Journal of Multivariate Analysis 190 (2022) 104987

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Quadratic discriminant analysis by projection

Ruiyang Wu, Ning Hao *

Department of Mathematics, The University of Arizona, 617 N Santa Rita Ave, Tucson, AZ 85721-0089, USA

ARTICLE INFO

Article history:
Received 19 August 2021
Received in revised form 2 March 2022
Accepted 3 March 2022
Available online 10 March 2022

AMS 2020 subject classifications:
primary 62H30
secondary 62H12
62F12

Keywords:
Classification
Consistency
Heteroscedasticity
Invariance
Normality

ABSTRACT

Discriminant analysis, including linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA), is a popular approach to classification problems. It is well known that LDA is suboptimal to analyze heteroscedastic data, for which QDA would be an ideal tool. However, QDA is less helpful when the number of features in a data set is moderate or high, and LDA and its variants often perform better due to their robustness against dimensionality. In this work, we introduce a new dimension reduction and classification method based on QDA. In particular, we define and estimate the optimal one-dimensional (1D) subspace for QDA, which is a novel hybrid approach to discriminant analysis. The new method can handle data heteroscedasticity with number of parameters equal to that of LDA. Therefore, it is more stable than the standard QDA and works well for data in moderate dimensions. We show an estimation consistency property of our method, and compare it with LDA, QDA, regularized discriminant analysis (RDA) and a few other competitors by simulated and real data examples.

1. Introduction

Discriminant analysis is a standard tool for classification. For example, LDA and QDA aim to find hyperplanes and quadratic hypersurfaces, respectively, to separate the data points. LDA is one of the most popular techniques for classification because of its simplicity and robustness against growing dimensionality. Nevertheless, the performance of LDA relies on the equal covariance assumption. In contrast, QDA allows data heteroscedasticity. The cost of the flexibility is to estimate more parameters of the QDA model, which requires a large sample size. To make the QDA approach more robust, Friedman [8] proposed regularized discriminant analysis (RDA), which shrinks the separate covariances of different classes toward a common pooled covariance that can be further shrunken to a diagonal matrix when necessary. The level of shrinkage is controlled by tuning parameters, which are often tuned by cross-validation. As a compromise between LDA and QDA, RDA is a successful classification tool which has been further developed in Guo et al. [10].

Based on Fisher's original idea [7], LDA aims to find a 1D projection which best separates the data. Fisher suggested the direction that maximizes the ratio of between-class variance to within-class variance. Under the Gaussian and equal covariance assumption, the population version of LDA rule, or PoLDA for short, is the optimal classification rule. This implies two facts. First, there is no information loss to project the data onto the PoLDA direction. Second, PoLDA minimizes classification error. These properties of PoLDA do not hold under data heteroscedasticity. In general, it is impossible to project the data to a 1D subspace without loss of information. Even if a good projection exists, QDA might be a better choice than LDA to separate the projected data. This motivates us to study the optimal 1D projection for heteroscedastic data. To elaborate, we will define an optimal direction in which the projected data are separated by QDA with least

* Corresponding author.
E-mail address: nhao@math.arizona.edu (N. Hao).

https://doi.org/10.1016/j.jmva.2022.104987
0047-259X/

classification error. We propose an algorithm to approximate this optimal direction and show its consistency. With strong heteroscedasticity, our method can outplay LDA methods. Because the number of parameters in our algorithm is similar to that of LDA, our method needs a smaller sample size than QDA does.

In this work, we are mainly concerned with data sets with $p < n < p^2$ where n is the sample size and p is the number of features. In this situation, LDA may be seriously biased because of data heteroscedasticity, and QDA is not stable numerically due to dimensionality. Our method offers an alternative classification tool for practitioners. We have to point out that, in the last 20 years, there are many works on discriminant analysis for high and ultra-high dimensional data, Li and Shao [15], Jiang et al. [12], Wu et al. [23], Gaynanova and Wang [9], just to name a few closely related to QDA. We refer readers to two review papers [17,19] for more comprehensive summaries of recent developments. A majority of these works are based on various sparsity assumptions. In contrast, we do not impose sparsity assumptions and our method shares invariance property with the classical LDA and QDA methods. Therefore, we do not suggest to apply our method to high dimensional data directly.

The rest of this paper is organized as follows. In Section 2, we define the optimal 1D projection for heteroscedastic Gaussian data, based on which we propose a new classification rule, QDA by Projection (QDAP). In Section 3, we show that the direction of optimal 1D projection is estimated consistently by our plug-in estimator, and that our algorithm of QDAP is invariant under invertible linear transformations. In Section 4, we present numerical results from both simulated and real data examples. In Section 5, we discuss briefly a few related works and possible extensions.

2. Classification by 1D projection

2.1. Optimal 1D projection for heteroscedastic Gaussian data

Let $\mathbf{X}$ be a p-dimensional random vector, and $Y \in \{0, 1\}$ be its class label with $\mathbf{X}|\{Y = k\} \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, and $\Pr(Y = k) = \pi_k$, $k = 0, 1$, where $\boldsymbol{\mu}_k$'s are p-dimensional vectors and $\boldsymbol{\Sigma}_k$'s are p by p symmetric positive definite matrices. Define $\boldsymbol{\Sigma} = \pi_0\boldsymbol{\Sigma}_0 + \pi_1\boldsymbol{\Sigma}_1$, which is the weighted average of within-class covariances. Note that $\boldsymbol{\Sigma}$ is the common within-class covariance for homoscedastic data, and when $\boldsymbol{\Sigma}_0 \neq \boldsymbol{\Sigma}_1$, it is the expectation of estimated within-class covariance under a misspecified homoscedastic model. We assume $\pi_0 = 1/2$ in this paper for easy presentation.

For a heteroscedastic Gaussian model with known parameters, the QDA rule is optimal in a sense that it minimizes the classification error for any $\mathbf{X} \in \mathbb{R}^p$. It labels an observation $\mathbf{X} = \mathbf{x}^*$ by class 1 when

$$\mathbf{x}^{*\top}\left(\boldsymbol{\Sigma}_0^{-1} - \boldsymbol{\Sigma}_1^{-1}\right)\mathbf{x}^* - 2\mathbf{x}^{*\top}\left(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 - \boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1\right) + \boldsymbol{\mu}_0^\top\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1^\top\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 + \ln\left(|\boldsymbol{\Sigma}_0|/|\boldsymbol{\Sigma}_1|\right) > 0. \tag{1}$$

The discriminant boundary of the QDA rule is a quadratic hypersurface, determined by $p(p + 3)/2$ parameters. In practice, when p is moderate or large, it is difficult to estimate the quadratic boundary accurately due to the large parameter space. While dimension reduction is a plausible approach to reduce the number of parameters, it is impossible to reduce the sample space without loss of classification power for general covariances $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\Sigma}_1$. In contrast, LDA assumes equal covariance $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_1$, under which the QDA rule (1) reduces to the LDA rule

$$-\mathbf{x}^{*\top}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1) + \frac{1}{2}(\boldsymbol{\mu}_0 + \boldsymbol{\mu}_1)^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1) > 0. \tag{2}$$

In this special case, the optimal discriminant boundary is a hyperplane with the normal vector

$$\boldsymbol{\beta} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1). \tag{3}$$

The classification error of the optimal rule is

$$\Phi\left(-\sqrt{(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)}/2\right) = \Phi\left(-\sqrt{\boldsymbol{\beta}^\top\boldsymbol{\Sigma}\boldsymbol{\beta}}/2\right),$$

where $\Phi(\cdot)$ is the cumulative distribution function (CDF) of a standard normal random variable. Note that for any nonzero vector $\boldsymbol{\alpha}$, the distribution of $\boldsymbol{\alpha}^\top\mathbf{X}|\{Y = k\}$ is $\mathcal{N}(\boldsymbol{\alpha}^\top\boldsymbol{\mu}_k, \boldsymbol{\alpha}^\top\boldsymbol{\Sigma}\boldsymbol{\alpha})$. It is straightforward to derive that the LDA rule in the direction $\boldsymbol{\alpha}$ can achieve classification error $\Phi(-|\boldsymbol{\alpha}^\top(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)|/(2\sqrt{\boldsymbol{\alpha}^\top\boldsymbol{\Sigma}\boldsymbol{\alpha}}))$, with a minimal value $\Phi(-\sqrt{\boldsymbol{\beta}^\top\boldsymbol{\Sigma}\boldsymbol{\beta}}/2)$ when $\boldsymbol{\alpha} = c\boldsymbol{\beta}$ for any $c \neq 0$. In summary, the direction of $\boldsymbol{\beta}$ is the optimal direction to project the data to achieve the best classification accuracy. More importantly, we will not lose any classification power after dimension reduction to this 1D subspace. This is one of the reasons that the LDA-based approach is more popular than QDA in data analysis. For the downside, LDA is suboptimal when the data is heteroscedastic. First of all, the LDA direction, calculated by the same formula $\boldsymbol{\beta} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)$, with $\boldsymbol{\Sigma} = \pi_0\boldsymbol{\Sigma}_0 + \pi_1\boldsymbol{\Sigma}_1$, is not the best direction to project the data onto. In a special case when $\boldsymbol{\mu}_0 = \boldsymbol{\mu}_1$ and $\boldsymbol{\Sigma}_0 = \mathbf{I}$ and $\boldsymbol{\Sigma}_1 = \mathrm{diag}\{2, 1, \ldots, 1\}$, the direction $\mathbf{e}_1 = (1, 0, \ldots, 0)^\top$ is the best, but $\boldsymbol{\beta} = \mathbf{0}$. An estimator to $\boldsymbol{\beta}$ would give a random and uninformative direction. Second, even if the best direction is known, the 1D LDA is outperformed by the 1D QDA after the projection, when the marginal variances are not equal. While the second issue is minor and easy to fix, in this paper, we aim to define and estimate the optimal 1D projection for heteroscedastic Gaussian data.

For a classification rule $\Psi : \mathbb{R}^p \to \{0, 1\}$, the classification error is defined by $\Pr(\Psi(\mathbf{X}) \neq Y)$. Let E_0 be the classification error of the QDA rule defined in (1), and E_{LDA} be the classification error of the LDA rule (2) with $\boldsymbol{\Sigma} = \pi_0\boldsymbol{\Sigma}_0 + \pi_1\boldsymbol{\Sigma}_1$ under

heteroscedasticity. Under a projection of $\mathbf{X}$ to a 1D subspace spanned by $\boldsymbol{\alpha}$, define $E(\boldsymbol{\alpha})$ and $E_{LDA}(\boldsymbol{\alpha})$ by the classification errors of the QDA and LDA rules for the projected data. Then we have

$$E_0 \le \min_{\boldsymbol{\alpha}\neq\mathbf{0}} E(\boldsymbol{\alpha}) \le \min_{\boldsymbol{\alpha}\neq\mathbf{0}} E_{LDA}(\boldsymbol{\alpha}) \le E_{LDA}. \tag{4}$$

The equal signs in (4) hold in the special case when $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_1$. In general cases, it is impossible to approach E_0 empirically if $p^2 > n$. Nevertheless, it is easier to estimate the direction that minimizes $E(\boldsymbol{\alpha})$. We show an explicit formula for $E(\boldsymbol{\alpha})$ in the following theorem.

Theorem 1. *Assume* $\mathbf{X}|\{Y=k\} \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, $\pi_k = P(Y=k) = 1/2$, $k = 0, 1$. *Let* $m_k = \boldsymbol{\alpha}^\top\boldsymbol{\mu}_k$, $\sigma_k^2 = \boldsymbol{\alpha}^\top\boldsymbol{\Sigma}_k\boldsymbol{\alpha}$, *where* $k = 0, 1$, $\boldsymbol{\alpha} \in \mathbb{R}^p \setminus \{\mathbf{0}\}$. *Then the classification error function for 1D QDA* $E : \mathbb{R}^p \setminus \{\mathbf{0}\} \to \mathbb{R}$ *in* (4) *satisfies*

$$E(\boldsymbol{\alpha}) = \begin{cases} \Phi\Big(-\dfrac{|m_0 - m_1|}{2\sigma}\Big), & \sigma_0 = \sigma_1 := \sigma, \\ \dfrac{1}{2} + \dfrac{1}{2}\Phi\Big(\dfrac{\sigma_1(m_1-m_0)-\sigma_0\sqrt{\Delta}}{\sigma_0^2-\sigma_1^2}\Big) - \dfrac{1}{2}\Phi\Big(\dfrac{\sigma_1(m_1-m_0)+\sigma_0\sqrt{\Delta}}{\sigma_0^2-\sigma_1^2}\Big) & \\ \quad + \dfrac{1}{2}\Phi\Big(\dfrac{\sigma_0(m_1-m_0)+\sigma_1\sqrt{\Delta}}{\sigma_0^2-\sigma_1^2}\Big) - \dfrac{1}{2}\Phi\Big(\dfrac{\sigma_0(m_1-m_0)-\sigma_1\sqrt{\Delta}}{\sigma_0^2-\sigma_1^2}\Big), & \sigma_0 \neq \sigma_1, \end{cases} \tag{5}$$

where $\Delta = (m_0 - m_1)^2 + (\sigma_0^2 - \sigma_1^2)\ln(\sigma_0^2/\sigma_1^2)$.

We define a direction $\boldsymbol{\alpha}_0 \in \operatorname{argmin}_{\boldsymbol{\alpha}\neq\mathbf{0}} E(\boldsymbol{\alpha})$ an optimal direction for 1D QDA. The following proposition summarizes two well-known special cases when close-form solutions can be derived to minimize $E(\boldsymbol{\alpha})$.

Proposition 1. *Under the assumptions in Theorem 1, the following results hold.*

(i) *If* $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}$,

$$\boldsymbol{\alpha}_0 = \operatorname*{argmin}_{\boldsymbol{\alpha}\neq\mathbf{0}} E(\boldsymbol{\alpha}) = \operatorname*{argmax}_{\boldsymbol{\alpha}\neq\mathbf{0}} \Big(\frac{\boldsymbol{\alpha}^\top(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^\top\boldsymbol{\alpha}}{\boldsymbol{\alpha}^\top\boldsymbol{\Sigma}\boldsymbol{\alpha}}\Big) = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0). \tag{6}$$

(ii) *If* $\boldsymbol{\mu}_0 = \boldsymbol{\mu}_1$,

$$\boldsymbol{\alpha}_0 = \operatorname*{argmin}_{\boldsymbol{\alpha}\neq\mathbf{0}} E(\boldsymbol{\alpha}) = \operatorname*{argmax}_{\boldsymbol{\alpha}\neq\mathbf{0}} \left(\max\{\frac{\boldsymbol{\alpha}^\top\boldsymbol{\Sigma}_1\boldsymbol{\alpha}}{\boldsymbol{\alpha}^\top\boldsymbol{\Sigma}_0\boldsymbol{\alpha}}, \frac{\boldsymbol{\alpha}^\top\boldsymbol{\Sigma}_0\boldsymbol{\alpha}}{\boldsymbol{\alpha}^\top\boldsymbol{\Sigma}_1\boldsymbol{\alpha}}\}\right). \tag{7}$$

If there is a unique maximum among all the eigenvalues of $\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\Sigma}_1$ *and* $\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_0$, *then* $\boldsymbol{\alpha}_0$ *is the eigenvector corresponding to the greatest eigenvalue.*

When the number of features is moderate, e.g., $p < n < p^2$, the standard QDA is not stable empirically. As an alternative approach, we attempt to estimate the best 1D subspace for dimension reduction before conducting QDA. Intuitively, this approach is more robust than the standard QDA because much fewer parameters are needed. In particular, it requires $p-1$ parameters for direction estimation and two more parameters for the quadratic boundary after projection. Thus the total number of parameters is similar to that of LDA. As a result, our method performs similarly to LDA for homoscedastic data, and it is more sensitive to data heteroscedasticity than the LDA approach.

The LDA direction in (6) is well-defined and unique up to a scalar when $\boldsymbol{\mu}_0 \neq \boldsymbol{\mu}_1$. However, the optimal direction to 1D QDA might not be unique, especially when some symmetric structure occurs in the model. For example, in (7), if $\boldsymbol{\Sigma}_0 = c_0\mathbf{I}$ and $\boldsymbol{\Sigma}_1 = c_1\mathbf{I}$ with $c_0 \neq c_1$, every direction is an optimal direction because of symmetry. In general, the optimal direction would be unique up to a scalar although it is difficult to specify the exact conditions on uniqueness.

2.2. Method and computation

Let $\{\mathbf{x}_k^i : 1 \le i \le n_k\}$ be i.i.d. observations from $\mathbf{X}|\{Y=k\}$, $k = 0, 1$. With Theorem 1, we can approximate the classification error $E(\boldsymbol{\alpha})$ with $\hat{E}(\boldsymbol{\alpha})$, which is equation (5) plugged in by sample means $\hat{\boldsymbol{\mu}}_0$ $\hat{\boldsymbol{\mu}}_1$ and sample covariance matrices $\hat{\boldsymbol{\Sigma}}_0$ $\hat{\boldsymbol{\Sigma}}_1$. We then find the minimizing direction $\hat{\boldsymbol{\alpha}}_0$ of $\hat{E}(\boldsymbol{\alpha})$. After projecting all the training data and test data to $\hat{\boldsymbol{\alpha}}_0$, predictions are made based on the 1D QDA rule of the projected data. We call this procedure QDA by projection (QDAP), which is summarized in Algorithm 1.

Algorithm 1 QDA by Projection (QDAP)

1: $\hat{\boldsymbol{\alpha}}_0 \leftarrow \operatorname{argmin}_{\boldsymbol{\alpha}\neq\mathbf{0}} \hat{E}(\boldsymbol{\alpha})$
2: $x_k^i \leftarrow \hat{\boldsymbol{\alpha}}_0^\top\mathbf{x}_k^i$ for $1 \le i \le n_k$, $k = 0, 1$
3: $\hat{\phi}(x) \leftarrow$ 1D QDA rule derived with projected data $\{x_k^i : 1 \le i \le n_k, k = 0, 1\}$
4: **return** $\hat{y} \leftarrow \hat{\phi}(x)$ for any new observation $\mathbf{x}$, where $x = \hat{\boldsymbol{\alpha}}_0^\top\mathbf{x}$

By Corollaries 2 and 3 in the appendix, $\hat{E}(\boldsymbol{\alpha})$ is smooth almost everywhere, and it is uniformly continuous when viewed as a function defined on the unit sphere. Thus the existence of minimizer is guaranteed by the compactness of the unit sphere. We implemented a coordinate descent algorithm to conduct the optimization. Proposition 1 provides two good initial directions, i.e., (6) and (7) to warm start the coordinate descent algorithm. The implementation details are illustrated in Appendix B.

3. Theoretical properties

Proposition 1 shows that LDA is a special case of our method in the population level. Thanks to the explicit formula (3), it is straightforward to see that the LDA direction can be consistently estimated. The following theorem shows a counterpart result for the 1D QDA. As a by-product, it implies our method performs similar to LDA under the equal covariance assumption.

Theorem 2. *Assume that* $\mathbf{X}|\{Y=k\} \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, $k=0,1$. *Let* $\{\mathbf{x}_k^i : i \geq 1\}$ *be a sequence of i.i.d. observations from* $\mathbf{X}|\{Y=k\}$, $\hat{\boldsymbol{\mu}}_0^n, \hat{\boldsymbol{\mu}}_1^n, \hat{\boldsymbol{\Sigma}}_0^n, \hat{\boldsymbol{\Sigma}}_1^n$ *be sample means and sample covariance matrices calculated with first n observations in each class, and* $\hat{E}^n(\boldsymbol{\alpha})$ *be the empirical classification error, i.e., (5) with the previous estimates plugged in. If* $E(\boldsymbol{\alpha})$ *has a unique minimizer* $\boldsymbol{\alpha}_0 = \operatorname{argmin}_{\boldsymbol{\alpha}\in\mathbb{P}^{p-1}} E(\boldsymbol{\alpha})$ *and assume that* $\hat{\boldsymbol{\alpha}}_0^n \in \operatorname{argmin}_{\boldsymbol{\alpha}\in\mathbb{P}^{p-1}} \hat{E}^n(\boldsymbol{\alpha})$, *then*

$$\hat{\boldsymbol{\alpha}}_0^n \xrightarrow{a.s.} \boldsymbol{\alpha}_0 \quad \text{as } n \to \infty.$$

Since the classification error function E depends only on the direction of vectors in $\mathbb{R}^p \setminus \{\mathbf{0}\}$, it is essentially a function defined on the $p-1$ dimensional real projective space $\mathbb{P}^{p-1}$, which consists of all one dimensional subspaces of $\mathbb{R}^p$ (see Corollaries 3 and 4 in appendix for details). Practically, we may simply view $\boldsymbol{\alpha}$ as a unit vector up to a sign. To make the theorem mathematically rigorous, we use $\mathbb{P}^{p-1}$ as the domain of $\boldsymbol{\alpha}$. It is standard in mathematics to denote the one dimensional subspace spanned by a vector $\boldsymbol{\alpha}$ by equivalent class $[\boldsymbol{\alpha}]$. But we will omit the brackets for easy presentation whenever there is no ambiguity.

LDA and QDA share an invariance property, which ensures that the classification result is unaffected by any invertible affine transformation of the data. To elaborate, if we apply the same nonsingular linear transformation to the training data and future test data, the prediction results of LDA and QDA will not change. The following proposition indicates that the invariance property also holds for our method.

Proposition 2. *For* $k=0,1$, *let* $\{\mathbf{x}_k^i : 1 \leq i \leq n_k\}$ *be i.i.d observations from* $\mathbf{X}|\{Y=k\}$, *and* $\tilde{\mathbf{x}}_k^i = \mathbf{b} + \mathbf{A}\mathbf{x}_k^i$, *where* $\mathbf{b} \in \mathbb{R}^p$, $\mathbf{A}$ *is a p by p full rank matrix. Let* $\hat{\boldsymbol{\alpha}}_0$ ($\hat{\tilde{\boldsymbol{\alpha}}}_0$) *be the unique (up to a scalar) minimizer in step 1 of Algorithm 1, with* $\hat{E}$ ($\hat{\tilde{E}}$) *derived from training data* $\{\mathbf{x}_k^i\}$ ($\{\tilde{\mathbf{x}}_k^i\}$). *Then the following equation holds:*

$$\hat{\tilde{\boldsymbol{\alpha}}}_0 = c\left(\mathbf{A}^\top\right)^{-1}\hat{\boldsymbol{\alpha}}_0,$$

where c is a nonzero constant.

This implies $\tilde{x}_k^i = \hat{\tilde{\boldsymbol{\alpha}}}_0^\top \tilde{\mathbf{x}}_k^i = c\hat{\boldsymbol{\alpha}}_0^\top \mathbf{A}^{-1}\mathbf{b} + c\hat{\boldsymbol{\alpha}}_0^\top \mathbf{x}_k^i = c\hat{\boldsymbol{\alpha}}_0^\top \mathbf{A}^{-1}\mathbf{b} + cx_k^i$, where x_k^i is the projected data defined in Algorithm 1, step 2. That is, the projected data before and after transformation, x_k^i and $\tilde{x}_k^i$ are up to an affine transformation. It implies

Corollary 1. *Algorithm 1 is invariant under invertible affine transformations.*

Here is a remark on the Gaussian assumption before we move on to the numerical studies. The formulation (5) of the classification error of QDA with respect to direction $\boldsymbol{\alpha}$ relies on the Gaussian distribution. As a consequence, the definition of the optimal projection, $\boldsymbol{\alpha}_0$, depends on the Gaussian assumption. Without the Gaussian assumption, the direction $\boldsymbol{\alpha}_0$ is still defined as the minimizer of (5), although it might not the be the optimal projection in the sense of minimizing expected classification error. This is analogous to the story for LDA. Without the Gaussian assumption, LDA still works and is consistent to its population version, although the population version of LDA is not the Bayesian or optimal rule any more. In our case, the main theoretical results, i.e., consistency (Theorem 2) and invariance (Proposition 2) still hold without the Gaussian assumption.

4. Numerical studies

4.1. Method for comparison

In this section, we compare our method, Algorithm 1 (QDAP), with LDA, DSDA [18], QDA, DAP [9], and RDA [10] by both simulated and real data examples. Besides the classical methods LDA and QDA, RDA is a well known regularization approach which works well for moderate and high dimensional data. DSDA and DAP are two representatives of modern high dimensional classification tools. For DSDA, DAP and RDA, we used the `R` packages provided by the authors with default settings. For LDA and QDA, we used functions from `R` recommended package `MASS`. In simulated data examples, the oracle method that employs the true model for prediction is included for comparison as a benchmark.

Table 1
Average classification errors of QDAP (proposed method) and other classification methods in percentage (with standard errors in parentheses) for Model 1 over 100 replications. For each replication, different sizes of training sets are generated, with $n/2$ samples in each class. A test set with 500 observations in each class is used for calculating classification errors.

n	LDA	QDA	RDA	DSDA	DAP	QDAP	Oracle
200	17.41 (0.18)	35.79 (0.25)	14.22 (0.16)	17.59 (0.17)	19.04 (0.17)	17.46 (0.18)	11.89 (0.10)
300	15.37 (0.14)	31.37 (0.20)	13.41 (0.13)	15.68 (0.14)	16.74 (0.14)	15.42 (0.14)	11.93 (0.10)
400	14.63 (0.13)	28.79 (0.19)	13.08 (0.10)	15.00 (0.13)	15.95 (0.15)	14.65 (0.13)	11.86 (0.11)
500	14.07 (0.11)	26.39 (0.18)	12.84 (0.09)	14.29 (0.12)	15.04 (0.12)	14.06 (0.11)	11.72 (0.10)
600	13.64 (0.12)	24.63 (0.17)	12.74 (0.11)	13.92 (0.12)	14.52 (0.13)	13.67 (0.12)	11.90 (0.11)

Table 2
Average classification errors of QDAP (proposed method) and other classification methods in percentage (with standard errors in parentheses) for Model 2 over 100 replications. For each replication, different sizes of training sets are generated, with $n/2$ samples in each class. A test set with 500 observations in each class is used for calculating classification errors.

n	LDA	QDA	RDA	DSDA	DAP	QDAP	Oracle
200	9.11 (0.14)	26.61 (0.27)	9.49 (0.16)	9.66 (0.17)	26.52 (0.44)	9.24 (0.14)	5.31 (0.07)
300	7.67 (0.10)	20.22 (0.22)	7.85 (0.10)	8.04 (0.11)	20.06 (0.35)	7.69 (0.10)	5.27 (0.06)
400	6.98 (0.09)	16.85 (0.17)	7.18 (0.11)	7.28 (0.10)	17.30 (0.32)	6.99 (0.09)	5.30 (0.07)
500	6.59 (0.08)	14.91 (0.15)	6.71 (0.07)	6.80 (0.08)	15.29 (0.26)	6.57 (0.08)	5.24 (0.06)
600	6.30 (0.08)	13.47 (0.14)	6.41 (0.08)	6.53 (0.08)	13.72 (0.22)	6.32 (0.08)	5.32 (0.06)

Table 3
Average classification errors of QDAP (proposed method) and other classification methods in percentage (with standard errors in parentheses) for Model 3 over 100 replications. For each replication, different sizes of training sets are generated, with $n/2$ samples in each class. A test set with 500 observations in each class is used for calculating classification errors.

n	LDA	QDA	RDA	DSDA	DAP	QDAP	Oracle
200	36.92 (0.26)	28.38 (0.23)	18.12 (0.15)	24.66 (0.27)	15.20 (0.18)	17.16 (0.26)	7.94 (0.08)
300	35.02 (0.25)	25.29 (0.20)	18.23 (0.13)	23.29 (0.24)	12.34 (0.16)	11.76 (0.16)	8.10 (0.09)
400	33.14 (0.28)	23.62 (0.19)	18.27 (0.13)	22.70 (0.21)	11.31 (0.15)	10.41 (0.13)	8.20 (0.10)
500	31.44 (0.24)	21.80 (0.14)	18.12 (0.13)	21.92 (0.15)	10.65 (0.12)	9.63 (0.09)	8.21 (0.09)
600	30.60 (0.24)	20.43 (0.14)	18.16 (0.13)	21.41 (0.16)	10.02 (0.10)	9.18 (0.08)	8.06 (0.07)

4.2. Simulated data

We illustrate seven data generation settings as follows. In the first five models, the data are generated from Gaussian distributions with parameters specified below.

- Model 1: $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_1 = \mathbf{I}_p$. $\boldsymbol{\mu}_0 = \mathbf{0}_p$, $\boldsymbol{\mu}_1 = \mathbf{1}_p/3$.
- Model 2: $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_1 = \mathbf{B}^\top\mathbf{B} + \mathrm{diag}(\mathbf{v})$, where $\mathbf{B}$ is a $p \times p$ matrix with IID entries from $\mathcal{N}(0, 1)$ distribution, and $\mathbf{v}$ is a $p \times 1$ vector with IID entries from $\mathcal{U}(0, 1)$ distribution. $\boldsymbol{\mu}_0 = \mathbf{0}_p$, $\boldsymbol{\mu}_1 = \mathbf{1}_p$.
- Model 3: $\boldsymbol{\Sigma}_0 = \mathbf{I}_p$, $\boldsymbol{\Sigma}_1 = (\sigma_{ij})$, where $\sigma_{ii} = 3$ and $\sigma_{ij} = 2$ for $i \neq j$. $\boldsymbol{\mu}_0 = \mathbf{0}_p$, $\boldsymbol{\mu}_1 = \mathbf{1}_p$.
- Model 4: Same settings as Model 3 except that $\boldsymbol{\mu}_1 = \mathbf{0}_p$.
- Model 5: Same settings as Model 3 except that $\boldsymbol{\Sigma}_0 = \mathrm{diag}(10, \mathbf{1}_{p-1})$, and $\boldsymbol{\mu}_1$ has IID entries from $\mathcal{N}(0, 1/p)$ distribution.

In the next two models, the data are from multivariate t-distributions with 3 degrees of freedom $t_3(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, [1].

- Model 6: Same $\boldsymbol{\Sigma}_k$'s and $\boldsymbol{\mu}_k$'s as Model 2.
- Model 7: Same $\boldsymbol{\Sigma}_k$'s and $\boldsymbol{\mu}_k$'s as Model 5.

The number of features is set to $p = 50$. In each model, sample sizes are set to $n \in \{200, 300, 400, 500, 600\}$ for training, with $n/2$ samples in each class. A test set with 500 observations in each class is used for calculating classification errors.

In Tables 1–7, we report the average classification errors (in percentage) with standard errors, based on 100 replicates for each scenario. In Models 2, 5, 6 and 7, the model parameters are generated once, and all replicates are independently generated from the same model.

For Models 1 and 2, the LDA assumption of equal covariance matrices is satisfied. LDA performs well, and our method performs similarly to LDA. RDA performs better than LDA for Model 1, due to the diagonal covariance structure. For Models 3 and 4, the data are heteroscedastic, and there is only one useful direction for classification. As a result, our method (QDAP) performs the best. The LDA-based methods performs much worse due to the unequal covariance structure. The

Table 4
Average classification errors of QDAP (proposed method) and other classification methods in percentage (with standard errors in parentheses) for Model 4 over 100 replications. For each replication, different sizes of training sets are generated, with $n/2$ samples in each class. A test set with 500 observations in each class is used for calculating classification errors.

n	LDA	QDA	RDA	DSDA	DAP	QDAP	Oracle
200	49.88 (0.16)	30.59 (0.22)	46.44 (0.48)	49.61 (0.17)	25.02 (0.77)	19.53 (0.26)	10.10 (0.08)
300	50.42 (0.15)	27.56 (0.20)	46.33 (0.44)	49.80 (0.17)	20.39 (0.80)	13.93 (0.16)	9.91 (0.08)
400	50.17 (0.18)	25.82 (0.16)	46.23 (0.43)	49.88 (0.18)	18.45 (0.72)	12.41 (0.12)	9.93 (0.09)
500	49.95 (0.17)	24.23 (0.17)	47.41 (0.36)	49.67 (0.15)	18.02 (0.84)	11.71 (0.11)	10.16 (0.10)
600	50.09 (0.16)	23.03 (0.13)	47.73 (0.32)	50.04 (0.15)	19.90 (1.15)	11.18 (0.10)	9.96 (0.09)

Table 5
Average classification errors of QDAP (proposed method) and other classification methods in percentage (with standard errors in parentheses) for Model 5 over 100 replications. For each replication, different sizes of training sets are generated, with $n/2$ samples in each class. A test set with 500 observations in each class is used for calculating classification errors.

n	LDA	QDA	RDA	DSDA	DAP	QDAP	Oracle
200	35.82 (0.22)	22.36 (0.19)	35.61 (0.27)	35.98 (0.26)	23.07 (0.62)	19.72 (0.25)	7.31 (0.08)
300	34.69 (0.22)	18.69 (0.15)	34.44 (0.24)	34.83 (0.25)	20.33 (0.66)	14.03 (0.17)	7.29 (0.08)
400	32.98 (0.17)	16.82 (0.13)	32.81 (0.19)	33.21 (0.19)	18.63 (0.65)	12.53 (0.12)	7.14 (0.08)
500	32.65 (0.16)	15.58 (0.13)	32.39 (0.17)	32.39 (0.16)	19.06 (0.62)	11.89 (0.11)	7.42 (0.08)
600	32.00 (0.13)	14.69 (0.11)	31.79 (0.15)	32.07 (0.16)	19.57 (0.70)	11.37 (0.10)	7.25 (0.07)

Table 6
Average classification errors of QDAP (proposed method) and other classification methods in percentage (with standard errors in parentheses) for Model 6 over 100 replications. For each replication, different sizes of training sets are generated, with $n/2$ samples in each class. A test set with 500 observations in each class is used for calculating classification errors.

n	LDA	QDA	RDA	DSDA	DAP	QDAP	Oracle
200	8.69 (0.12)	23.54 (0.26)	9.07 (0.15)	9.09 (0.15)	24.06 (0.37)	8.78 (0.12)	5.22 (0.07)
300	7.40 (0.10)	19.23 (0.19)	7.67 (0.11)	7.72 (0.11)	18.36 (0.29)	7.45 (0.10)	5.09 (0.06)
400	6.86 (0.09)	16.44 (0.14)	7.03 (0.10)	7.16 (0.11)	14.81 (0.26)	6.90 (0.09)	5.06 (0.07)
500	6.25 (0.08)	14.78 (0.16)	6.49 (0.08)	6.46 (0.09)	13.43 (0.24)	6.28 (0.08)	4.96 (0.07)
600	6.23 (0.08)	13.57 (0.14)	6.37 (0.09)	6.37 (0.09)	11.68 (0.19)	6.23 (0.08)	5.13 (0.06)

Table 7
Average classification errors of QDAP (proposed method) and other classification methods in percentage (with standard errors in parentheses) for Model 7 over 100 replications. For each replication, different sizes of training sets are generated, with $n/2$ samples in each class. A test set with 500 observations in each class is used for calculating classification errors.

n	LDA	QDA	RDA	DSDA	DAP	QDAP	Oracle
200	32.36 (0.20)	22.57 (0.24)	31.60 (0.25)	32.08 (0.24)	23.72 (0.55)	22.54 (0.26)	6.27 (0.09)
300	30.48 (0.18)	19.62 (0.20)	29.79 (0.18)	30.19 (0.19)	22.05 (0.48)	17.72 (0.16)	6.28 (0.08)
400	29.57 (0.17)	17.74 (0.20)	28.97 (0.19)	29.50 (0.18)	21.30 (0.45)	16.08 (0.14)	6.36 (0.07)
500	28.52 (0.16)	16.48 (0.18)	27.97 (0.15)	28.19 (0.16)	21.10 (0.44)	15.43 (0.13)	6.22 (0.08)
600	28.08 (0.14)	16.03 (0.21)	27.52 (0.14)	27.99 (0.14)	20.87 (0.45)	14.77 (0.11)	6.18 (0.07)

standard QDA suffers from small sample sizes. DAP method performs reasonably well and ranks in the second place. Model 5 represents a more general heteroscedastic setting. In this case, our method is suboptimal to QDA if the sample size is big enough. However, our method could outperform QDA when the sample size is moderate, due to the bias–variance trade-off. As a result, our method performs best in Table 5 for all sample sizes in the given range. To demonstrate the robustness of our algorithm for non-Gaussian data, we consider Models 6 and 7, which are similar to Models 2 and 5 except that multivariate t distributions with 3 degrees of freedom are employed. In Model 6, LDA performs the best, while our method performs similarly to LDA in terms of both classification error and its standard error. For Model 7, three QDA-based methods are better than LDA-based methods and our method achieves the best accuracy. We conclude from these two examples that our method is similar to LDA and other methods in terms of robustness to heavy-tailed data.

4.3. Real data

In this subsection, five real data sets are used to compare these classification methods. In each real data experiment, we randomly assigned 60% of the observations into the training set and the rest into the test set. We randomly split each real data set 300 times, and calculated average classification error along with its standard error.

Table 8
Average classification errors of QDAP (proposed method) and other classification methods in percentage (with standard errors in parentheses) over 300 replications. Data set 1: Breast cancer Wisconsin data set. Data set 2: Ultrasonic flowmeter diagnostics data set. Data set 3: Heart disease data set. Data set 4: Image segmentation data set. Data set 5: Satellite data set.

	LDA	QDA	RDA	DSDA	DAP	QDAP
Data set 1	4.62 (0.06)	5.02 (0.07)	4.23 (0.06)	4.87 (0.06)	4.24 (0.06)	3.30 (0.04)
Data set 2	1.58 (0.11)	NA	34.05 (0.38)	2.94 (0.26)	15.52 (0.41)	0.89 (0.08)
Data set 3	17.81 (0.17)	20.86 (0.18)	17.56 (0.18)	18.00 (0.17)	18.43 (0.19)	17.48 (0.17)
Data set 4	0.72 (0.02)	NA	0.78 (0.03)	0.84 (0.03)	1.64 (0.04)	0.69 (0.02)
Data set 5	1.37 (0.02)	1.79 (0.03)	1.38 (0.02)	1.39 (0.02)	1.54 (0.02)	1.32 (0.02)

4.3.1. Breast cancer Wisconsin data set

The breast cancer data set, created by Dr. WIlliam H. Wolberg [21], is available on the UCI Machine Learning Repository [6]. There are $n = 699$ instances of patients from Dr. Wolberg's clinical cases. 10 features are recorded for each patient, $p = 9$ of which are the explanatory variables. The 10th feature assigns the patients into two classes — "benign" and "malignant".

4.3.2. Ultrasonic flowmeter diagnostics data set

This data set, provided by Gyamfi et al. [11], is available on the UCI Machine Learning Repository [6]. The goal of this data set is to predict the health status of some flowmeters installed at UK using diagnostic data. There are $n = 87$ instances of diagnosed flowmeters and the diagnostic data comes in $p = 36$ dimensions. Two classes are either "Healthy" or "Installation effects".

4.3.3. Heart disease data set

This data set, provided by Andras Janosi, William Steinbrunn, Matthias Pfisterer and Robert Detrano, is available on the UCI Machine Learning Repository [6]. There are $n = 303$ patients in total. $p = 13$ different attributes are used to predict the patients' angiographic disease status, which could be either 0 ($< 50\%$ diameter narrowing) or 1 ($> 50\%$ diameter narrowing).

4.3.4. Image segmentation data

This data set, created by Vision Group, University of Massachusetts, is available on the UCI Machine Learning Repository [6]. There are 2310 total images in 7 different classes, with 330 images each. To make this a binary classification problem, we only include class 1 (brickface) and 4 (cement) for analysis. There are 19 features in total. Features 1, 3, 4, 5 are almost constants within the chosen classes, so they were removed from the data, leaving $p = 15$ features for classification.

4.3.5. Satellite data set

This data set, provided by Ashwin Srinivasan, is available on the UCI Machine Learning Repository [6]. Satellite images are labeled into 9 classes. Only class 1 (red soil) and class 3 (gray soil) are considered for our analysis, where there are 1072 images in class 1 and 961 images in class 3. $p = 36$ attributes (9 pixels times 4 spectral bands) are used for classification.

4.3.6. Results

Average classification errors (in percentage) for these experiments are summarized in Table 8. LDA performs reasonably well for all data sets, but our method outplays LDA with a margin, especially in the first two data sets. To better understand the result, we performed classical Box's M test [2] and a modern high dimensional two-sample covariance test proposed by Cai et al. [3]. All the p values for the 5 data sets are below 2.68×10^{-8}, indicating strong evidence of heteroscedasticity. Nevertheless, the original QDA suffers from low sample sizes, and in particular, fails to work in data sets 2 and 4. As a QDA based method, our method is more versatile and gives better classification results. It outperforms both LDA and QDA. RDA performs well except in data set 2. DSDA and DAP, as representatives of sparse methods for high dimensional data, produce slightly worse results than LDA and our method. Overall, our method performs the best among the algorithms in comparison.

5. Discussion

In this work, we propose a new dimension reduction and classification method based on QDA. The empirical studies show that our algorithm performs well for data sets with moderate dimensions and unequal covariance structures. An `R` package `QDAP` implementing our algorithm is available on GitHub [22]. Note that we assume equal prior probability in this paper for easy presentation, without which all theoretical results still hold with minor modifications. Moreover, the implementation in our `R` package does not rely on this assumption.

We discuss here briefly a few related works in the literature. In particular, Gaynanova and Wang [9] propose a quadratic classification rule via linear dimension reduction called DAP, which works for high dimensional classification with unequal covariances. Roughly speaking, DAP estimates simultaneously two directions $\boldsymbol{\psi}_0 = \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\delta}$ and $\boldsymbol{\psi}_1 = \boldsymbol{\Sigma}_1^{-1}\boldsymbol{\delta}$ where $\boldsymbol{\delta} = \boldsymbol{\mu}_0 - \boldsymbol{\mu}_1$, and then employs QDA for classification after projecting the data to these two directions. Empirically, a sparse method is used for estimating $\boldsymbol{\psi}_0$ and $\boldsymbol{\psi}_1$. In the population level, the space spanned by $\boldsymbol{\psi}_0$ and $\boldsymbol{\psi}_1$ can be very different from or even orthogonal to our 1D optimal subspace spanned by $\boldsymbol{\alpha}_0 = \arg\min_{\boldsymbol{\alpha}} E(\boldsymbol{\alpha})$. In short, DAP does not aim to find such an optimal projection. An advantage of DAP is that it conducts variable selection and works for high dimensional data. It is an interesting research direction to extend our method in a sparse high dimensional setting. Some recent works [4,20] propose to ensemble classifiers on random subspaces. Instead of searching for an optimal projection, these works employ and combine a collection of classifiers on subspaces, which may perform better when a single optimal projection does not exist. In practice, an asymptotic expansion of the classification error would be helpful to decide sample sizes for training [13]. It is an interesting research direction to study such an expansion for our method. Last but not least, it is momentous to study classification with dependent observations, for example, time series data [14], spatially correlated data [16], and clipping of random field [5].

CRediT authorship contribution statement

Ruiyang Wu: Conceptualization, Methodology, Software, Formal analysis Writing – original draft. **Ning Hao:** Conceptualization, Methodology, Writing – review & editing, Supervision, Funding acquisition.

Acknowledgments

The authors are grateful to the Editor, Associate Editor and two referees for helpful comments. This work was supported by the National Science Foundation, United States Grant DMS-1722691 and CCF-1740858; and Simons Foundation, United States Grant 524432.

Appendix A. Proofs and auxiliary results

Proof of Theorem 1. Let $\Psi_{\boldsymbol{\alpha}}$ be the 1D Bayesian rule for $(\boldsymbol{\alpha}^\top\mathbf{X}, Y)$. Clearly, $\boldsymbol{\alpha}^\top\mathbf{X}|\{Y=k\} \sim \mathcal{N}(m_k, \sigma_k^2)$. We prove by 2 cases:

(i) $\sigma_0 \neq \sigma_1$. Without loss of generality, we may assume $\sigma_0 > \sigma_1$. In this case,

$$\Psi_{\boldsymbol{\alpha}}(x) = 1_{\{x:\, q(x)>0\}}(x) = 1_{(r_1, r_2)}(x),$$

where

$$q(x) = \Big(\frac{1}{\sigma_0^2} - \frac{1}{\sigma_1^2}\Big)x^2 - 2\Big(\frac{m_0}{\sigma_0^2} - \frac{m_1}{\sigma_1^2}\Big)x + \Big(\frac{m_0^2}{\sigma_0^2} - \frac{m_1^2}{\sigma_1^2}\Big) + \ln\Big(\frac{\sigma_0^2}{\sigma_1^2}\Big) > 0$$

is the 1D version of QDA rule (1), and $r_1, r_2 = ((m_1\sigma_0^2 - m_0\sigma_1^2) \pm \sigma_0\sigma_1\sqrt{\Delta})/(\sigma_0^2 - \sigma_1^2)$ with $\Delta = (m_0 - m_1)^2 + (\sigma_0^2 - \sigma_1^2)\ln(\sigma_0^2/\sigma_1^2)$ are the roots of $q(x)$.
The classification error is calculated as follows.

$$\begin{aligned}
E(\boldsymbol{\alpha}) =& \frac{1}{2}\Pr\Big(\Psi_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}^\top\mathbf{X}) = 1|Y=0\Big) + \frac{1}{2}\Pr\Big(\Psi_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}^\top\mathbf{X}) = 0|Y=1\Big)\\
=& \frac{1}{2}\Pr\Big(r_1 < \boldsymbol{\alpha}^\top\mathbf{X} < r_2|Y=0\Big) + \frac{1}{2}\Pr\Big(\boldsymbol{\alpha}^\top\mathbf{X} < r_1 \text{ or } \boldsymbol{\alpha}^\top\mathbf{X} > r_2|Y=1\Big)\\
=& \frac{1}{2}\Pr\Big(\frac{r_1 - m_0}{\sigma_0} < \frac{\boldsymbol{\alpha}^\top\mathbf{X} - m_0}{\sigma_0} < \frac{r_2 - m_0}{\sigma_0}\Big|Y=0\Big) + \frac{1}{2}\Pr\Big(\frac{\boldsymbol{\alpha}^\top\mathbf{X} - m_1}{\sigma_1} < \frac{r_1 - m_1}{\sigma_1} \text{ or } \frac{\boldsymbol{\alpha}^\top\mathbf{X} - m_1}{\sigma_1} > \frac{r_2 - m_1}{\sigma_1}\Big|Y=1\Big)\\
=& \frac{1}{2}\Phi\Big(\frac{r_2 - m_0}{\sigma_0}\Big) - \frac{1}{2}\Phi\Big(\frac{r_1 - m_0}{\sigma_0}\Big) + \frac{1}{2}\Phi\Big(\frac{r_1 - m_1}{\sigma_1}\Big) + \frac{1}{2}\Big(1 - \Phi\Big(\frac{r_2 - m_1}{\sigma_1}\Big)\Big).
\end{aligned}$$

This is exactly the expression of $E(\boldsymbol{\alpha})$ in Theorem 1 when $\sigma_0 \neq \sigma_1$.

(ii) $\sigma_0 = \sigma_1 = \sigma$. In this case $\Psi_{\boldsymbol{\alpha}}$ reduces to the 1D LDA rule. Assuming $m_0 > m_1$, $\Psi_{\boldsymbol{\alpha}}(x) = 1_{(-\infty, (m_0+m_1)/2)}(x)$. So

$$\begin{aligned}
E(\boldsymbol{\alpha}) =& \frac{1}{2}\Pr\Big(\Psi_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}^\top\mathbf{X}) = 1|Y=0\Big) + \frac{1}{2}\Pr\Big(\Psi_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}^\top\mathbf{X}) = 0|Y=1\Big)\\
=& \frac{1}{2}\Pr\Big(\boldsymbol{\alpha}^\top\mathbf{X} < \frac{m_0 + m_1}{2}\Big|Y=0\Big) + \frac{1}{2}\Pr\Big(\boldsymbol{\alpha}^\top\mathbf{X} > \frac{m_0 + m_1}{2}\Big|Y=1\Big)\\
=& \frac{1}{2}\Pr\Big(\frac{\boldsymbol{\alpha}^\top\mathbf{X} - m_0}{\sigma} < \frac{m_1 - m_0}{2\sigma}\Big|Y=0\Big) + \frac{1}{2}\Pr\Big(\frac{\boldsymbol{\alpha}^\top\mathbf{X} - m_1}{\sigma} > \frac{m_0 - m_1}{2\sigma}\Big|Y=1\Big)
\end{aligned}$$

$$=\frac{1}{2}\Phi\Big(\frac{m_1-m_0}{2\sigma}\Big)+\frac{1}{2}\Big(1-\Phi\Big(\frac{m_0-m_1}{2\sigma}\Big)\Big)=\Phi\Big(-\frac{|m_1-m_0|}{2\sigma}\Big).$$

Similarly, we can show the same formula for $m_0 < m_1$. When $m_0 = m_1$, LDA becomes random guess so $E(\boldsymbol{\alpha}) = 1/2$, which is again the same as function value $\Phi(0)$. □

Continuity and analyticity of classification error function. We present a few properties of the classification error function $E(\boldsymbol{\alpha})$ which are helpful in the proof of Theorem 2.

Assuming $r(\boldsymbol{\alpha}) = (m_0 - m_1)/\sigma_1$ and $g(\boldsymbol{\alpha}) = \sigma_0/\sigma_1$, we can rewrite the classification error $E(\boldsymbol{\alpha})$ as the composition of $\mathcal{E}:\mathbb{R}\times\mathbb{R}_{>0}\to\mathbb{R}$ and $(r(\boldsymbol{\alpha}), g(\boldsymbol{\alpha}))$, where

$$\mathcal{E}(r,g)=\begin{cases}\Phi\Big(-\frac{|r|}{2}\Big), & g=1\\ \frac{1}{2}+\frac{1}{2}\Phi\Big(\frac{r-g\sqrt{\Delta}}{g^2-1}\Big)-\frac{1}{2}\Phi\Big(\frac{r+g\sqrt{\Delta}}{g^2-1}\Big) & \\ \quad+\frac{1}{2}\Phi\Big(\frac{rg+\sqrt{\Delta}}{g^2-1}\Big)-\frac{1}{2}\Phi\Big(\frac{rg-\sqrt{\Delta}}{g^2-1}\Big), & g\neq 1, g>0\end{cases} \tag{A.1}$$

$\Delta = r^2 + (g^2-1)\ln(g^2)$.

Proposition 3. *The following properties hold for $\mathcal{E}$:*

(i) $\forall (r,g)\in\mathbb{R}\times\mathbb{R}_{>0}$, $\mathcal{E}(r,g)\in(0,1/2]$,
(ii) $\mathcal{E}$ *is continuous,*
(iii) $\mathcal{E}$ *is analytic on* $\mathbb{R}\times(\mathbb{R}_{>0}\setminus\{1\})$.

Proof.

(i) We prove this by two cases:

(a) If $g = 1$, since $0 < \Phi(-|r|/2) \le \Phi(0) = 1/2$, $\mathcal{E}(r,g) = \Phi(-|r|/2)\in(0,1/2]$.
(b) If $g \neq 1$, we can rewrite $\mathcal{E}$ as

$$\mathcal{E}(r,g)=\frac{1}{2}+\frac{1}{2}(\Phi(c_1)-\Phi(c_2))+\frac{1}{2}(\Phi(d_1)-\Phi(d_2))=\frac{1}{2}+\frac{1}{2}(\Phi(c_1)-\Phi(d_2))+\frac{1}{2}(\Phi(d_1)-\Phi(c_2)),$$

where $c_1=(r-g\sqrt{\Delta})/(g^2-1)$, $c_2=(rg-\sqrt{\Delta})/(g^2-1)$, $d_1=(rg+\sqrt{\Delta})/(g^2-1)$, $d_2=(r+g\sqrt{\Delta})/(g^2-1)$. Since $\sqrt{\Delta}=\sqrt{r^2+(g^2-1)\ln(g^2)}>\sqrt{r^2}=|r|$, we have $c_1-c_2=-(\sqrt{\Delta}+r)/(g+1)<0$, $d_1-d_2=-(\sqrt{\Delta}-r)/(g+1)<0$, which implies $\Phi(c_1)-\Phi(c_2)<0$ and $\Phi(d_1)-\Phi(d_2)<0$. Thus, $\mathcal{E}(r,g)<1/2$.
To prove $\mathcal{E}(r,g)>0$, we investigate separately for $0<g<1$ and $g>1$. When $0<g<1$, $c_1-d_2=-2g\sqrt{\Delta}/(g^2-1)>0$, so $\Phi(c_1)-\Phi(d_2)>0$, and $\mathcal{E}(r,g)>1/2+(1/2)0+(1/2)(0-1)=0$. When $g>1$, we can prove $\Phi(d_1)-\Phi(c_2)>0$ and get $\mathcal{E}(r,g)>0$ as well.
Combining these two inequalities, we have $\mathcal{E}(r,g)\in(0,1/2)$.

(ii) Let $U=\mathbb{R}\times(\mathbb{R}_{>0}\setminus\{1\})$, then $U^c=\mathbb{R}\times\{1\}$. $\mathcal{E}$ restricted on U is continuous because it is a composition of continuous functions. Similarly, $\mathcal{E}$ restricted on U^c is also continuous. Since U is an open subset of $\mathbb{R}\times\mathbb{R}_{>0}$, $\mathcal{E}$ is continuous at every point of U. Thus, we only need to prove $\mathcal{E}$ is continuous at every point of U^c.
For any $(\rho,1)\in U^c$, it suffices to show $\lim_{U\ni(r,g)\to(\rho,1)}\mathcal{E}(r,g)=\mathcal{E}(\rho,1)$. There are three cases:

(a) If $\rho=0$, then for any $(r,g)\in U$

$$\begin{aligned}\Big|\mathcal{E}(r,g)-\frac{1}{2}\Big|&=\Big|\frac{1}{2}(\Phi(c_1)-\Phi(c_2))+\frac{1}{2}(\Phi(d_1)-\Phi(d_2))\Big|\le\Big|\frac{1}{2}(\Phi(c_1)-\Phi(c_2))\Big|+\Big|\frac{1}{2}(\Phi(d_1)-\Phi(d_2))\Big|\\&\le\frac{L}{2}|c_1-c_2|+\frac{L}{2}|d_1-d_2|\end{aligned}$$

The last inequality holds because Φ is Lipschitz continuous. Since $|c_1-c_2|=|\sqrt{\Delta}+r|/(g+1)\to 0$ and $|d_1-d_2|=|\sqrt{\Delta}-r|/(g+1)\to 0$ as $(r,g)\to(0,1)$ in U, we have $\lim_{U\ni(r,g)\to(0,1)}\mathcal{E}(r,g)=1/2=\mathcal{E}(0,1)$.
(b) If $\rho>0$, as $(r,g)\to(\rho,1)$ in U,

$$|d_1-d_2|=\frac{|\sqrt{\Delta}-r|}{g+1}=\frac{(g^2-1)\ln(g^2)}{(g+1)|\sqrt{\Delta}+r|}\to 0,$$

so $\lim_{U\ni(r,g)\to(\rho,1)}|\Phi(d_1)-\Phi(d_2)|\to 0$ by Lipschitz continuity of Φ.
For any $(r,g)\in U$,

$$\Phi(c_1)=\Phi\Big(\frac{r-g\sqrt{\Delta}}{g^2-1}\Big)=\Phi\Big(\frac{r-rg}{g^2-1}+\frac{rg-g\sqrt{\Delta}}{g^2-1}\Big)=\Phi\Big(-\frac{r}{g+1}-g\frac{\Delta-r^2}{(g^2-1)\Big(\sqrt{\Delta}+r\Big)}\Big)$$

$$=\Phi\Big(-\frac{r}{g+1}-g\frac{\ln(g^2)}{\sqrt{\Delta}+r}\Big)\to\Phi\Big(-\frac{|\rho|}{2}\Big)$$

when $(r,g)\to(\rho,1)$. Similar arguments yield $\lim_{U\ni(r,g)\to(\rho,1)}\Phi(c_2)=\Phi(|\rho|/2)$. As a result,

$$\begin{aligned}\lim_{U\ni(r,g)\to(\rho,1)}\mathcal{E}(r,g)&=\lim_{U\ni(r,g)\to(\rho,1)}\frac{1}{2}(\Phi(c_1)-\Phi(c_2))+\frac{1}{2}(\Phi(d_1)-\Phi(d_2))+\frac{1}{2}\\&=\frac{1}{2}\Big(\Phi\Big(-\frac{|\rho|}{2}\Big)-\Phi\Big(\frac{|\rho|}{2}\Big)\Big)+\frac{1}{2}=\Phi\Big(-\frac{|\rho|}{2}\Big)=\mathcal{E}(\rho,1)\end{aligned}$$

(c) For $\rho<0$, by a similar argument to the last case, we have $\lim_{U\ni(r,g)\to(\rho,1)}\mathcal{E}(r,g)=\mathcal{E}(\rho,1)$.

(iii) Clearly, $\mathbb{R}\times(\mathbb{R}_{>0}\setminus\{1\})$ is an open subset of $\mathbb{R}\times\mathbb{R}_{>0}$. $\mathcal{E}$ is analytic on $\mathbb{R}\times(\mathbb{R}_{>0}\setminus\{1\})$ because it is a composition of analytic functions. □

The properties of $\mathcal{E}$ have direct implications on the properties of E. The next corollary presents a few of them.

Corollary 2. *The following results hold for $E:\mathbb{R}^p\setminus\{\mathbf{0}\}\to\mathbb{R}$:*

(i) $\forall\boldsymbol{\alpha}\neq 0, E(\boldsymbol{\alpha})\in(0,1/2]$,
(ii) *E is continuous,*
(iii) *E is analytic Lebesgue a.e.*

One important property of E is homogeneity of degree 0, i.e., $E(c\boldsymbol{\alpha})=E(\boldsymbol{\alpha})$ for any $c\neq 0$, which is easy to see by definition (5). This allows us to characterize E with function $E':\mathbb{P}^{p-1}\to\mathbb{R}$ through the factorization $E=E'\circ Q$, where E' is defined as $E'([\boldsymbol{\alpha}])=E(\boldsymbol{\alpha})$ and $Q:\mathbb{R}^p\setminus\{\mathbf{0}\}\to\mathbb{P}^{p-1}$ is the canonical projection $Q(\boldsymbol{\alpha})=[\boldsymbol{\alpha}]$.

Corollary 3. *E' is a well-defined uniformly continuous function.*

Proof. If $[\boldsymbol{\alpha}]=[\boldsymbol{\beta}]$, then $\boldsymbol{\alpha}=c\boldsymbol{\beta}$ for some $c\neq 0$. Thus, $E'([\boldsymbol{\alpha}])=E(\boldsymbol{\alpha})=E(c\boldsymbol{\beta})=E(\boldsymbol{\beta})=E'([\boldsymbol{\beta}])$. This proves E' is well-defined.

$\mathbb{P}^{p-1}$ is endowed with the quotient topology induced by Q, that is, $U\subseteq\mathbb{P}^{p-1}$ is open iff $Q^{-1}(U)\subseteq\mathbb{R}^p\setminus\{\mathbf{0}\}$ is open. For any $V\subseteq\mathbb{R}$, $E^{-1}(V)=Q^{-1}(E'^{-1}(V))$ is open since E is continuous. As a result, $E'^{-1}(V)$ must be open as well. This proves E' is continuous.

Since $\mathbb{P}^{p-1}$ is compact, we conclude E' is uniformly continuous by Heine–Cantor theorem. □

With the help of E' we can prove the following property of E:

Corollary 4. $\operatorname{argmin}_{\boldsymbol{\alpha}}E(\boldsymbol{\alpha})$ *is non-empty.*

Proof. Since $E':\mathbb{P}^{p-1}\to\mathbb{R}$ is continuous and its domain is compact, $\operatorname{argmin}E'$ is non-empty by Extreme Value Theorem.

Assume $[\boldsymbol{\beta}]\in\operatorname{argmin}E'$ and $\boldsymbol{\gamma}$ is arbitrary element of $\mathbb{R}^p\setminus\{\mathbf{0}\}$, then $E(\boldsymbol{\beta})=E'([\boldsymbol{\beta}])\leq E'([\boldsymbol{\gamma}])=E(\boldsymbol{\gamma})$. So $\boldsymbol{\beta}\in\operatorname{argmin}E$. □

The proof of Corollary 4 shows how we can translate a property of E' directly to a property of E. In practice, this is often possible. With some abuse of notation, it is beneficiary to identify E with E', and write $[\boldsymbol{\alpha}]$ just as $\boldsymbol{\alpha}$. With this in mind, we can think of E as a uniformly continuous function defined on projective space $\mathbb{P}^{p-1}$.

Some lemmas for the proof of Theorem 2. Denote by $f_n\rightrightarrows f$ if f_n is uniformly convergent to f.

Lemma 1. *Let S be a set, X, Y be metric spaces. Assume $f, f_n: S\to X$, $g: X\to Y$, $f_n\rightrightarrows f$. If g is uniformly continuous, then $g\circ f_n\rightrightarrows g\circ f$.*

Proof. Let d_X and d_Y be the metrics on X and Y respectively. For any $\epsilon>0$, there exists a $\delta>0$, such that whenever $d_X(x_1,x_2)\leq\delta$, $d_Y(g(x_1),g(x_2))\leq\epsilon$. For this δ, there exists an $N>0$, such that whenever $n\geq N$, $d_X(f(s),f_n(s))\leq\delta$ for all $s\in S$, thus $d_Y(g\circ f(s),g\circ f_n(s))\leq\epsilon$ for all $s\in S$. □

Lemma 2. *Let X, Y, Z be metric spaces. Assume $f, f_n: X\to Y$, $g: Y\to Z$, $f_n\rightrightarrows f$. If X is compact, Y is complete, f, f_n and g are all continuous, then $g\circ f_n\rightrightarrows g\circ f$.*

Proof. Let $I=f(X)\cup(\bigcup_{n=1}^{\infty}f_n(X))$, we first show I is totally bounded.

Since X is compact, f is continuous, it must also be uniformly continuous. For any $\epsilon>0$, there exists $\delta>0$, such that whenever $d_X(x,x')<\delta$, $d_Y(f(x),f(x'))<\epsilon/2$. Let $B_x(\delta)$ be open balls centered at x with radius δ, then $X\subset\bigcup_{x\in X}B_x(\delta)$. By compactness of X, X is covered by finite number of those balls, say, $X\subset\bigcup_{i=1}^{N_1}B_{x_i}(\delta)$. Because $f_n\rightrightarrows f$, there exists $N_2>0$, such that when $n>N_2$, $d_Y(f(x),f_n(x))<\epsilon/2$ for any $x\in X$.

We now claim that $f(X) \cup (\bigcup_{n=N_2+1}^{\infty} f_n(X)) \subset \bigcup_{i=1}^{N_1} B_{f(x_i)}(\epsilon)$. To see this, for any $x \in X$, there is an $i_0 \in \{1, \ldots, N_1\}$ such that $d_X(x, x_{i_0}) < \delta$, thus $d_Y(f(x), f(x_{i_0})) < \epsilon/2$. Moreover, if $n > N_2$, $d_Y(f(x), f_n(x)) < \epsilon/2$, so $d_Y(f_n(x), f(x_{i_0})) \le d_Y(f(x), f(x_{i_0})) + d_Y(f(x), f_n(x)) < \epsilon$. This proves that $f(X) \cup (\bigcup_{n=N_2+1}^{\infty} f_n(X))$ is covered by finite ϵ-balls.

$\bigcup_{n=1}^{N_2} f_n(X)$ is compact and totally bounded because it is finite union of compact sets. As a result, it can also be covered by finite ϵ-balls. Combining these two collections of ϵ-balls, we have found a finite cover of I. Thus I is totally bounded.

Since Y is complete, $\bar{I}$, the closure of I, must be complete and totally bounded, and thus compact. We can restrict g to $\bar{I}$ such that $g|_{\bar{I}}$ becomes uniformly continuous. Obviously, $g \circ f_n = g|_{\bar{I}} \circ f_n$, $g \circ f = g|_{\bar{I}} \circ f$. By Lemma 1, we have $g \circ f_n \rightrightarrows g \circ f$. □

Lemma 3. *f and $\{f_n\}_{n=1}^{\infty}$ are functions on a compact metric space X. Assume that f is continuous, and has a unique minimizer $x_* = \operatorname{argmin}_X f$. If $f_n \rightrightarrows f$, then $x_n \to x_*$, where $x_n \in \operatorname{argmin}_X f_n$.*

Proof. Suppose $x_n \not\to x_*$, then there exists an open ball B centered at x_*, and a subsequence $x_{n(m)} \subset B^c$. Since X is compact, we can further find a subsequence $x_{n(m(l))}$ and $t \in X$ such that $x_{n(m(l))} \to t$. B^c is closed, thus $t \in B^c$ and $t \neq x_*$. For any $\epsilon > 0$, there is $l_0 > 0$, such that whenever $l \ge l_0$, $|f(x) - f_{n(m(l))}(x)| \le \epsilon/2$ for all $x \in X$. So $f(x_{n(m(l))}) \le f_{n(m(l))}(x_{n(m(l))}) + \epsilon/2 \le f_{n(m(l))}(x_*) + \epsilon/2 \le f(x_*) + \epsilon$. This yields $f(t) = f(\lim_l x_{n(m(l))}) = \lim_l f(x_{n(m(l))}) \le f(x_*)$, which contradicts with the uniqueness of global minimizer of f. Thus, we can conclude $x_n \to x_*$. □

Proof of Theorem 2. By strong Law of Large Numbers, we have $\hat{\boldsymbol{\mu}}_k^n \xrightarrow{a.s.} \boldsymbol{\mu}_k$ and $\hat{\boldsymbol{\Sigma}}_k^n \xrightarrow{a.s.} \boldsymbol{\Sigma}_k$. By Egorov's theorem, for any $i \in \mathbb{N}$, there exists an event Ω_i such that $P(\Omega_i^c) < 1/i$, and $\hat{\boldsymbol{\mu}}_k^n(\omega) \rightrightarrows \boldsymbol{\mu}_k$ and $\hat{\boldsymbol{\Sigma}}_k^n(\omega) \rightrightarrows \boldsymbol{\Sigma}_k$ for $k = 0, 1$ on Ω_i, where $\mathbb{R}^p$ is equipped with Euclidean norm $\|\cdot\|_2$ and $\mathbb{R}^{p\times p}$ is equipped with Frobenius norm $\|\cdot\|_F$. Let $\lambda_k^{\min}$ be the smallest eigenvalue of $\boldsymbol{\Sigma}_k$, and $\lambda = \min\{\lambda_0^{\min}, \lambda_1^{\min}\} > 0$. There exists an integer $N > 0$ such that whenever $n \ge N$, $\|\boldsymbol{\Sigma}_k - \hat{\boldsymbol{\Sigma}}_k^n(\omega)\|_F \le \lambda/2$ for any $\omega \in \Omega_i$. From now on, we shall fix an $\omega \in \Omega_i$, and omit "ω" for easy presentation.

Consider the following subsequences $\hat{\boldsymbol{\mu}}_k^{n(l)} = \hat{\boldsymbol{\mu}}_k^{N+l}$ and $\hat{\boldsymbol{\Sigma}}_k^{n(l)} = \hat{\boldsymbol{\Sigma}}_k^{N+l}$, $k = 0, 1$. We want to show that $\hat{E}^{n(l)}(\boldsymbol{\alpha}) = \mathcal{E}((\hat{m}_0^{n(l)} - \hat{m}_1^{n(l)})/\hat{\sigma}_1^{n(l)}, \hat{\sigma}_0^{n(l)}/\hat{\sigma}_1^{n(l)})$ converges uniformly to $E(\boldsymbol{\alpha}) = \mathcal{E}((m_0 - m_1)/\sigma_1, \sigma_0/\sigma_1)$ on $\boldsymbol{\alpha} \in \mathbb{S}^{p-1}$, where $m_k(\boldsymbol{\alpha}) = \boldsymbol{\alpha}^\top \boldsymbol{\mu}_k$, $\sigma_k(\boldsymbol{\alpha}) = \sqrt{\boldsymbol{\alpha}^\top \boldsymbol{\Sigma}_k \boldsymbol{\alpha}}$, $\hat{m}_k^n(\boldsymbol{\alpha}) = \boldsymbol{\alpha}^\top \hat{\boldsymbol{\mu}}_k^n$, $\hat{\sigma}_k^n(\boldsymbol{\alpha}) = \sqrt{\boldsymbol{\alpha}^\top \hat{\boldsymbol{\Sigma}}_k^n \boldsymbol{\alpha}}$, and $\mathcal{E}$ is defined as in (A.1). We also use $\|\cdot\|_2$ to denote the matrix operator norm induced by Euclidean norm. For any $\boldsymbol{\alpha} \in \mathbb{S}^{p-1}$,

$$\left|\boldsymbol{\alpha}^\top \boldsymbol{\Sigma}_k \boldsymbol{\alpha} - \boldsymbol{\alpha}^\top \hat{\boldsymbol{\Sigma}}_k^{n(l)} \boldsymbol{\alpha}\right| \le \|\boldsymbol{\alpha}\|_2 \left\|\left(\boldsymbol{\Sigma}_k - \hat{\boldsymbol{\Sigma}}_k^{n(l)}\right)\boldsymbol{\alpha}\right\|_2 \le \|\boldsymbol{\alpha}\|_2 \left\|\boldsymbol{\Sigma}_k - \hat{\boldsymbol{\Sigma}}_k^{n(l)}\right\|_2 \|\boldsymbol{\alpha}\|_2 = \left\|\boldsymbol{\Sigma}_k - \hat{\boldsymbol{\Sigma}}_k^{n(l)}\right\|_2 \le \left\|\boldsymbol{\Sigma}_k - \hat{\boldsymbol{\Sigma}}_k^{n(l)}\right\|_F,$$

which has the following consequences:

(i) Since $\|\boldsymbol{\Sigma}_k - \hat{\boldsymbol{\Sigma}}_k^{n(l)}\|_F \le \lambda/2$, $\boldsymbol{\alpha}^\top \hat{\boldsymbol{\Sigma}}_k^{n(l)} \boldsymbol{\alpha} \ge \boldsymbol{\alpha}^\top \boldsymbol{\Sigma}_k \boldsymbol{\alpha} - \lambda/2 \ge \lambda_k^{\min} - \lambda/2 \ge \lambda/2$. This implies $\hat{\sigma}_k^n(\boldsymbol{\alpha}) \ge \sqrt{\lambda/2}$.

(ii) Since $\|\boldsymbol{\Sigma}_k - \hat{\boldsymbol{\Sigma}}_k^{n(l)}\|_F \to 0$ as $l \to \infty$, $\boldsymbol{\alpha}^\top \hat{\boldsymbol{\Sigma}}_k^{n(l)} \boldsymbol{\alpha}$ converges to $\boldsymbol{\alpha}^\top \boldsymbol{\Sigma}_k \boldsymbol{\alpha}$ uniformly. $\sqrt{\cdot}$ is uniformly continuous, so $\hat{\sigma}_k^{n(l)}(\boldsymbol{\alpha})$ converges to $\sigma_k(\boldsymbol{\alpha})$ uniformly by Lemma 1.

Similarly, we can prove $\hat{m}_k^{n(l)}(\boldsymbol{\alpha})$ converges to $m_k(\boldsymbol{\alpha})$ uniformly.

Let $f = (m_0, m_1, \sigma_0, \sigma_1)$ and $f^l = (\hat{m}_0^{n(l)}, \hat{m}_1^{n(l)}, \hat{\sigma}_0^{n(l)}, \hat{\sigma}_1^{n(l)})$ be functions from compact $\mathbb{S}^{p-1}$ to complete $\mathbb{R}^2 \times [\sqrt{\lambda/2}, \infty)^2$. We have proved that $f^l \rightrightarrows f$ as $l \to \infty$, so we can apply Lemma 2 and conclude $\hat{E}^{n(l)} \rightrightarrows E$ on $\mathbb{S}^{p-1}$. Since $\sup_{[\boldsymbol{\alpha}]\in\mathbb{P}^{p-1}} |\hat{E}^{n(l)}([\boldsymbol{\alpha}]) - E([\boldsymbol{\alpha}])| = \sup_{\boldsymbol{\alpha}\in\mathbb{S}^{p-1}} |\hat{E}^{n(l)}(\boldsymbol{\alpha}) - E(\boldsymbol{\alpha})| \to 0$, we also have $\hat{E}^{n(l)} \rightrightarrows E$ as functions on $\mathbb{P}^{p-1}$. By Lemma 3, $\hat{\boldsymbol{\alpha}}_0^{n(l)} \to \boldsymbol{\alpha}_0$ and thus $\hat{\boldsymbol{\alpha}}_0^n(\omega) \to \boldsymbol{\alpha}_0$. Recall this is true for any $\omega \in \Omega_i$ and any $i \in \mathbb{N}$, so we have

$$\hat{\boldsymbol{\alpha}}_0^n(\omega) \to \boldsymbol{\alpha}_0, \ \forall \omega \in \bigcup_{i\in\mathbb{N}} \Omega_i.$$

Clearly, $P((\bigcup_{i\in\mathbb{N}} \Omega_i)^c) = 0$. As a result, $\hat{\boldsymbol{\alpha}}_0^n \xrightarrow{a.s.} \boldsymbol{\alpha}_0$. □

Proof of Proposition 2. For $k \in \{0, 1\}$, since $\tilde{\mathbf{x}}_k^i = \mathbf{b} + \mathbf{A}\mathbf{x}_k^i$, we have $\hat{\tilde{\boldsymbol{\mu}}}_k = (1/n_k)\sum_{i=1}^{n_k} \tilde{\mathbf{x}}_k^i = \mathbf{b} + (1/n_k)\mathbf{A}\sum_{i=1}^{n_k} \mathbf{x}_k^i = \mathbf{b} + \mathbf{A}\hat{\boldsymbol{\mu}}_k$, and $\hat{\tilde{\boldsymbol{\Sigma}}}_k = (1/(n_k-1))\sum_{i=1}^{n_k} (\tilde{\mathbf{x}}_k^i - \hat{\tilde{\boldsymbol{\mu}}}_k)(\tilde{\mathbf{x}}_k^i - \hat{\tilde{\boldsymbol{\mu}}}_k)^\top = (1/(n_k-1))\sum_{i=1}^{n_k} \mathbf{A}(\mathbf{x}_k^i - \hat{\boldsymbol{\mu}}_k)(\mathbf{x}_k^i - \hat{\boldsymbol{\mu}}_k)^\top \mathbf{A}^\top = \mathbf{A}\hat{\boldsymbol{\Sigma}}_k\mathbf{A}^\top$.

Given any directions $\boldsymbol{\alpha}$ and $\tilde{\boldsymbol{\alpha}} = (\mathbf{A}^\top)^{-1}\boldsymbol{\alpha}$, for $k = 0, 1$, $\hat{\tilde{m}}_k = \tilde{\boldsymbol{\alpha}}^\top \hat{\tilde{\boldsymbol{\mu}}}_k = \boldsymbol{\alpha}^\top \mathbf{A}^{-1}(\mathbf{b} + \mathbf{A}\hat{\boldsymbol{\mu}}_k) = \boldsymbol{\alpha}^\top \mathbf{A}^{-1}\mathbf{b} + \hat{m}_k$, $\hat{\tilde{\sigma}}_k = \tilde{\boldsymbol{\alpha}}^\top \hat{\tilde{\boldsymbol{\Sigma}}}_k \tilde{\boldsymbol{\alpha}} = \boldsymbol{\alpha}^\top \mathbf{A}^{-1}(\mathbf{A}\hat{\boldsymbol{\Sigma}}_k\mathbf{A}^\top)(\mathbf{A}^\top)^{-1}\boldsymbol{\alpha} = \hat{\sigma}_k$. Thus, $\hat{\tilde{m}}_0 - \hat{\tilde{m}}_1 = \hat{m}_0 - \hat{m}_1$, and this implies that $\hat{E}(\boldsymbol{\alpha}) = \hat{\tilde{E}}(\tilde{\boldsymbol{\alpha}})$ by Eq. (5). In other words, $\hat{E}$ and $\hat{\tilde{E}}$ only differ by a nonsingular linear transformation of the domain, defined by $(\mathbf{A}^\top)^{-1}$.

By assumption, $[\hat{\boldsymbol{\alpha}}_0]$ and $[\hat{\tilde{\boldsymbol{\alpha}}}_0]$ are unique minimizers of $\hat{E}$ and $\hat{\tilde{E}}$ respectively, so we have $[\hat{\boldsymbol{\alpha}}_0] = [(\mathbf{A}^\top)^{-1}\hat{\tilde{\boldsymbol{\alpha}}}_0]$. As a result, there exists a constant $c \neq 0$ such that $\hat{\boldsymbol{\alpha}}_0 = c(\mathbf{A}^\top)^{-1}\hat{\tilde{\boldsymbol{\alpha}}}_0$.

Appendix B. Coordinate descent algorithm

Assume f is a function on $\mathbb{R}^p$. Given an initial $\mathbf{x}_0 \in \mathbb{S}^{p-1} \subset \mathbb{R}^p$, a prefixed number of maximal iterations $m > 0$ and a tolerance level $\epsilon > 0$, the coordinate descent algorithm adapted for our method is described as the following:

Algorithm 2 Coordinate Descent.

procedure MAIN($f, \mathbf{x}_0, m, \epsilon$)
  $i \leftarrow 0$
  repeat
    $\mathbf{x}_{i+1} \leftarrow$ ONE_ITER_COORDINATE_DESCENT($f, \mathbf{x}_i$)
    $i \leftarrow i + 1$
  until $i = m$ **or** $|f(\mathbf{x}_i) - f(\mathbf{x}_{i-1})| \leq \epsilon$
  return $\mathbf{x}_i$ and $f(\mathbf{x}_i)$
end procedure

procedure ONE_ITER_COORDINATE_DESCENT($g, \mathbf{y}$)
  $p \leftarrow$ length of $\mathbf{y}$
  for $j \leftarrow 1, \ldots, p$ **do**
    $g_j(*) \leftarrow g(y_1, \ldots, y_{j-1}, *, y_{j+1}, \ldots, y_p)$
    $y_j \leftarrow$ ONE_DIM_COORDINATE_DESCENT(g_j, y_j)
  end for
  $\mathbf{y} \leftarrow \mathbf{y}/\|\mathbf{y}\|_2$
  return y
end procedure

procedure ONE_DIM_COORDINATE_DESCENT(h, t)
  $h_t \leftarrow$ a quadratic approximation of h at t
  if h_t is concave up **then**
    $t \leftarrow \operatorname{argmin} h_t$
  else if h is increasing at t **then**
    $t \leftarrow t - 0.1$
  else
    $t \leftarrow t + 0.1$
  end if
  return t
end procedure

Remark 1. Empirically, the quadratic approximation h_t is not always concave up when we update each coordinate. If it is concave down, we update the coordinate by adding or subtracting a fixed step size of 0.1 to avoid saddle points.

Remark 2. It is possible that the sample covariance matrices $\hat{\Sigma}_0$, $\hat{\Sigma}_1$ are singular. We add a small scalar matrix (e.g., $10^{-7}\mathbf{I}_p$) to $\hat{\Sigma}_0$ and $\hat{\Sigma}_1$.

References

[1] T.W. Anderson, An introduction to multivariate statistical analysis, third ed., in: Wiley Series in Probability and Statistics, Wiley-Interscience, Hoboken, N.J, 2003, p. 55, 289.
[2] G.E. Box, A general distribution theory for a class of likelihood criteria, Biometrika 36 (3/4) (1949) 317–346.
[3] T. Cai, W. Liu, Y. Xia, Two-sample covariance matrix testing and support recovery in high-dimensional and sparse settings, J. Amer. Statist. Assoc. 108 (501) (2013) 265–277.
[4] T.I. Cannings, R.J. Samworth, Random-projection ensemble classification, J. R. Stat. Soc. Ser. B Stat. Methodol. 79 (4) (2017) 959–1035.
[5] V. De Oliveira, Bayeslan prediction of clipped Gaussian random fields, Comput. Statist. Data Anal. 34 (3) (2000) 299–314.
[6] D. Dua, C. Graff, UCI Machine Learning Repository, University of California, Irvine, School of Information and Computer Sciences, 2017, http://archive.ics.uci.edu/ml.
[7] R.A. Fisher, The use of multiple measurements in taxonomic problems, Ann. Eugen. 7 (2) (1936) 179–188.
[8] J.H. Friedman, Regularized discriminant analysis, J. Amer. Statist. Assoc. 84 (405) (1989) 165–175.
[9] I. Gaynanova, T. Wang, Sparse quadratic classification rules via linear dimension reduction, J. Multivariate Anal. 169 (2019) 278–299.
[10] Y. Guo, T. Hastie, R. Tibshirani, Regularized linear discriminant analysis and its application in microarrays, Biostatistics 8 (1) (2007) 86–100.
[11] K.S. Gyamfi, J. Brusey, A. Hunt, E. Gaura, Linear dimensionality reduction for classification via a sequential Bayes error minimisation with an application to flow meter diagnostics, Expert Syst. Appl. 91 (2018) 252–262.
[12] B. Jiang, X. Wang, C. Leng, A direct approach for sparse quadratic discriminant analysis, J. Mach. Learn. Res. 19 (31) (2018) 1–37.
[13] Y. Kharin, Robustness in Statistical Pattern Recognition, in: Mathematics and Its Applications, vol. 380, Springer, Dordrecht, 1996.
[14] R.T. Krafty, Discriminant analysis of time series in the presence of within-group spectral variability, J. Time Series Anal. 37 (4) (2016) 435–450.
[15] Q. Li, J. Shao, Sparse quadratic discriminant analysis for high dimensional data, Statist. Sinica 25 (2) (2015) 457–473.
[16] Y. Li, L. Zhang, T. Maiti, High dimensional classification for spatially dependent data with application to neuroimaging, Electron. J. Stat. 14 (2) (2020) 3429–3486.
[17] Q. Mai, A review of discriminant analysis in high dimensions, Wiley Interdiscip. Rev. Comput. Stat. 5 (3) (2013) 190–197.
[18] Q. Mai, H. Zou, M. Yuan, A direct approach to sparse discriminant analysis in ultra-high dimensions, Biometrika 102 (1) (2012) 33–45.
[19] Y. Qin, A review of quadratic discriminant analysis for high-dimensional data, Wiley Interdiscip. Rev. Comput. Stat. 10 (4) (2018) e1434.

[20] Y. Tian, Y. Feng, RaSE: Random subspace ensemble classification, J. Mach. Learn. Res. 22 (45) (2021) 1–93.
[21] W.H. Wolberg, O.L. Mangasarian, Multisurface method of pattern separation for medical diagnosis applied to breast cytology, Proc. Natl. Acad. Sci. 87 (23) (1990) 9193–9196.
[22] R. Wu, QDA by projection – R package, 2021, GitHub Repository, GitHub. https://github.com/ywwry66/QDA-by-Projection-R-Package.
[23] Y. Wu, Y. Qin, M. Zhu, Quadratic discriminant analysis for high-dimensional data, Statist. Sinica 29 (2) (2019) 939–960.