Submitted to the Annals of Statistics

HIGH DIMENSIONAL QUADRATIC DISCRIMINANT ANALYSIS: OPTIMALITY AND PHASE TRANSITIONS

BY WANJIE WANG[1,†,*], JINGJING WU[2,‡] AND ZHIGANG YAO[1,§,†]

[1]*Department of Statistics and Data Science, National University of Singapore,* [*]*wanjie.wang@nus.edu.sg;* [†]*zhigang.yao@nus.edu.sg*

[2]*Department of Mathematics and Statistics, University of Calgary, jinwu@ucalgary.ca*

Consider a two-class classification problem where we observe samples (X_i, Y_i) for $i = 1, \cdots, n$, $X_i \in \mathcal{R}^p$ and $Y_i \in \{0, 1\}$. Given $Y_i = k$, X_i is assumed to follow a multivariate normal distribution with mean $\mu_k \in \mathcal{R}^k$ and covariance matrix Σ_k, $k = 0, 1$. Supposing a new sample X from the same mixture is observed, our goal is to estimate its class label Y. Such a high-dimensional classification problem has been studied thoroughly when $\Sigma_0 = \Sigma_1$. However, the discussions over the case $\Sigma_0 \neq \Sigma_1$ are much less over the years.

This paper presents the quadratic discriminant analysis (QDA) for the weak signals (QDAw) algorithm, and the QDA with feature selection (QDAfs) algorithm. QDAfs applies Partial Correlation Screening in [17] to estimate $\hat{\Omega}_0$ and $\hat{\Omega}_1$, and then applies a hard-thresholding on the diagonals of $\hat{\Omega}_0 - \hat{\Omega}_1$. QDAfs further includes the linear term $d^\top X$, where d is achieved by a hard-thresholding on $\hat{\Omega}_1\hat{\mu}_1 - \hat{\Omega}_0\hat{\mu}_0$. QDAfs achieves theoretical optimality and outperforms recent works on the linear discriminant analysis of high-dimensional data on a real data set.

We further propose the rare and weak model to model the signals in $\Omega_0 - \Omega_1$ and $\mu_0 - \mu_1$. Based on the signal weakness and sparsity in $\mu_0 - \mu_1$, we propose two ways to estimate labels: 1) QDAw for weak but dense signals; 2) QDAfs for relatively strong but sparse signals. We figure out the classification boundary on the 4-dim parameter space: 1) Region of possibility, where either QDAw or QDAfs will achieve a mis-classification error rate of 0; 2) Region of impossibility, where all classifiers will have a constant error rate. The numerical results from real datasets support our theories and demonstrate the necessity and superiority of using QDA over LDA for classification.

1. Introduction. Consider a two-class classification problem, where we have n labeled training samples (X_i, Y_i), $i = 1, \ldots, n$. Here, X_i's are p-dimensional feature vectors and $Y_i \in \{0, 1\}$ are the corresponding class labels. X_i is assumed to have mean μ_k and covariance matrix Σ_k where $k = Y_i \in \{0, 1\}$. The goal is to estimate the label of a new observation X. A significant amount of work has been done in this field; see [1, 16, 24].

Fisher's Linear Determinant Analysis (LDA) in [14] utilizes a weighted average of the features of the test sample to make a prediction. The optimal weight vector for LDA where the two classes are assumed to share the same correlation structure $\Omega = \Sigma_0^{-1} = \Sigma_1^{-1}$ satisfies

$$d \propto \Omega(\mu_1 - \mu_0). \tag{1.1}$$

[*]Equal contribution.

[†]Supported by MOE Start-up grant 155-000-173-133 and Tier 1 grant 0004813-00-00.

[‡]Supported by NSERC Discovery Grants RGPIN-2018-04328.

[§]Supported by MOE Tier 2 grant 155-000-184-114 and Tier 1 grant 0004813-00-00.

MSC2020 subject classifications: Primary 62H30; secondary 62F05.

Keywords and phrases: High-dimensional classification, quadratic discriminant analysis, phase transitions, rare and weak signals.

When $n \gg p$, the mean vectors μ_0 and μ_1, and the precision matrix Ω can be easily estimated. Therefore Fisher's LDA is approachable.

In modern analytical approaches, high-dimensional data have flooded, where a prominent number of measurements of features, often in the millions, are gathered for a single subject ([8]). Although the number of features is huge, usually only a small portion of them are regarded as relevant to the classification decision, but these are not known in advance. In this sense, the traditional classification methods will lose power because of the large amount of noise. Methods have been proposed to reduce the noises in methods; see [9, 10, 12, 18, 30]. For such high dimensional classification problems, recent developments provide methods to estimate the precision matrix Ω in LDA; see [5, 17].

However, LDA still faces two problems:

- It does not account for the information from Ω_0 and Ω_1. The features in two classes may not share the same conditional independence structure, however, LDA doesn't take this information into consideration. Such difference will impact the distribution of the new variable X and the estimation of d.
- The theoretical analysis for the case $\Omega_0 \neq \Omega_1$ is short of discussion. Actually, the analysis of error rates is quite complicated for the high dimensional model, even in the simplest case where $\Omega_0 = \Omega_1 = I_p$, as the signals are rare and weak. An extensive discussion may be found in [6, 7, 12, 20, 22, 23].

These two problems motivates us to derive an algorithm and relative theoretical analysis for the case $\Omega_0 \neq \Omega_1$.

Inherited from the previous studies, many applications in high-dimensional classifier problems share the same aspects: (1) The signals in the mean vector are comparatively rare and weak; and (2) the precision matrices Ω_0 and Ω_1 are sparse. Such sparsity properties guide us to propose rare and weak model about them and solve the problem.

In this paper, we propose a rare and weak model for the high-dimensional classification problem. In the new model, we allow $\Omega_0 \neq \Omega_1$ under both the diagonals and off-diagonals with different parameterizations. It is a more reasonable fit with real data than the LDA model. To parameterize (μ_0, μ_1) and (Ω_0, Ω_1), we normalize the data by centering the mean vector and scaling features to have unit variance. These are standard processing steps in the real data analysis, which will be discussed later with more details. In theoretical analysis, we will see that using $\Omega_0 - \Omega_1$ improves the classification accuracy.

This paper further explores the high-dimensional classification problem in the following aspects:

- We propose a rare and weak model to account for both (μ_0, μ_1) and (Ω_0, Ω_1). It models the weakness and sparsity of $\mu_0 - \mu_1$ and the diagonals and off-diagonals of $\Omega_0 - \Omega_1$. We tie all the parameters to p, which allows us to explore the phase diagram of the classification problem.
- We propose several algorithms: the Quadratic Discriminant Analysis with Feature Selection (QDAfs) algorithm that works when the signals in $\mu_0 - \mu_1$ are relatively strong and sparse; and the Quadratic Discriminant Analysis for Weak signals (QDAw) algorithm that works when the signals in $\mu_0 - \mu_1$ are weak and relatively sparse.
- We derive the phase diagram for the high-dimensional classification problem when $\Omega_0 \neq \Omega_1$. We find that the region that QDAfs and QDAw will give satisfactory classification results. We also find out the region that no classifier will succeed, i.e. region of impossibility. Under the rare and weak model, the success region of QDAw/QDAfs and the region of impossibility can form the whole phase diagram, which proves the optimality of QDAw/QDAfs.

Our method performs better than the optimal high-dimensional LDA method on a real data set in the numerical analysis section, which suggests that the second-order information should be incorporated to improve the classification results.

1.1. *Quadratic Discriminant Analysis on high-dimensional data.* Consider the two-class classification problem mentioned above, where we observe n training samples (X_i, Y_i), $i = 1, \ldots, n$. Given $Y_i = k$, we assume the feature vector $X_i \in \mathcal{R}^p$ follows a multivariate normal distribution with mean μ_k and covariance matrix $\Sigma_k = \Omega_k^{-1}$. Let X denote an independent test sample from the same population; then,

$$X|Y \sim (1-Y)N(\mu_0, \Omega_0^{-1}) + YN(\mu_1, \Omega_1^{-1}). \tag{1.2}$$

We would like to classify X as being from either $Y = 0$ or $Y = 1$.

For two-population classification problems, the QDA method is commonly used to exploit both the mean and covariance information; see [16, 27]. Consider the ideal case that both μ_k and Ω_k are known, $k = 0, 1$. The ratio of the likelihood functions in two classes gives the optimal classifier, which results in the QDA method, that

$$\begin{aligned} \hat{Y} = I\Big\{ & X^\top (\Omega_0 - \Omega_1) X - 2\big(\mu_0^\top \Omega_0 - \mu_1^\top \Omega_1\big) X \\ & + \big(\mu_0^\top \Omega_0 \mu_0 - \mu_1^\top \Omega_1 \mu_1 + \ln|\Omega_1| - \ln|\Omega_0|\big) \ > \ 0 \Big\}, \end{aligned} \tag{1.3}$$

where $I(A)$ is the indicator function of event A. If $P(Y_i = 1) \neq 0.5$, an additional term $2\ln\frac{P(Y_i=1)}{1-P(Y_i=1)}$ on the right-hand side of the inequality will improve the accuracy. However, as such term does not have effects on the possibility and impossibility regions, our analysis applied to the balanced case (i.e. $P(Y_i = 1) = 0.5$) will suffice.

For real data, the parameters μ_0, μ_1, Ω_0 and Ω_1 are all unknown. Estimation of them is challenging in the high-dimensional setting with $p \gg n$. There are various extensions of QDA for the high-dimensional data; see [2, 31, 32]. When the signals are sparse, it is modified accordingly with sparsity assumptions on Σ_1, Σ_0 and $\mu_1 - \mu_0$; see [13, 19, 26]. Since sparsity assumptions on the precision matrices Ω_0 and Ω_1 are more commonly seen in applications ([25, 34, 35]), which means sparse conditional dependency between features, we propose an approach based on the sparsity of Ω_0, Ω_1 and μ.

We begin with estimating these parameters in the high-dimensional setting.

- Updated Partial Correlation Screening (PCS) approach on precision matrices Ω_0 and Ω_1.
  Step 1. Estimate Ω_0 and Ω_1 by PCS in [17], denoted as $\hat{\Omega}_0$ and $\hat{\Omega}_1$.
  Step 2. Let $\hat{\Omega}_{\text{diff}} = \hat{\Omega}_0 - \hat{\Omega}_1$. For each diagonal $\hat{\Omega}_{\text{diff}}(i,i)$, update it as $\hat{\Omega}_{\text{diff}}(i,i) = \hat{\Omega}_{\text{diff}}(i,i) 1\{|\hat{\Omega}_{\text{diff}}(i,i)| \leq 2\sqrt{2\ln p/n}\}$.
- Estimation of the linear component $(\Omega_0\mu_0 - \Omega_1\mu_1)^\top X$.
  Weak signals: Let $\hat{\mu} = a * \mathbf{1}$, where $\mathbf{1}$ is a vector with all ones and a is a constant. Estimate it by $g_w(X; \hat{\mu}, \hat{\Omega}_0, \hat{\Omega}_1) = \hat{\mu}^\top(\hat{\Omega}_0 + \hat{\Omega}_1)X$.
  Strong signals: Let $d = \hat{\Omega}_1\hat{\mu}_1 - \hat{\Omega}_0\hat{\mu}_0$, where $\hat{\mu}_1$ and $\hat{\mu}_0$ are the sample mean vectors of class 1 and 0. Let $d^{(t)}(j) = 1\{|d(j) \geq t|\}$. Estimate it by $g_s(X; \hat{\mu}, \hat{\Omega}_0, \hat{\Omega}_1) = (d \circ d^{(t)})^\top X$.

With the estimates $\hat{\Omega}_0$, $\hat{\Omega}_1$ and $g(X; \hat{\mu}, \hat{\Omega}_0, \hat{\Omega}_1)$, the high-dimensional QDA is proposed in Table 1. We call the algorithm with $g_w(X; \hat{\mu}, \hat{\Omega}_0, \hat{\Omega}_1)$ as QDA with weak signals (QDAw), and the algorithm with $g_s(X; \hat{\mu}, \hat{\Omega}_0, \hat{\Omega}_1)$ as QDA with a feature-selection step (QDAfs).

There are multiple high-dimensional precision matrix estimation methods; see [5, 11, 17]. We estimate Ω_0 and Ω_1 with the Partial Correlation Screening (PCS) approach in [17] for this algorithm. PCS has good control on the Frobenius norm of $\hat{\Omega}_k - \Omega_k$, which is the main

TABLE 1
Algorithm 1: Pseudocode for QDA on high-dimensional data

	Input: data points (X_i, Y_i), $1 \le i \le n$; threshold $t > 0$; new data point X.
	Output: label $\hat{Y}$.
1.	Parameter Estimation: Estimate $\hat{\Omega}_0$, $\hat{\Omega}_1$, $\hat{\Omega}_{\text{diff}}$, $\hat{\mu}$ and d according to the procedure as above.
2.	Define a constant C according to the parameters; details in the algorithms in later sections.
3.	Let $g(X) = g_s(X)$ if $\mu_1 - \mu_0$ has strong signals or $g(X) = g_w(X)$ if $\mu_1 - \mu_0$ has only weak signals.
4.	QDA Score: Calculate the QDA score $Q(X) = X^\top \Omega_{\text{diff}} X + 2g(X) + C$.
5.	Prediction: Predict $\hat{Y} = I\{Q(X) > 0\}$.

factor in the error analysis of QDA. The thresholding step on the diagonals of $\hat{\Omega}_0 - \hat{\Omega}_1$ is as an adjustment on the element-wise error of PCS, at the order of $\sqrt{\ln p/n}$. Without the thresholding step, the random error in $X^\top \hat{\Omega}_{\text{diff}} X$ is large enough to cover the truth if signals in Ω_{diff} and $\mu_0 - \mu_1$ are weak. The theoretical limit of QDA with PCS can be found in Proposition 2.2, Theorems 2.3 and 2.4.

When the signals are strong, we propose a feature selection step on $d = \hat{\Omega}_1\hat{\mu}_1 - \hat{\Omega}_0\hat{\mu}_0$ instead of $\hat{\mu}_1 - \hat{\mu}_0$. The inclusion of precision matrix in the feature selection step has been shown optimality in the linear classifier case where $\Omega_0 = \Omega_1 = \Omega$. In [12], it has been proved such innovated thresholding is better than the thresholding on $\hat{\mu}$ or $\Omega^{-1/2}\hat{\mu}$. When it comes to quadratic forms, we borrowed this idea.

When the signals are weak, we suggest estimating $\hat{\mu}$ as a constant vector. It can be seen as a simple aggregation of all the features. The constants of it reduce the random error and hence achieve the optimal boundary; see Theorem 1.2. In the supplementary material [29, Section A], we have explored the performance of the original QDA in the region where the signals in μ are weak. We have proved two theorems about the phase transition phenomenon of QDA/QDAfs when Ω_1 is known or unknown. By QDA/QDAfs, there must be $\max\{\|\mu_0 - \mu_1\|^2, \|\Omega_1 - I\|^2\} \gg \sqrt{p/n}$ for successful classification when the signals in $\mu_0 - \mu_1$ are weak. There is a gap between this upper bound and the statistical lower bound in Theorem 1.1. By QDAw, this gap will be overcome.

Finally, in Step 2 we find the constant C by minimizing the training error for real data sets. Actually, in our detailed algorithms in Sections 2, we define C clearly based on the scenarios in concern.

1.2. *Asymptotic rare and weak signal model.* We propose a rare and weak signal model for both mean vectors and precision matrices.

The two classes have mean vectors $\mu_0 \in \mathcal{R}^p$ and $\mu_1 \in \mathcal{R}^p$. With a location shifting of the distance $\frac{1}{2}|\mu_1 - \mu_0|$, we can take $\mu_1 = \mu$ and $\mu_0 = -\mu$. Hence, the signals in μ_i are the non-zeros in μ. We model μ as

$$\mu_i \overset{i.i.d.}{\sim} (1-\epsilon_p)\mathcal{M}_0 + \epsilon_p \mathcal{H}, \quad i = 1, \ldots, p, \tag{1.4}$$

where $\mathcal{M}_0$ is the point mass at 0 and $\mathcal{H}$ is a distribution that concentrates at τ_p and has no point mass at 0. Hence, the density of the signals can be captured by ϵ_p and the strength can be captured by τ_p. To model the sparsity and weakness, we assume that when $p \to \infty$,

$$\epsilon_p \to 0, \quad \tau_p \to 0. \tag{1.5}$$

When it comes to delicate theoretical analysis, we assume the signals have the same signs and strengths, which means $\mathcal{H} = \mathcal{M}_\tau$, the point mass at τ. Such assumption is generally used in high-dimensional applications to facilitate the theoretical analysis; see [6, 12, 23].

The covariance matrix of class k is Σ_k. Say for each feature j, its conditional variances given $Y = 0$ and $Y = 1$ share the same main term; otherwise the signal in the variances

are strong enough. Let D_0 be the diagonal matrix where the diagonals are the variances of features in Class 0. We normalize Σ_k by $D_0^{-1/2}\Sigma_k D_0^{-1/2}$ so that Σ_0 have diagonals as 1 and Σ_1 have diagonals as $1+o(1)$. So we suppose the diagonals of Ω_0 and Ω_1 are around 1 without loss of generality. Let $D_\Omega^{(k)} = Diag(\Omega_k)$. We model it as follows:

$$D_\Omega^{(k)}(i,i) \overset{i.i.d.}{\sim} 1+\mathcal{D}_p, \quad 1\le i\le p, \ \ k=0,1. \tag{1.6}$$

Here, $\mathcal{D}_p$ is a distribution with the magnitude concentrating at ξ_p.

In many applications, Ω_k's, instead of Σ_k's, have comparatively small number of non-zero entries in each row. We model the non-zeros on the off-diagonals as V, where

$$V_{ij}^{(k)} = V_{ij}^{(k)} \overset{i.i.d.}{\sim} (1-\nu_p)\mathcal{M}_0 + \frac{\nu_p}{2}\mathcal{M}_{\eta_p} + \frac{\nu}{2}\mathcal{M}_{-\eta_p}, \ \ k=0,1, \ \ 1\le i<j\le p, \tag{1.7}$$

where $\mathcal{M}_{\eta_p}$ and $\mathcal{M}_{-\eta_p}$ are the point mass at η_p and $-\eta_p$ respectively. The analysis still holds when M_{η_p} and $M_{-\eta_p}$ are replaced by a symmetric distribution that concentrates on η_p and $-\eta_p$ with no point mass at 0. The information on diagonals and off-diagonals are modelled by two separate parameters because they have different effects on the clustering results; see Theorems 2.3 and 2.4.

Combine the modelling on the diagonals and off-diagonals, the precision matrices follow

$$\Omega_k \ = \ \Sigma_k^{-1} \ = \ D_\Omega^{(k)} + V^{(k)}, \quad k=0,1. \tag{1.8}$$

To model the sparsity and weakness of signals in Ω_0 and Ω_1, we assume

$$\nu_p \to 0, \qquad \eta_p \to 0, \qquad \xi_p \to 0. \tag{1.9}$$

In our analysis, we consider the case when Ω_0 is known and Ω_0 is unknown. In the former case, we can set $X = \Omega_0^{1/2}X$ and update Ω_1 and μ accordingly. Because of the sparsity in Ω_0, the updated Ω_1 and μ are still sparse. Hence, without loss of generality, we assume $\Omega_0 = I$ and assume Ω_1 follows (1.7) and (1.8). The model that satisfies (1.4) – (1.9) is called rare and weak model.

One aim of this paper is to derive the statistical limits for the classification problem on the phase diagram. To derive it, we should tie all the parameters to p by some constant parameters. The sample size n goes to infinity at a slower rate than p, so we tie the sample size n to p by

$$n = n_p = p^\delta, \qquad 0<\delta<1. \tag{1.10}$$

For μ, we define the signal sparsity parameter ϵ and the weakness parameter τ as

$$\epsilon_p \ = \ p^{-\zeta}, \quad \tau_p \ = p^{-\theta}, \quad 0<\zeta,\theta<1. \tag{1.11}$$

For the precision matrices Ω_i, $i=0,1$, we similarly define the parameters as

$$\eta_p = p^{-\alpha}, \quad \nu_p = p^{-\beta}, \quad \xi_p = p^{-\gamma}, \quad 0<\alpha,\gamma<1, \ \ 0<\beta<2. \tag{1.12}$$

Here, α, β, γ, ζ, θ, and δ are all constants. The regions of interest can be interpreted as the regions on the space formed by these parameters.

Finally, to guarantee that $\Omega_k = D_\Omega^{(k)} + V^{(k)}$ is a positive definite matrix, $V^{(k)}$ must be weak enough so that Ω_k is positive definite. According to Lemma 4.3, this requirement is satisfied with high probability under the condition

$$\beta > 1-2\alpha. \tag{1.13}$$

Hence, we discuss the regions under this condition only. The model that satisfies (1.4) – (1.13) is called asymptotic rare and weak model for classification.

1.3. *Phase transitions.* Under the ARW model, we analyze the regions of possibility and impossibility for any classifiers. In detail, we calibrate the impact of quadratic terms on classification in the following terms:

- the possibility and impossibility regions for the classification problem under the ideal case and the case that Ω_0 and Ω_1 are known.
- the possibility and impossibility regions for the classification problem when Ω_0 and Ω_1 are unknown but some sparsity conditions of them are satisfied.

We summarize our results about the first part here. When all parameters are unknown, we present the conditions and results by Theorems 2.3 and 2.4 in Section 2.

Define a function

$$\rho_\delta(\zeta) = \begin{cases} 1/2 - \zeta, & 0 < \zeta \le (1-\delta)/2, \\ \delta/2, & (1-\delta)/2 < \zeta \le 1-\delta, \\ (1-\zeta)/2, & 1-\delta < \zeta < 1. \end{cases} \tag{1.14}$$

Such a function can be found in multiple works about high-dimensional problems in the analysis of lower bounds; see [21, 23]. In different settings, the meaning of this function is different.

THEOREM 1.1. [Lower bound] Under the ARW model with $\Omega_0 = I$, if $\|\Omega_1 - I\|_F^2 \to 0$ and $\theta > \rho_\delta(\zeta)$, then for any classifier L, when $p \to \infty$, there is

$$\text{Mis-classification Rate of } L \ge 1/2.$$

Let QDAw be the QDA classifier with $\hat{\mu} = a * \mathbf{1}$ and $g(X) = g_w(X)$ as the estimation in weak signal case; and let QDAfs be the QDA classifier with $g(X) = g_s(X; \hat{\mu}, \Omega_1, I)$ as the strong signal case. Details of the two algorithms in Algorithms 3 and 4. They give the matching upper bound as the following theorem.

THEOREM 1.2. [Upper Bound] Under the ARW model with $\Omega_0 = I$,

(i) the mis-classification rate of $QDAw$ with $a = p^c$ for an arbitrary constant $0 < c < 1$ goes to 0 when $p \to \infty$, if $\theta \ge \delta/2$ and one of the following conditions hold
  (a) $\|\Omega_1 - I\|_F^2 \gg p^c$; or,
  (b) $\theta < \rho_\delta(\zeta)$;

(ii) the mis-classification rate of $QDAfs$ goes to 0 when $p \to \infty$, if $\theta < \delta/2$ and one of the following conditions hold
  (a) $\|\Omega_1 - I\|_F^2 \to \infty$; or,
  (b) $\theta < \rho_\delta(\zeta)$.

For QDAw, the constant c can be chosen arbitrarily. When $\|\Omega_1 - I\| \to \infty$, i.e. $\max\{1 - 2\gamma, 2 - 2\alpha - \beta\} > 0$, we can always choose $c = \frac{1}{2}\max\{1 - 2\gamma, 2 - 2\alpha - \beta\}$ so that the inequality holds. Hence, the two boundaries match.

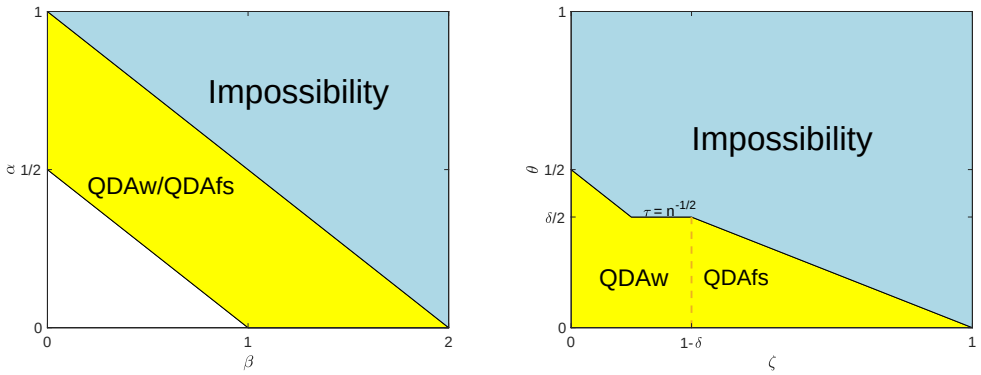
Theorems 1.1–1.2 show the determining factor of the classification problem contains two parts, $\|\Omega_1 - I\|_F^2$ and μ. Since we discuss the case $\Omega_0 = I$ here, $\|\Omega_1 - I\|_F^2$ can be regarded as $\|\Omega_1 - \Omega_0\|_F^2$, which is the effect of quadratic term. For μ, we have to consider two cases:

(i) When $\theta < \delta/2$, the sample size is large enough so that the signals in the mean vector can be almost perfectly recovered. With the feature selection step, the QDAfs achieves an asymptotic mis-classification rate of 0 when $\max\{\|\Omega_1 - I\|_F^2, \|\mu\|^2\} \to \infty$, and $1/2$ otherwise. In addition, the latter region is proven to be a failure region for all classifiers, which is referred to as the region of impossibility.

(ii) When $\theta > \delta/2$, the sample size is insufficient for the signal recovery, and the feature selection step is ineffective. We use $g_w(X)$ to aggregate all the information in μ to do estimation, which is $O(\|\mu\|_1)$. The QDAw mis-classification rate converges to 0 when $\max\{\|\Omega_1 - I\|_F^2, \|\mu\|_1\} \to \infty$, and $1/2$ otherwise. Again, the latter region is proven to be a failure region for all classifiers.

Figure 1 provides a sense of the relationship between the sparsity and weakness parameters of the mean and covariance matrix. Subfigure (a) is on the α-β plane to present results about $\|\Omega_1 - I\|_F^2$. Subfigure (b) is on the θ-ζ plane to present the regions on $\|\mu\|^2$. To see the effects clearly, we assume the information from the other part is insufficient for each subfigure. We can see QDAw and QDAfs are the optimal methods.

In Subfigure (a), we suggest QDAw and QDAfs in the region of possibility, instead of only one method. The reason is although the contribution from Ω_1 and μ seems independent of each other, the performance of QDAfs still relies on the signal strength in μ. When the signals in μ cannot be successfully recovered, QDAfs requires $\|\Omega_1 - I\|_F^2 \gg \sqrt{p/n}$ to success, which cannot achieve the bound $\|\Omega_1 - I\|_F^2 \gg 0$ by QDAw. It comes from the effects of μ on $X^\top \Omega_{\text{diff}} X$. This effect is rarely discussed in previous literature.



(a) Phase transition on Ω_1 (b) Phase transition on μ

FIG 1. *The possibility/impossibility regions derived in Theorems 1.1 and 1.2 when δ and part of the rest parameters are fixed: (a) δ, θ and ζ are fixed, $\gamma > 1/2$ and $\theta < \rho_\delta(\zeta)$; (b) δ, α and β are fixed, $\gamma > 1/2$ and $2 - 2\alpha - \beta < 0$.*

The methods employed and results obtained in this work are unique compared with other literature on QDA methods for high dimensional data with sparse signals ([26, 13, 31]). We propose QDAw and QDAfs for different types of mean vectors and show they match the statistical lower bound, which is rarely discussed in other works.

1.4. *A Real Data Example.* We use a quick example to demonstrate how this works on the real data. We consider the rats dataset with summaries given in Table 2. This dataset consists of 181 samples measured on the same set of 8491 genes, with 61 samples labeled by [33] as toxicants and the other 120 as drugs. The original rats dataset was collected in a study of gene expressions of live rats in response to different drugs and a toxicant; we use the cleaned version by [33].

This dataset has been carefully studied in [17], with the performance of the two-class classification compared among a sequence of popular classifiers, including SVM in [4], Random Forest in [3], and HCT-PCS. The HCT-PCS, which achieves optimal classification when it

TABLE 2
A gene-expression microarray rats dataset.

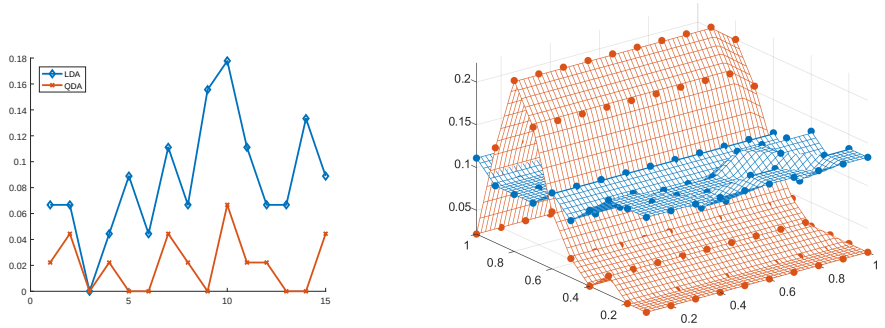
Data Name	Source	n (# of subjects)	p (# of genes)
Rats	Yousefi et al. (2010)	181	8491

adapts LDA [6, 12] in the rare and weak signal setting, was shown to have very promising classification results with this data.

That said, in HCT-PCS, all samples of the two classes are assumed to share the same precision matrix, leaving room for improvement. We now apply QDAfs with data normalization (details in Table 7) to this data set and compare the results with those from the LDA with HCT-PCS approach. Here, we leave out all the implementation details, which will be introduced in Section 5, and only highlight our findings for this rats data:

- QDA further outperforms LDA with HCT-PCS, and produces better results than those other methods in [17], including SVM and Random Forest, suggesting that QDA gives a better separation by taking into account the second-order difference between the two classes.

We record 15 random splits of the rats data for the training data and test data. The test error is illustrated in Figure 2 (below, left). We can see that the test error of LDA are all above those of QDA at every data splitting, given that all the tuning parameters are selected in the same way. Figure 2 (below, right) demonstrates the surface of the test error between LDA and QDA, by varying the tuning parameters in the precision-matrix estimation. This Zoom-in plot shows that the QDA does bring necessary improvement over LDA when the precision matrices are appropriately estimated.



(a) (b)

FIG 2. *Comparison of testing errors for the rats data: (a) error rates (y-axis) of LDA (blue) and QDA (red) for 15 data splittings (x-axis) at a certain sparsity of the precision matrix; (b) Zoom-in errors for the rats data for varying choices of parameters in estimating the precision matrices for one splitting of 15 splittings.*

1.5. *Content and Notations.* The main results for the phase transitions under various scenarios are discussed in Section 2. Proofs of lower bounds are given in Section 3 and that of upper bounds are given in Section 4. In Section 5, we present numerical results of the proposed methods and algorithms on real data. In Section 6, some concluding remarks and potential directions of future work are discussed. The details of the proofs are provided in the supplementary materials.

Here we list the notations used throughout the paper. Let the eigenvalues of W be denoted by $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$. For a matrix M, we use $\|M\|$ and $\|M\|_F$ to denote its spectral normal and Frobenius norm, respectively, $det(M)$ to denote the determinant of M and $Tr(M)$ to

denote its trace which equals the summation of the eigenvalues of M. We use $diag(c_1, \ldots, c_p)$ to denote a diagonal matrix with diagonal elements $c_1, \ldots, c_p$, and use $I(A)$ to denote an indicator function over event A. For two vectors or matrices a and b of same dimension, $a \circ b$ denotes the Hadamard (entrywise) product.

2. Phase transition for the classification problem. Throughout the whole paper, we consider the mixture model

$$X|Y \sim (1-Y)N(-\mu, \Omega_0^{-1}) + YN(\mu, \Omega_1^{-1}), \tag{2.15}$$

and the mis-classification rate as

$$MR = [P_{\epsilon,\tau,\eta,\nu,\xi}(\hat{Y} = 0|Y = 1) + P_{\epsilon,\tau,\eta,\nu,\xi}(\hat{Y} = 1|Y = 0)]/2, \tag{2.16}$$

where X is a fresh data vector with Y as the true label and $\hat{Y}$ being the estimated label. We refer to the error rate by a classifier L as $MR(L)$. Since we consider $P(Y = 0) = q = 1/2$, so MR is the average of two types of errors. For a general q, we should update MR as $MR = (1-q)P(\hat{Y} = 0|Y = 1) + qP(\hat{Y} = 1|Y = 0)$ and the results still hold.

2.1. *New QDA approaches.* When all the parameters are known, the classifier is in (1.3), where we calculate

$$Q(X) = X^\top(\Omega_0 - \Omega_1)X + 2\mu^\top(\Omega_0 + \Omega_1)X + \mu^\top(\Omega_0 - \Omega_1)\mu + \ln|\Omega_1| - \ln|\Omega_0|, \tag{2.17}$$

and estimate $\hat{Y} = I(Q(X) > 0)$.

We have presented the estimation of $X^\top(\Omega_0 - \Omega_1)X$ and $2\mu^\top(\Omega_0 - \Omega_1)X$ for the unknown parameter case in Section 1.1. Here we define two functions to better present the estimates in the algorithms. For a symmetric matrix M and threshold t, define

$$T(M; t) = M - diag(diag(M) \circ I\{|diag(M)| <= t\}). \tag{2.18}$$

By $T(M; t)$, the diagonals of M that are smaller than t are truncated to be 0 and the off-diagonals do not change. Given matrices A, B, vector d and threshold t, define a vector $d^{(t)}$ with $d^{(t)}(j) = I\{|d(j)| \geq t\}$ and the functions

$$g_w(X; \hat{\mu}, A, B) = \hat{\mu}^\top(A + B)X, \quad g_s(X; d, t) = (d \circ d^{(t)})^\top X. \tag{2.19}$$

With all the preparations, we present the algorithms for various scenarios. Tables 3 and 4 are the algorithms employed in Theorem 1.2, for the case that both Ω_i's are known. The algorithm for that Ω_0 is known and Ω_1 is unknown is in Table 5, and the most general case is in Table 6. For previous cases, the constants are clearly stated.

TABLE 3
Algorithm QDAw: weak and relatively dense signals in μ, Ω_1 is known, $\Omega_0 = I$.

	Input: data points (X_i, Y_i), $1 \leq i \leq n$; constant $0 < c < 1$; new data point X; true precision matrix Ω_1.
	Output: label $\hat{Y}$.
1.	Parameter Estimation: Let $\hat{\mu}_0 = \frac{1}{n_0}\sum_{i:Y_i=0} X_i$, where $n_0 = n - n_1$. Let $\hat{\mu} = p^{(c-1)/2} * \mathbf{1}$, where $\mathbf{1} \in \mathcal{R}^p$ is the vector of ones.
2.	Let $C = \hat{\mu}_0^\top(\Omega_1 - I)\hat{\mu}_0 + \ln|\Omega_1| + \frac{1}{n_0}Tr(\Omega_1 - I)$.
3.	QDA Score: Calculate the QDA score $Q = X^\top(I - \Omega_1)X + 2g_w(X; \hat{\mu}, I, \Omega_1) + C$.
4.	Prediction: Predict $\hat{Y} = I\{Q > 0\}$.

TABLE 4
Algorithm QDAfs: moderately strong and sparse signals in μ, Ω_1 is known, $\Omega_0 = I$.

Input: data points (X_i, Y_i), $1 \le i \le n$; threshold $t > 0$; new data point X. true precision matrix Ω_1.
Output: label $\hat{Y}$.

1. Parameter Estimation: Let $\hat{\mu}_0 = \frac{1}{n_0}\sum_{i:Y_i=0} X_i$ and $\hat{\mu}_1 = \frac{1}{n_1}\sum_{i:Y_i=1} X_i$, where $n_1 = \sum_{i=1}^n Y_i$ and $n_0 = n - n_1$.
2. Let $d_0 = \hat{\mu}_0$, $d_1 = \Omega_1\hat{\mu}_1$ and $d = d_1 - d_0 = (d(1), \ldots, d(p))^\top$.
3. Thresholding: Let $d^{(t)}$ denote the indicator vector of feature selection, i.e. $d^{(t)}(j) = 1\{|d(j)| \ge t\}$, for $j = 1, \ldots, p$.
   Let $\Omega_1^{(d)}$ be the sub-matrix of Ω_1 constrained on rows and columns that $d^{(t)} = 1$.
4. Let $C = (\hat{\mu}_0 \circ d^{(t)})^\top (I - \Omega_1)(\hat{\mu}_0 \circ d^{(t)}) + \ln|\Omega_1| + \frac{1}{n_0} Tr(\Omega_1^{(d)} - I)$.
5. QDA Score: Calculate the QDA score $Q = X^\top (I - \Omega_1)X + 2g_s(X; d, t) + C$.
6. Prediction: Predict $\hat{Y} = I\{Q > 0\}$.

TABLE 5
Algorithm QDAfs/QDAw-PCS: $\Omega_0 = I$.

Input: data points (X_i, Y_i), $1 \le i \le n$; threshold $t > 0$; new data point X.
Output: label $\hat{Y}$.

1. Parameter Estimation: Let $\hat{\Omega}_1$ be the estimation from PCS.
2. Thresholding: Update $\hat{\Omega}_1$: $\hat{\Omega}_1 = T(\hat{\Omega}_1 - I; \sqrt{2\ln p/n}) + I$.
3. Apply QDAw in Table 3 or QDAfs in Table 4 with the input precision matrix as $\hat{\Omega}_1$.

TABLE 6
Algorithm QDAfs-PCS: all unknown.

Input: data points (X_i, Y_i), $1 \le i \le n$; threshold $t > 0$; new data point X.
Output: label $\hat{Y}$.

1. Parameter Estimation: Let $\hat{\Omega}_0$ and $\hat{\Omega}_1$ be the estimation from PCS.
2. Thresholding: Let $\hat{\Omega}_{\text{diff}} = T(\hat{\Omega}_0 - \hat{\Omega}_1; 2\sqrt{2\ln p/n})$.
3. Parameter Estimation: Let $\hat{\mu}_0 = \frac{1}{n_0}\sum_{i:Y_i=0} X_i$ and $\hat{\mu}_1 = \frac{1}{n_1}\sum_{i:Y_i=1} X_i$, where $n_1 = \sum_{i=1}^n Y_i$ and $n_0 = n - n_1$.
4. Let $d_0 = \hat{\Omega}_0\hat{\mu}_0$, $d_1 = \hat{\Omega}_1\hat{\mu}_1$ and $d = d_1 - d_0 = (d(1), \ldots, d(p))^\top$.
5. Let $C = (\hat{\mu}_0 \circ d^{(t)})^\top \hat{\Omega}_{\text{diff}}(\hat{\mu}_0 \circ d^{(t)}) + \ln|\hat{\Omega}_0 - diag(\hat{\Omega}_0) + I| - \ln|\hat{\Omega}_1 - diag(\hat{\Omega}_1) + I|$.
6. QDA Score: Calculate the QDA score $Q(X) = X^\top \hat{\Omega}_{\text{diff}} X + 2g_s(X; d, t) + C$.
7. Prediction: Predict $\hat{Y} = I\{Q(X) > 0\}$.

2.2. *Ideal case.* When all the parameters are known, the classical QDA classifier provides the optimal results; see Proposition 2.1.

PROPOSITION 2.1. [Phase transition for the ideal case.] Consider the rare and weak signal model (1.4) – (1.9).

(i) The QDA classifier (2.17) has a mis-classification rate $MR(QDA) \to 0$ as $p \to \infty$, if $\|\Omega_1 - \Omega_0\|_F^2 + 8\|\mu\|^2 \to \infty$.

If further (1.10)–(1.13) are satisfied, then $MR(QDA) \to 0$ if one of the following conditions is satisfied,

(1) $\beta < 2 - 2\alpha$; or
(2) $\gamma < 1/2$; or
(3) $\zeta < 1 - 2\theta$.

(ii) The mis-classification rate $MR(L)$ of any classifier L converges to $1/2$ when $p \to \infty$, if $\|\Omega_1 - \Omega_0\|_F^2 + 8\|\mu\|^2 \to 0$.

Remark 1. Proposition 2.1 describes an exact phase diagram of the classification problem. When $\|\Omega_1 - I\|_F^2 + 8\|\mu\|^2 \to \infty$, the QDA method achieves a mis-classification rate of 0 asymptotically. On the complement region, all classifiers fail. In this sense, Proposition 2.1 demonstrates QDA succeeds in the whole possibility region and thus is optimal.

Remark 2. It can be observed the contribution of μ is $\|\mu\|^2$ and the contribution of Ω_1 is $\|\Omega_1 - I\|_F^2$. There is no intersection between them because all the parameters are known. When we consider data-trained classifiers, the interaction may happen to be the choice of algorithms (see Theorem 1.2), or the two possibility regions will depend on the parameter from the other part (see Theorems 2.3 and 2.4).

Remark 3. Since the contribution of Ω_1 is $\|\Omega_1 - I\|_F^2$, the diagonals and off-diagonals perform independently. The off-diagonals of $\Omega - I$ is modelled in (1.7) and (1.12) by α and β to measure the signal strength and sparsity. The diagonals only have one signal strength parameter $\xi = p^{-\gamma}$. Hence, condition (1) is an inequality between α and β while condition (2) is about γ solely.

We do not consider the sparsity in the diagonals of $\Omega - I$ due to model complexity. In that sense, the model thus raised will have 6 sparsity and weakness indices in total: 2 for means, 2 for diagonals, and 2 for off-diagonals of Ω_1. We can readily obtain the possibility region for this model, but it will be difficult to visualize and is thus omitted here.

2.3. *Phase transitions with partial information.* With Theorems 1.1 and 1.2 in Section 1.3, we have discussed the phase transition phenomenon when μ is unknown but Ω_0 and Ω_1 are known. Here we further show the results when $\Omega_0 = I$ without loss of generality, and Ω_1 is unknown.

When Ω_1 is unknown, the first problem is to estimate it. It has been discussed in numerous publications in the literature, such as [5, 12, 15, 17]. However, estimation of high-dimensional precision matrix is restricted to the case that Ω_1 is sparse. Here, we consider the case that the signals in Ω_1 are sparse and strong, that

$$\alpha < \delta/2, \quad 1 - \delta/2 < \beta < 2. \tag{2.20}$$

Under (2.20), with high probability, the number of non-zero entries in each row of Ω_1 is $o(n)$ and the signal strength $\eta \gg 1/\sqrt{n}$.

Under (2.20), we suggest to estimate Ω_1 by the Partial Correlation Screening (PCS) approach in [17]. The PCS approach has a good control on $\|\hat{\Omega}_1 - \Omega_1\|_{\max}$, and hence $\|\hat{\Omega}_1 - \Omega_1\|_F^2$ is under control. With the PCS estimation, we further apply a truncation on the diagonals of $\hat{\Omega}_1$, where we assign $\hat{\Omega}_1(i,i)$ to be 1 if it is close to 1, i.e. $|\hat{\Omega}_1(i,i) - 1| \leq \sqrt{2\ln p/n})$. The truncation step helps to remove the noise on diagonals, but suffers a loss on the diagram of γ.

With $\hat{\Omega}_1$, we apply QDAw and QDAfs to estimate Y; details in Table 5. We call it as QDAw-PCS or QDAfs-PCS. To find the optimality of it, we first find the phase diagram of the classification problem when μ is given; see the following proposition. When μ is given, we do not need QDAw or QDAfs. The classical QDA in (2.17) with $\hat{\Omega}_1$, called QDA-PCS, will work.

PROPOSITION 2.2. Consider model (2.15) with the parameterizations (1.4)–(1.13) and (2.20). Suppose μ is given and $\Omega_0 = I$.

(i) the QDA classification rule (2.17) with $\Omega_1 = \hat{\Omega}_1$ as the truncated PCS estimate of Ω_1 has a mis-classification rate $MR(QDA\text{-}PCS) \to 0$ as $p \to \infty$, if

(i) $\gamma < \delta/2$; or
(ii) $\gamma > (2\alpha+\beta-1)/2$ and $2\alpha+\beta-2<0$; or
(iii) $\gamma > \frac{1}{2}\min\{1, 2\alpha+\beta-1\}$ and $2\theta+\zeta-1<0$.

(ii) $MR(L) \geq 1/2$ for any classifier L when $p \to \infty$, if $\gamma > 1/2$, $2\alpha+\beta-2>0$ and $2\theta+\zeta-1>0$.

When $\gamma > 1/2$, then QDA-PCS achieves the optimal boundary. When $\gamma < \delta/2$, then QDA-PCS also provides satisfactory classification results. However, when $2\alpha+\beta-1>\delta$, then on $\delta/2 < \gamma < (2\alpha+\beta-1)/2$, QDA-PCS suffers a power loss because of the truncation on the diagonals.

Now we consider the case μ and Ω_1 are both unknown. Similar as QDA-PCS in Proposition 2.2, here we find QDAw-PCS and QDAfs-PCS can match the lower bound.

THEOREM 2.3. Consider model (2.15) with the parameterizations (1.4)–(1.13) and (2.20). Suppose $\Omega_0 = I$.

(i) MR(QDAw-PCS) with arbitrary constant c goes to 0 when $p \to \infty$, if $\theta \geq \delta/2$ and one of the following conditions hold
  (a) $\gamma < \delta/2$; or
  (b) $\gamma > (2\alpha+\beta-1)/2$ and $2\alpha+\beta-2<c$; or
  (c) $\gamma > \frac{1}{2}\min\{1, 2\alpha+\beta-1\}$ and $\theta < \rho_\delta(\zeta)$.
(ii) MR(QDAfs-PCS) goes to 0 when $p \to \infty$, if $\theta < \delta/2$ and one of the following conditions hold
  (a) $\gamma < \delta/2$; or
  (b) $\gamma > (2\alpha+\beta-1)/2$ and $2\alpha+\beta-2<0$; or
  (c) $\gamma > \frac{1}{2}\min\{1, 2\alpha+\beta-1\}$ and $\theta < \rho_\delta(\zeta)$.
(iii) $MR(L) \geq 1/2$ for any classifier L when $p \to \infty$, if $\gamma > 1/2$, $2\alpha+\beta-2>0$, and $\theta > \rho_\delta(\zeta)$ in (1.14).

Remark 1. The upper bound of QDAfs-PCS or QDAw-PCS matches the lower bound when $\gamma > \min\{1, 2\alpha+\beta-1\}/2$ or $\gamma < \delta/2$. When the diagonals have parameter $\delta/2 < \gamma < \min\{1, 2\alpha+\beta-1\}/2$, the random error is too large to recognize the truth and do successful classification.

Remark 2. Even for the case that $\theta < \rho_\delta(\zeta)$ and μ performs the main role in classification, we still need the condition that $\gamma > \frac{1}{2}\min\{1, 2\alpha+\beta-1\}$ so that $\|\hat{\Omega}_1 - I\|_F^2$ is under control. This can be seen as the intervene between the quadratic and the linear terms.

2.4. *Phase transitions with unknown parameters.* The most generalized case is that all the parameters are unknown. According to Theorem 2.3, we consider both Ω_1 and Ω_0 have sparse and strong off-diagonal signals as (2.20) and very weak diagonal signals that

$$\gamma > 1/2. \tag{2.21}$$

Under this condition, the diagonals are too weak to do successful classification.

THEOREM 2.4. Consider the ARW model (2.15) with the parameterizations (1.4)–(1.13), (2.20) and (2.21).

(i) MR(QDAfs-PCS) in Table 6 goes to 0 when $p \to \infty$, if $\theta < \delta/2$ and one of the following conditions hold
  (a) $2-2\alpha-\beta>0$; or
  (b) $1-2\theta-\zeta>0$.

(ii) $MR(L) \geq 1/2$ for any classifier L when $p \to \infty$, if $\theta < \delta/2$, $2 - 2\alpha - \beta < 0$ and $1 - 2\theta - \zeta < 0$.

The theorem suggests QDAfs-PCS is optimal if Ω_i and μ have strong and sparse signals.

3. Proof of lower bounds. We present the lower bound when μ and Ω_i's are all known in Proposition 2.1, only μ and Ω_0 are known in Theorem 1.1, only Ω_0 is known in Theorem 2.3 and all are unknown in Theorem 2.4. When the signal strength and sparsity parameters falls below the lower bound, any classifier L will fail. In this section we will prove these results.

3.1. *Proof of lower bound in Proposition 2.1.* In this ideal case both μ and Ω_i's are known. Let f be the density function of $X \sim N(-\mu, \Omega_0^{-1})$ and g be the density function of $X \sim N(\mu, \Omega_1^{-1})$. The Hellinger affinity between f and g is defined as $H(f,g) = \int \sqrt{f(x)g(x)}dx$.

LEMMA 3.1. *For any classifier $L = L(X|\mu, \Omega_0, \Omega_1)$,*

$$|MR(L) - 1/2| \leq C(1 - H(f,g))^{1/2}.$$

This lemma is well known, and so we omit the proof. According to this lemma, $H(f,g) = 1 + o(1)$ suffices to prove the impossibility. Introduce the normal density into $H(f,g)$, with basic calculations we have

$$H(f,g) = \exp\{-\frac{1}{2}\big[\|\mu\|^2(1+o(1)) + \|\Omega_0 - \Omega_1\|_F^2/8\big]\}. \tag{3.22}$$

Therefore, when $\|\mu\|^2 + \|\Omega_0 - \Omega_1\|_F^2/8 = o(1)$, $H(f,g) = 1 + o(1)$ and the mis-classification error from any classifier will be close to 1/2.

Under (1.4)–(1.9), with high probability, $\|\Omega_0 - \Omega_1\|_F^2 \leq 4(p\xi_p^2 + \eta_p^2 p^2 \nu_p)(1 + o(1))$ and $\|\mu\|^2 = \tau_p^2 p \epsilon_p^2 (1 + o(1))$. In the region of impossibility, both terms converge to 0, and then $H(f,g) = 1 + o(1)$. As a result, $MR(L) \to 1/2$ for any classifier L. The lower bound in Theorem 2.1 is proved.

3.2. *Proof of Theorem 1.1.* Here we consider the case μ and Ω_1 are unknown, and $\Omega_0 = I$ without loss of generality. Since we have to use training data to estimate μ, the density functions are updated to be $f = f(X, X_1, \cdots, X_n; \Omega_1)$ and $g = g(X, X_1, \cdots, X_n; \Omega_1)$, where $X_i \sim N((2Y_i - 1)\mu, I + Y_i(\Omega_1^{-1} - I))$ with known Y_i's for both cases. The new data point is assumed to be $X \sim N(-\mu, I)$ for f and $X \sim N(\mu, \Omega_1^{-1})$ for g. We want to prove $H(f,g) = 1 + o(1)$.

Here, f and g differ at both mean and covariance matrix of X. We define $\tilde{f}$ to be a middle state, that $\tilde{f} = \tilde{f}(X, X_1, \cdots, X_n; \Omega_1)$, where $X \sim N(\mu, I)$ and others are the same. Hence, $\tilde{f}$ differs with f only on the mean vector of X, and differs with g only on the covariance matrix of X. When both both $\|f - \tilde{f}\|_1 = o(1)$ and $\|g - \tilde{f}\|_1 = o(1)$, there is $\|f - g\|_1 = o(1)$ and hence $H(f,g) = 1 + o(1)$.

Consider $\|f - \tilde{f}\|_1$ first. It comes to the classification problem with an identity covariance matrix. In [23], it has been proved that $\|f - \tilde{f}\|_1 = Cp\epsilon_p^2(e^{\tau_p^2} - 1)(1 + o(1))e^{n\tau_p^2}$ when $\zeta < 1 - \delta$, and $\|f - \tilde{f}\|_1 = C\sqrt{(e^{2p\epsilon_p\tau_p^2} - e^{-2p\epsilon_p\tau_p^2})/2}(1 + o(1))$ when $\zeta > 1 - \delta$. Introducing (1.11) that models ϵ_p and τ_p, $\|f - \tilde{f}\|_1 = o(1)$ when one of the following can be satisfied:

(a) $\theta \geq \delta/2$, $\zeta < 1 - \delta$, $\zeta + \theta < 1/2$; or

(b) $\zeta > 1-\delta, 1-\zeta-2\theta < 0$.

Consider $\|g-\tilde{f}\|_1$ where μ is unknown and (1.4) holds. With some calculations,

$$\|g-\tilde{f}\|_1 = \int\int \frac{1}{(2\pi)^{p/2}} e^{-\frac{1}{2}(X-\mu)^\top(X-\mu)} |1-det(\Omega_1)^{1/2} e^{-\frac{1}{2}(X-\mu)^\top(\Omega_1-I)(X-\mu)}| dX dF(\mu)$$

$$= \int \frac{1}{(2\pi)^{p/2}} e^{-\frac{1}{2}X^\top X} |1-det(\Omega_1)^{1/2} e^{-\frac{1}{2}X^\top(\Omega_1-I)X}| dX. \tag{3.23}$$

It equals to the L_1 distance between $f_s \sim N(0,I)$ and $g_s \sim N(0,\Omega_1^{-1})$. Therefore, to show $\|g-\tilde{f}\|_1 = o(1)$, it is to prove $\|f_s - g_s\| = o(1)$, which is equivalent with $H(f_s,g_s) = 1+o(1)$. For $H(f_s,g_s)$, there is no training data and we can calculate the Hellinger distance directly, which is

$$H(f_s,g_s) = \frac{det(\Omega_1)^{1/4}}{det((\Omega_1+I)/2)^{1/2}} = \|\Omega_1 - I\|_F^2/8(1+o(1)). \tag{3.24}$$

As a conclusion, $\|g-\tilde{f}\|_1 = o(1)$ when $\|\Omega_1 - I\|_F^2 \to 0$.

Recall that $H(f,g) = 1+o(1)$ when both $\|f-\tilde{f}\|_1$ and $\|g-\tilde{f}\|_1$ are $o(1)$. Combine it with the results for $\|f-\tilde{f}\|_1$ and $\|g-\tilde{f}\|_1$. Therefore, $H(f,g) = 1+o(1)$ when $\|\Omega - I\|_F^2 \to 0$ and one of the following conditions can be satisfied:

(a) $\theta \geq \delta/2, \zeta < 1-\delta, 1-2\zeta-2\theta < 0$; or
(b) $\zeta > 1-\delta, 1-\zeta-2\theta < 0$.

Consider condition (a), $\zeta < 1-\delta$ always holds when $\theta \geq \delta/2$ and $\zeta+\theta < 1/2$, so the condition $\zeta < 1-\delta$ can be removed. Actually, when $\theta \geq \delta/2$, the region of impossibility will be decided by condition (a) because $1-\zeta-2\theta < 0$ in (b) always indicate $1-2\zeta-2\theta < 0$ in (a). So we only need to consider the case $\theta < \delta/2$ for condition (b). When $\theta < \delta/2$, $1-\zeta < 2\theta < \delta$, so the condition $\zeta > 1-\delta$ always hold. Hence, the conditions can be simplified as $\|\Omega - I\|_F^2 \to 0$ and one of the following conditions can be satisfied:

(a) $\theta > \delta/2, 1-2\zeta-2\theta > 0$; or
(b) $\theta < \delta/2, 1-\zeta-2\theta < 0$.

Theorem 1.1 is proved.

3.3. *Proof of lower bound in Theorem 2.3.* Consider the case $\Omega_0 = I$ without loss of generality. Because loss of information about Ω_1, the region of impossibility cannot be larger than that in the case Ω_1 is known in Theorem 1.1. Hence, $MR(L) \geq 1/2 + o(1)$ when $2-2\alpha-\beta < 0$, $1-2\gamma < 0$ and one of the following conditions are satisfied:

(a) $\theta > \delta/2, 1-2\zeta-2\theta > 0$; or
(b) $\theta < \delta/2, 1-\zeta-2\theta < 0$.

The region of impossibility in Theorem 2.3 is proved.

3.4. *Proof of lower bound in Theorem 2.4.* When both μ and Ω_i's are unknown, with the same analysis in Section 3.3, we have the region of impossibility in Theorem 2.4.

4. Proof of upper bounds. In this section, we present the proof of upper bounds in Theorem 1.2, Proposition 2.2 and Theorem 2.3. This section is structured as follows. In Section 4.1, we present some mathematical results as the preparations. In Section 4.2, we present the upper bounds of QDAw and QDAfs in Theorem 1.2. We prove the case that Ω_1 is unknown

in Section 4.3. All the proofs of the lemmas in this section can be found in the supplementary material [29]. In this section, we always use $\Omega = \Omega_1$ for simplification without confusion.

We begin with the expression of the mis-classification rate MR in terms of QDA. Given μ and Ω, the two types of mis-classification rates are defined as

$$p_{0,\mu,\Omega} = P_{Y=0}(Q > 0|\mu,\Omega), \quad p_{1,\mu,\Omega} = P_{Y=1}(Q < 0|\mu,\Omega). \tag{4.25}$$

Then, the population mis-classification rate (MR) of QDA is

$$MR(QDA) \ = \ [E[p_{0,\mu,\Omega}] + E[p_{1,\mu,\Omega}]/2. \tag{4.26}$$

Given a parameter set $(\alpha,\beta,\gamma,\zeta,\theta)$, if both $E[p_{0,\mu,\Omega}]$ and $E[p_{1,\mu,\Omega}]$ converge to 0, then $MR(QDA)$ converges to 0, which means QDA is successful.

4.1. *Preparations and notations.* To find the upper bounds, we should analyze the asymptotic distribution of the QDA score. In the analysis, we keep on using the quadratic terms of X, in the form of $X^\top AX + 2d^\top X$. The following lemma states the asymptotic distribution of such quadratic terms.

LEMMA 4.1. *[Quadratic functional of normal distributions] Consider $X \sim N(\mu,\Sigma)$ where Σ is positive definite. Let $S = X^\top AX + 2d^\top X$ with a symmetric matrix A and a vector d,*

$$E[S] = Tr(A\Sigma) + \mu^\top A\mu + 2d^\top\mu, \tag{4.27}$$

$$\mathrm{Var}(S) = 2Tr((A\Sigma)^2) + 4(\mu^\top A\Sigma A\mu + \mu^\top A\Sigma d + d^\top\Sigma d). \tag{4.28}$$

(a) $\frac{\sum_{i=1}^p |\lambda_i|^3(1+|\tilde{\mu}(i)|^3)}{(\sum_{i=1}^p \lambda_i^2(1+\tilde{\mu}_i^2))^{3/2}} \to 0$; *or*
(b) $\mathrm{Var}(S) = \sum_{i:\lambda_i=0} \tilde{d}^2(i)(1+o(1))$.

LEMMA 4.2. *Under current model and assumptions, for a given matrix A with spectrum in $(1-o(1), 1+o(1))$, there exists a constant $C > 0$, so that with probability $1-o(1)$,*

$$\left|\left[\hat{\mu}_0^\top(I-A)\hat{\mu}_0 - \mu^\top(I-A)\mu\right] + \frac{1}{n_0}Tr(A-I)\right| \le C\sqrt{\frac{\ln p}{n}}(\|A-I\|_F/\sqrt{n} + \|(I-A)\mu\|).$$

In the analysis, we have to relate the terms $\|\Omega_i - I\|$, $\|\Omega_i - I\|_F^2$ and $\|\mu\|^2$ to the constant parameters. The following two lemmas describe how these terms rely on the parameters.

LEMMA 4.3. *[Bounds on the signals in precision matrix] Under models (1.7) and (1.12), when $p \to \infty$, with probability $1-o(1)$,*

$$\|V^{(k)}\| \le \eta_p b(p,\beta) = \begin{cases} 3\eta_p\sqrt{p\nu} = 3\eta_p p^{(1-\beta)/2}, & 0 < \beta < 1, \\ 2\eta_p\sqrt{\ln p/\ln\ln p}, & \beta = 1, \\ 2\eta_p/(\beta-1), & 1 < \beta \le 2. \end{cases}$$

The results are summarized in in the following lemma.

LEMMA 4.4. *Consider model (2.15) with the parameterizations (1.4) and (1.8)–(1.13). With probability $1-o(1)$, we have $\|\Omega_k - I\| = o(1)$ and*

$$\begin{aligned} &\|V^{(k)}\|_F^2 = \eta_p^2 p^2\nu_p(1+o(1)), \quad \|\Omega_k - I\|_F^2 = p\xi_p^2 + \eta_p^2 p^2\nu_p(1+o(1)), \\ &\|\mu\|^2 = p\tau_p^2\epsilon_p(1+o(1)). \end{aligned} \tag{4.29}$$

4.2. *Proof of Theorem 1.2.* When μ is unknown, we propose two algorithms that work in different regions. When the non-zeros in μ are weak and relatively dense, then we apply QDAw which averages all the features; when the non-zeros in μ are relatively strong but sparse, we apply QDAfs to select features first. We find the upper bounds for both algorithms to prove Theorem 1.2.

4.2.1. *Performance of QDAw.* In QDAw, we estimate labels by $\hat{Y} = I(Q_w > 0)$. Here, $Q^w = S^w - T_S^w$, where

$$S^w = X^\top(I-\Omega)X + 2\hat{\mu}^\top(I+\Omega)X, \quad T_S^w = \hat{\mu}_0^\top(I-\Omega)\hat{\mu}_0 - \ln|\Omega| - \frac{1}{n_0}Tr(\Omega - I).$$

Here, $\hat{\mu}_0 = \frac{1}{\sum_i I\{Y_i=0\}}\sum_{i:Y_i=0} X_i$ as the average of training samples in Class 0, and $\hat{\mu} = a*\mathbf{1}$, a vector with all the entries as a. In the algorithm, we take $a = p^{(c-1)/2}$. The errors are $p_{i,\mu,\Omega} = P((-1)^i(S^w - T_S^w) > 0)$. We want to find the region that both $p_{i,\mu,\Omega} \to 0$.

Consider S^w, which is a quadratic term with $A = I - \Omega$ and $d = (I+\Omega)\hat{\mu}$. Apply Lemma 4.1 to S^2 with $\Sigma = I$ for the case $Y = 0$ and $\Sigma = \Omega^{-1}$ for the case $Y = 1$. There is

$$E[S^w|Y=0] = \mu^\top(I-\Omega)\mu + Tr(I-\Omega) - 2a\mu^\top(I+\Omega)\mathbf{1}, \tag{4.30}$$

$$E[S^w|Y=1] = \mu^\top(I-\Omega)\mu + Tr(\Omega^{-1}-I) + 2a\mu^\top(I+\Omega)\mathbf{1}, \tag{4.31}$$

and

$$\mathrm{Var}(S^w|Y=i) = 2\|\Omega - I\|_F^2 + (16pa^2 + \|(\Omega-I)\mu\|^2)(1+o(1)),\ i=0,1. \tag{4.32}$$

Given $Y = i$, we define $Z_i = [S^w - E[S^w|Y=i]]/\sqrt{\mathrm{Var}(S^w|Y=i)}$, then $\sup_{-\infty<x<\infty}|F_{Z_i}(x) - \Phi(x)| \to 0$. So the asymptotic distribution of S^w is clear. When $Y = 0$ and $Y = 1$, the mean of S^w differs in two parts, $\mu^\top(I+\Omega)\mathbf{1}$ and $\|\Omega - I\|_F^2$, with a shift that $\mu^\top(I-\Omega)\mu + \ln|\Omega|$.

Compare $E[S^w|Y]$ with T_S^w, we can see T_S^w mainly captures the shift. The difference is that T_S^w uses $\hat{\mu}_0$ instead of the true parameter μ. Consider the relative term $\hat{\mu}_0^\top(I-\Omega)\hat{\mu}_0$ in T_S^w. Apply Lemma 4.2 to it with $A = \Omega$ and we have

$$T_S^w = \mu^\top(I-\Omega)\mu - \ln|\Omega| + \Delta T, \tag{4.33}$$

where $|\Delta T| \le C\sqrt{\ln p}(\|\Omega - I\|_F/n + \|(I-\Omega)\mu\|/\sqrt{n})$.

Introduce the results about S^w and T_S^w into $p_{i,\mu,\Omega} = P((-1)^i(S^w - T_S^w) > 0|Y=i)$. By the asymptotic normality of S^w, the error is $\Phi(\frac{(-1)^i*(E[S^w|Y=i]-T_S^w)}{\sqrt{\mathrm{Var}(S^w|Y=i)}}) + o(1)$. Introduce in (4.30), (4.32) and (4.33) into $p_{0,\mu,\Omega}$, and we have

$$\begin{aligned} p_{0,\mu,\Omega} &= \Phi\Big(\frac{\ln|\Omega| + Tr(I-\Omega) - 2a\mu^\top(\Omega+I)\mathbf{1} + \Delta T}{\sqrt{2\|\Omega-I\|_F^2 + 16pa^2 + \|(\Omega-I)\mu\|^2}}\Big) + o(1) \\ &= \Phi\Big(\frac{-\|\Omega-I\|_F^2/2 - 4a\|\mu\|_1(1+o(1)) + \Delta T}{\sqrt{2\|\Omega-I\|_F^2 + 16pa^2 + \|(\Omega-I)\mu\|^2}}\Big) + o(1). \end{aligned}$$

Since $|\Delta T| \le C\sqrt{\ln p}(\|\Omega - I\|_F/n + \|(I-\Omega)\mu\|/\sqrt{n}) \ll \sqrt{2\|\Omega-I\|_F^2 + \|(\Omega-I)\mu\|^2}$ the denominator, so ΔT has negligible effects. Consider $\|(\Omega-I)\mu\|^2$. When $0 < \beta < 1$, then $\|(\Omega-I)\mu\| \ll \sqrt{p}\|\Omega - I\| \le \|\Omega - I\|_F$ by Lemma 4.3. When $1 \le \beta < 2$, there are at most constant non-zeros in each row of Ω. Hence, with probability $1 - o(1)$, $\|(\Omega-I)\mu\|^2 = \|V^{(1)}\mu + \xi\mu\|^2 \le p\xi^2 + (\eta^2p^2\nu)\tau^2\epsilon \ll \|\Omega - I\|_F^2$. In all, We only need to discuss

$$\frac{-\|\Omega-I\|_F^2/2 - 4a\|\mu\|_1}{\sqrt{2\|\Omega-I\|_F^2 + 16pa^2}} \le \frac{-\|\Omega-I\|_F^2/2 - 4a\|\mu\|_1}{2\max\{\sqrt{2}\|\Omega-I\|_F, 4a\sqrt{p}\}}.$$

Now we discuss two cases:

- Case 1. Suppose $\|\Omega - I\|_F^2 \gg a\sqrt{p} \to \infty$. In this case, both $\|\Omega - I\|_F^2/\|\Omega - I\|_F$ and $\|\Omega - I\|_F^2/a\sqrt{p}$ go to infinity, and the term of interest goes to negative infinity.
- Case 2. Suppose $\sqrt{p}\tau\epsilon \to \infty$. Then $\|\mu\|_1 \to \infty$ with probability $1 - o(1)$. If $\|\Omega - I\|_F^2 \gg a\sqrt{p}$, then it comes to case 1 which is solved. If $\|\Omega - I\|_F^2 \ll a\sqrt{p}$, then the term of interest comes to $a\|\mu\|_1/4a\sqrt{p} = \sqrt{p}\tau\epsilon(1 + o(1)) \to \infty$.

Therefore, $p_{i,\mu,\Omega} \to 0$ with probability $1 - o(1)$, and $E[p_{i,\mu,\Omega}] \to 0$. The same derivation holds for $p_{1,\mu,\Omega}$. As a conclusion, $MR(QDAw) \to 0$ in this region.

4.2.2. *Performance of QDAfs.* Now we consider the case $\tau \gg 1/\sqrt{n}$, i.e., $\theta < \delta/2$. The signals in μ are individually strong enough for successful recovery. Hence, we select features first, and then apply QDA on the post-selection data.

The feature selection step is as follows. Define d as

$$d \;=\; \Omega\hat{\mu}_1 - \hat{\mu}_0 \;\sim\; N\big((I+\Omega)\mu,\; \frac{1}{n_0}I + \frac{1}{n_1}\Omega\big). \tag{4.34}$$

When $\max_{1\le j\le p}|d_i| > 2\ln p/\sqrt{n}$, we let $d_j^{(t)} = I(|d_j| \ge t)$ with the threshold $t = 2\sqrt{\ln p}/\sqrt{n}$. Define $\hat{\mu}_0^{(t)} = \hat{\mu}_0 \circ d^{(t)}$ and $\hat{\mu}_d^{(t)} = d \circ d^{(t)}$ as the post-selection estimators. Define $\Omega^{(d)}$ as the sub-matrix of Ω consisting of rows and columns that $d^{(t)} = 1$. When $\theta < \delta/2$, this feature selection step happens with probability $1 - o(1)$. In supplementary materials [29], it is shown that the signals can be exactly recovered with probability $1 - o(1)$. Hence, we only consider the event that $\{t = \sqrt{2\ln p/n}\}$ and all the signals are exactly recovered.

In QDAfs, the criteria is updated as $Q^s = S^s - T_S^s$, where

$$S^s = X^\top(I-\Omega)X + 2\hat{\mu}_d^\top X, \quad T_S^s = (\hat{\mu}_0^{(t)})^\top(\Omega - I)\hat{\mu}_0^{(t)} - \ln|\Omega| - \frac{1}{n_0}Tr(\Omega^{(d)} - I).$$

Compare it with the ideal case that μ is known, the difference in the criteria is $\Delta Q = Q^s - Q(X,\mu,\Omega)$, where

$$\Delta Q = 2(\hat{\mu}_d^{(t)} - (I+\Omega)\mu)^\top X + \big[(\hat{\mu}_0^{(t)})^\top(I-\Omega)\hat{\mu}_0^{(t)} - \mu^\top(I-\Omega)\mu + \frac{1}{n_0}Tr(\Omega^{(d)} - I)\big].$$

In Supplementary Materials [29], we prove that, $\frac{Q(X,\mu,\Omega)}{\sqrt{2\|\Omega-I\|_F^2+16\|\mu\|^2}}$ is asymptotically normal distributed with mean $(-1)^{Y+1}\sqrt{\|\Omega - I\|_F^2/8 + \|\mu\|^2}$ and variance 1. Therefore, the mis-classification rate by $I\{Q(X,\mu,\Omega) > 0\}$ converges to 0 when the mean diverges.

When μ is unknown, the classification rule is $I\{Q^s = Q(X,\mu,\Omega) + \Delta Q > 0\}$. The error rate can be bounded by

$$\begin{aligned} p_{i,\mu,\Omega} &= P((-1)^i(Q(X,\mu,\Omega) + \Delta Q) > 0) \\ &= P\Big(\frac{(-1)^i(\|\Omega - I\|_F^2/2 + 4\|\mu\|^2) + \Delta Q}{\sqrt{2\|\Omega - I\|_F^2 + 16\|\mu\|^2}} > 0\Big), \quad i = 0, 1. \end{aligned} \tag{4.35}$$

Therefore, $|\Delta Q| \le \sqrt{2\|\Omega - I\|_F^2 + 16\|\mu\|^2}$ with probability $1 + o(1)$ suffices to show the success of QDAfs.

LEMMA 4.5. *Under the model assumptions and the definition of ΔQ, with probability $1 - o(1)$, there is*

$$|\Delta Q| \le O(\sqrt{p\epsilon_p(\xi_p^2 + p\epsilon_p\eta_p^2\nu_p)}/n) + \sqrt{p\epsilon_p}\tau_p \ln p(1 + o(1)). \tag{4.36}$$

By Lemma 4.5 about the magnitude of ΔQ, when $p\xi_p^2 + \eta_p^2 p^2 \nu_p \to \infty$ or $\tau_p^2 p\epsilon_p \to \infty$,

$$|\Delta Q| \ll \sqrt{p\xi_p^2/8 + \eta_p^2 p^2 \nu_p/8 + \tau_p^2 p\epsilon_p}(1+o(1)) = \sqrt{2\|\Omega - I\|_F^2 + 16\|\mu\|^2}.$$

Therefore, in the region of possibility identified by part (ii) of Theorem 1.2, $MR(QDAfs)$ converges to 0. □

4.3. *Proof of Theorem 2.3.* To prove Theorem 4.3, we start with the proof of Proposition 2.2 when μ is known and Ω is estimated by PCS in Section 4.3.1. The effects of estimated Ω can be found. Then we use the result to prove Theorem 2.3.

4.3.1. *Proof of Proposition 2.2.* When μ is known and Ω is estimated by PCS, we classify by $\hat{Y} = I(Q(X, \mu, \hat{\Omega}) > 0)$, where

$$Q(X, \mu, \hat{\Omega}) = X^\top(I - \hat{\Omega})X + 2\mu^\top(I + \hat{\Omega})X + \mu^\top(I - \hat{\Omega})\mu + \ln|\hat{\Omega}|.$$

We do not need to consider QDAw or QDAfs, and the focus is on $\hat{\Omega}$ by PCS only.

Let $Q(X, \mu, \hat{\Omega}) = S^{PCS} - T_S^{PCS}$, where $S^{PCS} = X^\top(I - \hat{\Omega})X + 2\mu^\top(I + \hat{\Omega})X$, $T_S = \mu^\top(\hat{\Omega} - I)\mu - \ln|\hat{\Omega}|$. Note that X and $\hat{\Omega}$ are independent. Given $\hat{\Omega}$, we derive the asymptotic distribution of S^{PCS} by Lemma 4.1. In details, the expectations and variances are

- $E[S^{PCS}|Y=0] = T_S^{PCS} - 4\mu^\top\hat{\Omega}\mu + \ln|\hat{\Omega}| + Tr(I - \hat{\Omega})$;
- $E[S^{PCS}|Y=1] = T_S^{PCS} + 4\mu^\top\mu + \ln|\hat{\Omega}| + Tr(\Omega^{-1}(I - \hat{\Omega}))$;
- $\mathrm{Var}(S^{PCS}|Y=0) = 2Tr((\hat{\Omega} - I)^2) + 16\mu^\top\hat{\Omega}^2\mu$;
- $\mathrm{Var}(S^{PCS}|Y=1) = 2Tr((\Omega^{-1} - \hat{\Omega}\Omega^{-1})^2) + 16\mu^\top\Omega^{-1}\mu$.

Define $Z_i = [S^{PCS} - E[S^{PCS}|Y=i]]/\sqrt{\mathrm{Var}(S^{PCS}|Y=i)}$ and $F_{Z_i}(x) = P(Z_i \le x)$, then $\sup_{-\infty<x<\infty}|F_{Z_i}(x) - \Phi(x)| \to 0$.

The mean and variance for $S^{PCS}|Y=0$ are similar as those of the ideal case, except all Ω are replaced by $\hat{\Omega}$. With similar derivations, we have that

$$p_{0,\mu,\Omega} = \Phi\Big(\frac{E[S^{PCS}|Y=0] - T_S^{PCS}}{\sqrt{\mathrm{Var}(S^{PCS}|Y=0)}}\Big) = \Phi\Big(-\sqrt{\|\hat{\Omega} - I\|_F^2/8 + \|\mu\|^2}\Big) + o(1). \tag{4.37}$$

The derivation for $p_{1,\mu,\Omega}$ is more complicated. Both the mean and variance of $S^{PCS}|Y=1$ involves the term $Tr(\Omega^{-1}(I-\hat{\Omega}))$, which is related to both $\hat{\Omega}$ and Ω. To bound it, we compare the term with $Tr(\hat{\Omega}^{-1}(I-\hat{\Omega}))$. The difference between them is $\Delta V = Tr((\Omega^{-1} - \hat{\Omega}^{-1})(I - \hat{\Omega}))$. The goal is to bound $|\Delta V|$.

For any square matrices A and B with ordered singular values as α_i and β_i, respectively. By Von Neuman's trace inequality in [28], $|Tr(AB)| \le \sum\alpha_i\beta_i \le \sqrt{\sum\alpha_i^2\sum\beta_i^2} = \sqrt{Tr(A^\top A)}\sqrt{Tr(B^\top B)}$. Apply this result to ΔV and recall that both Ω and $\hat{\Omega}$ has eigenvalues at $1 + o(1)$. Then we have

$$\begin{aligned}
|\Delta V| &\le \sqrt{Tr((\Omega^{-1} - \hat{\Omega}^{-1})^2)}\sqrt{Tr((I - \hat{\Omega})^2)} \\
&\le \sqrt{Tr(\Omega^{-2}(\Omega - \hat{\Omega})^2\hat{\Omega}^{-2})}\|I - \hat{\Omega}\|_F(1 + o(1)). \\
&\le \|\hat{\Omega} - \Omega\|_F\|I - \hat{\Omega}\|_F(1 + o(1)).
\end{aligned}$$

Introduce the bound of ΔV into $p_{1,\mu,\Omega}$,

$$p_{1,\mu,\Omega} = \Phi\Big(\frac{T_S^{PCS} - E[S^{PCS}|Y=1]}{\sqrt{\mathrm{Var}(S^{PCS}|Y=1)}}\Big)$$

$$\leq \Phi\Big(\frac{-4\|\mu\|^2 - \|\hat{\Omega}-I\|_F^2/2 + \|\hat{\Omega}-\Omega\|_F\|I-\hat{\Omega}\|_F}{\sqrt{2\|\hat{\Omega}-I\|_F^2 + 16\|\mu\|^2}}\Big) + o(1). \tag{4.38}$$

For $p_{i,\mu,\Omega}$, now we only need to consider $\|\hat{\Omega}-I\|_F^2$ and $\|\hat{\Omega}-\Omega\|_F^2$. According to Theorem 2.3 in [17], when $1-\delta/2 < \beta < 2$ and $\eta \gg 1/\sqrt{n}$, PCS recovers the exact support with probability $1-o(1/p^2)$, and $\max_{i,j}|\Omega(i,j)-\hat{\Omega}(i,j)| \leq C\sqrt{\ln p/n}$. Since $|\Omega(i,j)| \gg \sqrt{\ln p/n}$ on the off-diagonals, the estimation error is at a smaller order than the off-diagonal signals in Ω. On the diagonals, we have to consider several cases.

- Case 1. $\xi_p \gg 1/\sqrt{n}$. With probability $1-o(1)$, $|\hat{\Omega}(i,i)-\Omega(i,i)| \leq \sqrt{\ln p/n}$, which is at a smaller order than $\xi_p = |\Omega(i,i)-1|$, for all i. Therefore, $\|\hat{\Omega}-I\|_F^2 = \|\Omega-I\|_F^2(1+o(1))$ and ΔV is negligible compared to $\|\hat{\Omega}-I\|_F^2$. Since $\xi \gg 1/\sqrt{n}$, $p\xi_p^2 \to \infty$, therefore $\|\Omega - I\|_F^2 \to \infty$ and $p_{1,\mu,\Omega} \to 0$.
- Case 2. $\xi_p \ll \max\{\eta_p\sqrt{p\nu_p}, 1/\sqrt{p}\}$. When $\xi_p \ll 1/\sqrt{n}$, with probability $1-o(1)$, $|\hat{\Omega}(i,i)-1| \leq \ln p/\sqrt{n}$ for all i and therefore the diagonals of $\hat{\Omega}$ will be updated to 1. Hence, $\|I-\hat{\Omega}\|_F^2 = \eta_p^2 p^2 \nu(1+o(1))$ and $\|\Omega-\hat{\Omega}\|_F^2 = p\xi_p^2 + p^2\nu_p \ln p/n(1+o(1))$. When $\xi_p \ll \max\{\eta_p\sqrt{p\nu_p}, 1/\sqrt{p}\}$, $\|\hat{\Omega}-I\|_F^2 = \|\Omega-I\|_F^2(1+o(1)) + o(1)$ and ΔV is either $o(1)$ or negligible compared to $\|\hat{\Omega}-I\|_F^2$.

Introduce these terms into (4.37) and (4.38), we can see $p_{i,\mu,\Omega} \to 0$ when a) $\xi \gg 1/\sqrt{n}$, or b) $\xi \ll \max\{\eta\sqrt{p\nu}, 1/\sqrt{p}\}$ and $\|\Omega-I\|_F^2 \to \infty$ or $\|\mu\|^2 \to \infty$. Proposition 2.2 is proved.

4.3.2. *Proof of Theorem 2.3.* We examine the performance of QDAw with PCS for the region $\tau_p \ll 1/\sqrt{n}$ and that of QDAfs with PCS for the region $\tau_p \gg 1/\sqrt{n}$.

We first consider the weak signal region that $\tau \ll 1/\sqrt{n}$. Here we use adjusted PCS to estimate Ω and a constant vector $\hat{\mu} = a * \mathbf{1}$ to estimate the mean vector. We classify X to be in class 0 if $Q(X,\hat{\mu},\hat{\Omega}) < 0$, where

$$Q(X,\hat{\mu},\hat{\Omega}) = X^\top(I-\hat{\Omega})X + 2\hat{\mu}^\top(I+\hat{\Omega})X - \hat{\mu}_0^\top(I-\hat{\Omega})\hat{\mu}_0 + \ln|\hat{\Omega}| + \frac{1}{n_0}Tr(\hat{\Omega}-I).$$

We rewrite it as $Q(X,\hat{\mu},\hat{\Omega}) = S^{w,pcs} - T_S^{w,pcs}$, where $S^{w,pcs} = X^\top(I-\hat{\Omega})X + 2\hat{\mu}^\top(I+\hat{\Omega})X$ and $T_S^{w,pcs} = \hat{\mu}_0^\top(I-\hat{\Omega})\hat{\mu}_0 - \ln|\hat{\Omega}| - \frac{1}{n_0}Tr(\hat{\Omega}-I)$.

Apply Lemma 4.1 to $S^{w,pcs}$ and we can prove that,

- the expectations are

$$E[S^{w,pcs}|Y=0] = \mu^\top(I-\hat{\Omega})\mu + Tr(I-\hat{\Omega}) - 2a\mu^\top(I+\hat{\Omega})\mathbf{1},$$
$$E[S^{w,pcs}|Y=1] = \mu^\top(I-\hat{\Omega})\mu + Tr(\Omega^{-1}(I-\hat{\Omega})) + 2a\mu^\top(I+\hat{\Omega})\mathbf{1}.$$

- when $p \to \infty$, the asymptotic variances are

$$\mathrm{Var}(S^{w,pcs}|Y=i) = 2\|\hat{\Omega}-I\|_F^2 + (16pa^2 + \|(\hat{\Omega}-I)\mu\|^2)(1+o(1)), \qquad i=0,1.$$

Further, $S^{w,pcs}|Y=i$ normalized by mean and variance converges to normal distribution when $p \to \infty$.

Therefore, the error rates $p_{i,\mu,\Omega}$ can be approximated by

$$p_{0,\mu,\Omega} = \Phi\Big(\frac{E[S^{w,pcs}|Y=0] - T_S^{w,pcs}}{\sqrt{\mathrm{Var}(S^{w,pcs}|Y=0)}}\Big) + o(1)$$
$$\leq \Phi\Big(\frac{-\|\hat{\Omega}-I\|_F^2/2 - 4a\|\mu\|_1(1+o(1)) + \Delta T}{\sqrt{2\|\hat{\Omega}-I\|_F^2 + 16pa^2 + \|(\hat{\Omega}-I)\mu\|^2}}\Big) + o(1),$$

where $\Delta T = \hat{\mu}_0^\top(I-\hat{\Omega})\hat{\mu}_0 - \mu^\top(I-\hat{\Omega})\mu + \frac{1}{n_0}Tr(\hat{\Omega}-I)$.

Apply Lemma 4.2 to ΔT with $A=\hat{\Omega}$, $|\Delta T| \le C\ln p(\|\hat{\Omega}-I\|_F/n + \|(I-\hat{\Omega})\mu\|/\sqrt{n}) \ll \sqrt{2\|\hat{\Omega}-I\|_F^2 + \|(\hat{\Omega}-I)\mu\|^2}$, so ΔT has negligible effects. In Section 4.2.1, we found $\|(\hat{\Omega}-I)\mu\| \ll \|\hat{\Omega}-I\|_F$ holds with probability $1-o(1)$. Hence, we only need to discuss

$$\frac{-\|\hat{\Omega}-I\|_F^2/2 - 4a\|\mu\|_1}{\sqrt{2\|\hat{\Omega}-I\|_F^2 + 16pa^2}} \le \frac{-\|\hat{\Omega}-I\|_F^2/2 - 4a\|\mu\|_1}{2\max\{\sqrt{2}\|\hat{\Omega}-I\|_F, 4a\sqrt{p}\}}.$$

In Section 4.3.1, we have found $\|\hat{\Omega}-I\|_F = \|\Omega-I\|_F(1+o(1))+o(1)$ in the current region of interest. Hence, it comes back to the equation when Ω is known. In the region of possibility identified by part (i) of Theorem 2.3, $MR(QDAw) \to 0$.

Now we consider the case $\tau_p \gg 1/\sqrt{n}$, i.e., $\theta < \delta/2$. The signals in μ are individually strong enough for successful recovery. Hence, we estimate Ω by PCS, then threshold on $d = \hat{\Omega}\hat{\mu}_1 - \hat{\mu}_0$. QDA is applied to the post-selection data.

In [29, Appendix C.1], it is shown that the signals can be exactly recovered with probability $1-o(1)$. Hence, we only consider the event that $\{t = \sqrt{2\ln p/n}\}$ and all the signals are exactly recovered.

By Proposition 2.2, we analyze the performance of $Q(X,\mu,\hat{\Omega}) = S^{PCS} - T_S^{PCS}$. In QDAfs, the criteria is updated as

$$Q(X,\hat{\mu},\hat{\Omega}) = Q(X,\mu,\hat{\Omega}) + \Delta Q, \tag{4.39}$$

where $\Delta Q = 2(\hat{\mu}_d^{(t)} - (I+\hat{\Omega})\mu)^\top X + \big[(\hat{\mu}_0^{(t)})^\top(I-\hat{\Omega})\hat{\mu}_0^{(t)} - \mu^\top(I-\hat{\Omega})\mu + \frac{1}{n_0}Tr(\hat{\Omega}^{(d)}-I)\big]$. The following lemma bounds $|\Delta Q|$.

LEMMA 4.6. *Under the model assumptions and the definition of ΔQ, there is*

$$|\Delta Q| \le \eta\tau\max\{p\epsilon\nu, 1\}\ln p + O_p(\sqrt{4n^{-1}p\epsilon}). \tag{4.40}$$

Combining Lemma 4.6 with Section 4.3.1 about Q, the errors are

$$p_{i,\mu,\Omega} = P((-1)^i * (S^{PCS} - T_S^{PCS} + \Delta Q) > 0)$$
$$= \Phi\Big(\frac{(-1)^i * (T_S^{PCS} - E[S^{PCS}|Y=i])}{\sqrt{\mathrm{Var}(S^{PCS}|Y=i)}} + \frac{(-1)^i * \Delta Q}{\sqrt{\mathrm{Var}(S^{PCS}|Y=i)}}\Big) + o(1), \quad i=0,1. \tag{4.41}$$

The first term $\frac{(-1)^i*(T_S - E[S^{PCS}|Y=i])}{\sqrt{\mathrm{Var}(S^{PCS}|Y=i)}} = -\sqrt{\|\Omega-I\|_F^2/8 + \|\mu\|^2}(1+o(1)) + o(1)$ in Section 4.3.1. The second term can be bounded by

$$\frac{|\Delta Q|}{\sqrt{\mathrm{Var}(S^{PCS}|Y=i)}} \le \frac{\eta\tau\max\{p\epsilon\nu,1\}\ln p + O_p(\sqrt{4n^{-1}p\epsilon})}{\sqrt{p\xi^2/8 + \eta^2p^2\nu/8 + \tau^2p\epsilon}(1+o(1))}.$$

It goes to 0 in the region of possibility identified in part (ii) of Theorem 2.3.

Therefore, in the region of possibility identified by part (ii) of Theorem 2.3, $MR(QDAfs)$ converges to 0. □

5. Real Data Analysis. In this paper, we consider the rats dataset present in [33]. As we introduced in Section 1.4, this data set record the gene expressions of live rats in response to different drugs and toxicant. There are 181 samples and 8491 genes, where 61 samples are labeled as toxicant and the other 120 are labeled as other drugs. We compare QDA with LDA, where the latter one is shown to enjoy the best performance compared to classifiers such as SVM, RandomForest, GLasso and FoBa. The QDA with feature selection for the real data is discussed in Section 5.1 and the implementation details and results are in Section 5.2.

5.1. *Procedure for the real data.* Here, we present a procedure for the classification based on QDA for the real data. For the real data, we have to estimate Ω_0, Ω_1, μ_0 and μ_1 separately. Further, we need to eliminate the effect of the feature variances. Hence, there is an additional scaling step in the following algorithm.

TABLE 7
Algorithm 2: Pseudocode for QDA with feature selection on real data

Input: data points (X_i, Y_i), $1 \le i \le n$; threshold $t > 0$; new data point X; tuning parameters: C, t.
Output: label $\hat{Y}$.

1. Find $\hat{\Omega}_0$ and $\hat{\Omega}_1$ by PCS. Let $\hat{\mu}_0 = \frac{1}{n_0}\sum_{i:Y_i=0} X_i$ and $\hat{\mu}_1 = \frac{1}{n_1}\sum_{i:Y_i=1} X_i$, where $n_1 = \sum_{i=1}^n Y_i$ and $n_0 = n - n_1$.
2. Let $\hat{\Omega}_{\text{diff}} = \hat{\Omega}_0 - \hat{\Omega}_1 - diag(\hat{\Omega}_0 - \hat{\Omega}_1)$.
3. Let $d_0 = \hat{\Omega}_0\hat{\mu}_0/s_0$, $d_1 = \hat{\Omega}\hat{\mu}_1/s_1$ and $d = d_1 - d_0 = (d(1), \ldots, d(p))^\top$. Here s_i are the standard deviation vector of the train data from class i, $i = 0, 1$. The division means element-wise division.
4. Thresholding: Let $d^{(t)}$ denote the indicator vector of feature selection, i.e. $d^{(t)}(j) = 1\{|d(j)| \ge t\}$, for $j = 1, \ldots, p$. Let $\hat{\mu}_d^{(t)}$ be the hard-thresholded $\hat{\mu}_d^{(t)} = d \circ d^{(t)}$.
5. Scale X as $x_j = [X_j - \bar{\mu}_j]/s_j$, where $\bar{\mu} = (\hat{\mu}_1 + \hat{\mu}_0)/2$ and s is the standard error of the pooled data $s_j = \sqrt{[(n_0 - 1)((s_0)_j)^2 + (n_1 - 1)((s_1)_j)^2]/(n_0 + n_1 - 2)}$.
6. QDA Score: Calculate the QDA score $Q = x^\top \hat{\Omega}_{\text{diff}} x + 2(\hat{\mu}_d^{(t)})^\top X + C$.
7. Prediction: Predict $\hat{Y} = I\{Q > 0\}$.

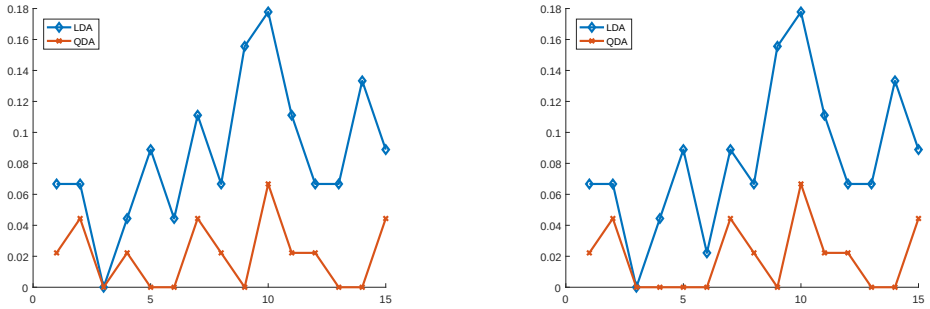
Here are two tuning parameters, t and C. In the implementations, we use a grid search to find the optimal values of them. Details in Section 5.2.

5.2. *Implementation and Results.* Following the setup of the data analysis in [17], we apply 4-fold data splitting to the sample. For each class, we randomly draw one fourth of the samples, and then combine them to be the test data while using the leftover to be the training data. We do the splitting for 15 times independently and record the error with QDA and LDA for each splitting. The data (sample indices) for the 15 splittings is available upon request.

In the real data analysis section, we focus on comparing QDA and LDA. The LDA is implemented within the setting of QDA, where in Step (3) of the algorithm in Section 5.1 we use clipping thresholding instead of hard thresholding, and in Step (5) we set $\hat{\Omega}_{\text{diff}} = 0$ for LDA. The clipping threshold is employed since it gives much more satisfactory results than hard thresholding for LDA; details in [17]. For QDA, the two ways give similar results. Since the calculation of $\hat{d}$ involves the calculation of $\hat{\Omega}_0$ and $\hat{\Omega}_1$ and the thresholding, LDA algorithm has exactly the same tuning parameters with QDA. The procedure of determining these tuning parameters are the same for both algorithms, so that the results are comparable.

For PCS, there are four tuning parameters (q_1, q_2, δ, L). Here we use the same set of tuning parameters for the estimation of both $\hat{\Omega}_0$ and $\hat{\Omega}_1$, since the two classes are from the same data set and the performance of PCS is not sensitive to the choice of these parameters ([17]). Following the setting in [17], we set $(\delta, L) = (.1, 30)$, and also tried $(\delta, L) = (.1, 50)$. For (q_1, q_2), we consider $.1 \le q_k \le 1$, with an increment of .1, $k = 0, 1$. The selection is done by grid search.

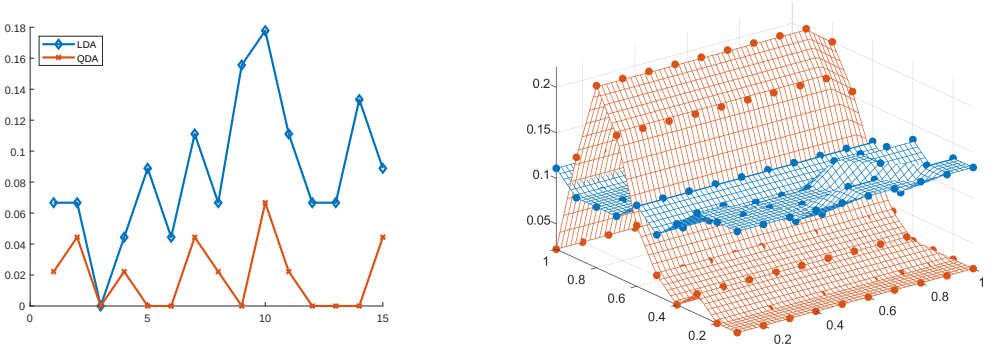
For Algorithm 2, there are two tuning parameters t and C. We set the ranges $[t_{\min}, t_{\max}] = [0, \max_{1\le j\le p}|d_j|]$ with an increment of .1 and $[C_{\min}, C_{\max}] = [-50, 50]$ or $[C_{\min}, C_{\max}] = [-100, 100]$ with an increment of 1. The smallest error is obtained over a grid search of t, C, and (q_1, q_2). This step is the same for both QDA and LDA to be fair. We compare the smallest error that LDA and QDA can achieve.



(a) $(C_{\min}, C_{\max}) = (-50, 50)$ (b) $(C_{\min}, C_{\max}) = (-100, 100)$

FIG 3. *Comparison of testing error rate (y-axis) of LDA and QDA for the rats data with* $(\delta_1, L_1) = (\delta_2, L_2) = (.1, 30)$ *and 15 data splittings.*

Both the LDA test error (the best error) and the QDA test error (the best error) over all 15 data splittings are reported in Figure 3. In the left panel of Figure 3, we can see that the error rates of LDA are all above QDA at every data splitting. To better show the difference between them, we also plot the testing error rate in the right panel of Figure 3 for a wider grid-search range that $[C_{\min}, C_{\max}] = [-100, 100]$.



(a) $L_1 = L_2 = 50$ (b) $(q_1, q_2) \in (.1, 1) \times (.1, 1)$

FIG 4. *Comparison of testing error rate of LDA and QDA for the rats data on: (a)* $(\delta_1, L_1) = (\delta_2, L_2) = (.1, 50)$ *among 15 data splittings (*x*-axis); (b) varying choices of* $(q_1, q_2) \in (.1, 1) \times (.1, 1)$ *for one splitting of 15 splittings in Figure 4(a).*

The impact of the tuning parameters in PCS is presented in Figure 4. When L changes from 30 to 50, the results are summarized in subfigure (a), which is similar. This comparison clearly demonstrates the expected superiority of QDA over LDA. When (q_1, q_2) changes, the results for one splitting are presented in subfigure (b). It suggests a proper choice of the tuning parameters will largely improve the QDA results, and overcome the LDA classifier.

As a conclusion, the results suggest that, for rats data, QDA outperforms LDA in terms of both best error rate and average error; with the results in [17] for other methods, where the authors have shown that HCT-based LDA significantly outperforms all other HCT-based methods as well as SVM and RF, our findings also suggest that the QDA gives a better separation than the LDA by taking into account the second order difference between the two classes.

6. Discussion. This paper focuses on the classification problem associated with the use of QDA and feature selection for data of rare and weak signals. We derived the successful and unsuccessful classification regions, by using first the case of a known mean vector and covariance matrix, then the case of an unknown mean vector but known covariance matrix, and finally the case in which both mean vector and covariance matrix were unknown. We also proved that these regions were actually the possibility and impossibility regions under the same modeling, which indicates that QDA achieves the optimal classification results in this manner. In addition, we developed computing and classification algorithms that incorporated feature selection for rare and weak data. With these algorithms, our real data analysis showed that QDA had much-improved performance over LDA.

Our theoretical results showed that the two sets of signal weakness and sparsity parameters, one set from the mean vector and the other set from the covariance matrix, influence the possibility/impossibility regions or QDA successful/unsuccessful regions almost independently (except for a $\max$ operator over the two sets of parameters) when the covariance matrix is known. When both the mean vector and covariance matrix are unknown, the two sets of parameters interact with each other as indicated in Theorem 2.3. For the latter case, the analysis of the mis-classification rate is very complicated and we only obtained partial results for this most general case; further study is therefore warranted. Also, for the precision matrix Ω given in (1.8), we can introduce sparsity and weakness in the diagonal elements of $I - \Omega$, the difference in precision matrices, instead of using a constant $\xi = 1 - c$ for all diagonal elements.

SUPPLEMENTARY MATERIAL

Supplememtary Material for "High Dimensional Quadratic Discriminant Analysis: Optimality and Phase Transitions"
(http://www.e-publications.org/ims/support/dowload/imsart-ims.zip). Owing to space constraints, some technical proofs are relegated a supplementary document [29]. It contains proofs of Theorem 2.4 and lemmas mentioned in this document.

REFERENCES

[1] ANDERSON, T. W. (2003). *An introduction to multivariate statistical analysis*. Wiley, New York.

[2] AOSHIMA, M. and YATA, K. (2019). High-dimensional quadratic classifiers in non-sparse settings. *Methodology and Computing in Applied Probability* **21** 663–682.

[3] BREIMAN, L. (2001). Random forests. *Mach. Learn.* **24** 5–32.

[4] BURGES, C. (1998). A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Discov.* **2** 121–167.

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

[6] DONOHO, D. and JIN, J. (2008). Higher Criticism Thresholding: Optimal feature selection when useful features are rare and weak. *Proceedings of the National Academy of Sciences* **105** 14790–14795.

[7] DONOHO, D. and JIN, J. (2015). Higher Criticism for large-scale inference, especially for rare and weak effects. *Statistical Science* **30** 4427–4448.

[8] EFRON, B. (2011). *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Cambridge Univ. Press, Cambridge.

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

[10] FAN, J., FENG, Y. and TONG, X. (2012). A road to classification in high dimension space: the regularized optimal affine discriminant. *J. Roy. Statist. Soc.* **74** 745-771.

[11] FAN, J., LIAO, Y. and LIU, H. (2016). An overview of the estimation of large covariance and precision matrices. *The Econometrics Journal* **19** C1–C32.

[12] FAN, Y., JIN, J. and YAO, Z. (2013). Optimal classification in sparse Gaussian graphic model. *Annals of Statistics* **41** 2537–2571.

[13] FAN, Y., KONG, Y., LI, D. and ZHENG, Z. (2015). Innovated interaction screening for high-dimensional nonlinear classification. *The Annals of Statistics* **43** 1243–1272.
[14] FISHER, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of eugenics* **7** 179–188.
[15] FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2007). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9** 432–441.
[16] FRIEDMAN, J. H. (1989). Regularized discriminant analysis. *Journal of the American statistical association* **84** 165–175.
[17] HUANG, S., JIN, J. and YAO, Z. (2016). Partial Correlation Screening for estimating large precision matrices, with applications to classification. *Annals of Statistics* **44** 2018-2057.
[18] INGSTER, Y., POUET, C. and TSYBAKOV, A. (2009). Classification of sparse high-dimensional vectors. *Phil. Trans. R. Soc. A* **367** 4427–4448.
[19] JIANG, B., WANG, X. and LENG, C. (2018). A Direct Approach for Sparse Quadratic Discriminant Analysis. *Journal of Machine Learning Research* **19** 1–37.
[20] JIN, J. (2009). Impossibility of successful classification when useful features are rare and weak. *Proc. Natl. Acad. Sci.* **106** 8859–8864.
[21] JIN, J. (2009). Impossibility of successful classification when useful features are rare and weak. *Proceedings of the National Academy of Sciences* **106** 8859–8864.
[22] JIN, J. and KE, Z. (2016). Rare and weak effects in large-scale inference: methods and phase diagrams. *Statistica Sinica* **26** 1–34.
[23] JIN, J., KE, Z. T. and WANG, W. (2017). Phase transitions for high dimensional clustering and related problems. *The Annals of Statistics* **45** 2151–2189.
[24] LACHENBRUCH, P. A. and GOLDSTEIN, M. (1979). Discriminant analysis. *Biometrics* 69–85.
[25] LAURITZEN, S. L. (1996). *Graphical models* **17**. Clarendon Press.
[26] LI, Q. and SHAO, J. (2015). Sparse quadratic discriminant analysis for high dimensional data. *Statistica Sinica* 457–473.
[27] MCLACHLAN, G. J. (2004). *Discriminant analysis and statistical pattern recognition* **544**. John Wiley & Sons.
[28] VON NEUMANN, J. (1937). *Some matrix-inequalities and metrization of matric space* **1**.
[29] WANG, W., WU, J. and YAO, Z. (2021). Supplememtary Material for "High Dimensional Quadratic Discriminant Analysis: Optimality and Phase Transitions". *Manuscript.*
[30] WASSERMAN, L. and ROEDER, K. (2009). High dimensional variable selection. *Ann. Statist.* **37** 2178–2201.
[31] WU, Y., QIN, Y. and ZHU, M. (2019). Quadratic discriminant analysis for high-dimensional data. *Statistica Sinica* **29** 939–960.
[32] XIONG, C., ZHANG, J. and LUO, X. (2016). Ridge-forward quadratic discriminant analysis in high-dimensional situations. *Journal of Systems Science and Complexity* **29** 1703–1715.
[33] YOUSEFI, M., HUA, J., SIMA, C. and DOUGHERTY, E. (2010). Reporting bias when using real data sets to analyze classification performance. *Bioinformatics* **26** 68-76.
[34] YU, G. and BIEN, J. (2017). Learning local dependence in ordered data. *The Journal of Machine Learning Research* **18** 1354–1413.
[35] YUAN, M. and LIN, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika* **94** 19–35.