Statistica Sinica **25** (2015), 457-473
doi:http://dx.doi.org/10.5705/ss.2013.150

SPARSE QUADRATIC DISCRIMINANT ANALYSIS FOR HIGH DIMENSIONAL DATA

Quefeng Li and Jun Shao

University of Wisconsin-Madison and East China Normal University

Abstract: Many contemporary studies involve the classification of a subject into two classes based on n observations of the p variables associated with the subject. Under the assumption that the variables are normally distributed, the well-known linear discriminant analysis (LDA) assumes a common covariance matrix over the two classes while the quadratic discriminant analysis (QDA) allows different covariance matrices. When p is much smaller than n, even if they both diverge, the LDA and QDA have the smallest asymptotic misclassification rates for the cases of equal and unequal covariance matrices, respectively. However, modern statistical studies often face classification problems with the number of variables much larger than the sample size n, and the classical LDA and QDA can perform poorly. In fact, we give an example in which the QDA performs as poorly as random guessing even if we know the true covariances. Under some sparsity conditions on the unknown means and covariance matrices of the two classes, we propose a sparse QDA based on thresholding that has the smallest asymptotic misclassification rate conditional on the training data. We discuss an example of classifying normal and tumor colon tissues based on a set of $p = 2,000$ genes and a sample of size $n = 62$, and another example of a cardiovascular study for $n = 222$ subjects with $p = 2,434$ genes. A simulation is also conducted to check the performance of the proposed method.

Key words and phrases: Classification, high dimensionality, normality, smallest asymptotic misclassification rate, sparsity estimates, unequal covariance matrices.

1. Introduction

Consider the problem of classifying a p-dimensional normally distributed vector $\boldsymbol{x}$ into one of two classes represented by two p-dimensional normal distributions, $N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, where $\boldsymbol{\mu}_k$'s are mean vectors and $\boldsymbol{\Sigma}_k$'s are positive definite covariance matrices. If $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$, $k = 1, 2$, are known, then an optimal classification rule having the smallest possible misclassification rate can be constructed. However, $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$, $k = 1, 2$, are usually unknown and a classification rule has to be constructed using a training sample to estimate unknown parameters. In the traditional setup where the dimension p of $\boldsymbol{x}$ is fixed, the well-known linear discriminant analysis (LDA) for the case of $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$ or quadratic discriminant analysis (QDA) for the case of $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$ has the smallest asymptotic misclassification rate in the sense that its misclassification rate converges to that of the optimal rule as the training sample size $n \to \infty$. In fact, Shao et al. (2011) showed that the LDA still has the smallest asymptotic misclassification rate when p diverges to infinity at a rate slower than $\sqrt{n}$ as $n \to \infty$. A similar result for the QDA is established in this paper.

Nowadays, many more characteristics are collected simultaneously, which results in a high dimensional $\boldsymbol{x}$. In many recent applications, p is much larger than the training sample size n, referred to as the large-p-small-n problem, or ultra-high dimension problem when p is of the order $e^{n^{\nu}}$ with a constant $\nu \in (0,1)$. An example is a study with genetic or microarray data. In one of our examples presented in Section 4, to classify tumor and normal colon tissues by Oligonucleotide microarray technique, $p = 2,000$ genes are involved whereas the size of the sample is only $n = 62$. Other examples include data from radiology, biomedical imaging, signal processing, climate, and finance. When $p > n$, Bickel and Levina (2004) and Shao et al. (2011) showed that the LDA may be asymptotically as bad as random guessing.

Some improvements over the LDA for large p problems have been made in recent years. See, for example, Bickel and Levina (2004), Fan and Fan (2008), Guo, Hastie, and Tibshirani (2007), Clemmensen, Hastie, and Ersbøll (2008), Qiao, Zhou, and Huang (2009), and Zhang and Wang (2011). Moreover, Shao et al. (2011) proposed a sparse LDA (SLDA) by thresholding and showed that it has the smallest asymptotic misclassification rate under some sparsity conditions on unknown parameters. Another attempt to improve LDA is to directly find a linear rule that minimize the misclassification rate. Under normality, Fan, Feng, and Tong (2012) proposed a Regularized Optimal Affine Discriminant (ROAD) that directly minimizes the misclassification rate and showed consistency of their method to the Bayes rule. Han, Zhao, and Liu (2012) relaxed the normality assumption and extended the linear rule to a copula model, reaching a similar consistency result. Cai and Liu (2011) built a linear rule by finding a sparse classification direction and showed the consistency of their method.

To the best of our knowledge, most theoretical work on the asymptotic misclassification rate of discriminant analysis assumes a common covariance matrix. Very little has been done for the QDA, even for the case where $p < n$. Cheng (2004) established some asymptotic results for the QDA, but assumed that $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ are diagonal. Simon and Tibshirani (2011) proposed a regularization-based algorithm to estimate parameters in the QDA for high-dimensional settings, but didn't discuss asymptotical optimality.

The purpose of this paper is to construct a sparse QDA (SQDA) and establish its asymptotic optimality under some sparsity conditions on $\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$, $\boldsymbol{\Sigma}_1$, and $\boldsymbol{\Sigma}_2$. Although our proposed SQDA is based on the well-known thresholding methodology, the study of asymptotic properties of the SQDA is much more complicated and difficult than that for the SLDA studied in Shao et al. (2011). First, the misclassification rate of the LDA has a closed form, but the misclassification rate of the QDA does not, since it involves a probability related to a complicated quadratic form of $\boldsymbol{x}$. Second, for a good performance of the QDA, we need sparsity conditions on each of $\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$, $\boldsymbol{\Sigma}_1$, $\boldsymbol{\Sigma}_2$, and the difference $\boldsymbol{\Sigma}_2 - \boldsymbol{\Sigma}_1$ lest the QDA be asymptotically as bad as random guessing. This is quite different from the LDA, in which we only need sparsity of $\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$ and $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$. To accommodate this, we construct mean estimators by thresholding and covariance matrix estimators by double thresholding, one for the covariance matrices and another for their differences. Finally, because of the existence of quadratic forms of $\boldsymbol{x}$, we have to handle convergence of estimated covariance matrices in terms of not only the usual L_2 norm, but also L_1 norm, the Frobenius norm, and another norm defined in Section 2. For the SLDA, however, only L_2 norm is needed. As by-products, we derived some results on convergence of estimated covariance matrices in terms of several norms that may be useful in other studies.

The rest of this paper is organized as follows. In Section 2, we introduce some notation and preliminary results, including a result showing that the classical QDA has the smallest asymptotic misclassification rate when $p \to \infty$, but at a rate much slower than n, and an example indicating that it is necessary to regulate the difference of covariance matrices. The main results are presented in Section 3, where we first state some sparsity conditions on $\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$ and $\boldsymbol{\Sigma}_k$ and construct sparse estimators of $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ based on the training data, which results in our proposed SQDA classification rule. Asymptotic properties of sparse estimators and the SQDA are established under the sparsity conditions and some conditions on the divergence rate of p. Section 4 contains a simulation comparison of the SQDA, SLDA, and ROAD. It also presents two data example, in which we compare the SQDA with the SLDA, ROAD, and some other popular classifiers in the literature. Proofs are given in a supplementary document.

2. Preliminary Results

For vector $\boldsymbol{a}$, $\boldsymbol{a}'$ denotes its transpose and $\|\boldsymbol{a}\|$ denotes its L_2 norm. For symmetric $p \times p$ matrix $\boldsymbol{A}$ whose (i,j)th element is a_{ij}, we take $\|\boldsymbol{A}\|_G = \sum_{i=1}^p \sum_{j=1}^p |a_{ij}|$, $\|\boldsymbol{A}\|_F = (\sum_{i=1}^p \sum_{j=1}^p a_{ij}^2)^{1/2}$, $\|\boldsymbol{A}\|_1 = \max_{1 \le i \le p} \sum_{j=1}^p |a_{ij}|$, and $\|\boldsymbol{A}\|_2 = \max_{1 \le j \le p} |\lambda_{p,j}(\boldsymbol{A})|$, where $\lambda_{p,j}(\boldsymbol{A})$ is the jth smallest eigenvalue of $\boldsymbol{A}$. Here, $\|\boldsymbol{A}\|_F$ is the Frobenius norm related to the L_2 norm, and $\|\boldsymbol{A}\|_G$ is a counterpart of $\|\boldsymbol{A}\|_F$ related to the L_1 norm: $\|\boldsymbol{A}\|_2 \le \|\boldsymbol{A}\|_1 \le \|\boldsymbol{A}\|_G$ and $\|\boldsymbol{A}\|_2 \le \|\boldsymbol{A}\|_F$.

Lemma 1. *If $\boldsymbol{L}$, $\boldsymbol{C}$, and $\boldsymbol{R}$ are symmetric $p \times p$ matrices,*

$$\|\boldsymbol{LCR}\|_G \le \|\boldsymbol{L}\|_1 \|\boldsymbol{C}\|_G \|\boldsymbol{R}\|_1 \quad \textit{and} \quad \|\boldsymbol{LCR}\|_F \le \|\boldsymbol{L}\|_2 \|\boldsymbol{C}\|_F \|\boldsymbol{R}\|_2.$$

Let $\boldsymbol{\mu}_k$ be the mean and $\boldsymbol{\Sigma}_k$ be the covariance matrix of the p-dimensional normal distribution, $k = 1, 2$, $\boldsymbol{I}$ be the identity matrix of order p, and

$$\boldsymbol{\delta} = \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1, \quad \boldsymbol{\Delta} = \boldsymbol{\Sigma}_2 - \boldsymbol{\Sigma}_1, \quad \boldsymbol{\nabla} = \boldsymbol{\Sigma}_2^{-1} - \boldsymbol{\Sigma}_1^{-1}, \quad \boldsymbol{\Lambda} = \boldsymbol{\Sigma}_1^{1/2}\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1^{1/2} - \boldsymbol{I}.$$

Throughout, we assume the following conditions on $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$: there are positive constants m and M (not depending on p) such that

(C1) all absolute values of components of $\boldsymbol{\mu}_k \leq M$;

(C2) $m \leq$ all eigenvalues of $\boldsymbol{\Sigma}_k \leq M$;

(C3) $m \leq \liminf_{p\to\infty} D_p$, where $D_p = \sqrt{\|\boldsymbol{\Delta}\|_F^2 + \|\boldsymbol{\delta}\|^2}$.

Condition (C3) avoids the trivial case where the two classes are the same as $p \to \infty$.

When $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ are known, the optimal classification rule, the Bayes rule, classifies $\boldsymbol{x}$ to class 2 if and only if

$$(\boldsymbol{x} - \boldsymbol{\mu}_1)'\boldsymbol{\nabla}(\boldsymbol{x} - \boldsymbol{\mu}_1) - 2\boldsymbol{\delta}'\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_1) + \boldsymbol{\delta}'\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\delta} - \log(|\boldsymbol{\Sigma}_1|/|\boldsymbol{\Sigma}_2|) < 0. \quad (2.1)$$

It has a misclassification rate of

$$R_B = \frac{R_{B1} + R_{B2}}{2}, \qquad R_{Bk} = P\,(\text{incorrectly classify } \boldsymbol{x} \text{ to class } k)\,. \quad (2.2)$$

If $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$, then the probabilities in R_B are related to normal distributions. Otherwise, these probabilities have no known form and we need the following.

Lemma 2. *Suppose that* (C1)−(C2) *hold. Let* $\boldsymbol{z} \sim N_p(\boldsymbol{0}, \boldsymbol{I})$ *and* $T_p = \boldsymbol{z}'\boldsymbol{\Lambda}\boldsymbol{z} - 2\boldsymbol{\delta}'\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1^{1/2}\boldsymbol{z}$*. If* $D_p \to \infty$ *as* $p \to \infty$*, then* $[T_p - \mathrm{E}(T_p)]/\sqrt{\mathrm{Var}(T_p)} \xrightarrow{D} N(0,1)$*, where* $\xrightarrow{D}$ *denotes convergence in distribution.*

When $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ are unknown, the optimal rule cannot be used. To estimate $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$, we assume that there is a training sample $\boldsymbol{X} = \{\boldsymbol{x}_{ki}, i = 1, \ldots, n_k, k = 1, 2\}$, where n_k is the sample size for class k, $\boldsymbol{x}_{ki} \sim N_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, $k = 1, 2$, all $\boldsymbol{x}_{ki}$'s are independent, and $\boldsymbol{X}$ is independent of $\boldsymbol{x}$ to be classified. For any unknown $\boldsymbol{a}$ or $\boldsymbol{A}$, let $\hat{\boldsymbol{a}}$ and $\hat{\boldsymbol{A}}$ be their estimators based on the training sample $\boldsymbol{X}$. Then the sample analog of the optimal rule classifies $\boldsymbol{x}$ to class 2 if and only if

$$(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_1)'\hat{\boldsymbol{\nabla}}(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_1) - 2\hat{\boldsymbol{\delta}}'\hat{\boldsymbol{\Sigma}}_2^{-1}(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_1) + \hat{\boldsymbol{\delta}}'\hat{\boldsymbol{\Sigma}}_2^{-1}\hat{\boldsymbol{\delta}} - \log(|\hat{\boldsymbol{\Sigma}}_1|/|\hat{\boldsymbol{\Sigma}}_2|) < 0. \quad (2.3)$$

Its conditional misclassification rate, given $\boldsymbol{X}$, is

$$R(\boldsymbol{X}) = \frac{R_1(\boldsymbol{X}) + R_2(\boldsymbol{X})}{2},$$

where

$$R_k(\boldsymbol{X}) = P\left(\text{incorrectly classify } \boldsymbol{x} \text{ to class } k \mid \boldsymbol{X}\right), \tag{2.4}$$

and the probability is with respect to $\boldsymbol{x}$ conditional on $\boldsymbol{X}$. Unlike the LDA case where $R(\boldsymbol{X})$ has a simple explicit form, the probability $R_k(\boldsymbol{X})$ is complicated and does not have an explicit form.

The limiting process we consider has $n = n_1 + n_2 \to \infty$ with $n_1/n \to$ a constant strictly between 0 and 1. Throughout, p is considered as a function of n and p may diverge to ∞ at a certain rate as $n \to \infty$.

Theorem 1. *Suppose that conditions* (C1)−(C3) *hold.*
(i) *When D_p is bounded as $p \to \infty$, if $p = o(n^{1/5})$, and*
(C4) *the density function of T_p is bounded by a constant not depending on p, then*

$$R_{\text{QDA}}(\boldsymbol{X}) - R_B \xrightarrow{P} 0, \tag{2.5}$$

where $R_{\text{QDA}}(\boldsymbol{X})$ is the conditional misclassification rate of the QDA given the training data $\boldsymbol{X}$, R_B is the optimal rate of the Bayes rule, and $\xrightarrow{P}$ denotes convergence in probability.
(ii) *When $D_p \to \infty$ as $p \to \infty$ and $p < n$, if $p^2/(nD_p^2) \to 0$, then* (2.5) *holds.*

(C4) holds when $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$, or when $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}_1^{1/2}\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1^{1/2} - \boldsymbol{I}$ has some eigenvalues that are always equal to 0; (C4) also holds if $\boldsymbol{\Lambda}$ has at least two eigenvalues of in $(-\infty, m]$ or $[m, \infty)$ (see the proof of Theorem 3).

When $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$ and $p > n$, the results in Bickel and Levina (2004) and Shao et al. (2011) indicated that some sparsity conditions on $\boldsymbol{\delta}$ and $\boldsymbol{\Sigma}$ are necessary in order to obtain an asymptotically optimal classification rule. When $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$ and $p > n$, we need sparsity conditions on $\boldsymbol{\delta}$ and $\boldsymbol{\Sigma}_k$, $k = 1, 2$. Further, some condition on $\boldsymbol{\Delta} = \boldsymbol{\Sigma}_2 - \boldsymbol{\Sigma}_1$ is necessary.

We consider the case where $p/n \to \infty$, $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ are known, but $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are unknown. In this case, the QDA classifies $\boldsymbol{x}$ to class 2 if and only if

$$\hat{T}_p - \mathrm{E}(\hat{T}_p|\boldsymbol{X}) < -\hat{\phi}_p,$$

where $\hat{T}_p = (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_1)'\boldsymbol{\nabla}(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_1) - 2\hat{\boldsymbol{\delta}}'\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_1)$, $\mathrm{E}(\hat{T}_p|\boldsymbol{X}) = \mathrm{tr}(\boldsymbol{\Lambda}) + (\hat{\boldsymbol{\mu}}_1 - \boldsymbol{\mu}_1)'\boldsymbol{\nabla}(\hat{\boldsymbol{\mu}}_1 - \boldsymbol{\mu}_1) - 2\hat{\boldsymbol{\delta}}'\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_1 - \hat{\boldsymbol{\mu}}_1)$ when $\boldsymbol{x} \sim N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$, and $\hat{\phi}_p = \mathrm{tr}(\boldsymbol{\Lambda}) - \log(|\boldsymbol{\Sigma}_1|/|\boldsymbol{\Sigma}_2|) + (\hat{\boldsymbol{\mu}}_1 - \boldsymbol{\mu}_1)'\boldsymbol{\nabla}(\hat{\boldsymbol{\mu}}_1 - \boldsymbol{\mu}_1) - 2\hat{\boldsymbol{\delta}}'\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_1 - \hat{\boldsymbol{\mu}}_1) + \hat{\boldsymbol{\delta}}'\boldsymbol{\Sigma}_2^{-1}\hat{\boldsymbol{\delta}}$. We now show that the conditional misclassification rate of QDA converges to 1/2 when $\|\boldsymbol{\Sigma}_2 - \boldsymbol{\Sigma}_1\|_F \to \infty$ and $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$.

Following the proof of Lemma 2, we can show that

$$[\hat{T}_p - \mathrm{E}(\hat{T}_p|\boldsymbol{X})]/[\mathrm{Var}(\hat{T}_p|\boldsymbol{X})]^{1/2} \xrightarrow{D|\boldsymbol{X}} N(0, 1),$$

where $\mathrm{Var}(\hat{T}_p|\boldsymbol{X}) = 2\|\boldsymbol{\Lambda}\|_F^2 + 4\hat{\boldsymbol{\delta}}'\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2^{-1}\hat{\boldsymbol{\delta}}$ and the convergence is with respect to the distribution of the new observation $\boldsymbol{x}$, conditioned on $\boldsymbol{X}$. Under (C2),

$$-1 < \frac{m}{M} - 1 \leq \lambda_{p,j} \leq \frac{M}{m} - 1, \quad j = 1, \ldots, p, \tag{2.6}$$

which implies

$$\left|\mathrm{tr}(\boldsymbol{\Lambda}) - \log\left(\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_2|}\right)\right| \leq \frac{M^2}{2m^2}\|\boldsymbol{\Lambda}\|_F^2.$$

Then, $|\hat{\phi}_p|/[\mathrm{Var}(\hat{T}_p|\boldsymbol{X})]^{1/2}$ is bounded by

$$\frac{\frac{M^2}{2m^2}\|\boldsymbol{\Lambda}\|_F^2 + |2\hat{\boldsymbol{\delta}}'\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_1 - \hat{\boldsymbol{\mu}}_1) - \hat{\boldsymbol{\delta}}'\boldsymbol{\Sigma}_1^{-1}\hat{\boldsymbol{\delta}}| + |(\hat{\boldsymbol{\mu}}_1 - \boldsymbol{\mu}_1)'\boldsymbol{\nabla}(\hat{\boldsymbol{\mu}}_1 - \boldsymbol{\mu}_1)|}{\sqrt{2\|\boldsymbol{\Lambda}\|_F^2 + 4\hat{\boldsymbol{\delta}}'\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2^{-1}\hat{\boldsymbol{\delta}}}}. \tag{2.7}$$

We consider $\boldsymbol{\Sigma}_1 = \boldsymbol{I}$, a diagonal $\boldsymbol{\Sigma}_2$ with jth diagonal $\sigma_{2j}^2 = 2$ for $j = 1, \ldots, K$, $\sigma_{2j}^2 = (\sqrt{17} - 3)/2$ for $j = K+1, \ldots, 2K$, $\sigma_{2j}^2 = 1$ for $j = 2K+1, \ldots, p$, and $n_1 = n_2 = n/2$. Then, $\hat{\boldsymbol{\mu}}_1 - \boldsymbol{\mu}_1 \sim N(0, (2/n)\boldsymbol{I})$ and $\hat{\boldsymbol{\mu}}_2 - \boldsymbol{\mu}_2 \sim N(0, (2/n)\boldsymbol{\Sigma}_2)$. Let ϵ_{1j} and ϵ_{2j} be independent standard normal random variables. Then,

$$\begin{aligned}\hat{\boldsymbol{\delta}}'\boldsymbol{\Sigma}_1^{-1}\hat{\boldsymbol{\delta}} - 2\hat{\boldsymbol{\delta}}'\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_1 - \hat{\boldsymbol{\mu}}_1) &= \frac{2}{n}\sum_{j=1}^{p}\left[(\sigma_{2j}\epsilon_{2j} - \epsilon_{1j})^2 + 2\frac{\epsilon_{1j}(\sigma_{2j}\epsilon_{2j} - \epsilon_{1j})}{\sigma_{2j}^2}\right] \\ &= \frac{2}{n}\sum_{j=1}^{p}\left[\left(1 - \frac{2}{\sigma_{2j}^2}\right)\epsilon_{1j}^2 + 2\left(\frac{1}{\sigma_{2j}} - \sigma_{2j}\right)\epsilon_{1j}\epsilon_{2j} + \sigma_{2j}^2\epsilon_{2j}^2\right],\end{aligned}$$

which has mean 0 for the particular set of σ_{2j}^2's we have chosen. Hence, by the Central Limit Theorem, $\hat{\boldsymbol{\delta}}'\boldsymbol{\Sigma}_1^{-1}\hat{\boldsymbol{\delta}} - 2\hat{\boldsymbol{\delta}}'\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_1 - \hat{\boldsymbol{\mu}}_1) = O_P\left(\sqrt{p}/n\right)$. Also, $4\hat{\boldsymbol{\delta}}'\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2^{-1}\hat{\boldsymbol{\delta}} = O_P\left(p/n\right)$, $\|\boldsymbol{\Lambda}\|_F^2 = (11 - \sqrt{17})K/8 = O(K)$, and $(\hat{\boldsymbol{\mu}}_1 - \boldsymbol{\mu}_1)'\boldsymbol{\nabla}(\hat{\boldsymbol{\mu}}_1 - \boldsymbol{\mu}_1) = O_P\left(K/n\right)$. Therefore, the quantity in (2.7) is bounded by

$$\frac{|\hat{\phi}_p|}{\sqrt{\mathrm{Var}(\hat{T}_p|\boldsymbol{X})}} \leq \frac{O(K) + O_P(\sqrt{p}/n) + O_P(K/n)}{\sqrt{O(K) + O_P(p/n)}},$$

which is $o(1/\sqrt{n})$ if we choose $K = o(\sqrt{p}/n)$. This together with the asymptotic normality of $\hat{T}_p$ shows that the conditional misclassification rate of the QDA converges to 1/2, provided that $K = o(\sqrt{p}/n)$.

3. Sparse Estimators and SQDA

For $\boldsymbol{\delta}$, we adopt the sparsity measure in Shao et al. (2011),

$$d_p = \sum_{j=1}^{p}|\delta_j|^{2g}, \tag{3.1}$$

where δ_j is the jth component of $\boldsymbol{\delta}$ and g is a constant in $[0, 1)$. As $n \to \infty$, d_p may diverge to ∞, but if its divergence rate is much slower than p, then $\boldsymbol{\delta}$ is sparse. If $g = 0$, then d_p is the maximum of the numbers of non-zero components of $\boldsymbol{\delta}$. Similarly, we consider a sparsity measure for covariance matrices:

$$c_p = \max_{k=1,2} \max_{i=1,\ldots,p} \sum_{j=1}^{p} |\sigma_{kij}|^h, \tag{3.2}$$

where σ_{kij} is the (i, j)th element of $\boldsymbol{\Sigma}_k$ and h is a constant in $[0, 1)$. When c_p is much smaller than p, $\boldsymbol{\Sigma}_k$'s are sparse in terms of off-diagonal values, but the diagonal elements of $\boldsymbol{\Sigma}_k$'s are not sparse.

We need to regulate the magnitude of $\boldsymbol{\Delta} = \boldsymbol{\Sigma}_2 - \boldsymbol{\Sigma}_1$ in some sense. We consider the sparsity measure

$$c_{1p} = \sum_{1\le i,\, j\le p} |\Delta_{ij}|^\eta, \tag{3.3}$$

where Δ_{ij} is the (i, j)th element of $\boldsymbol{\Delta}$ and η is a constant in $[0, 1)$. If c_{1p} is much smaller than p, then $\boldsymbol{\Delta}$ is sparse. Unless otherwise mentioned, we eliminate the case of $c_{1p} = 0$.

We allow $p > n$ to be ultra-high, but assume that, as $n \to \infty$

(C5) $n^{-1} \log p \to 0$.

Condition (C5) allows that p diverges at the rate e^{n^ν} for some $\nu \in (0, 1)$.

We need sparse estimators of $\boldsymbol{\delta}$, $\boldsymbol{\Sigma}_1$, $\boldsymbol{\Sigma}_2$, and $\boldsymbol{\Delta}$ that asymptotically valid in terms of several measures.

A sparse estimator of $\boldsymbol{\delta}$ is obtained by thresholding the MLE $\bar{\boldsymbol{x}}_2 - \bar{\boldsymbol{x}}_1$ at

$$t_n = M_0 \left(n^{-1} \log p\right)^\alpha \tag{3.4}$$

for some constants $\alpha \in (0, 1/2)$ and $M_0 > 0$, where $\bar{\boldsymbol{x}}_k = n_k^{-1} \sum_{i=1}^{n_k} \boldsymbol{x}_{ki}$. The thresholded estimator of $\boldsymbol{\delta}$ is $\hat{\boldsymbol{\delta}}$ with jth component $(\bar{x}_{2j} - \bar{x}_{1j}) I(|\bar{x}_{2j} - \bar{x}_{1j}| > t_n)$, where $I(A)$ is the indicator function of the event A and $\bar{x}_{kj}$ is the jth component of $\bar{\boldsymbol{x}}_k$. The parameter $\boldsymbol{\mu}_k$ in (2.3) is estimated by $\hat{\boldsymbol{\mu}}_k = \bar{\boldsymbol{x}}_k$ without thresholding.

From the proof of Theorem 3 in Shao et al. (2011), if

$$b_n = d_p t_n^{2(1-g)} \to 0, \tag{S1}$$

where d_p is given by (3.1), then

$$\|\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}\|^2 = O_P\left(b_n\right). \tag{3.5}$$

The estimation of $\boldsymbol{\Sigma}_k$ is more complicated, since we need estimators of $\boldsymbol{\Sigma}_k$'s to be sparse in terms of off-diagonal elements as well as a sparse estimator of $\boldsymbol{\Delta}$.

We propose an estimator with two steps of thresholding. Let $\boldsymbol{S}_k$ be the MLE of $\boldsymbol{\Sigma}_k$ based on $\{\boldsymbol{x}_{ki}, i = 1, \ldots, n_k\}$ and let s_{kij} be the (i, j)th element of $\boldsymbol{S}_k$, $k = 1, 2$. $\boldsymbol{S}_2 - \boldsymbol{S}_1$ is a natural estimator of $\boldsymbol{\Delta}$, but it is not sparse. In the first step, small elements of $\boldsymbol{S}_2 - \boldsymbol{S}_1$ are thresholded to 0. That is, we replace s_{1ij} and s_{2ij} by $\bar{s}_{ij} = (n_1 s_{1ij} + n_2 s_{2ij})/n$ whenever $|s_{1ij} - s_{2ij}|$ is less than or equal to the threshold value

$$t_{1n} = M_1 \left(n^{-1} \log p\right)^{1/2}, \tag{3.6}$$

where M_1 is a constant. This produces an estimator of $\boldsymbol{\Sigma}_k$, $\tilde{\boldsymbol{\Sigma}}_k$, whose (i, j)th element $\tilde{s}_{kij} = \bar{s}_{ij}$ when $|s_{1ij} - s_{2ij}| \le t_{1n}$ and $\tilde{s}_{kij} = s_{kij}$ otherwise, $k = 1, 2$. Although $\tilde{\boldsymbol{\Sigma}}_2 - \tilde{\boldsymbol{\Sigma}}_1$ is sparse, each $\tilde{\boldsymbol{\Sigma}}_k$ may not be sparse in terms of its off-diagonal elements. Hence, we apply the second step of thresholding to the elements of $\tilde{\boldsymbol{\Sigma}}_k$, which results in the estimator $\hat{\boldsymbol{\Sigma}}_k$ whose (i, j)th element is $\tilde{s}_{kij} I(|\tilde{s}_{kij}| > t_{2n})$, $k = 1, 2$, $i \neq j$, where t_{2n} is given by (3.6) with M_1 replaced by a possibly different constant M_2. The resulting estimator $\hat{\boldsymbol{\Sigma}}_k$ is sparse in terms of its off-diagonal elements and $\hat{\boldsymbol{\Sigma}}_2 - \hat{\boldsymbol{\Sigma}}_1$ is sparse. Here

$$\begin{aligned}
\max_{i,j} |\tilde{s}_{1ij} - \sigma_{1ij}| &= \max_{i,j} \{|s_{1ij} - \sigma_{1ij}| I(|s_{1ij} - s_{2ij}| \ge t_{1n}) \\
&\quad + |\bar{s}_{ij} - \sigma_{1ij}| I(|s_{1ij} - s_{2ij}| < t_{1n})\} \\
&\le \max_{i,j} \{|s_{1ij} - \sigma_{1ij}| + |s_{1ij} - s_{2ij}| I(|s_{1ij} - s_{2ij}| < t_{1n})\} \\
&\le \max_{i,j} |s_{1ij} - \sigma_{1ij}| + M_1 (n^{-1} \log p)^{1/2} \\
&= O_P\left((n^{-1} \log p)^{1/2}\right),
\end{aligned}$$

where the last equality follows from (12) of Bickel and Levina (2008). Similarly,

$$\max_{i,j} |\tilde{s}_{2ij} - \sigma_{2ij}| = O_P\left((n^{-1} \log p)^{1/2}\right).$$

Following the proof of Theorem 1 in Bickel and Levina (2008), we have that, if (C2) and (C5) hold and

$$\text{(S2)} \qquad a_n = c_p (n^{-1} \log p)^{(1-h)/2} \to 0,$$

where c_p is given by (3.2), then

$$\|\hat{\boldsymbol{\Sigma}}_k - \boldsymbol{\Sigma}_k\|_2 = O_P(a_n), \qquad k = 1, 2. \tag{3.7}$$

Hence, $\hat{\boldsymbol{\Sigma}}_1$ and $\hat{\boldsymbol{\Sigma}}_2$ are asymptotically invertible and (3.7) also holds with $\|\hat{\boldsymbol{\Sigma}}_k - \boldsymbol{\Sigma}_k\|_2$ replaced by $\|\hat{\boldsymbol{\Sigma}}_k^{-1} - \boldsymbol{\Sigma}_k^{-1}\|_2$. In fact, from the proof in Bickel and Levina (2008), (3.7) still holds with $\|\hat{\boldsymbol{\Sigma}}_k - \boldsymbol{\Sigma}_k\|_2$ replaced by $\|\hat{\boldsymbol{\Sigma}}_k - \boldsymbol{\Sigma}_k\|_1$.

Lemma 3. *Under* (C2) *and* (C5), *if* $a_n v_p \to 0$, *where* $v_p = \max\{\|\boldsymbol{\Sigma}_1^{-1}\|_1, \|\boldsymbol{\Sigma}_2^{-1}\|_1\}$ *and* a_n *is as in* (S2),

$$\|\hat{\boldsymbol{\Sigma}}_k^{-1} - \boldsymbol{\Sigma}_k^{-1}\|_1 = O_P\left(a_n v_p^2\right), \qquad k = 1, 2.$$

We estimate $\boldsymbol{\Delta}$ by $\hat{\boldsymbol{\Delta}} = \hat{\boldsymbol{\Sigma}}_2 - \hat{\boldsymbol{\Sigma}}_1$ and $\boldsymbol{\nabla}$ by $\hat{\boldsymbol{\nabla}} = \hat{\boldsymbol{\Sigma}}_2^{-1} - \hat{\boldsymbol{\Sigma}}_1^{-1}$.

Theorem 2. *Assume that* (C2)−(C3) *and* (C5) *hold.*
(i) *If* $a_{1n} = c_{1p}(n^{-1}\log p)^{(1-\eta)/2} \to 0$, *then* $\|\hat{\boldsymbol{\Delta}} - \boldsymbol{\Delta}\|_G = O_P(a_{1n})$.
(ii) *If*

$$\text{(S3)} \qquad \tau_n = c_{1p} c_p v_p^3 (n^{-1}\log p)^{(1-\max\{h,\eta\})/2} \to 0,$$

where v_p *is as defined in Lemma* 3, *then* $\|\hat{\boldsymbol{\nabla}} - \boldsymbol{\nabla}\|_G = O_P(\tau_n)$.

We define the SQDA to be the classification rule (2.3) with the sparse estimators $\hat{\boldsymbol{\delta}}$, $\hat{\boldsymbol{\Sigma}}_k$, and $\hat{\boldsymbol{\Delta}}$ previously described, and $\hat{\boldsymbol{\mu}}_k = \bar{\boldsymbol{x}}_k$. We allow the number of non-zero estimators (for the mean differences or covariances) to be much larger than n; this differs from variable selection and is necessary when there are many components of $\boldsymbol{x}$ that have no mean effects for classification but are correlated with those having mean effects.

The tuning parameters M_0, M_1, and M_2 can be selected by searching the optimal thresholds for $\bar{\boldsymbol{x}}_2 - \bar{\boldsymbol{x}}_1$, $\boldsymbol{S}_2 - \boldsymbol{S}_1$, and $\tilde{\boldsymbol{\Sigma}}_k$ that minimize the leave-one-out cross-validation estimate of the misclassification rate. We propose the following "bisection" strategy. Set the search intervals for thresholds of $\bar{\boldsymbol{x}}_2 - \bar{\boldsymbol{x}}_1$, $\boldsymbol{S}_2 - \boldsymbol{S}_1$, and $\tilde{\boldsymbol{\Sigma}}_k$ as $[0, H_1]$, $[0, H_2]$, and $[0, H_3]$, respectively, where $H_1 = \max_j |\bar{x}_{2j} - \bar{x}_{1j}|$, $\bar{x}_{kj}$ is the jth element of $\bar{\boldsymbol{x}}_k$, $H_2 = \max_{ij} |s_{2ij} - s_{1ij}|$, and $H_3 = \max_k \max_{i \neq j} |s_{kij}|$. Consider thresholds that are the end points of search intervals, eight possible threshold combinations, and find the best thresholds by minimizing the leave-one-out cross-validation over them. If the best choice is $(0, H_2, H_3)$, set the search intervals to be $[0, H_1/2]$, $[H_2/2, H_2]$ and $[H_3/2, H_3]$, and repeat the search procedure, iterating until the maximal length of all three search intervals is less than a pre-defined small positive number.

If H_2 is finally chosen to be the optimal threshold for $\boldsymbol{S}_2 - \boldsymbol{S}_1$, then $\hat{\boldsymbol{\Sigma}}_1 = \hat{\boldsymbol{\Sigma}}_2$ and the quadratic term in (2.3) disappears so that the SQDA becomes the SLDA in Shao et al. (2011).

Under some conditions, we now establish that the conditional misclassification rate of the SQDA converges to the same limit as R_B, the misclassification rate of the Bayes rule. To this end, we study the difference between the left hand sides of (2.1) and (2.3).

Lemma 4. *Assume sparsity conditions* (S1), (S2), (C2)−(C3), (C5), *and* $c_p q_n / n \to 0$, *where* q_n *is the number of components of* $\boldsymbol{\delta}$ *whose absolute values are larger than* t_n / r *with a constant* $r > 1$. *If* $\|\boldsymbol{\delta}\|$ *is bounded, then, when* $\boldsymbol{x} \sim N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$,

$$\left|\hat{\boldsymbol{\delta}}'\hat{\boldsymbol{\Sigma}}_2^{-1}(\boldsymbol{x}-\hat{\boldsymbol{\mu}}_1)-\boldsymbol{\delta}'\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)\right| = O_P\left(\max\left\{\sqrt{b_n}, a_n, \sqrt{\frac{c_p q_n}{n}}\right\}\right).$$

Lemma 5. *Under sparsity conditions* (S1), (S2), (C2)−(C3), *and* (C5), *if* $\|\boldsymbol{\delta}\|$ *is bounded, then,*

$$|\hat{\boldsymbol{\delta}}'\hat{\boldsymbol{\Sigma}}_2^{-1}\hat{\boldsymbol{\delta}}-\boldsymbol{\delta}'\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\delta}| = O_P\left(\max\{\sqrt{b_n}, a_n\}\right).$$

Lemma 6. *Under sparsity conditions* (S1), (S3), (C2)−(C3), *and* (C5), *when* $\boldsymbol{x} \sim N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$,

$$\left|(\boldsymbol{x}-\hat{\boldsymbol{\mu}}_1)'\hat{\boldsymbol{\nabla}}(\boldsymbol{x}-\hat{\boldsymbol{\mu}}_1)-(\boldsymbol{x}-\boldsymbol{\mu}_1)'\boldsymbol{\nabla}(\boldsymbol{x}-\boldsymbol{\mu}_1)\right| = O_P(\tau_n).$$

Lemma 7. *Under sparsity conditions* (S1), (S3), (C2)−(C3), *and* (C5),

$$\left|\mathrm{tr}(\hat{\boldsymbol{\Lambda}})-\mathrm{tr}(\boldsymbol{\Lambda})\right| = O_P(\tau_n) \quad \textit{and} \quad \left|\log(|\hat{\boldsymbol{\Sigma}}_1|/|\hat{\boldsymbol{\Sigma}}_2|)-\log(|\boldsymbol{\Sigma}_1|/|\boldsymbol{\Sigma}_2|)\right| = O_P(\tau_n).$$

Theorem 3. *Suppose that conditions* (C1)−(C3) *and* (C5) *hold.*

(i) *When* D_p *is bounded as* $p \to \infty$, *if* (C4) *holds and*

$$\max\left\{\sqrt{b_n}, a_n, \tau_n, \sqrt{\frac{c_p q_n}{n}}\right\} \to 0, \tag{3.8}$$

then

$$R_{\mathrm{SQDA}}(\boldsymbol{X}) - R_B \xrightarrow{P} 0, \tag{3.9}$$

where $R_{\mathrm{SQDA}}(\boldsymbol{X})$ *is the conditional misclassification rate of the SQDA given* $\boldsymbol{X}$ *and* R_B *is the optimal misclassification rate of the Bayes rule in* (2.1).

(ii) *When* $D_p \to \infty$ *as* $p \to \infty$, *if* $a_n \to 0$ *and*

$$\frac{\max\{b_n, a_{1n}\}}{D_p^2} \to 0, \tag{3.10}$$

then (3.9) *holds.*

(1) *Condition* (3.8) *for the case of bounded* D_p.

If we assume that both $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are sparse instead of a sparse $\boldsymbol{\delta}$, then we can replace condition (3.8) in Theorem 3 by the weaker condition that $\max\{\sqrt{b_n}, a_n, \tau_n\} \to 0$. In view of (3.5) and (3.7), $a_n \to 0$ and $b_n \to 0$ in (3.8) ensures that $\hat{\boldsymbol{\delta}}$ and $\hat{\boldsymbol{\Sigma}}_k$ are consistent estimators of $\boldsymbol{\delta}$ and $\boldsymbol{\Sigma}_k$, respectively. These conditions are similar to those for the SLDA in Shao et al. (2011) and for the ROAD in Fan, Feng, and Tong (2012). The condition $\sqrt{c_p q_n/n} \to 0$ is also required by Shao et al. (2011), but it is not needed if both $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are sparse. The extra requirement by the SQDA is $\tau_n \to 0$ for the quadratic and the nonrandom terms. To illustrate, we consider $g = 0$, $h = 0$, and $\eta = 0$, so the sparsity of

$\boldsymbol{\Sigma}_1$, $\boldsymbol{\Sigma}_2$, and $\boldsymbol{\Sigma}_2 - \boldsymbol{\Sigma}_1$ is measured by their numbers of non-zero elements. Then $\tau_n = c_{1p} v_p^3 a_n$, where $v_p = \max\{\|\boldsymbol{\Sigma}_1^{-1}\|_1, \|\boldsymbol{\Sigma}_2^{-1}\|_1\}$. To the best of our knowledge, there is no explicit results on the bound of L_1 norm of the inverse of a sparse matrix. However, we have through numerical studies that, if $\boldsymbol{\Sigma}_k$ is sparse, $\|\boldsymbol{\Sigma}_k^{-1}\|_1$ is usually small. Under sparsity assumption on $\boldsymbol{\Sigma}_2 - \boldsymbol{\Sigma}_1$, c_{1p} is also small. Hence, $\tau_n \to 0$ is slightly stronger than $a_n \to 0$.

To check condition (3.8) we can represent d_p, q_n, c_{1p}, c_p, and v_p in terms of powers of n, as $d_p = O(n^{\alpha_d})$, $q_n = O(n^{\alpha_q})$, $c_{1p} = O(n^{\alpha_{c_1}})$, $c_p = O(n^{\alpha_c})$, and $v_p = O(n^{\alpha_v})$. Then, (3.8) is implied by $\alpha_q + \alpha_c < 1$, $6\alpha_v + 2\alpha_c + 2\alpha_{c_1} < 1$, and $\log p = O(n^\gamma)$, where $\gamma < 1 - 6\alpha_v - 2\alpha_c - 2\alpha_{c_1}$. We can choose the threshold t_n in (3.4) with any $\alpha > \alpha_d/\{2(1-\gamma)\}$.

(2) *Condition* (3.10) *when* $D_p \to \infty$.

The convergence in (3.10) depends on D_p. Again, consider $D_p = O(n^{\alpha_D})$ and $g = 0$, $h = 0$, and $\eta = 0$. Then, (3.10) is implied by $\alpha_c < 1/2$, $2\alpha_{c_1} - 4\alpha_D < 1$, and $\log p = O(n^\gamma)$, where $\gamma < \min\{1 - 2\alpha_c, 1 - 2\alpha_{c_1} + 4\alpha_D\}$. We can choose the threshold t_n in (3.4) with any $\alpha > (\alpha_d - 2\alpha_D)/\{2(1-\gamma)\}$. Compared with the case of bounded D_p, the conditions for unbounded D_p are much relaxed and the dimension of p is allowed to be higher. Also, we have larger feasible regions for the choice of the threshold t_n. In general, the faster D_p diverges, the easier that (3.10) holds and the larger p is allowed. This is intuitively correct, since a larger value of D_p means that the two populations are more separated.

(3) *Sparsity conditions.*

We impose sparsity conditions on $\boldsymbol{\delta}$, $\boldsymbol{\Sigma}_1$, $\boldsymbol{\Sigma}_2$, and $\boldsymbol{\Delta} = \boldsymbol{\Sigma}_2 - \boldsymbol{\Sigma}_1$. Alternatively, we may impose sparsity conditions on the inverses of covariance matrices. From the form of the Bayes rule in (2.1), we only need to assume sparsity for $\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\delta}$ and $\boldsymbol{\nabla} = \boldsymbol{\Sigma}_2^{-1} - \boldsymbol{\Sigma}_1^{-1}$. Indeed, Cai and Liu (2011) established asymptotic results by assuming that $\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}$ is sparse in the case where $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$ ($\boldsymbol{\nabla} = 0$). However, the sparsity of $\boldsymbol{\delta}$ and covariance matrices and the sparsity of $\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\delta}$ are not comparable conditions, and their interpretations differ.

When $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$, the SQDA is asymptotically the same as the SLDA in Shao et al. (2011). On the other hand, if $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$, the SQDA may be much better than the SLDA in terms of the asymptotic misclassification rate. In Shao et al. (2011), an estimator of the covariance matrix (assuming that $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$) is obtained by thresholding $\boldsymbol{S} = (n_1\boldsymbol{S}_1 + n_2\boldsymbol{S}_2)/n$. It converges in L_2 norm to $\boldsymbol{\Sigma}^* = \gamma\boldsymbol{\Sigma}_1 + (1-\gamma)\boldsymbol{\Sigma}_2$ when actually $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$, where $n_1/n \to \gamma \in (0,1)$. Let

s_{ij} and σ^*_{ij} be the (i,j)th element of $\boldsymbol{S}$ and $\boldsymbol{\Sigma}^*$, respectively. Then,

$$\begin{aligned}
\max_{i,j}|s_{ij}-\sigma^*_{ij}| &\le \max_{i,j}\Big(\Big|\frac{n_1 s_{1ij}}{n}-\gamma\sigma_{1ij}\Big|\Big)+\Big|\frac{n_2 s_{2ij}}{n}-(1-\gamma)\sigma_{2ij}\Big| \\
&\le \max_{i,j}[|s_{1ij}-\sigma_{1ij}|+|s_{2ij}-\sigma_{2ij}|+|\frac{n_1}{n}-\gamma||\sigma_{1ij}| \\
&\quad +|\frac{n_2}{n}-(1-\gamma)||\sigma_{2ij}|] \\
&= O_P\Big([n^{-1}\log p]^{1/2}\Big)+O(|\frac{n_1}{n}-\gamma|).
\end{aligned}$$

If $n_2/n-\gamma = O((n^{-1}\log p)^{1/2})$, then $\max_{i,j}|s_{ij}-\sigma^*_{ij}| = O_P\left([n^{-1}\log p]^{1/2}\right)$. With this result, we can show that $\|\tilde{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}^*\|_2 = O_P(a_n)$, where $\tilde{\boldsymbol{\Sigma}}$ is $\boldsymbol{S}$ thresholded at $M_3(\log p/n)^{1/2}$ with a constant $M_3>0$. Then, under the regularity conditions stated in Theorem 3 of Shao et al. (2011),

$$R_{\text{SLDA}}(\boldsymbol{X})-\Phi\Big(-\frac{\sqrt{\boldsymbol{\delta}'\boldsymbol{\Sigma}^*\boldsymbol{\delta}}}{2}\Big)\xrightarrow{P}0.$$

If $\|\boldsymbol{\delta}\|$ is bounded as $p\to\infty$, then

$$\liminf_{p\to\infty}\Phi\Big(-\frac{\sqrt{\boldsymbol{\delta}'\boldsymbol{\Sigma}^*\boldsymbol{\delta}}}{2}\Big)>\liminf_{p\to\infty}R_B\ge 0$$

since R_B is the misclassification rate of the Bayes rule. These results, together with Theorem 3, imply that

$$\lim_{n\to\infty}P\Big(R_{\text{SLDA}}(\boldsymbol{X})>R_{\text{SQDA}}(\boldsymbol{X})+\epsilon_0\Big)=1$$

for some fixed $\epsilon_0>0$.

If $\|\boldsymbol{\delta}\|\to\infty$, then R_B, $R_{\text{SLDA}}(\boldsymbol{X})$, and $R_{\text{SQDA}}(\boldsymbol{X})$ all converge to 0, and the asymptotic relative performance between the SLDA and SQDA depends on $\|\boldsymbol{\delta}\|$ and $\|\boldsymbol{\Delta}\|_F$ in a complicated manner. We compare the SLDA and SQDA in a simulation study in the next section.

4. Numerical Work

We first compare the SQDA with the SLDA and ROAD in a simulation study. Then, we apply the SQDA to two data sets and compare it with other popular classifiers.

4.1. A simulation comparison of the SQDA, SLDA and ROAD

We considered two scenarios for the mean vectors:

$$\begin{aligned}
&A.\quad \boldsymbol{\mu}_1=(1,\mathbf{0}_{p-1})',\quad \boldsymbol{\mu}_2=(2,\mathbf{0}_{p-1})',\quad \|\boldsymbol{\delta}\|=1,\\
&B.\quad \boldsymbol{\mu}_1=(\boldsymbol{e}_5,\mathbf{0}_{p-5})',\quad \boldsymbol{\mu}_2=(3\boldsymbol{e}_5,\mathbf{0}_{p-5})',\quad \|\boldsymbol{\delta}\|=4.47,
\end{aligned}$$

Table 1. Misclassification Rate (in %) and Simulation Standard Error (in parenthesis) for balanced design: $n_1 = 20$ and $n_2 = 20$.

		mean scenario A				mean scenario B			
		SLDA	SQDA	ROAD	Bayes	SLDA	SQDA	ROAD	Bayes
V1	$p = 50$	46.6(8.1)	44.7(7.6)	44.2(9.1)	39.8	22.2(6.3)	24.6(6.9)	26.4(8.1)	18.3
	$p = 200$	45.9(9.9)	46.6(9.1)	45.2(8.9)	39.8	22.4(8.6)	22.9(7.5)	26.8(8.8)	18.3
	$p = 1,000$	48.3(8.3)	44.6(9.7)	46.1(9.0)	39.8	23.5(9.9)	21.8(7.9)	24.6(7.9)	18.3
V2	$p = 50$	45.0(7.9)	23.2(7.4)	41.7(9.3)	14.1	11.7(5.3)	10.9(5.2)	18.1(7.2)	5.4
	$p = 200$	45.7(9.7)	23.9(9.6)	41.8(10.8)	14.1	12.7(6.1)	11.4(6.3)	16.5(6.8)	5.4
	$p = 1,000$	47.1(8.8)	28.1(10.7)	45.4(9.4)	14.1	14.2(8.0)	12.6(5.3)	17.9(6.7)	5.4
V3	$p = 50$	44.2(8.3)	9.6(5.6)	42.9(8.8)	6.2	18.2(6.4)	8.8(5.7)	22.2(6.9)	3.7
	$p = 200$	46.6(9.6)	11.6(6.8)	42.6(10.4)	6.2	17.2(7.4)	10.4(6.3)	22.4(7.5)	3.7
	$p = 1,000$	48.1(9.3)	13.2(8.2)	44.5(10.5)	6.2	20.5(9.6)	11.9(6.2)	22.4(6.8)	3.7

where $\boldsymbol{e}_t$ denotes a t-dimensional vector of 1's and $\mathbf{0}_t$ denotes a t-dimensional vector of 0's. For the covariance matrices, we considered three cases:

$$V1.\quad \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2,\quad \boldsymbol{\Sigma}_2 = \begin{pmatrix} \boldsymbol{B} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{I}_{p-5} \end{pmatrix},\qquad \|\boldsymbol{\Delta}\|_F = 0,$$

$$V2.\quad \boldsymbol{\Sigma}_1 = \boldsymbol{I}_p,\quad \boldsymbol{\Sigma}_2 = \begin{pmatrix} \boldsymbol{B} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{I}_{p-5} \end{pmatrix},\qquad \|\boldsymbol{\Delta}\|_F = 8.92,$$

$$V3.\quad \boldsymbol{\Sigma}_1 = \boldsymbol{I}_p,\quad \boldsymbol{\Sigma}_2 = \begin{pmatrix} 2\boldsymbol{B} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{I}_{p-5} \end{pmatrix},\quad \|\boldsymbol{\Delta}\|_F = 16.82,$$

where $\boldsymbol{I}_t$ denotes the identity matrix of order t and

$$\boldsymbol{B} = \begin{pmatrix} 4 & 1 & 0.5 & 0 & 0 \\ 1 & 4 & 1 & 0.5 & 0 \\ 0.5 & 1 & 4 & 1 & 0.5 \\ 0 & 0.5 & 1 & 4 & 1 \\ 0 & 0 & 0.5 & 1 & 4 \end{pmatrix}.$$

In each scenario, we took $p = 50, 200$, 1,000, a balanced design with $n_1 = n_2 = 20$, and an unbalanced design with $n_1 = 10$ and $n_2 = 30$. The tuning parameters M_0, M_1 and M_2 were chosen by the bisection procedure described in Section 3. The same procedure was used to choose thresholds in the SLDA. In practice, if the resulting $\hat{\boldsymbol{\Sigma}}_k$ is not invertible, we used $\hat{\boldsymbol{\Sigma}}_k + \rho\boldsymbol{I}$ (e.g., $\rho = \sqrt{\log p/n}$) instead. The ROAD was implemented using the authors' Matlab code with tuning parameters chosen by their algorithm. We ran 100 simulations for each setting. The misclassification rates of the three methods are shown in Table 1 and Table 2, along with the optimal rates of the Bayes rule.

The following is a summary of the results in Table 1 and Table 2.

1. Mean scenario A. With V1 ($\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$), all three methods do not perform well, since the signal strength $\|\boldsymbol{\delta}\| = 1$ is low and $\|\boldsymbol{\Delta}\|_F = 0$. Even the Bayes

Table 2. Misclassification Rate (in %) and Simulation Standard Error (in parenthesis) for unbalanced design: $n_1 = 10$ and $n_2 = 30$.

		mean scenario A				mean scenario B			
		SLDA	SQDA	ROAD	Bayes	SLDA	SQDA	ROAD	Bayes
V1	$p = 50$	47.7(8.1)	43.2(7.4)	41.6(8.8)	39.8	30.3(9.6)	28.1(7.4)	24.9(8.8)	18.3
	$p = 200$	49.9(8.5)	47.5(3.9)	40.2(8.8)	39.8	31.3(8.9)	26.9(5.7)	24.3(8.1)	18.3
	$p = 1,000$	49.0(8.9)	45.2(0.7)	37.2(8.8)	39.8	32.8(9.7)	28.4(6.5)	24.2(7.9)	18.3
V2	$p = 50$	47.5(8.7)	21.6(6.4)	41.5(7.6)	14.1	17.6(8.3)	10.8(5.5)	23.4(8.0)	5.4
	$p = 200$	50.3(7.7)	22.4(5.5)	38.9(7.6)	14.1	19.3(8.8)	11.0(4.5)	22.1(7.4)	5.4
	$p = 1,000$	48.5(8.5)	24.9(2.2)	35.4(7.1)	14.1	20.9(9.3)	13.3(4.2)	20.9(8.3)	5.4
V3	$p = 50$	48.9(8.0)	10.1(5.1)	43.0(8.5)	6.2	24.8(9.0)	10.7(5.3)	29.4(9.1)	3.7
	$p = 200$	49.7(8.1)	13.3(6.3)	40.1(8.3)	6.2	24.8(8.3)	11.3(5.5)	29.2(8.1)	3.7
	$p = 1,000$	49.2(8.7)	22.4(4.7)	39.6(9.2)	6.2	27.3(9.5)	12.0(5.5)	28.7(7.6)	3.7

rule has a high misclassification rate. Under V2 or V3, $\|\boldsymbol{\Delta}\|_F$ is much larger than that in V1, although $\|\boldsymbol{\delta}\|$ is small. The SQDA is clearly better than the SLDA and ROAD, both of which are linear classifiers that cannot capture the differences between covariances.

2. Mean scenario B. SLDA and ROAD are substantially better when $\|\boldsymbol{\delta}\|$ is larger. For V1 ($\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$), the SQDA has a similar performance to those of SLDA and ROAD. Under V2 or V3, the SQDA is again much better than the SLDA and ROAD, due to the difference between covariances. The covariance difference in V3 is larger than that in V2 and, hence, the SQDA performs better in V3. On the other hand, the SLDA and ROAD both perform better than in mean scenario A, because $\|\boldsymbol{\delta}\|$ is larger.
3. Overall, the performance of three methods in the unbalanced design is similar to or worse than that in the balanced design. But the patterns are the same.

To conclude, a large difference in $\boldsymbol{\mu}_k$'s is needed to have good performance of the SLDA and ROAD. The same is true for the SQDA, but the SQDA may still be good when there is a large difference in covariance matrices. The SQDA has similar performance to the SLDA and ROAD when $\|\boldsymbol{\Delta}\|_F$ is small, but substantially outperforms them when $\|\boldsymbol{\Delta}\|_F$ is large.

The SQDA requires more computational time than the linear rules. In the simulation study, the CPU time for the SQDA was 50−60 seconds for $p = 200$ and 1,600−2,400 seconds for $p = 1,000$, whereas the CPU time for the SLDA was 10−15 seconds for $p = 200$ and 400−1,000 seconds for $p = 1,000$, and the CPU time for the ROAD was about 4 seconds for $p = 200$ and about 20 seconds for $p = 1,000$. A long CPU time in SQDA (and sometimes in the SLDA) results from the computation of the inverse of very large covariance matrices. Research on short cuts in the computation of SQDA is desired, especially when we want to handle the case with an even larger p.

Table 3. Average Misclassification Rates (in %) of 9 Classifiers for Colon Data

BagBoost	RF	SVM	kNN	DLDA	Boosting	PAM	SLDA	SQDA
16.10	14.86	15.05	16.38	12.86	19.14	11.90	12.20	10.40

Table 4. Quantiles of Misclassified Objects by the SLDA and SQDA for Colon Data

	Min.	25%	Median	75%	Max.
SLDA	0	1.25	2.5	3	6
SQDA	1	1	2	2	5

4.2. The example of colon tissues

Alon et al. (1999) studied gene expression difference between tumor and normal colon tissues using the Oligonucleotide microarray technique. The dataset contains $n_1 = 20$ observations from normal tissues and $n_2 = 42$ observations from tumor tissues. A total of $p = 2,000$ genes with highest minimal intensity is included in the study. Dettling (2004) used this dataset to compare the performance of seven different classifiers: the Boosting, Bagging and boosting (BagBoost), Support Vector Machine (SVM), random forest (RF), the k nearest neighbor (kNN), the nearest shrunken centroid classifier (PAM), and diagonal LDA (DLDA), which applies the LDA by assuming that $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$ is a diagonal matrix. The dataset was randomly split into a training set of 13 observations from normal tissues and 29 observations from tumor tissues and a test set of 7 observations from normal tissues and 13 observations from tumor tissues. For each classifier, a misclassification rate was calculated by classifying observations in the test set using the rule constructed based on the training set. To reduce variability, Dettling (2004) independently repeated this process 50 times and reported the average misclassification rates of the seven classifiers over the 50 random splitting. The results are listed in our Table 3.

To compare, we added the average misclassification rates of the SLDA and SQDA calculated using the same random splitting process but a necessarily different random seed. We used the same procedure as in the simulation study to choose the tuning parameters in the SQDA and SLDA. The results are given in Table 3. In this example, the SQDA is the best among all classifiers. The SLDA, slightly behind the PAM, is actually the third winner.

The absolute gain in misclassification rate for the SQDA over the SLDA was 1.8%, a relative gain of 14.8%. Table 4 lists some quantiles of numbers of misclassified subjects by the SLDA and SQDA in 50 replications.

Table 5. Mean(Standard Error) of Misclassification Rates (in %) of Seven Classifiers for GSE12288 Data.

SQDA	SLDA	ROAD	SVM	kNN	Boosting	RF
24.3(5.2)	27.6(5.1)	43.0(5.9)	33.1(4.7)	46.4(4.9)	38.5(5.3)	37.8(4.5)

4.3. A cardiovascular study example

Sinnaeve et al. (2009) studied the relationship between gene expression patterns and atherosclerotic coronary artery disease (CAD). Patients were selected according to their Duke CAD index (CADi), a validated angiographical measure of the extent of coronary atherosclerosis. 110 patients were collected with CADi > 23 and 112 persons without CAD were formed as a control group. Their gene expression was assessed using Affymetrix U133A chips. A total number of 19,940 genes were collected. The raw data can be accessed from Gene Expression Omnibus under the name of “GSE12288”. Since many genes in Affymetrix U133A platform are not related to CAD, we performed two sample t-tests to each gene and only kept genes with p-value < 0.1. As a result, $p = 2,434$ genes were used in our analysis.

We compared our method with six other classifiers: SLDA, ROAD, Support Vector Machine (SVM), k-nearest neighborhood (kNN), Boosting, and Random Forest (RF), in terms of the misclassification rate. The dataset was randomly split into a set of 73 patients and 75 healthy persons as the training set (about 2/3 of the entire data set) and a set of 37 patients and 37 healthy persons as the test set. The SVM, kNN, Boosting, and Random Forest were implemented by the R packages of “e1071”, “class”, “ada” and “randomForest” with default settings, respectively. The ROAD was implemented by Matlab code on the authors’ website. For SLDA and SQDA, tuning parameters were chosen by the same scheme as discussed in Section 3. We repeated the random splitting 200 times. The mean misclassification percentage of each method is listed in Table 5 with standard error in parenthesis.

The SQDA performed significantly better than the other six classifiers. Compared to SLDA, the SQDA had a relative gain of 12.0%.

Acknowledgements

We would like to thank three referees and an associate editor for providing helpful comments and suggestions.

References

Alon, U., Barkai, N., Notterman, D., Gish, K., Ybarra, S., Mack, D. and Levine, A. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Nat. Acad. Sci.* **96**, 6745.

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

Bickel, P. and Levina, E. (2008). Covariance regularization by thresholding. *Ann. Statist.* **36**, 2577-2604.

Cai, T. and Liu, W. (2011). A direct estimation approach to sparse linear discriminant analysis. *J. Amer. Statist. Assoc.* **106**, 1566–1577.

Cheng, Y. (2004). Asymptotic probabilities of misclassification of two discriminant functions in cases of high dimensional data. *Statist. Probab. Lett.* **67**, 9-17.

Clemmensen, L., Hastie, T. and Ersbøll, B. (2008). Sparse discriminant analysis. Technical report, DTU Informatics Lyngby.

Dettling, M. (2004). Bagboosting for tumor classification with gene expression data. *Bioinformatics* **20**, 3583-3593.

Fan, J. and Fan, Y. (2008). High dimensional classification using features annealed independence rules. *Ann. Statist.* **36**, 2605-2637.

Fan, J., Feng, Y. and Tong, X. (2012). A road to classification in high dimensional space. *J. Roy. Statist. Soc. Ser. B* **74**, 745-711.

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

Han, F., Zhao, T. and Liu, H. (2012). Coda: High dimensional copula discriminant analysis. *J. Mach. Learn. Res.* **14**, 629-671.

Qiao, Z., Zhou, L. and Huang, J. (2009). Sparse linear discriminant analysis with applications to high dimensional low sample size data. *IAENG Internat. J. Appl. Math.* **39**, 1-13.

Shao, J., Wang, Y., Deng, X. and Wang, S. (2011). Sparse linear discriminant analysis by thresholding for high dimensional data. *Ann. Statist.* **39**, 1241-1265.

Simon, N. and Tibshirani, R. (2011). Discriminant analysis with adaptively pooled covariance. arXiv:1111.1687 .

Sinnaeve, P. R., Donahue, M. P., Grass, P., Seo, D., Vonderscher, J., Chibout, S.-D., Kraus, W. E., Sketch Jr, M., Nelson, C., Ginsburg, G. S., et al. (2009). Gene expression patterns in peripheral blood correlate with the extent of coronary artery disease. *PLoS One* **4**, e7037.

Zhang, Q. and Wang, H. (2011). On bic's selection consistency for discriminant analysis. *Statist. Sinica* **21**, 731-740.

Department of Statistics, University of Wisconsin–Madison, 1300 University Avenue, Madison, WI 53705, USA.

E-mail: quefeng@stat.wisc.edu

Department of Statistics, University of Wisconsin–Madison, 1300 University Avenue, Madison, WI 53705, USA.

E-mail: shao@stat.wisc.edu

(Received May 2013; accepted March 2014)