-2

[41] TSYBAKOV, A. B. (2009). *Introduction to Nonparametric Estimation*. *Springer Series in Statistics*. Springer, New York. MR2724359 https://doi.org/10.1007/b13794

[42] VERSHYNIN, R. (2010). *Introduction to the non-asymptotic analysis of random matrices*, Cambridge University Press, New York.

[43] WRIGHT, J., YANG, A. Y., GANESH, A., SASTRY, S. S. and MA, Y. (2009). Robust face recognition via sparse representation. *IEEE Trans*. *Pattern Anal*. *Mach*. *Intell*. **31** 210–227.

[44] XIA, Y., CAI, T. and CAI, T. T. (2015). Testing differential networks with applications to the detection of gene-gene interactions. *Biometrika* **102** 247–266. MR3371002 https://doi.org/10.1093/biomet/asu074

[45] XU, M., ZHANG, D. and WU, W. B. (2014). L^2 asymptotics for high-dimensional data. Preprint. Available at arXiv:1405.7244.

[46] XUE, L. and ZOU, H. (2012). Regularized rank-based estimation of high-dimensional nonparanormal graphical models. *Ann*. *Statist*. **40** 2541–2571. MR3097612 https://doi.org/10.1214/12-AOS1041

[47] ZHAO, S. D., CAI, T. T. and LI, H. (2014). Direct estimation of differential networks. *Biometrika* **101** 253–268. MR3215346 https://doi.org/10.1093/biomet/asu009

[48] ZHAO, Y. and WEGKAMP, M. (2014). Semiparametric Gaussian copula classification. Preprint. Available at arXiv:1411.2944.