

Statistica Sinica Preprint No: SS-2016-0034.R2

Title	Quadratic Discriminant Analysis for High-Dimensional Data
Manuscript ID	SS-2016-0034.R2
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202016.0034
Complete List of Authors	Yilei Wu Yingli Qin and Mu Zhu
Corresponding Author	Yingli Qin
E-mail	yingli.qin@uwaterloo.ca
Notice: Accepted version subject to English editing.	

Quadratic Discriminant Analysis for High-Dimensional Data

Yilei Wu, Yingli Qin, Mu Zhu

University of Waterloo

Abstract: High-dimensional classification is an important and challenging statistical problem. We develop a set of quadratic discriminant rules by simplifying the structure of the covariance matrices instead of imposing sparsity assumptions — either on the covariance matrices themselves (or their inverses), or on the standardized between-class distance. Under moderate conditions on the population covariance matrices, our specialized quadratic discriminant rules enjoy good asymptotic properties. Computationally, they are easy to implement and do not require large-scale mathematical programming. Numerically, they perform well in finite dimensions and with finite sample sizes. We also present real-data analyses of several classic micro-array data sets.

1. Introduction

In this paper, we study discriminant analysis in high dimensions. Suppose a random vector $\mathbf{x} \in \mathbb{R}^p$, where p is very large, comes from either class 1 (\mathcal{C}_1) or class 2 (\mathcal{C}_2). On the training data, the class memberships of these vectors are labelled. The goal is to classify an unlabelled observation using a *discriminant rule* that is learned from the training data. To focus on the main issues, we shall assume that the unconditional prior probabilities of both classes are equal to $1/2$. Otherwise, all discriminant rules mentioned in this paper can be modified simply by adding a constant to correct for the class imbalance.

For $i = 1, 2$, let $\boldsymbol{\mu}_i$ and Σ_i be the class mean and class covariance matrix, respectively. To determine the class label of \mathbf{x} , Fisher's *linear* discriminant rule (see, e.g., Anderson et al., 1958), which assumes $\Sigma_1 = \Sigma_2 = \Sigma$, classifies \mathbf{x} to class 1 if

$$(\mathbf{x} - \boldsymbol{\mu})' \Sigma^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \geq 0, \quad (1.1)$$

where $\boldsymbol{\mu} = (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)/2$, and to class 2 otherwise. If the two covariance matrices cannot be taken to be identical, then the *quadratic* discriminant rule can be used, which classifies \mathbf{x} to class 1 if

$$\ln(|\Sigma_1|/|\Sigma_2|) + (\mathbf{x} - \boldsymbol{\mu}_1)' \Sigma_1^{-1} (\mathbf{x} - \boldsymbol{\mu}_1) - (\mathbf{x} - \boldsymbol{\mu}_2)' \Sigma_2^{-1} (\mathbf{x} - \boldsymbol{\mu}_2) \leq 0, \quad (1.2)$$

and to class 2 otherwise. Equation (1.2) is also the Bayes rule under the assumption that $\mathbf{x} \sim N(\boldsymbol{\mu}_i, \Sigma_i)$ if $\mathbf{x} \in \mathcal{C}_i$, and so is equation (1.1) when $\Sigma_1 = \Sigma_2$.

In practice, the parameters $\boldsymbol{\mu}_i$ and Σ_i are unknown and need to be estimated from training data. Let $\hat{\boldsymbol{\mu}}_i$ and $\hat{\Sigma}_i$ be the sample mean and sample covariance matrix of class i . They are conventionally used as estimators of $\boldsymbol{\mu}_i$ and Σ_i . The common covariance matrix in (1.1) is estimated by the pooled sample covariance matrix, $\hat{\Sigma} = (n_1 + n_2 - 2)^{-1}[(n_1 - 1)\hat{\Sigma}_1 + (n_2 - 1)\hat{\Sigma}_2]$. When the dimension is high and the number of covariates p is close to or larger than the number of observations n , the sample covariance matrix is well-known to be a poor estimate of its population counterpart; in fact, it is often singular and cannot be directly plugged into the discriminant rules.

1.1 Linear discriminant analysis (LDA)

In recent years, many methods have been proposed in the literature for performing linear discriminant analysis (LDA) in high dimensions. For example, one can ignore the covariance terms and use just a diagonal matrix in (1.1) — these are referred to as “independence rules”. Bickel and Levina (2004) showed that, if one simply uses the Moore-Penrose inverse of $\widehat{\Sigma}$, then the misclassification error of (1.1) converges to $1/2$ as $p/n \rightarrow \infty$, whereas the independence rule is at least as good. These “independence rules” can, and often should, be applied in conjunction with feature selection. For instance, Fan and Fan (2008) pointed out that they can perform poorly by themselves due to noise accumulation in estimating the population centroids, $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, in high-dimensional spaces. They proposed to select a subset of important features by performing two-sample t-tests before applying the independence rule. Based on similar considerations, Tibshirani et al. (2002) shrunk class centroids towards the overall center of the data in order to reduce noise, and also estimated Σ with a diagonal matrix.

Another popular approach in the literature is to impose sparsity assumptions. For example, Shao et al. (2011) assumed both Σ and the mean difference vector, $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$, to be sparse, and estimated them by thresholding. Fan et al. (2013) performed variable selection by “innovated thresholding” and “higher criticism thresholding” before carrying out LDA with the selected set of features. Hao et al. (2015) rotated the data to create sparsity prior to applying existent classifiers. Witten and Tibshirani (2011) applied a sparsity penalty in seeking out a projection direction that maximized the between-class variance. Notice that, for LDA, the (pooled) covariance matrix Σ affects classification only through the discriminant direction, $\Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$. Thus, various methods have been pro-

posed to avoid the estimation of Σ itself — e.g., Fan et al. (2012) solved for the discriminant direction directly by minimizing the misclassification rate under a sparsity constraint; Mai et al. (2012) found the direction by solving a penalized linear regression problem; see also Cai and Liu (2011).

1.2 Quadratic discriminant analysis (QDA)

The LDA rule (1.1) assumes that two classes share the same covariance matrix, which is challenging to test in high dimensions (see, e.g., Li et al., 2012; Cai et al., 2013, and many others). If the null hypothesis, $H_0 : \Sigma_1 = \Sigma_2$, cannot be accepted for sound reasons, it may become necessary to consider quadratic discriminant analysis (QDA). However, because there are many more unknown parameters to estimate, QDA is much more challenging than LDA, especially in high dimensions, and much less work has been done about it.

As in the case of LDA, it is also natural to use just diagonal covariance matrices or to impose some sparsity conditions in order to regularize QDA. For example, diagonal quadratic discriminant analysis (DQDA) was studied by Dudoit et al. (2002), whereas Li and Shao (2014) suggested a sparse QDA (SQDA) procedure by thresholding not only the mean difference vector $\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2$, but also the covariance matrices $\hat{\Sigma}_i$ and their difference $\hat{\Sigma}_1 - \hat{\Sigma}_2$. A more recent work on sparse QDA rule is based on the dimension reduction method, QUADRO, proposed by Fan et al. (2015). QUADRO constructs a quadratic projection $f(\mathbf{x}) = \mathbf{x}'\Omega\mathbf{x} - 2\boldsymbol{\delta}'\mathbf{x}$ by maximizing the Rayleigh quotient of f , which is the ratio of the variance explained by the class label to the remaining variance. The parameters, Ω and $\boldsymbol{\delta}$, are encouraged to be sparse by ℓ_1 penalties. The estimated

projection can then be used for classification. For example, the class label can be decided by the sign of $\mathbf{x}'\widehat{\Omega}\mathbf{x} - 2\widehat{\delta}'\mathbf{x} - c$ for some thresholding constant c .

Friedman (1989) proposed regularized discriminant analysis (RDA) as a way to compromise between LDA and QDA. In particular, his proposal shrinks the sample class covariance matrix $\widehat{\Sigma}_i$ twice — once toward the pooled sample covariance matrix, $\widehat{\Sigma}$, and once again toward the diagonal matrix, $p^{-1}tr(\widehat{\Sigma}_i)I_p$, where $tr(\cdot)$ denotes the trace of a matrix and I_p is $p \times p$ identity matrix.

We shall refer to the quantity, $p^{-1}tr(\widehat{\Sigma}_i)I_p$, simply as the “trace estimator”. It has been used in the literature for high-dimensional hypothesis testing and classification problems, and is closely related to our methods. One reason why the trace estimator is useful is that, under some mild conditions, $p^{-1}tr(\widehat{\Sigma}_i)$ can be shown to be a consistent estimator of $p^{-1}tr(\Sigma_i)$ even as $p \rightarrow \infty$.

For classification, Friedman’s RDA (Friedman, 1989) clearly uses the trace estimator, as it shrinks the sample covariance matrix $\widehat{\Sigma}_i$ towards both the pooled covariance estimator $\widehat{\Sigma}$ and the trace estimator. Shrinking toward the trace estimator is one way to overcome the well-known bias in the sample covariance matrix, which inflates large eigenvalues and deflates smaller ones. The two directions of shrinkage are controlled by two separate tuning parameters, λ and γ , as follows:

$$\widehat{\Sigma}_i(\lambda) = \frac{(1 - \lambda)(n_i - 1)\widehat{\Sigma}_i + \lambda(n_1 + n_2 - 2)\widehat{\Sigma}}{(1 - \lambda)(n_i - 1) + \lambda(n_1 + n_2 - 2)},$$

and

$$\widehat{\Sigma}_i(\lambda, \gamma) = (1 - \gamma)\widehat{\Sigma}_i(\lambda) + \gamma \left[p^{-1}tr(\widehat{\Sigma}_i(\lambda))I_p \right].$$

There are four extreme cases. When $\lambda = 0$ and $\gamma = 0$, RDA reduces to vanilla

QDA. When $\lambda = 1$ and $\gamma = 0$, RDA amounts to LDA. When $\lambda = 1$ and $\gamma = 1$, RDA is equivalent to replacing $\widehat{\Sigma}$ in LDA with just the identity matrix — in this case, classification is based on comparing Euclidean distances $\|\mathbf{x} - \widehat{\boldsymbol{\mu}}_i\|^2$ instead of Mahalanobis distances $(\mathbf{x} - \widehat{\boldsymbol{\mu}}_i)' \widehat{\Sigma}^{-1} (\mathbf{x} - \widehat{\boldsymbol{\mu}}_i)$, for $i = 1, 2$. When $\lambda = 0$ and $\gamma = 1$, RDA is equivalent to replacing $\widehat{\Sigma}_i$ in the QDA rule (1.2) with the trace estimator, $p^{-1} \text{tr}(\widehat{\Sigma}_i) I_p$.

For hypothesis testing, Bai and Saranadasa (1996) proposed a test statistic, which replaces the pooled sample covariance matrix $\widehat{\Sigma}$ in Hotelling's two-sample T^2 -statistic with the identity matrix I_p and uses just the squared Euclidean distance (rather than Mahalanobis distance) between the sample means for high-dimensional problems. However, to do so, a bias-correction term must be added that depends on $\text{tr}(\widehat{\Sigma})$. Chen and Qin (2010) generalized this to the case where $\Sigma_1 \neq \Sigma_2$ so using the pooled estimate $\widehat{\Sigma}$ is no longer appropriate.

Aoshima and Yata (2014) then followed up on these ideas and used them for classification. In particular, they substituted the identity matrix I_p for the sample covariance matrix $\widehat{\Sigma}$ in the LDA rule (1.1), and used the trace estimator in place of each $\widehat{\Sigma}_i$ in the QDA rule (1.2). These two rules are similar to two of the four extreme cases in Friedman's RDA, corresponding to $(\lambda, \gamma) = (1, 1)$ and $(\lambda, \gamma) = (0, 1)$, except for the aforementioned bias-correction terms involving $\text{tr}(\widehat{\Sigma})$ and $\text{tr}(\widehat{\Sigma}_i)$. They also investigated a few variants of their quadratic rule (Aoshima and Yata, 2015).

1.3 Handling nonnormal data

Compared with LDA, QDA is more sensitive to deviations from normality (Fried-

man, 1989). A common approach for relaxing the normality assumption is to assume that there exists a strictly monotone transformation for each dimension such that the transformed vector \mathbf{x} follows a multivariate normal distribution given its class label (e.g., Lin and Jeon, 2003; Liu et al., 2009; Mai and Zou, 2015). After first estimating and then applying these transformations, Lin and Jeon (2003) performed classic LDA and QDA; Liu et al. (2009) estimated undirected graphical models; and Mai and Zou (2015) applied their direct method for sparse discriminant analysis (DSDA). In this paper, we will also rely on this idea to generalize our methods.

1.4 Outline and summary of this paper

One can view the trace estimator as the result of two operations: pooling the diagonal elements of each sample covariance matrix, and ignoring its off-diagonal elements. In this paper, we take the idea of the trace estimator one step further, and introduce an estimator that also pools the off-diagonal elements. We will refer to the resulting QDA rule as ppQDA (for having performed two pooling operations), and the QDA rule with the trace estimator as pQDA — a special case of our more general method. We will study their asymptotic performances (Section 2), and also generalize them to handle nonnormal data (Section 3). Our generalization is based on first estimating a set of nonparametric data transformations and then applying our methods to the transformed data. As such, we will refer to these generalized QDA rules as semiparametric ppQDA (Se-pQDA) and semiparametric pQDA (Se-pQDA), respectively. We will prove a result for Se-pQDA, but only demonstrate the performance of Se-ppQDA empirically; the

proof of a similar result for Se-ppQDA is more complicated, and will be left for future research.

Here is a summary of our main contributions. First, while most existing high-dimensional discriminant analysis methods focus on LDA, we fill this gap by focusing on QDA. Second, the sample covariance matrix is inconsistent when the dimension is high but, instead of making sparsity assumptions, we reduce the number of unknown parameters by simplifying the matrix structure in a different way. Third, using more than just the trace estimator in the QDA rule, our proposed ppQDA rule allows us to make use of information about the correlations among different dimensions. Fourth, we relax the normality assumption for both ppQDA and pQDA, and establish theoretical results for all of them except Se-ppQDA, the semiparametric extension of ppQDA. Finally, because our methods are based on using a very simple matrix structure, all our methods are computationally feasible and easy to apply in practice.

We proceed as follows. In Section 2, we introduce our notation, and describe our main methods, ppQDA and pQDA. In Section 3, we propose semiparametric generalizations of our main methods for nonnormal data. Section 4 contains extensive numerical experiments. Then, in Section 5, we provide some important discussions about the relative performance of our ppQDA rule to that of the Bayes decision rule, before we close with some concluding remarks in Section 6.

2. QDA by pooling elements of covariance matrices

Let $\{\mathbf{y}_{1k} : 1 \leq k \leq n_1\}$ and $\{\mathbf{y}_{2k} : 1 \leq k \leq n_2\}$ be training samples from p -dimensional normal distributions $N(\boldsymbol{\mu}_1, \Sigma_1)$ and $N(\boldsymbol{\mu}_2, \Sigma_2)$, respectively. That

is, $\mathbf{y}_{1k} \in \mathcal{C}_1$ and $\mathbf{y}_{2k} \in \mathcal{C}_2$. In addition, all \mathbf{y}_{ik} s are assumed to be independent. Let y_{ijk} to denote the j th dimension of \mathbf{y}_{ik} , for $j = 1, \dots, p$. In what follows, $\mathbf{x} \in \mathbb{R}^p$ is used to denote a generic feature vector observation *without* a class label, and our target is to classify \mathbf{x} based on an rule learned from the training samples. The sample version of the QDA rule (1.2) is to classify \mathbf{x} to class 1 if

$$\ln(|\widehat{\Sigma}_1|/|\widehat{\Sigma}_2|) + (\mathbf{x} - \widehat{\boldsymbol{\mu}}_1)' \widehat{\Sigma}_1^{-1} (\mathbf{x} - \widehat{\boldsymbol{\mu}}_1) - (\mathbf{x} - \widehat{\boldsymbol{\mu}}_2)' \widehat{\Sigma}_2^{-1} (\mathbf{x} - \widehat{\boldsymbol{\mu}}_2) \leq 0, \quad (2.1)$$

and to class 2 otherwise, but this does not work when p is larger than or close to n . We propose to replace the sample covariance matrices in (2.1) with simpler alternatives. Our main idea is to simplify the matrix structure in order to reduce the number of unknown parameters. When there are fewer parameters, we can expect to estimate them consistently.

2.1 Some basic conditions

Before introducing the special matrix structure that we propose to use, we first describe some common conditions on the covariance matrices and sample sizes.

Let $\Sigma_{j_1 j_2}$ be the element of Σ in the j_1 th row and j_2 th column.

(C.1) For a constant $c > 0$, $|\Sigma_{j_1 j_2}| < c$ for $j_1 = 1, \dots, p$ and $j_2 = 1, \dots, p$.

Condition (C.1) places a bound on all the elements of Σ . Throughout the paper, we shall assume that both Σ_1 and Σ_2 satisfy condition (C.1).

Let $\mathbf{1}_p = (1, 1, \dots, 1)' \in \mathbb{R}^p$ and $Su(\Sigma) = \mathbf{1}_p' \Sigma \mathbf{1}_p$ be the summation of all elements in Σ . Condition (C.1) implies (C.1') below.

(C.1') For both $i = 1, 2$, $tr(\Sigma_i) = O(p)$, $tr(\Sigma_i^2) = O(p^2)$ and $Su(\Sigma_i) = O(p^2)$.

We also assume that the sample sizes n_i for the two classes are close.

(C.2) There exist $n > 0$ and constants $0 < c_1 < c_2 < +\infty$ such that $c_1 < n_i/n < c_2$ as $n \rightarrow \infty$ for both $i = 1, 2$.

Condition (C.2) is equivalent to saying that $n_1 \asymp n_2$. The value n has the same order as n_1 and n_2 ; it will be used later in our theoretical statements, where we will often refer to the sample size in general, without specifying the classes.

2.2 Main method: ppQDA

We now describe our main idea. Given Σ_i , let

$$a_i = p^{-1}tr(\Sigma_i) \quad \text{and} \quad r_i = (p(p-1))^{-1}(Su(\Sigma_i) - tr(\Sigma_i)),$$

be the average of its diagonal elements and the average of its off-diagonal elements, respectively. Our main idea is to use the structured matrix,

$$A_i = \begin{pmatrix} a_i & r_i & \cdots & r_i \\ r_i & a_i & \cdots & r_i \\ \vdots & \vdots & \ddots & \vdots \\ r_i & r_i & \cdots & a_i \end{pmatrix} = (a_i - r_i)I_p + r_i\mathbf{1}_p\mathbf{1}'_p,$$

which has uniform diagonal elements and uniform off-diagonal elements, in place of Σ_i , for $i = 1, 2$, in the quadratic discriminant rule (1.2).

Estimators of a_i and r_i , and hence of A_i as well, are based on the sample covariance matrix, i.e.,

$$\hat{a}_i = p^{-1}tr(\hat{\Sigma}_i), \quad \hat{r}_i = (p(p-1))^{-1}(Su(\hat{\Sigma}_i) - tr(\hat{\Sigma}_i)),$$

and

$$\hat{A}_i = (\hat{a}_i - \hat{r}_i)I_p + \hat{r}_i \mathbf{1}_p \mathbf{1}_p'.$$

As both a_i and r_i are scalar parameters, their estimators \hat{a}_i and \hat{r}_i are consistent even when p is large.

Using \hat{A}_i to replace $\hat{\Sigma}_i$, for $i = 1, 2$, in (2.1), we call the resulting decision rule the ‘‘ppQDA rule’’, where each ‘‘p’’ is short for ‘‘pooling’’ as constructing \hat{A}_i involves pooling both the diagonal and the off-diagonal elements of $\hat{\Sigma}_i$. Specifically, the ppQDA rule classifies \mathbf{x} to class 1 if

$$\hat{Q} = \ln(|\hat{A}_1|/|\hat{A}_2|) + (\mathbf{x} - \hat{\boldsymbol{\mu}}_1)' \hat{A}_1^{-1} (\mathbf{x} - \hat{\boldsymbol{\mu}}_1) - (\mathbf{x} - \hat{\boldsymbol{\mu}}_2)' \hat{A}_2^{-1} (\mathbf{x} - \hat{\boldsymbol{\mu}}_2) \leq 0, \quad (2.2)$$

and to class 2 otherwise. Due to its special structure, the inverse of A_i can be directly calculated:

$$\hat{A}_i^{-1} = (\hat{a}_i - \hat{r}_i)^{-1} I_p - \hat{r}_i (\hat{a}_i - \hat{r}_i)^{-1} (\hat{a}_i + (p-1)\hat{r}_i)^{-1} \mathbf{1}_p \mathbf{1}_p'. \quad (2.3)$$

Hence, we see that no matrix inversion is required, which is also a highly desirable property, especially for large p .

Theoretically, we are able to establish that our simplified ppQDA rule has good classification performance under (C.1)-(C.2) and some additional conditions on the matrices, A_i for $i = 1, 2$, given below:

$$(A.1) \quad a_i - r_i > \delta_i > 0, p[a_i + (p-1)r_i] > \delta_i' > 0;$$

$$(A.2) \quad |(a_1 - r_1) - (a_2 - r_2)| > \delta_0 > 0;$$

$$(A.3) \quad \text{tr}((A_i - \Sigma_i)^2) = o(p^2);$$

$$(A.4) \quad \sum_{j=1}^p (v_{ij} - \bar{v}_i)^2 = o(p^2), \text{ where } (v_{i1}, v_{i2}, \dots, v_{ip}) = \mathbf{1}_p' \Sigma_i \text{ — i.e., } v_{ij} \text{ is } j\text{th}$$

column-sum of Σ_i — and $\bar{v}_i = p^{-1} \sum_{j=1}^p v_{ij}$.

Theorem 1. *If conditions (C.1), (C.2) and (A.1) – (A.4) hold, the misclassification probability of the ppQDA rule (2.2) is asymptotically 0, i.e.,*

$$\lim_{p \rightarrow \infty, n \rightarrow \infty} \widehat{R}_{n,p} = \mathbb{P}(\widehat{Q} > 0 | \mathbf{x} \in \mathcal{C}_1) + \mathbb{P}(\widehat{Q} \leq 0 | \mathbf{x} \in \mathcal{C}_2) \rightarrow 0.$$

Notice that, in Theorem 1, we do not need to restrict the rate with which p approaches infinity relative to how fast the sample size n increases, a common requirement for high-dimensional problems. This is because the ppQDA rule, in effect, reduces each covariance matrix to just two scalar parameters, a_i and r_i , which can be consistently estimated regardless of how big the dimension p is. However, we will require a restriction of the aforementioned kind later in Section 3 as we extend our basic ideas to a semiparametric setting (see Remark 6 below).

While Theorem 1 establishes conditions under which the ppQDA rule can be nearly perfect asymptotically, we will also discuss in more detail below (Section 5) the factors that control how close the ppQDA rule can approach the Bayes decision rule when nearly perfect classification is not achievable.

We make some important remarks about the conditions (A.1) – (A.4).

Remark 1. As long as Σ_i is a positive definite matrix, the inequalities, $a_i - r_i > 0$ and $a_i + (p - 1)r_i > 0$, in (A.1) always hold (see Lemma 1, Supplement). In addition, the condition (A.1) requires that both $a_i - r_i$ and $p[a_i + (p - 1)r_i]$ be bounded away from 0, a degeneracy, even as the dimension gets high.

Remark 2. Condition (A.2) essentially requires that there is some difference between the two class covariance matrices, Σ_1 and Σ_2 , so that the two classes

can be separated. Generally for multivariate normal distributions, there are two sources of information that make classification possible: differences between the mean vectors (locations), and differences between the covariance matrices. Condition (A.2) is sufficient but not necessary, since it only requires some difference between the covariance matrices. If there is adequate signal in the mean vectors, e.g., if $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ is fairly large, then condition (A.2) can be relaxed. This will be discussed in more detail in the Supplement, after the proof of Lemma 2. We choose to use a condition that is solely focused on the covariance matrices for two reasons. First, there are already many papers in the literature (see Section 1) about discriminant analysis and classification based on signals from the mean vectors alone. Second, our main idea — that of replacing Σ_i with A_i — is about dealing with large covariance matrices (by introducing a structural simplification). As a result, our condition (A.2) actually makes classification possible even if there is no location separation at all ($\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = 0$).

Remark 3. Both conditions (A.3) and (A.4) place a bound on the difference between the true covariance matrix Σ_i and its structural simplification A_i . Naturally, if the true covariance matrix Σ_i really does have the simplified structure A_i , then our proposed ppQDA rule will be trivially optimal. What makes our proposal useful and interesting, of course, is that it can perform well even when the true covariance matrix does not have exactly the special structure. Conditions (A.3)-(A.4) make it precise how much Σ_i can deviate from the structure that would be “ideal” for our proposal. In particular, condition (A.3) means that the average of squared elementwise difference between Σ_i and A_i is $o(1)$. Condition (A.4) is similar to (A.3) except it is about the column sums of Σ_i , v_{i1}, \dots, v_{ip} ,

instead of about its individual elements. Notice that the average column sum, \bar{v}_i , can be expressed as $Su(\Sigma_i)/p = a_i + (p-1)r_i$, which is also equal to the uniform column sum of A_i for every column. Thus, condition (A.4) also means that the average squared difference between the column sums of Σ_i and those of A_i is $o(p)$. Here, it is important to note that some commonly used covariance structures do, in fact, satisfy these two conditions, including the autoregressive matrix such as $\Sigma_i(j_1, j_2) = \sigma_i^2 \rho_i^{|j_1 - j_2|}$ and the block diagonal matrix — provided that the block size q is $o(p)$. Of course, if Σ_i deviates a lot from the structural simplification, then both of these conditions can be violated. For example, if half of the off-diagonal entries in Σ_i are zero and the other half are 0.2, then it easily can be derived that $tr((A_i - \Sigma_i)^2) \geq 0.01p(p-1)$, so $tr((A_i - \Sigma_i)^2) \neq o(p^2)$ and (A.3) no longer holds.

2.3 Special case: pQDA

We also consider a special case, which uses just the trace estimator, $\hat{a}_i I_p$, to replace $\hat{\Sigma}_i$, $i = 1, 2$. We call this rule “pQDA” because only the diagonal elements of $\hat{\Sigma}_i$ are pooled and the off-diagonal elements are simply “ignored”. This rule classifies \mathbf{x} to class 1 if

$$\hat{Q}_0 = p \ln(\hat{a}_1/\hat{a}_2) + \hat{a}_1^{-1}(\mathbf{x} - \hat{\boldsymbol{\mu}}_1)'(\mathbf{x} - \hat{\boldsymbol{\mu}}_1) - \hat{a}_2^{-1}(\mathbf{x} - \hat{\boldsymbol{\mu}}_2)'(\mathbf{x} - \hat{\boldsymbol{\mu}}_2) \leq 0, \quad (2.4)$$

and to class 2 otherwise.

Clearly, the trace estimator, $\hat{a}_i I_p$ is a special case of \hat{A}_i . But we can take advantage of the added special structure and derive a stronger and more interpretable result under a different set of assumptions:

(B.1) there exist positive constants c_3 and c_4 such that $c_3 < \lambda_{ij} < c_4$ for $i = 1, 2$ and $j = 1, \dots, p$, where λ_{ij} is the j th eigenvalue of Σ_i ;

(B.2) there exists some positive constant c_5 such that $(a_{i_1}/a_{i_2} - \ln(a_{i_1}/a_{i_2}) - 1) + p^{-1}a_{i_2}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) > c_5$ for $(i_1, i_2) = (1, 2)$ and $(2, 1)$.

Theorem 2. *If conditions (C.1), (C.2), and (B.1) – (B.2) hold, the misclassification probability of the pQDA rule (2.4) is asymptotically 0, i.e.,*

$$\lim_{p \rightarrow \infty, n \rightarrow \infty} \widehat{R}_{0,n,p} = \mathbb{P}(\widehat{Q}_0 > 0 | \mathbf{x} \in \mathcal{C}_1) + \mathbb{P}(\widehat{Q}_0 \leq 0 | \mathbf{x} \in \mathcal{C}_2) \rightarrow 0.$$

The proof of Theorem 2 is, by and large, similar to that of Theorem 1 and the details will be omitted. Below, we make some important remarks about the conditions (B.1) – (B.2).

Remark 4. Condition (B.1) requires that the Σ_i s have bounded eigenvalues in order for pQDA to work. The reason why ppQDA does not require bounded eigenvalues is that, although both A_i and $a_i I_p$ have a similar structure (uniform diagonal elements and uniform off-diagonal elements), A_i has a spiked eigenvalue spectrum (provided that r_i does not degenerate to 0, the case of pQDA), whereas $a_i I_p$ has uniform eigenvalues. Boundedness can also be thought of as a different way of stating closeness. In addition, as $a_i I_p$ has uniform eigenvalues, it is intuitive that our pQDA rule will perform better if the true covariance matrix Σ_i has eigenvalues that are closer to each other.

Remark 5. As we mentioned before (Remark 2), in quadratic discriminant analysis, there are two sources of information that are useful for class separation. One is the difference in the mean vectors, and the other is the difference in the covariance matrices. In our pQDA rule, these two sources of information are

parameterized by $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ and a_1/a_2 or a_2/a_1 , respectively. Condition (B.2) simply requires that there is sufficient combined information for class separation from both sources. Note that the expression $a_{i_1}/a_{i_2} - \ln(a_{i_1}/a_{i_2}) - 1$ achieves its minimum value of 0 when $a_{i_1} = a_{i_2}$. Hence, classification becomes easier the larger the difference is between a_1 and a_2 — regardless of whether $a_1 > a_2$ or $a_2 > a_1$.

3. Generalization to deal with nonnormal data

As we briefly mentioned in Section 1, QDA often is more sensitive to violations of the normality assumption than is LDA. In this section, we investigate a semiparametric method to relax the normality assumption for the pQDA rule. The ppQDA rule can be generalized similarly, but the theoretical justification is much more tedious, although it requires no additional technique (more on this in Remark 9, Supplement). Thus, we will state generalized versions of both the ppQDA rule and the pQDA rule, as well as include both of them in our empirical studies (Sections 4), but we will only develop the theory for generalized pQDA.

For non-normal data, we follow a common approach in the literature (e.g., Lin and Jeon, 2003; Liu et al., 2009; Mai and Zou, 2015) and assume that

(D.1) there exist a set of strictly monotonic transformations

$$h(\mathbf{y}) \equiv (h_1(y_1), h_2(y_2), \dots, h_p(y_p))'$$

such that $h(\mathbf{y}_{ik}) \sim N(\boldsymbol{\mu}_i, \Sigma_i)$ for $k = 1, \dots, n_i$ and $i = 1, 2$.

This assumption is equivalent to using a Gaussian copula model to describe the dependence structure of multivariate observation \mathbf{y}_{ik} (Lin and Jeon, 2003).

To test the validity of (D.1), any high-dimensional normality test can be applied to the transformed data. However, testing normality in high dimensions is another complex research problem in itself. According to Lin and Jeon (2003), an alternative may be to check the classification results directly, as it is possible for a classification rule to work reasonably well even if the underlying normality assumption is violated.

Under this assumption, the generalization of ppQDA and pQDA is straightforward. First, we obtain a nonparametric estimate of the transformations, say

$$\widehat{h}(\cdot) \equiv (\widehat{h}_1(\cdot_1), \widehat{h}_2(\cdot_2), \dots, \widehat{h}_p(\cdot_p))',$$

from the training sample. Then, we apply ppQDA and pQDA to the transformed data, $\{\widehat{h}(\mathbf{y}_{ik}) : k = 1, \dots, n_i; i = 1, 2\}$ and $\widehat{h}(\mathbf{x})$. We refer to these procedures as Se-ppQDA and Se-pQDA, respectively, where ‘‘Se’’ is short for ‘‘semiparametric’’.

In what follows, we will use the same notations as before to denote various distributional parameters and their estimates for the *transformed* data. For example, $\boldsymbol{\mu}_i$ and Σ_i will now denote the mean vector and covariance matrix of the transformed sample $\{h(\mathbf{y}_{ik}) : k = 1, \dots, n_i\}$, while

$$\widehat{\boldsymbol{\mu}}_i = n_i^{-1} \sum_{k=1}^{n_i} \widehat{h}(\mathbf{y}_{ik}) \quad \text{and} \quad \widehat{\Sigma}_i = (n_i - 1)^{-1} \sum_{k=1}^{n_i} (\widehat{h}(\mathbf{y}_{ik}) - \widehat{\boldsymbol{\mu}}_i)(\widehat{h}(\mathbf{y}_{ik}) - \widehat{\boldsymbol{\mu}}_i)'$$

will denote the corresponding sample quantities based on the estimated transformation, \widehat{h} . Similarly, a_i, r_i (likewise $\widehat{a}_i, \widehat{r}_i$) will continue to denote, respectively, the average of the diagonal and off-diagonal elements of Σ_i (likewise $\widehat{\Sigma}_i$) — except Σ_i and $\widehat{\Sigma}_i$ are now covariance and sample covariance matrices of the *transformed*

data.

3.1 Estimation of h

Let F_{ij} be the class- i marginal cumulative distribution function (CDF) for the j -th dimension. Let σ_{ij}^2 be the variance of $h_j(y_{ij})$, i.e., σ_{ij}^2 is the j -th diagonal element of Σ_i . Notice that each of the assumed transformations $h_j(\cdot)$ in (D.1) must satisfy the following: if $u \sim F_{1j}$ and $v \sim F_{2j}$, then after transformation the marginal distributions of $h_j(u)$ and $h_j(v)$ can differ only up to a location- and-scale transform. Thus, we can set $\mu_{1j} = 0$ and $\sigma_{1j}^2 = 1$ for all $j = 1, \dots, p$, without loss of generality. This, in turn, means that each h_j can be equivalently expressed as

$$h_j = \Phi^{-1} \circ F_{1j} \quad \text{or} \quad h_j = \sigma_{2j} (\Phi^{-1} \circ F_{2j}) + \mu_{2j}, \quad (3.1)$$

where Φ denotes the CDF of the standard normal.

This means the transformation h_j can be estimated using training samples from either class. For this paper, we will estimate it using data from class 1, i.e.,

$$\hat{h}_j = \Phi^{-1} \circ \hat{F}_{1j},$$

where \hat{F}_{1j} is an “edge-smoothed” version of the empirical CDF (e.g., Mai and Zou, 2015),

$$\hat{F}_{1j}(t) = \begin{cases} 1 - \frac{1}{n_1^2}, & \text{if } \tilde{F}_{1j}(t) > 1 - \frac{1}{n_1^2}; \\ \tilde{F}_{1j}(t), & \text{if } \frac{1}{n_1^2} \leq \tilde{F}_{1j}(t) \leq 1 - \frac{1}{n_1^2}; \\ \frac{1}{n_1^2}, & \text{if } \tilde{F}_{1j}(t) < \frac{1}{n_1^2}, \end{cases}$$

and \tilde{F}_{1j} is the actual empirical CDF, $\tilde{F}_{1j}(t) = n_1^{-1} \sum_{k=1}^{n_1} \mathbf{1}\{y_{1jk} \leq t\}$. But our choice of using data from class 1 is entirely arbitrary. In practice, we recommend using data from the larger class in order to maximize estimation accuracy.

It is also possible to estimate the transformation h_j by making use of data from both classes. For example, Mai and Zou (2015) proposed such a pooled estimator for the special case in which the class covariance matrices are identical. A closer look at (3.1) suggests that a potential generalization of their pooled, two-sample estimator could be to take a weighted average of two different, one-sample estimators of h_j , e.g.,

$$\hat{h}_j = \frac{n_1}{n}(\Phi^{-1} \circ \hat{F}_{1j}) + \frac{n_2}{n} \left[\hat{\sigma}_{2j}(\Phi^{-1} \circ \hat{F}_{2j}) + \hat{\mu}_{2j} \right],$$

where \hat{F}_{2j} is defined similarly as \hat{F}_{1j} above. To take full advantage of pooled estimation, one could obtain $\hat{\sigma}_{2j}$ and $\hat{\mu}_{2j}$ with a pooled method as well, as there is information about them not only in the transformed sample $\{\Phi^{-1}[\hat{F}_{1j}(y_{2jk})]\}_{k=1}^{n_2}$ but also in $\{\Phi^{-1}[\hat{F}_{2j}(y_{1jk})]\}_{k=1}^{n_1}$. However, since this is not the main focus of our study, we will not pursue this more complicated strategy in the current paper.

3.2 Se-ppQDA and Se-pQDA

Since our estimated transformations $\hat{h}_1, \dots, \hat{h}_p$ automatically make $\hat{\mu}_1 = \mathbf{0}$, the Se-ppQDA rule classifies \mathbf{x} to class 1 if

$$\hat{Q}_{\hat{h}} = \ln(|\hat{A}_1|/|\hat{A}_2|) + \hat{h}(\mathbf{x})' \hat{A}_1^{-1} \hat{h}(\mathbf{x}) - (\hat{h}(\mathbf{x}) - \hat{\mu}_2)' \hat{A}_2^{-1} (\hat{h}(\mathbf{x}) - \hat{\mu}_2) \leq 0, \quad (3.2)$$

and to class 2 otherwise. Similarly, that $\hat{\sigma}_{1j}^2 = 1$ for all $j = 1, \dots, p$ implies

$\hat{a}_1 = p^{-1}tr(\hat{\Sigma}_1) = 1$, so the Se-pQDA rule classifies \mathbf{x} to class 1 if

$$\hat{Q}_{\hat{h},0} = p \ln(1/\hat{a}_2) + \hat{h}(\mathbf{x})'\hat{h}(\mathbf{x}) - \hat{a}_2^{-1}(\hat{h}(\mathbf{x}) - \hat{\boldsymbol{\mu}}_2)'(\hat{h}(\mathbf{x}) - \hat{\boldsymbol{\mu}}_2) \leq 0, \quad (3.3)$$

and to class 2 otherwise.

We are now ready to establish some theoretical results about the asymptotic performance of the Se-pQDA rule. While the idea behind Se-pQDA — first estimating the transformations and then applying pQDA to transformed data — is straight-forward, its performance is much more intricate to analyze than that of pQDA, being affected by not only the structural simplifications of the pQDA rule itself, but also the estimation quality of the p univariate transformations and that of the key model parameters for the transformed data.

Theorem 3. *Under (D.1), if (C.1), (C.2), and (B.1) – (B.2) hold for the transformed data, then, the misclassification probability of the Se-pQDA rule (3.3) is asymptotically 0, i.e.,*

$$\lim_{p \rightarrow \infty, n \rightarrow \infty} \hat{R}_{\hat{h},0,n,p} = \mathbb{P}(\hat{Q}_{\hat{h},0} > 0 | \mathbf{x} \in \mathcal{C}_1) + \mathbb{P}(\hat{Q}_{\hat{h},0} \leq 0 | \mathbf{x} \in \mathcal{C}_2) \rightarrow 0,$$

provided that $p \exp(-Cn^{1/3-\theta}) \rightarrow 0$ for some $C > 0$ and $0 < \theta < 1/3$, and that there exists some constant $c_6 > 0$ such that $|\mu_{2j}| < c_6$ for all $j = 1, \dots, p$.

Compared with Theorem 2 and aside from the obvious additional condition of (D.1), Theorem 3 requires two more assumptions, about which we will make some remarks below.

Remark 6. Recall that, previously for ppQDA and pQDA, we did not need to control the rate with which p goes to infinity relative to that of n , but we do

now for Se-pQDA. This is because we must now estimate p univariate transformations. To ensure that we can estimate these transformations reasonably well, the dimension p cannot grow too fast relative to the overall sample size n . More precisely, we require $p \exp(-Cn^{1/3-\theta}) \rightarrow 0$ for some $C > 0$ and $0 < \theta < 1/3$ as both p and n tend to infinity.

Remark 7. The additional assumption in Theorem 3 — that every $|\mu_{2j}|$ is bounded — is introduced to avoid some unnecessary technical difficulties in our proof. This added assumption does not really weaken our result. If $|\mu_{2j}|$ is very large, it only makes classification easier, and the more challenging (and hence more interesting) problem in practice occurs when the marginal signals are relatively weak. This is especially relevant as we have not made any sparsity assumptions about $\mu_1 - \mu_2$. Situations in which signals from the mean vectors are relatively dense (see, e.g., Fan et al., 2013) are only interesting when those signals are marginally faint.

4. Numerical studies

In this section, we demonstrate the performance of pQDA, ppQDA, Se-pQDA and Se-ppQDA by simulation. Two real-data examples are provided in a supplementary section.

Three other methods — DSDA (Mai et al., 2012), SSDA (Mai and Zou, 2015) and random forest (Breiman, 2001) — are included for comparison purpose. Both DSDA and SSDA are penalized linear discriminant rules, and the latter deals with nonnormal data; for these methods we used the R package `dsda`, provided by the authors of the methods. For random forest, we used the R package `randomForest`

with a forest size of 1000; for all other parameters, we simply used their default values as further adjustments did not noticeably affect the performance.

We also include a benchmark classifier, in which the true covariance matrices (Σ_1, Σ_2) and the *sample* means $(\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2)$ are plugged into the QDA rule. Note that we used only the true covariance matrices — but *not* the true mean vectors — in the benchmark classifier, because we would like to focus on the effect of using our structured covariance matrices for classification, and to avoid letting the estimation of the mean vectors $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ (an intricate problem on its own in high dimensions) unduly confound our performance evaluation.

For all our QDA procedures (i.e., pQDA, ppQDA, Se-pQDA, Se-ppQDA), we standardized the variance of each dimension j by the larger of the two within-class standard deviations, i.e., $\max\{\hat{\sigma}_{1j}, \hat{\sigma}_{2j}\}$. In the case of Se-pQDA and Se-ppQDA, such standardization was performed after first estimating and then applying the transformation h_j .

4.1 Simulated examples

We considered nine different types of covariance matrix structures (e.g., partly autoregressive, block-diagonal, and so on), the details of which are described in a supplementary section. Based on these nine structures, we created ten simulated examples, setting either $p = 400$ or $p = 800$. In all of them, the means of the two classes were taken to be $\boldsymbol{\mu}_1 = \mathbf{0}_p$ and $\boldsymbol{\mu}_2 = (3.5p^{-1/2}\mathbf{1}'_{0.6p}, \mathbf{0}'_{0.4p})'$. That is, the signal was spread out evenly among the first $0.6p$ dimensions. The magnitude of the signal in each dimension was controlled so that the between-class Euclidean distance did not change with p . The ten examples differed mostly

by the covariance matrices of the two classes. In all cases, we also controlled the difference between the two within-class covariance matrices by a parameter $s \equiv 3p^{-1/2}$ (see below).

Example 1: $\Sigma_1 = M_1$, partly autoregressive, and $\Sigma_2 = \Sigma_1 + sI_p$.

Example 2: $\Sigma_1 = M_3$, block diagonal, and $\Sigma_2 = \Sigma_1 + sI_p$.

Example 3: $\Sigma_1 = M_4$, modified version of M_1 , and $\Sigma_2 = \Sigma_1 + sI_p$. This example is designed to investigate a case in which the covariance matrices have eigenvalues that are quite close to each other — one in which pQDA is expected to perform well (see Remark 4).

Example 4: $\Sigma_1 = M_1$, partly autoregressive, and $\Sigma_2 = M_2 + sI_p$, also partly autoregressive, but with some elements (both diagonal and off-diagonal ones) being different from those in Σ_1 .

Example 5: $\Sigma_1 = \Sigma_2 = M_1$, partly autoregressive, and identical between the two classes. This example is designed to test the performance of our QDA rules when the LDA rule is optimal.

Example 6: $\Sigma_1 = M_5$, compound symmetry, and $\Sigma_2 = \Sigma_1 + sI_p$. In this example ppQDA is expected to have an advantage over pQDA.

Example 7: $\Sigma_1 = M_6$, also compound symmetry, and $\Sigma_2 = \Sigma_1 + sI_p$. The matrix M_6 is different from M_5 in that it has negative off-diagonal elements that are close to 0 and is almost not positive definite.

Example 8: $\Sigma_1 = M_7$, compound symmetry with off-diagonal perturbations, and $\Sigma_2 = \Sigma_1 + sI_p$.

Example 9: $\Sigma_1 = M_8$, compound symmetry with diagonal perturbations, and $\Sigma_2 = \Sigma_1 + sI_p$.

Example 10: $\Sigma_1 = M_9$, unstructured, and $\Sigma_2 = \Sigma_1 + sI_p$.

4.2 Results

For all of our simulated examples, we used $n_i = 100$ training samples, and 1000 independent testing samples, from $N(\boldsymbol{\mu}_i, \Sigma_i)$, $i = 1, 2$. All simulations were repeated for 100 times, and the average misclassification rates on the testing samples were recorded, together with their standard errors.

Table 1 shows how the methods compared on the ten examples. Our suite of methods were generally better than DSDA, SSDA and random forest. This is not surprising as both DSDA and SSDA assume sparsity and identical within-class covariance matrices, and the random forest does not make (or take advantage of) any specific distribution assumption. In each example, the best method statistically matched the benchmark classifier. Recall that, for the benchmark classifier, we used only the true covariance matrices but still kept using the sample rather than the population mean vectors, so it was possible sometimes for other methods to outperform it.

In Examples 1-4, the covariance matrices are better approximated by diagonal ones, so pQDA is expected to perform well, but we see that ppQDA performed reasonably well, too. This indicates that, whenever pQDA works, ppQDA is only slightly worse than, if not as good as, pQDA.

In Example 5, the two within-class covariance matrices are the same, so LDA is actually optimal, but we see that both pQDA and ppQDA still continued to perform well.

In Examples 6-7, the covariance matrices have exactly the compound sym-

metry structure, so naturally in these cases we see that ppQDA performed considerably better than all other methods.

In Examples 8-9, the covariance matrices no longer have exactly the compound symmetry structure, due to perturbations to the various off-diagonal (M_7 , Example 8) and diagonal (M_8 , Example 9) elements. In Example 10, the covariance matrices are largely unstructured, except that a few randomly selected entries are much larger than others. These examples were designed to test the robustness and sensitivity of ppQDA. In all of these cases, ppQDA maintained good performance and sometimes still commanded a considerable advantage over all other methods.

In Table 1, we see that both Se-pQDA and Se-ppQDA performed slightly worse than their counterparts without any nonlinear transformations. Clearly, estimating these extra transformations when they were unnecessary introduced additional errors. We also transformed data from these ten examples to be non-normally distributed and repeated our experiments. The details of these experiments and their results are described in a supplementary section. When the data were non-normal, the advantages of Se-pQDA and Se-ppQDA over other methods became clear.

5. Discussion

So far our theoretical results have focused on establishing conditions under which our proposed methods (e.g., ppQDA, pQDA, Se-pQDA) can have nearly perfect performance asymptotically. In reality, of course, perfect classification is not always possible, in which case we would like to know how well our methods can

Table 1: Average misclassification rates (%) and their standard errors. Data are generated from $N(\mu_1, \Sigma_1)$, $N(\mu_2, \Sigma_2)$.

Example	pQDA	ppQDA	Se-pQDA	Se-ppQDA	DSDA	SSDA	RF	Benchmark	
$p = 400$	1	13.5(0.11)	14.3(0.12)	14.1(0.12)	15.3(0.13)	32.3(0.26)	34.7(0.26)	24.6(0.13)	13.7(0.11)
	2	13.7(0.11)	14.7(0.12)	14.2(0.12)	15.6(0.13)	32.4(0.21)	35.1(0.24)	25.1(0.12)	14.1(0.11)
	3	20.8(0.12)	21.0(0.14)	21.2(0.11)	22.0(0.11)	35.6(0.31)	38.5(0.34)	30.4(0.14)	20.5(0.13)
	4	13.6(0.09)	14.5(0.11)	14.3(0.10)	15.5(0.12)	32.0(0.20)	34.9(0.30)	24.8(0.15)	13.5(0.10)
	5	20.5(0.11)	22.3(0.15)	21.9(0.12)	24.7(0.15)	31.7(0.26)	34.5(0.28)	26.8(0.13)	24.9(0.14)
	6	38.3(0.41)	14.0(0.10)	36.5(0.40)	16.4(0.11)	38.8(0.26)	38.4(0.25)	36.4(0.28)	13.0(0.08)
	7	13.4(0.10)	0.00(0.00)	15.7(0.11)	2.70(0.06)	28.0(0.19)	33.1(0.30)	25.9(0.14)	0.00(0.00)
	8	33.8(0.46)	16.7(0.12)	30.1(0.46)	17.7(0.13)	38.4(0.29)	38.8(0.25)	34.9(0.23)	6.50(0.07)
	9	39.3(0.35)	26.1(0.14)	37.1(0.35)	25.7(0.12)	42.4(0.23)	42.4(0.23)	39.0(0.17)	24.8(0.12)
	10	23.1(0.36)	9.40(0.09)	18.4(0.30)	11.0(0.11)	35.5(0.26)	36.1(0.25)	28.2(0.15)	5.50(0.06)
$p = 800$	1	16.7(0.11)	17.8(0.13)	17.1(0.12)	18.6(0.14)	36.8(0.21)	40.1(0.30)	29.7(0.12)	17.4(0.10)
	2	17.2(0.12)	18.2(0.14)	17.7(0.14)	19.2(0.15)	37.4(0.30)	40.5(0.27)	29.9(0.13)	17.8(0.11)
	3	25.6(0.13)	26.1(0.15)	25.1(0.14)	26.6(0.15)	40.8(0.33)	43.6(0.25)	35.5(0.12)	24.4(0.12)
	4	16.6(0.11)	17.7(0.11)	17.1(0.10)	18.7(0.11)	36.7(0.20)	39.7(0.25)	29.5(0.13)	17.4(0.11)
	5	24.3(0.14)	26.0(0.16)	26.2(0.15)	29.7(0.16)	36.5(0.35)	40.0(0.3)	31.7(0.13)	28.7(0.13)
	6	41.7(0.34)	18.2(0.12)	40.5(0.30)	20.0(0.12)	42.7(0.25)	43.1(0.25)	40.4(0.26)	17.0(0.10)
	7	18.9(0.12)	0.00(0.00)	20.2(0.13)	3.80(0.07)	37.2(0.33)	41.3(0.31)	32.1(0.13)	0.00(0.00)
	8	36.7(0.44)	22.0(0.13)	32.9(0.48)	21.8(0.13)	43.1(0.23)	43.1(0.21)	39.1(0.18)	8.40(0.08)
	9	42.9(0.25)	30.5(0.14)	40.8(0.30)	29.5(0.14)	45.4(0.20)	45.5(0.21)	42.7(0.18)	29.4(0.12)
	10	28.0(0.38)	16.0(0.12)	22.6(0.32)	16.6(0.12)	40.6(0.23)	40.8(0.23)	34.0(0.17)	0.60(0.02)

perform relative to the Bayes decision rule. In this section, we will provide some answers to this question for ppQDA.

To do so, we further simplify the situation by focusing on a special case where there is no signal for classification in the class means, i.e., $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \mathbf{0}$. As we have already stated earlier (see Remark 2), since there are already many papers in the literature about classification based on signals from the mean vectors alone, and since our main idea of replacing Σ_i with A_i is “only” about dealing with large covariance matrices, we think it actually makes things clearer if we concentrate on just the covariance matrices and ignore the mean vectors.

We will also focus on the population version of the ppQDA rule. This is because, in our proof of Theorem 1 (see Supplementary Materials, Section S4), we already established that the quantity $\widehat{Q} - Q$ is dominated by the population quantity Q as $p, n \rightarrow \infty$. However, our proof has assumed (A.1)-(A.4), but this section is primarily concerned with situations in which asymptotically perfect classification is not achievable, so it would be desirable if this dominance could be established without (A.2). Indeed, this is possible, provided that some mild modifications are made to (A.3) and (A.4). Specifically, instead of the difference between A_i and Σ_i being simply $o(p^2)$, now its order must also depend on how much signal there is for classification, as measured by $(a_{i_1} - r_{i_1})/(a_{i_2} - r_{i_2})$ for $(i_1, i_2) = (1, 2)$ and $(2, 1)$. A detailed proof is omitted, as the technique is similar to that used in the proof of Theorem 1.

Let A_1, A_2, Σ_1 and Σ_2 be defined as in Section 2. Under the assumption that $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \mathbf{0}$, the quantity that drives (population) ppQDA, using the true

(as opposed to estimated) parameters, is

$$Q = \ln(|A_1|/|A_2|) + \mathbf{x}'A_1^{-1}\mathbf{x} - \mathbf{x}'A_2^{-1}\mathbf{x},$$

whereas the Bayes decision rule is driven by

$$Q_B = \ln(|\Sigma_1|/|\Sigma_2|) + \mathbf{x}'\Sigma_1^{-1}\mathbf{x} - \mathbf{x}'\Sigma_2^{-1}\mathbf{x}.$$

Clearly, the performance of ppQDA will be close to that of the Bayes rule if $\Sigma_i \approx A_i$ for both $i = 1, 2$, but we will argue below that this need not necessarily be the case.

To see this, suppose first that $\mathbf{x} \in \mathcal{C}_1$. Then, for any matrix B , we have

$$\begin{aligned} \mathbb{E}(\mathbf{x}'B\mathbf{x}|\mathbf{x} \in \mathcal{C}_1) &= \mathbb{E}[tr(\mathbf{x}'B\mathbf{x})|\mathbf{x} \in \mathcal{C}_1] = \mathbb{E}[tr(B\mathbf{x}\mathbf{x}')|\mathbf{x} \in \mathcal{C}_1] \\ &= tr[B\mathbb{E}(\mathbf{x}\mathbf{x}'|\mathbf{x} \in \mathcal{C}_1)] = tr(B\Sigma_1), \end{aligned}$$

which immediately implies

$$\mathbb{E}(Q_B|\mathbf{x} \in \mathcal{C}_1) = \ln|\Sigma_2^{-1}\Sigma_1| + p - tr(\Sigma_2^{-1}\Sigma_1), \quad (5.1)$$

and

$$\mathbb{E}(Q|\mathbf{x} \in \mathcal{C}_1) = \ln|A_2^{-1}A_1| + tr(A_1^{-1}\Sigma_1) - tr(A_2^{-1}\Sigma_1). \quad (5.2)$$

But the inverse formula for \hat{A}_i , given in equation (2.3), applies to A_i as well, so

$$tr(A_i^{-1}\Sigma_1) = [(a_i - r_i)^{-1}]tr(\Sigma_1) - [r_i(a_i - r_i)^{-1}(a_i + (p - 1)r_i)^{-1}]tr(\mathbf{1}_p\mathbf{1}_p'\Sigma_1).$$

However, the definition of A_1 implies $tr(\Sigma_1) = tr(A_1)$ and

$$tr(\mathbf{1}_p \mathbf{1}_p' \Sigma_1) = tr(\mathbf{1}_p' \Sigma_1 \mathbf{1}_p) = Su(\Sigma_1) = Su(A_1) = tr(\mathbf{1}_p' A_1 \mathbf{1}_p) = tr(\mathbf{1}_p \mathbf{1}_p' A_1).$$

This means $tr(A_i^{-1} \Sigma_1) = tr(A_i^{-1} A_1)$ so that (5.2) can be further reduced to

$$\mathbb{E}(Q|\mathbf{x} \in \mathcal{C}_1) = \ln |A_2^{-1} A_1| + p - tr(A_2^{-1} A_1). \quad (5.3)$$

Together, equations (5.3) and (5.1) are highly suggestive of the possibility that, given $\mathbf{x} \in \mathcal{C}_1$, the performance of ppQDA can be close to that of the Bayes rule as long as $A_2^{-1} A_1$ is close to $\Sigma_2^{-1} \Sigma_1$ in the sense that

$$tr(A_2^{-1} A_1) \approx tr(\Sigma_2^{-1} \Sigma_1) \quad \text{and} \quad |A_2^{-1} A_1| \approx |\Sigma_2^{-1} \Sigma_1|,$$

whereas each A_i need not be close to Σ_i in itself.

Moreover, for two $p \times p$, symmetric, positive-definite matrices U, V , we can define the function,

$$\phi(U, V) = \left| \ln |V^{-1} U| + p - tr(V^{-1} U) \right|,$$

as one way to measure their difference — notice that $\phi(U, V) = 0$ if $U = V$, and the absolute value is needed because, for any $p \times p$, symmetric, positive-definite matrix M with eigenvalues $\lambda_1, \dots, \lambda_p$, the function $\ln |M| + p - tr(M) = \sum (\ln \lambda_j + 1 - \lambda_j) \leq 0$ with equality only when $\lambda_j = 1$ for all j ; see also Remark 5. For $\mathbf{x} \in \mathcal{C}_1$, our analysis above shows that, on average, the Bayes rule and the ppQDA rule are simply using the same $\phi(\cdot, \cdot)$ function to measure the differences between a different set of matrices — (Σ_1, Σ_2) for the Bayes rule and (A_1, A_2)

for ppQDA.

Combined with arguments similar to those we used to prove Theorem 1 (see Section S 4, Supplement), our analysis above also suggests that, for $\mathbf{x} \in \mathcal{C}_1$, the performance of ppQDA can be asymptotically close to that of the Bayes rule if

$$\frac{\phi(\Sigma_1, \Sigma_2) - \phi(A_1, A_2)}{\phi(\Sigma_1, \Sigma_2)} \sim o(1) \quad \text{as } p \rightarrow \infty.$$

The same argument applies to the case of $\mathbf{x} \in \mathcal{C}_2$, except that, in this case, the differences are measured by $\phi(A_2, A_1)$ and $\phi(\Sigma_2, \Sigma_1)$ instead of by $\phi(A_1, A_2)$ and $\phi(\Sigma_1, \Sigma_2)$. Thus, we define the symmetric difference measure,

$$\varphi(U, V) = \phi(U, V) + \phi(V, U),$$

and conjecture that the relative performance of our ppQDA rule to that of the Bayes rule depends very much on the quantity,

$$\Delta \equiv \frac{\varphi(\Sigma_1, \Sigma_2) - \varphi(A_1, A_2)}{\varphi(\Sigma_1, \Sigma_2)}, \quad (5.4)$$

and whether $\Delta \rightarrow 0$ as $p \rightarrow \infty$. In a supplementary section, we present some empirical evidence to support this observation.

6. Conclusion

We have proposed two simple rules — namely, ppQDA and pQDA — to perform quadratic discriminant analysis for high-dimensional data, and generalized both rules by using a semiparametric transformation in order to handle data that do not necessarily follow the normal distribution. Desirable theoretical properties

have been established for ppQDA, pQDA, and Se-pQDA — the semiparametric extension of pQDA. The performances of our specialized quadratic discriminant rules are comparable to, if not better than, other high-dimensional discriminant analysis methods in many numerical experiments and several real-data examples.

Unlike many existing high-dimensional discriminant analysis methods that focus on LDA, our methods aim at performing QDA, which allows us to exploit the difference between covariance matrices from separate classes and use it for classification. The sample covariance matrix is inconsistent when the dimension is high. Whereas most methods address this difficulty by imposing sparsity conditions, we do so by simplifying the structure of covariance matrices while still trying to capture some subtle information from across all dimensions. The special matrix structure that we use can be viewed as a generalization of the trace estimator, which has been used in high-dimensional hypothesis-testing as well as classification problems. Specifically, we pool not only the diagonal elements but also the off-diagonal ones in each covariance matrix, so as to obtain some information about the correlations among different dimensions. As a result, our easy-to-apply discriminant rules enjoy very low computational costs. The sparsity approach can be quite unstable for weak signals, and is more suitable for dealing with cases with just a few strong signals. Our approach is more attractive for cases with many weak signals.

Because of the complexity of the problem, at this point it is difficult to imagine that there could be a universally optimal discriminant analysis method for high-dimensional data. Almost every method can enjoy some advantages under certain circumstances. Due to noise accumulation, the performance of our

methods could certainly deteriorate when there are a large number of useless covariates, but so would most methods. Due to the special matrix structure that we use, which has a common set of diagonal elements and a common set of off-diagonal ones, one may also expect that our discriminant rules may not perform too well if the marginal variances across different dimensions are vastly different, or if some dimensions are very highly correlated while others have little correlation. In practice, however, these two problems can be alleviated by pre-screening and properly preprocessing the data (see real-data examples in the Supplement). Our current main interest lies in the question of what other special matrix structures we can exploit for high-dimensional QDA. Prominent candidates must allow us to capture more information in each covariance matrix (than what can be captured by just two scalars a_i, r_i), but still have a relatively small number of “easily estimable” parameters.

Supplementary Materials

Supplementary materials are provided in five separate sections. Section S 1 provides more details and results for our numerical studies in Section 4. Section S 2 contains two real-data examples. Section S 3 provides empirical evidence to support observations made in Section 5. Section S 4 is a brief outline of the main proofs, while the detailed proofs are given in Section S 5.

Acknowledgement

Qin’s research was supported by Natural Sciences and Engineering Research Council of Canada (NSERC) RGPIN-2016-03890 and the University of Water-

loo Research Incentive Fund 115953. Zhu's research was supported by NSERC RGPIN-2016-03876. The authors thank the editor, associate editor and referees for their insightful and constructive comments and suggestions.

References

- Anderson, T. W., Anderson, T. W., Anderson, T. W., and Anderson, T. W. (1958). *An introduction to multivariate statistical analysis*, volume 2. Wiley New York.
- Aoshima, M. and Yata, K. (2014). A distance-based, misclassification rate adjusted classifier for multiclass, high-dimensional data. *Annals of the Institute of Statistical Mathematics*, 66(5):983–1010.
- Aoshima, M. and Yata, K. (2015). High-dimensional quadratic classifiers in non-sparse settings. *arXiv preprint arXiv:1503.04549*.
- Bai, Z. D. and Saranadasa, H. (1996). Effect of high dimension: by an example of a two sample problem. *Statistica Sinica*, 6(2):311–329.
- Bickel, P. J. and Levina, E. (2004). Some theory for fisher's linear discriminant function, 'naive bayes', and some alternatives when there are many more variables than observations. *Bernoulli*, pages 989–1010.
- Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.
- Cai, T. and Liu, W. (2011). A direct estimation approach to sparse linear discriminant analysis. *Journal of the American Statistical Association*, 106(496).
- Cai, T., Liu, W., and Xia, Y. (2013). Two-sample covariance matrix testing and support recovery in high-dimensional and sparse settings. *Journal of the American Statistical Association*, 108(501):265–277.
- Chen, S. X. and Qin, Y.-L. (2010). A two-sample test for high-dimensional data with applica-

- tions to gene-set testing. *The Annals of Statistics*, 38(2):808–835.
- Dudoit, S., Fridlyand, J., and Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American statistical association*, 97(457):77–87.
- Fan, J. and Fan, Y. (2008). High dimensional classification using features annealed independence rules. *Annals of statistics*, 36(6):2605.
- Fan, J., Feng, Y., and Tong, X. (2012). A road to classification in high dimensional space. *Journal of the Royal Statistical Society. Series B, Statistical methodology*, 74(4):745–771.
- Fan, J., Ke, Z. T., Liu, H., and Xia, L. (2015). Quadro: A supervised dimension reduction method via rayleigh quotient optimization. *Annals of statistics*, 43(4):1498.
- Fan, Y., Jin, J., Yao, Z., et al. (2013). Optimal classification in sparse gaussian graphic model. *The Annals of Statistics*, 41(5):2537–2571.
- Friedman, J. H. (1989). Regularized discriminant analysis. *Journal of the American statistical association*, 84(405):165–175.
- Hao, N., Dong, B., and Fan, J. (2015). Sparsifying the fisher linear discriminant by rotation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(4):827–851.
- Li, J., Chen, S. X., et al. (2012). Two sample tests for high-dimensional covariance matrices. *The Annals of Statistics*, 40(2):908–940.
- Li, Q. and Shao, J. (2014). Sparse quadratic discriminant analysis for high dimensional data. *Statistica Sinica*.
- Lin, Y. and Jeon, Y. (2003). Discriminant analysis through a semiparametric model. *Biometrika*, 90(2):379–392.
- Liu, H., Lafferty, J., and Wasserman, L. (2009). The nonparanormal: Semiparametric estima-

tion of high dimensional undirected graphs. *The Journal of Machine Learning Research*, 10:2295–2328.

Mai, Q. and Zou, H. (2015). Sparse semiparametric discriminant analysis. *Journal of Multivariate Analysis*, 135:175–188.

Mai, Q., Zou, H., and Yuan, M. (2012). A direct approach to sparse discriminant analysis in ultra-high dimensions. *Biometrika*, page asr066.

Shao, J., Wang, Y., Deng, X., Wang, S., et al. (2011). Sparse linear discriminant analysis by thresholding for high dimensional data. *The Annals of statistics*, 39(2):1241–1265.

Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences*, 99(10):6567–6572.

Witten, D. M. and Tibshirani, R. (2011). Penalized classification using fisher’s linear discriminant. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(5):753–772.

Department of Statistics and Actuarial Sciences, University of Waterloo, 200 University Avenue
W., Waterloo, ON N2L 3G1, CANADA

E-mail: y335wu@uwaterloo.ca

E-mail: yingli.qin@uwaterloo.ca

E-mail: mu.zhu@uwaterloo.ca