A Large Dimensional Study of Regularized Discriminant Analysis

Khalil Elkhalil, *Student Member, IEEE,* Abla Kammoun, *Member, IEEE*, Romain Couillet, *Senior Member, IEEE*, Tareq Y. Al-Naffouri, *Member, IEEE*, and Mohamed-Slim Alouini, *Fellow, IEEE*

***Abstract*—In this paper, we conduct a large dimensional study of regularized discriminant analysis classifiers with its two popular variants known as regularized LDA and regularized QDA. The analysis is based on the assumption that the data samples are drawn from a Gaussian mixture model with different means and covariances and relies on tools from random matrix theory (RMT). We consider the regime in which both the data dimension and training size within each class tends to infinity with fixed ratio. Under mild assumptions, we show that the probability of misclassification converges to a deterministic quantity that describes in closed form the performance of these classifiers in terms of the class statistics as well as the problem dimension. The result allows for a better understanding of the underlying classification algorithms in terms of their performances in practical large but finite dimensions. Further exploitation of the results permits to optimally tune the regularization parameter with the aim of minimizing the probability of misclassification. The analysis is validated with numerical results involving synthetic as well as real data from the USPS dataset yielding a high accuracy in predicting the performances and hence making an interesting connection between theory and practice.**

***Index Terms*—Linear discriminant analysis, quadratic discriminant analysis, classification, random matrix theory, consistent estimator.**

I. Introduction

Linear Discriminant analysis (LDA) is an old concept that dates back to Fisher that generalizes the Fisher discriminant [2], [3]. Given two statistically defined datasets, or classes, the Fisher discriminant analysis is designed to maximize the ratio of the variance between classes to the variance within classes and is useful for both classification and dimensionality reduction [4], [5]. LDA, on the other hand, relying merely on the concept of model based classification [4], is conceived so that the misclassification rate is minimized under a Gaussian assumption for the data. Interestingly, both ideas lead to the same classifier when the data of both classes share the same covariance matrix. Maintaining the Gaussian assumption but considering the general case of distinct covariance matrices, quadratic discriminant analysis (QDA) becomes the optimal classifier in terms of the minimization of the misclassification rate when both statistical means and covariances of the classes are known.

In practice, these parameters are rarely given and only estimated based on training data. Assuming the number of training samples is high enough, QDA and LDA should remain asymptotically optimal. It is however often the case in practice that the data dimension is large, if not larger, than the number of observations. In such circumstances, the covariance matrix estimate becomes ill-conditioned or even non invertible, which leads to poor classification performance.

To overcome this difficulty, many techniques can be considered. One can resort to dimensionality reduction so as to embed the data in a low-dimensional space that retains most of the useful information from a classification point of view [6], [7]. This ensures a higher number of training samples than the effective data size. Questions as to which dimensions to be selected or to what extent dimension should be reduced remain open. Another alternative involves the regularized versions of LDA and QDA denoted, throughout this paper, by R-LDA and R-QDA [5], [8]. Both approaches constitute the main focus of the article.

There exist many works on the performance analysis of discriminant analysis classifiers. In [9], an exact analysis of QDA is made by relying on properties of Wishart matrices. This allows for exact expressions of the probability misclassification rate for all sample size n and dimension p. This analysis is however only valid as long as $n \geq p$. Generalizing this analysis to regularized versions is however beyond analytical reach. This motivated further studies to consider asymptotic regimes. In [10], [11] the authors consider the large p asymptotics and observe that LDA and QDA fall short even when the exact covariance matrix is known. [10] thus proposed improved LDA and PCA that exploit sparsity assumptions on the difference of the statistical means, however not necessarily met in practice. This leads us to consider in the present work the *double asymptotic regime* in which both p and n tend to infinity with fixed ratio. This regime leverages results from random matrix theory [12], [13], [14], [15], [16]. For LDA analysis, this regime was first considered in [17] under the assumption of equal covariance matrices. It was extended to the analysis of R-LDA in [8] and to the Euclidean distance discriminant rule in [18]. To the best of the authors' knowledge, the general case in which the covariances across classes are different was never treated. As shown in the course of the paper, a major difficulty for the analysis resides in choosing the assumptions governing the growth rate of means

Part of this work related to the derivation of an asymptotic equivalent for the R-QDA probability of misclassification has been accepted for publication in the IEEE MLSP workshop 2017 [1]. The current paper extends the analyis in [1] to cover both R-LDA and R-QDA in addition to the derivation of consistent estimators for the classification risk. All proofs are provided in the arxiv version available at https://arxiv.org/pdf/1711.00382.pdf

K. Elkhalil, A. Kammoun, T. Y. Al-Naffouri and M.-S. Alouini are with the Electrical Engineering Program, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia; e-mails: {khalil.elkhalil, abla.kammoun, tareq.alnaffouri, slim.alouini}@kaust.edu.sa.

R. Couillet is with the CentraleSupélec, Paris, 91190 Châtenay-Malabry, France; e-mail: romain.couillet@centralesupelec.fr.

and covariances to avoid nontrivial asymptotic classification performances.

This motivates the present work. Particularly, we propose a large dimensional analysis of both R-LDA and R-QDA in the double asymptotic regime (discussed earlier) for general Gaussian assumptions. Precisely, under technical, yet mild, assumptions controlling the distances between the class means and covariances, we prove that the probability of misclassification converges to a non-trivial deterministic quantity that only depends on the class statistics as well as the ratio p/n. Interestingly, R-LDA and R-QDA require different growth regimes, reflecting a fundamental difference in the way they leverage the information about the means and covariances. Notably, R-QDA requires a minimal distance between class means of order $O(\sqrt{p})$ while R-LDA necessitates a difference in means of order $O(1)$. However, R-LDA does not seem to leverage the information about the distance in covariance matrices. The results of [8] are in particular recovered when the spectral norm of the difference of covariance matrices is small. These findings lead to insights into when LDA or QDA should be preferred in practical scenarios.

To sum up, our main results are as follows:

- Under mild assumptions, we establish the convergence of the misclassification rate for both R-LDA and R-QDA classifiers to a deterministic error as a function of the statistical parameters associated with each class.
- We design a consistent estimator for the misclassification rate for both R-LDA and R-QDA classifiers that allows to estimate the optimal regularization parameter.
- We validate our theoretical findings on both synthetic and real data drawn from the USPS dataset and illustrate the good accuracy of our results in both settings.

The remainder is organized as follows. We give an overview of discriminant analysis for binary classification in Section II. The main results are presented in Section III, the proofs of which are deferred to the Appendix. In Section IV, we design a consistent estimator of the misclassification error rate. We validate our analysis for real data in Section V and conclude the article in Section VI.

Notations: Scalars, vectors and matrices are respectively denoted by non-boldface, boldface lowercase and boldface uppercase characters. $\mathbf{0}_{p\times n}$ and $\mathbf{1}_{p\times n}$ are respectively the matrix of zeros and ones of size $p\times n$, $\mathbf{I}_p$ denotes the $p\times p$ identity matrix. The notation $\|.\|$ stands for the Euclidean norm for vectors and the spectral norm for matrices. $(.)^T$, $\mathrm{tr}\,(.)$ and $|.|$ stands for the transpose, the trace and the determinant of a matrix respectively. For two functionals f and g, we say that $f = O(g)$, if $\exists\, 0 < M < \infty$ such that $|f| \leq Mg$. $\mathbb{P}(.)$, $\to_d$, $\to_{prob.}$ and $\to_{a.s.}$ respectively denote the probability measure, the convergence in distribution, the convergence in probability and the almost sure convergence of random variables. $\Phi(.)$ denotes the cumulative density function (CDF) of the standard normal distribution, i.e. $\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$.

II. Discriminant Analysis for Binary Classification

This paper studies binary discriminant analysis techniques which employs a discriminant rule to assign for an input data vector the class to which it most likely belongs. The discriminant rule is designed based on n available training data with known class labels. In this paper, we consider the case in which a bayesian discriminant rule is employed. Hence, we assume that observations from class $\mathcal{C}_i$, $i \in \{0,1\}$ are independent and are sampled from a multivariate Gaussian distribution with mean $\boldsymbol{\mu}_i \in \mathbb{R}^{p\times 1}$ and non-negative covariance matrix $\boldsymbol{\Sigma}_i \in \mathbb{R}^{p\times p}$. Formally speaking, an observation vector $\mathbf{x} \in \mathbb{R}^{p\times p}$ is classified to $\mathcal{C}_i$, $i \in \{0,1\}$, if

$$\mathbf{x} = \boldsymbol{\mu}_i + \boldsymbol{\Sigma}_i^{1/2}\boldsymbol{z}, \quad \boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p). \tag{1}$$

Let

$$\begin{aligned} W^{QDA}(\mathbf{x}) = &-\frac{1}{2}\log\frac{|\boldsymbol{\Sigma}_0|}{|\boldsymbol{\Sigma}_1|} - \frac{1}{2}\mathbf{x}^T\left(\boldsymbol{\Sigma}_0^{-1} - \boldsymbol{\Sigma}_1^{-1}\right)\mathbf{x} \\ &+ \mathbf{x}^T\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 - \mathbf{x}^T\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 - \frac{1}{2}\boldsymbol{\mu}_0^T\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \frac{1}{2}\boldsymbol{\mu}_1^T\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 \\ &- \log\frac{\pi_1}{\pi_0}. \end{aligned} \tag{2}$$

As stated in [19], for distinct covariance matrices $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\Sigma}_1$, the discriminant rule is summarized as follows

$$\begin{cases} \mathbf{x} \in \mathcal{C}_0 & \text{if} \quad W^{QDA}(\mathbf{x}) > 0. \\ \mathbf{x} \in \mathcal{C}_1 & \text{otherwise.} \end{cases} \tag{3}$$

When the considered classes have the same covariance matrix, i.e., $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_1$, the discriminant function simplifies to [5], [4], [8]

$$W^{LDA}(\mathbf{x}) = \left(\mathbf{x} - \frac{\boldsymbol{\mu}_0 + \boldsymbol{\mu}_1}{2}\right)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1) - \log\frac{\pi_1}{\pi_0}. \tag{4}$$

Classification obeys hence the following rule:

$$\begin{cases} \mathbf{x} \in \mathcal{C}_0 & \text{if} \quad W^{LDA}(\mathbf{x}) > 0 \\ \mathbf{x} \in \mathcal{C}_1 & \text{otherwise.} \end{cases} \tag{5}$$

Since W^{LDA} is linear in $\mathbf{x}$, the corresponding classification method is referred to as linear discriminant analysis. As can be seen from (3) and (5), the classification rules assume the knowedge of the class statistics, namely their associated covariance matrices and mean vectors. In practice, these statistics can be estimated using the available training data. As such, we assume that n_i, $i \in \{0,1\}$ independent training samples $\mathcal{T}_0 = \{\mathbf{x}_l \in \mathcal{C}_0\}_{l=1}^{n_0}$ and $\mathcal{T}_1 = \{\mathbf{x}_l \in \mathcal{C}_1\}_{l=n_0+1}^{n_0+n_1=n}$ are respectively available to estimate the mean and the covariance matrix of each class i[1]. For that, we consider the following sample estimates

$$\begin{aligned} \hat{\boldsymbol{\mu}}_i &= \frac{1}{n_i}\sum_{l\in\mathcal{T}_i}\mathbf{x}_l, \quad i \in \{0,1\} \\ \widehat{\boldsymbol{\Sigma}}_i &= \frac{1}{n_i - 1}\sum_{l\in\mathcal{T}_i}(\mathbf{x}_l - \hat{\boldsymbol{\mu}}_i)(\mathbf{x}_l - \hat{\boldsymbol{\mu}}_i)^T, \quad i \in \{0,1\} \\ \widehat{\boldsymbol{\Sigma}} &= \frac{(n_0 - 1)\,\widehat{\boldsymbol{\Sigma}}_0 + (n_1 - 1)\,\widehat{\boldsymbol{\Sigma}}_1}{n - 2}, \end{aligned}$$

[1]We assume that $\frac{n_i}{n_0+n_1} \to \pi_i$ for $i \in \{0,1\}$ which is valid under random sampling. Therefore, we do not consider the problem of separate sampling when $\frac{n_0}{n_0+n_1}$ and $\frac{n_1}{n_0+n_1}$ are not chosen to converge to the priors π_0 and π_1 [20].

where $\widehat{\mathbf{\Sigma}}$ is the pooled sample covariance matrix for both classes. To avoid singularity issues when $n_i < p$, we use the ridge estimator of the inverse of the covariance matrix [5]

$$\mathbf{H} = \left(\mathbf{I}_p + \gamma\widehat{\mathbf{\Sigma}}\right)^{-1}, \tag{6}$$

$$\mathbf{H}_i = \left(\mathbf{I}_p + \gamma\widehat{\mathbf{\Sigma}}_i\right)^{-1}, \quad i \in \{0,1\} \tag{7}$$

where $\gamma > 0$ is a regularization parameter. Replacing $\mathbf{\Sigma}^{-1}$ and $\mathbf{\Sigma}_i$ for $i \in \{0,1\}$ by (6) and (7) into (4) and (2), we obtain the following discriminant rules

$$\widehat{W}^{R-LDA}(\mathbf{x}) = \left(\mathbf{x} - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2}\right)^T \mathbf{H}\left(\hat{\boldsymbol{\mu}}_0 - \hat{\boldsymbol{\mu}}_1\right) - \log\frac{\pi_1}{\pi_0}. \tag{8}$$

$$\begin{aligned}\widehat{W}^{R-QDA}(\mathbf{x}) = {} & \frac{1}{2}\log\frac{|\mathbf{H}_0|}{|\mathbf{H}_1|} - \frac{1}{2}\left(\mathbf{x} - \hat{\boldsymbol{\mu}}_0\right)^T \mathbf{H}_0\left(\mathbf{x} - \hat{\boldsymbol{\mu}}_0\right) \\ & + \frac{1}{2}\left(\mathbf{x} - \hat{\boldsymbol{\mu}}_1\right)^T \mathbf{H}_1\left(\mathbf{x} - \hat{\boldsymbol{\mu}}_1\right) - \log\frac{\pi_1}{\pi_0}.\end{aligned} \tag{9}$$

The corresponding classification methods will be denoted respectively by R-LDA and R-QDA [2]. Conditioned on the training samples $\mathcal{T}_i$, $i \in \{0,1\}$, the classification errors associated with R-LDA and R-QDA when $\mathbf{x}$ belongs to class $\mathcal{C}_i$ are given by

$$\epsilon_i^{R-LDA} = \mathbb{P}\left[(-1)^i \widehat{W}^{R-LDA}(\mathbf{x}) < 0 \,|\mathbf{x} \in \mathcal{C}_i, \mathcal{T}_0, \mathcal{T}_1\right]. \tag{10}$$

$$\epsilon_i^{R-QDA} = \mathbb{P}\left[(-1)^i \widehat{W}^{R-QDA}(\mathbf{x}) < 0 \,|\mathbf{x} \in \mathcal{C}_i, \mathcal{T}_0, \mathcal{T}_1\right]. \tag{11}$$

The total classification errors are respectively given by

$$\epsilon^{R-LDA} = \pi_0\epsilon_0^{R-LDA} + \pi_1\epsilon_1^{R-LDA}.$$
$$\epsilon^{R-QDA} = \pi_0\epsilon_0^{R-QDA} + \pi_1\epsilon_1^{R-QDA}.$$

In the following, we propose to analyze the asymptotic classication errors of both R-LDA and R-QDA when p, n_i grow large at the same rate. For R-LDA, our results cover a more general setting than the one studied in [8], in that they apply to the case where both classes have distinct covariance matrices.

III. Main Results

The main contributions of the present work are two fold. First, we carry out an asymptotic analysis of the probability of mis-classification for both R-LDA and R-QDA, showing that they converge to some deterministic quantities that depend solely on the observations statistics associated with each class. Such a result allows a better understanding of the impact of these parameters on the performances. Second, we build consistent estimates of the asymptotic misclassification error rates for both estimators. An estimator of the misclassification error rate has been provided in [8] but for the R-LDA when the classes are assumed to have identical covariance matrices. Our results regarding R-LDA in this respect extends the one in [8] when the covariance matrices are not equal. The treatment of R-QDA is however new and constitute the main contribution of the present work.

[2]Please note that as long as $\gamma > 0$, R-LDA and R-QDA are fundamentally different and as we will show in the course of the paper, different tools will be used to characterize their classification performance. The case where $\gamma = 0$ results in R-LDA and R-QDA being the same classifier implicitely assuming equal covariance given by the identity matrix. Therefore, this scenario is neither of theoretical nor practical importance.

A. Asymptotic Performance of R-LDA with Distinct Covariance Matrices

In this section, we present an asymptotic analysis of the R-LDA classifier. Our analysis is mainly based on recent results from RMT concerning some properties of Gram matrices of mixture models [16]. We recall that [8] made a similar analysis of R-LDA in the double asymptotic regime when both classes have a common covariance matrix, thereby not requiring these advanced tools. As such, our results can be viewed as a generalization of [8] when both classes have distinct covariance matrices. This permits to evaluate the performance of R-LDA in practical scenarios when the assumption of common covariance matrices cannot always be guaranteed. To allow derivations, we shall consider the following growth rate assumptions

Assumption. 1 (Data scaling)**.** $\frac{p}{n} \to c \in (0, \infty)$.

Assumption. 2 (Class scaling)**.** $\frac{n_i}{n} \to c_i \in (0, \infty)$, for $i \in \{0,1\}$.

Assumption. 3 (Covariance scaling)**.** $\limsup_p \|\mathbf{\Sigma}_i\| < \infty$, for $i \in \{0,1\}$.

Assumption. 4 (Mean scaling)**.** *Let* $\boldsymbol{\mu} = \boldsymbol{\mu}_0 - \boldsymbol{\mu}_1$. *Then,* $\limsup_p \|\boldsymbol{\mu}\| = \limsup_p \|\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\| < \infty$.

These assumptions are mainly considered to achieve an asymptotically non-trivial classification error. Assumption 3 is frequently met within the framework of random matrix theory [16]. Under the setting of Assumption 3, Assumption 4 ensures that a nontrivial classification rate is obtained: if $\|\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\|$ scales faster than $O(1)$, then perfect asymptotic classification is achieved; however, if $\|\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\|$ scales slower than $O(1)$, classification is asymptotically impossible. Assumptions 1 and 2 respectively control the growth rate in the data and the training. More precisely, Assumption 1 says that both the sample size and the data dimension are large with the same order of magnitude. The same thing can be inferred from Assumption 2 regarding the training size of both classes. These two assumptions allow to leverage results form random matrix theory. Regarding Assumption 4 and 3, note that these two assumptions have been used before to prove the results in [8] and thus they are standard assumptions. They are indeed carefully devised so that R-LDA do not present asymptotically trivial classification behavior.

1) Deterministic Equivalent: We are in a position to derive a deterministic equivalent of the misclassification error rate of the R-LDA. Indeed, conditioned on the training data $\mathbf{x}_1, \cdots, \mathbf{x}_n$, the probability of misclassification is given by:

[8]

$$\epsilon_i^{R-LDA} = \Phi\left(\frac{(-1)^{i+1} G\left(\boldsymbol{\mu}_i, \hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\mu}}_1, \mathbf{H}\right) + (-1)^i \log\frac{\pi_1}{\pi_0}}{\sqrt{D\left(\hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\mu}}_1, \mathbf{H}, \boldsymbol{\Sigma}_i\right)}}\right), \tag{12}$$

where

$$G\left(\boldsymbol{\mu}_i, \hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\mu}}_1, \mathbf{H}\right) = \left(\boldsymbol{\mu}_i - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2}\right)^T \mathbf{H}\left(\hat{\boldsymbol{\mu}}_0 - \hat{\boldsymbol{\mu}}_1\right). \tag{13}$$

$$D\left(\hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\mu}}_1, \mathbf{H}, \boldsymbol{\Sigma}_i\right) = \left(\hat{\boldsymbol{\mu}}_0 - \hat{\boldsymbol{\mu}}_1\right)^T \mathbf{H}\boldsymbol{\Sigma}_i\mathbf{H}\left(\hat{\boldsymbol{\mu}}_0 - \hat{\boldsymbol{\mu}}_1\right). \tag{14}$$

The total misclassification probability is thus given by

$$\epsilon^{R-LDA} = \pi_0\epsilon_0^{R-LDA} + \pi_1\epsilon_1^{R-LDA}. \tag{15}$$

Prior to stating the main result concerning R-LDA, we shall introduce the following quantities, which naturally appear, as a result of applying [16]. Let $\bar{\mathbf{Q}}\left(z\right)$ be the matrix defined as follows

$$\bar{\mathbf{Q}}\left(z\right) \triangleq -\frac{1}{z}\left(\mathbf{I}_p + c_0 g_0(z)\boldsymbol{\Sigma}_0 + c_1 g_1(z)\boldsymbol{\Sigma}_1\right)^{-1},\ z \in \mathbb{C}. \tag{16}$$

where $g_i(z)$, $i \in \{0,1\}$, satisfies the following fixed point equations

$$\frac{p}{n} g_i(z) = -\frac{1}{z}\frac{1}{1+\tilde{g}_i(z)},\ \ \tilde{g}_i(z) = \frac{1}{p}\operatorname{tr}\boldsymbol{\Sigma}_i\bar{\mathbf{Q}}\left(z\right). \tag{17}$$

Also define $\mathbf{A}_i = \boldsymbol{\Sigma}_i\bar{\mathbf{Q}}(z)$ and $\widetilde{\mathbf{Q}}_i\left(z\right)$ as

$$\widetilde{\mathbf{Q}}_i\left(z\right) \triangleq \bar{\mathbf{Q}}(z)\left(\mathbf{A}_i + R_0\mathbf{A}_0 + R_1\mathbf{A}_1\right), \tag{18}$$

where

$$R_i = \frac{z^2 c_i g_i^2(z)\frac{1}{n}\operatorname{tr}\mathbf{A}_0\mathbf{A}_1}{1 - \left(z^2 c_0 g_0^2(z) + z^2 c_1 g_1^2(z)\right)\frac{1}{n}\operatorname{tr}\mathbf{A}_0\mathbf{A}_1},\quad i \in \{0,1\}. \tag{19}$$

The quantities in (17) can be computed in an iterative fashion where convergence is guaranteed after few iterations (see [16] for more details). Moreover, define

$$\overline{G}_i\left(z\right) \triangleq \frac{(-1)^{i+1}}{2} z\boldsymbol{\mu}^T\bar{\mathbf{Q}}\left(z\right)\boldsymbol{\mu} + \frac{z}{2n_0}\operatorname{tr}\mathbf{A}_0 - \frac{z}{2n_1}\operatorname{tr}\mathbf{A}_1. \tag{20}$$

$$\overline{D}_i\left(z\right) \triangleq z^2\boldsymbol{\mu}^T\widetilde{\mathbf{Q}}_i\left(z\right)\boldsymbol{\mu} + \frac{z^2}{n_0}\operatorname{tr}\boldsymbol{\Sigma}_0\widetilde{\mathbf{Q}}_i\left(z\right) + \frac{z^2}{n_1}\operatorname{tr}\boldsymbol{\Sigma}_1\widetilde{\mathbf{Q}}_i\left(z\right), \tag{21}$$

where $\boldsymbol{\mu} = \boldsymbol{\mu}_0 - \boldsymbol{\mu}_1$. With these definitions at hand, we state the following theorem

Theorem. 1. *Under Assumptions 1-4, we have*

$$G\left(\boldsymbol{\mu}_i, \hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\mu}}_1, \mathbf{H}\right) - \overline{G}_i\left(-\frac{1}{c\gamma}\right) \to_{a.s.} 0. \tag{22}$$

$$D\left(\hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\mu}}_1, \mathbf{H}, \boldsymbol{\Sigma}_i\right) - \overline{D}_i\left(-\frac{1}{c\gamma}\right) \to_{a.s.} 0. \tag{23}$$

As a consequence, the conditional misclassification probability converges almost surely to a deterministic quantity $\bar{\epsilon}_i^{R-LDA}$

$$\epsilon_i^{R-LDA} - \bar{\epsilon}_i^{R-LDA} \to_{a.s.} 0, \tag{24}$$

where

$$\bar{\epsilon}_i^{R-LDA} = \Phi\left(\frac{(-1)^{i+1}\overline{G}_i + (-1)^i\log\left(\frac{\pi_0}{\pi_1}\right)}{\sqrt{\overline{D}_i}}\right). \tag{25}$$

Proof. See Appendix A. □

Remark. 1. *As stated earlier, if* $\|\boldsymbol{\mu}\|$ *scales faster than* $O\left(1\right)$*, perfect asymptotic classification is achieved. This can be seen by noticing that* $\frac{\overline{G}_i\left(-\frac{1}{c\gamma}\right)}{\sqrt{\overline{D}_i\left(-\frac{1}{c\gamma}\right)}}$ *would grow indefinitely large with* p*, thereby making the conditional error rates vanish.*

When $\|\boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_1\|$ converges to zero, the asymptotic misclassification error rate of each class coincides with the one derived in [8] obtained when $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_1$.

Corollary. 1. *In the case where* $\|\boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_1\| = o\left(1\right)$ *(including the common covariance case where* $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}$*), the conditional misclassification error rate converges almost surely to* $\bar{\bar{\epsilon}}_i^{R-LDA}$

$$\bar{\bar{\epsilon}}_i^{R-LDA} = \Phi\left(\frac{(-1)^{i+1}\overline{\overline{G}}_i(-\frac{1}{c\gamma}) + (-1)^i\log\frac{\pi_0}{\pi_1}}{\sqrt{\overline{\overline{D}}_i(-\frac{1}{c\gamma})}}\right),$$

where

$$\overline{\overline{G}}_i\left(-\frac{1}{c\gamma}\right) = \frac{(-1)^i}{2}\boldsymbol{\mu}^T\left(\mathbf{I}_p + \frac{\gamma}{1+\gamma\delta}\boldsymbol{\Sigma}\right)^{-1}\boldsymbol{\mu} - \frac{n\delta}{2}\left(\frac{1}{n_0} - \frac{1}{n_1}\right)$$

$$\overline{\overline{D}}_i\left(-\frac{1}{c\gamma}\right) = \frac{\left[\boldsymbol{\mu}^T\boldsymbol{\Sigma}\left(\mathbf{I}_p + \frac{\gamma}{1+\gamma\delta}\boldsymbol{\Sigma}\right)^{-2}\boldsymbol{\mu} + \left(\frac{1}{n_0} + \frac{1}{n_1}\right)\operatorname{tr}\boldsymbol{\Sigma}^2\left(\mathbf{I}_p + \frac{\gamma}{1+\gamma\delta}\boldsymbol{\Sigma}\right)^{-2}\right]}{1 - \frac{\gamma^2}{n(1+\gamma\delta)^2}\operatorname{tr}\boldsymbol{\Sigma}^2\left(\mathbf{I}_p + \frac{\gamma}{1+\gamma\delta}\boldsymbol{\Sigma}\right)^{-2}}.$$

where $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0$ *or* $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_1$*, and* δ *is the unique positive solution to the following equation:*

$$\delta = \frac{1}{n}\operatorname{tr}\boldsymbol{\Sigma}\left(\mathbf{I}_p + \frac{\gamma}{1+\gamma\delta}\boldsymbol{\Sigma}\right)^{-1}.$$

Proof. When $\|\boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_1\| = o\left(1\right)$, we first prove that up to an error $o(1)$, the key deterministic equivalents can be simplified to depend only on $\boldsymbol{\Sigma}_0$ (or $\boldsymbol{\Sigma}_1$). In the sequel, we take $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0$. As $\|\boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_1\| = o\left(1\right)$, we have

$$\tilde{g}_i\left(z\right) = \underbrace{\frac{1}{p}\operatorname{tr}\boldsymbol{\Sigma}\bar{\mathbf{Q}}\left(z\right)}_{\tilde{g}(z)} + o(1),\ \forall i \in \{0,1\}.$$

It follows that $g_i(z) = g(z) + o(1)$ where $g(z) = -\frac{1}{z}\frac{n}{p}\frac{1}{1+\tilde{g}(z)}$. The above relations allow to simplify functionals involving matrix $\bar{\mathbf{Q}}$. To see that, we decompose $\bar{\mathbf{Q}}$ as

$$\begin{aligned}
&\bar{\mathbf{Q}}(z) \\
&= -z^{-1}\left(\mathbf{I} + g(z)\boldsymbol{\Sigma}_0 + c_1 g(z)(\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_0)\right)^{-1} + o_{\|.\|}(1) \\
&= -z^{-1}(\mathbf{I} + g(z)\boldsymbol{\Sigma}_0)^{-1} - z^{-1}\Bigg[\left(\mathbf{I} + g(z)\boldsymbol{\Sigma}_0 + c_1 g(z)(\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_0)\right)^{-1} \\
&- (\mathbf{I} + g(z)\boldsymbol{\Sigma}_0)^{-1}\Bigg] + o_{\|.\|}(1) \\
&\overset{(a)}{=} -z^{-1}(\mathbf{I} + g(z)\boldsymbol{\Sigma}_0)^{-1} \\
&- z^{-1}\left(\mathbf{I} + g(z)\boldsymbol{\Sigma}_0 + c_1 g(z)(\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_0)\right)^{-1} c_1 g(z)(\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_0) \\
&\times (\mathbf{I} + g(z)\boldsymbol{\Sigma}_0)^{-1} + o_{\|.\|}(1)
\end{aligned}$$

where (a) follows from the resolvent identity and $o_{\|.\|}(1)$ denotes a matrix with spectral norm converging to zero. Define

$$\begin{aligned}
\boldsymbol{\Psi} &= z^{-1}\left(\mathbf{I} + g(z)\boldsymbol{\Sigma}_0 + c_1 g(z)(\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_0)\right)^{-1} c_1 g(z)(\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_0) \\
&\times (\mathbf{I} + g(z)\boldsymbol{\Sigma}_0)^{-1}.
\end{aligned}$$

Then, it can be shown using the inequality $\|\mathbf{AB}\| \leq \|\mathbf{A}\|\|\mathbf{B}\|$ for $\mathbf{A}$ and $\mathbf{B}$ two matrices in $\mathbb{R}^{p\times p}$ that:

$$\begin{aligned}
\|\boldsymbol{\Psi}\| &\overset{(b)}{\leq} z^{-1} c_1 g(z)\|\boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_1\| \\
&\times \left\|\left(\mathbf{I} + g(z)\boldsymbol{\Sigma}_0 + c_1 g(z)(\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_0)\right)^{-1}(\mathbf{I} + g(z)\boldsymbol{\Sigma}_0)^{-1}\right\| \\
&= o(1).
\end{aligned}$$

Hence, for $\mathbf{a}, \mathbf{b} \in \mathbb{R}^p$,

$$\boldsymbol{a}^T \bar{\mathbf{Q}}(z)\boldsymbol{b} = -z^{-1}\boldsymbol{a}^T(\mathbf{I} + g(z)\boldsymbol{\Sigma})^{-1}\boldsymbol{b} + o(1),$$

and $\frac{1}{p}\operatorname{tr}\mathbf{A}\bar{\mathbf{Q}} = \frac{-z^{-1}}{p}\operatorname{tr}\mathbf{A}(\mathbf{I} + g(z)\boldsymbol{\Sigma})^{-1} + o(1)$. Using the same notations as in [8] we have in particular for $z = -\frac{1}{c\gamma}$, $\tilde{g}(z) = \delta + o(1)$ and $g(z) = \frac{\gamma}{1+\gamma\delta} + o(1)$, where δ is the fixed-point solution in [8, Proposition 1]. Moreover,

$$\begin{aligned}
&\frac{z}{2n_0}\operatorname{tr}\boldsymbol{\Sigma}_0\bar{\mathbf{Q}}(z) - \frac{z}{2n_1}\operatorname{tr}\boldsymbol{\Sigma}_1\bar{\mathbf{Q}}(z) \\
&= \frac{z\operatorname{tr}\boldsymbol{\Sigma}_0\bar{\mathbf{Q}}(z)}{2}\left(\frac{1}{n_0} - \frac{1}{n_1}\right) + \underbrace{\frac{z}{2n_1}\operatorname{tr}(\boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_1)\bar{\mathbf{Q}}(z)}_{\leq\|\boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_1\|\frac{z}{2n_1}\operatorname{tr}\bar{\mathbf{Q}}(z)} \\
&= \frac{z\operatorname{tr}\boldsymbol{\Sigma}\bar{\mathbf{Q}}(z)}{2}\left(\frac{1}{n_0} - \frac{1}{n_1}\right) + o(1) \\
&= \frac{-\operatorname{tr}\boldsymbol{\Sigma}(\mathbf{I} + g(z)\boldsymbol{\Sigma})^{-1}}{2}\left(\frac{1}{n_0} - \frac{1}{n_1}\right) + o(1).
\end{aligned}$$

It follows that

$$\begin{aligned}
\overline{G}_i\left(-\frac{1}{c\gamma}\right) &= \frac{(-1)^i}{2}\boldsymbol{\mu}^T\left(\mathbf{I}_p + \frac{\gamma}{1+\gamma\delta}\boldsymbol{\Sigma}\right)^{-1}\boldsymbol{\mu} \\
&- \frac{n\delta}{2}\left(\frac{1}{n_0} - \frac{1}{n_1}\right) + o(1).
\end{aligned}$$

Using the same arguments, we can show that

$$\begin{aligned}
\overline{D}_i\left(-\frac{1}{c\gamma}\right) &= \left[1 - \frac{\gamma^2}{n(1+\gamma\delta)^2}\operatorname{tr}\boldsymbol{\Sigma}^2\left(\mathbf{I}_p + \frac{\gamma}{1+\gamma\delta}\boldsymbol{\Sigma}\right)^{-2}\right]^{-1} \\
&\times \Bigg[\boldsymbol{\mu}^T\boldsymbol{\Sigma}\left(\mathbf{I}_p + \frac{\gamma}{1+\gamma\delta}\boldsymbol{\Sigma}\right)^{-2}\boldsymbol{\mu} \\
&+ \left(\frac{1}{n_0} + \frac{1}{n_1}\right)\operatorname{tr}\boldsymbol{\Sigma}^2\left(\mathbf{I}_p + \frac{\gamma}{1+\gamma\delta}\boldsymbol{\Sigma}\right)^{-2}\Bigg] + o(1).
\end{aligned}$$

□

Corollary 1 is useful because it allows to specify the range of applications of Theorem 1 in which the information on the covariance matrix is essential for the classification task. Also, it shows how R-LDA is robust against small perturbations in the covariance matrix. Similar observations have been made in [21] where it was shown via a Monte Carlo study that LDA is robust against the modeling assumptions.

B. Asymptotic Performance of R-QDA

In this part, we state the main results regarding the derivation of deterministic approximations of the R-QDA classification error. Such results have been obtained by considering some specific assumptions, carefully chosen such that an asymptotically non-trivial classification error (i.e., neither 0 nor 1) is achieved. We particularly highlight how the provided asymptotic approximations depend on such statistical parameters as the means and covariances within classes, thus allowing a better understanding of the performance of the R-QDA classifier. Ultimately, these results can be exploited in order to improve the performances by allowing optimal setting of the regularization parameter.

1) Technical Assumptions: For R-QDA, we require stronger assumptions as compared to R-LDA. This is mainly due to the fact that R-QDA is highly sensitive to the estimation noise in the covariance matrix. Therefore, we require a good separation between the means and the covariance matrices. In fact, we consider the following double asymptotic regime in which n_i, $p \to \infty$ for $i \in \{0, 1\}$ with the following assumptions met

Assumption. 5 (Data scaling)**.** $n_0 - n_1 = o(1)$ *and* $\frac{p}{n} \to c \in (0, \infty)$.

Assumption. 6 (Mean scaling)**.** $\|\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\|^2 = O(\sqrt{p})$.

Assumption. 7 (Covariance scaling)**.** $\|\boldsymbol{\Sigma}_i\| = O(1)$.

Assumption. 8. *Matrix* $\boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_1$ *has exactly* $O(\sqrt{p})$ *eigenvalues of order* $O(1)$. *The remaining eigenvalues are of order* $O(\frac{1}{\sqrt{p}})$.

Assumption 5 implies also that $\pi_i \to \frac{1}{2}$ for $i \in \{0, 1\}$. As we shall see later, if this is not satisfied, the R-QDA perform asymptotically as the classifier that assigns all observations to the same class. The second assumption governs the distance between the two classes in terms of the Euclidean distance between the means. This is mandatory in order to avoid asymptotic perfect classification. This is a much stronger

assumption than Assumption 2 in R-LDA since we allow larger values for $\|\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\|$. This can be understood as R-QDA being subject to strong noise induced when estimating $\boldsymbol{\Sigma}_i$, $i \in \{0,1\}$ which requires a large value $\|\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\|$ so that it can play a role in classification. A similar assumption is required to control the distance between the covariance matrices. Particularly, the spectral norm of the covariance matrices are required to be bounded as stated in Assumption 7 while their difference should satisfy Assumption 8.The latter assumption implies that for any matrix $\mathbf{A}$ of bounded spectral norm,

$$\frac{1}{\sqrt{p}} \operatorname{tr} \mathbf{A}(\boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_1) = O(1).$$

2) Central Limit Theorem (CLT): It can be easily shown that the R-QDA conditional classification error in (11) can be expressed as

$$\epsilon_i^{R-QDA} = \mathbb{P}\left[\boldsymbol{z}^T \mathbf{B}_i \boldsymbol{z} + 2\boldsymbol{z}^T \mathbf{r}_i < \xi_i | \boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p), \mathcal{T}_0, \mathcal{T}_1\right], \tag{26}$$

where

$$\begin{aligned}
\mathbf{B}_i &= \boldsymbol{\Sigma}_i^{1/2} (\mathbf{H}_1 - \mathbf{H}_0) \boldsymbol{\Sigma}_i^{1/2}, \\
\mathbf{r}_i &= \boldsymbol{\Sigma}_i^{1/2} \left[\mathbf{H}_1 (\boldsymbol{\mu}_i - \hat{\boldsymbol{\mu}}_1) - \mathbf{H}_0 (\boldsymbol{\mu}_i - \hat{\boldsymbol{\mu}}_0)\right], \\
\xi_i &= -\log\left(\frac{|\mathbf{H}_0|}{|\mathbf{H}_1|}\right) + (\boldsymbol{\mu}_i - \hat{\boldsymbol{\mu}}_0)^T \mathbf{H}_0 (\boldsymbol{\mu}_i - \hat{\boldsymbol{\mu}}_0) \\
&\quad - (\boldsymbol{\mu}_i - \hat{\boldsymbol{\mu}}_1)^T \mathbf{H}_1 (\boldsymbol{\mu}_i - \hat{\boldsymbol{\mu}}_1) + 2\log\frac{\pi_1}{\pi_0}.
\end{aligned}$$

Computing ϵ_i^{R-QDA} amounts to the cumulative distribution function (CDF) of quadratic forms of Gaussian random vectors, and hence cannot be derived in closed form in general. However, it can be still approximated by considering asymptotic regimes that allow to exploit results about central limit theorem involving quadratic forms. Under Assumptions 5-8, a central limit theorem (CLT) on the random variable $\boldsymbol{z}^T \mathbf{B}_i \boldsymbol{z} + 2\boldsymbol{z}^T \mathbf{r}_i$ when $\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$ is established.

Proposition. 1 (CLT)**.** *Assume that assumptions 5-8 hold true. Assume also that for* $i \in \{0,1\}$

$$\lim_{p\to\infty} \frac{60 \operatorname{tr} \mathbf{B}_i^2 + 240 \operatorname{tr} \mathbf{B}_i^2 \|\mathbf{r}_i\|_2^2 + 48 \|\mathbf{r}_i\|_2^4}{\left(2 \operatorname{tr} \mathbf{B}_i^2 + 4\|\mathbf{r}_i\|_2^2\right)^2} \to 0. \tag{27}$$

Then,

$$\frac{\boldsymbol{z}^T \mathbf{B}_i \boldsymbol{z} + 2\boldsymbol{z}^T \mathbf{r}_i - \operatorname{tr} \mathbf{B}_i}{\sqrt{2 \operatorname{tr} \mathbf{B}_i^2 + 4\mathbf{r}_i^T \mathbf{r}_i}} \to_d \mathcal{N}(0,1). \tag{28}$$

Proof. The proof is mainly based on the application of the Lyapunov's CLT for the sum of independent but non identically distributed random variables [22]. The detailed proof is postponed to Appendix B. □

The condition in(27) will be proven to hold almost surely. Hence, as a by-product of the above Proposition, we obtain the following expression for the conditional classification error ϵ_i

Corollary. 2. *Under the setting of Proposition 1 , the conditional classification error in (11) satisfies*

$$\epsilon_i^{R-QDA} - \Phi\left((-1)^i \frac{\xi_i - \operatorname{tr} \mathbf{B}_i}{\sqrt{2 \operatorname{tr} \mathbf{B}_i^2 + 4\mathbf{r}_i^T \mathbf{r}_i}}\right) \to_{a.s.} 0. \tag{29}$$

As such an asymptotic equivalent of the conditional classification error can be derived. This is the subject of the next subsection.

3) Deterministic Equivalents: This part is devoted to the derivation of deterministic equivalents of some random quantities involved in the R-QDA conditional classification error. Before that, we shall introduce the following notations which basically arise as a result of applying standard results from random matrix theory. We define for $i \in \{0,1\}$, δ_i as the unique positive solution to the following fixed point equation[3]

$$\delta_i = \frac{1}{n_i} \operatorname{tr} \boldsymbol{\Sigma}_i \left(\mathbf{I}_p + \frac{\gamma}{1+\gamma\delta_i} \boldsymbol{\Sigma}_i\right)^{-1}.$$

Define $\mathbf{T}_i$ as

$$\mathbf{T}_i = \left(\mathbf{I}_p + \frac{\gamma}{1+\gamma\delta_i} \boldsymbol{\Sigma}_i\right)^{-1},$$

and the scalar ϕ_i and $\tilde{\phi}_i$ as

$$\phi_i = \frac{1}{n_i} \operatorname{tr} \boldsymbol{\Sigma}_i^2 \mathbf{T}_i^2, \qquad \tilde{\phi}_i = \frac{1}{(1+\gamma\delta_i)^2}.$$

Define $\overline{\xi}_i$, $\overline{b}_i$ and $\overline{B}_i$ as

$$\begin{aligned}
\overline{\xi}_i \triangleq \frac{1}{\sqrt{p}} \Bigg[&-\log\frac{|\mathbf{T}_0|}{|\mathbf{T}_1|} + \log\frac{(1+\gamma\delta_0)^{n_0}}{(1+\gamma\delta_1)^{n_1}} \\
&+ \gamma\left(\frac{n_1\delta_1}{1+\gamma\delta_1} - \frac{n_0\delta_0}{1+\gamma\delta_0}\right) + (-1)^{i+1} \boldsymbol{\mu}^T \mathbf{T}_{1-i} \boldsymbol{\mu} \Bigg].
\end{aligned} \tag{30}$$

$$\overline{b}_i = \frac{1}{\sqrt{p}} \operatorname{tr} \boldsymbol{\Sigma}_i (\mathbf{T}_1 - \mathbf{T}_0). \tag{31}$$

$$\begin{aligned}
\overline{B_i} \triangleq &\frac{\phi_i}{1-\gamma^2\phi_i\tilde{\phi}_i} \frac{n_i}{p} + \frac{1}{p} \operatorname{tr} \boldsymbol{\Sigma}_i^2 \mathbf{T}_{1-i}^2 \\
&+ \frac{n_i}{p} \frac{\gamma^2 \tilde{\phi}_{1-i}}{1-\gamma^2\phi_{1-i}\tilde{\phi}_{1-i}} \left(\frac{1}{n_i} \operatorname{tr} \boldsymbol{\Sigma}_i \boldsymbol{\Sigma}_{1-1} \mathbf{T}_{1-i}^2\right)^2 \\
&- \frac{2}{p} \operatorname{tr} \boldsymbol{\Sigma}_i \mathbf{T}_1 \boldsymbol{\Sigma}_i \mathbf{T}_0.
\end{aligned} \tag{32}$$

As shall be shown in Appendix C, these quantities are deterministic approximations in probability of ξ, b_i and B_i. We therefore get

Theorem. 2. *Under assumptions 5-8, the following convergence holds for* $i \in \{0,1\}$

$$\epsilon_i^{R-QDA} - \Phi\left((-1)^i \frac{\overline{\xi}_i - \overline{b_i}}{\sqrt{2\overline{B_i}}}\right) \to_{prob.} 0.$$

Proof. The proof is postponed to Appendix C. □

At first sight, quantity $\overline{\xi}_i - \overline{b}_i$ appears to be of order $O(\sqrt{p})$, since $\frac{1}{\sqrt{p}} \log|\mathbf{T}_i|$ and $\frac{1}{\sqrt{p}} \operatorname{tr} \boldsymbol{\Sigma}_i \mathbf{T}_i$ are $O(\sqrt{p})$. Following this line of thought, the asymptotic misclassification probability error is expected to converge to a trivial misclassification error. This statement is, hopefully false. Assumption 8 and 5 were carefully designed so that $\frac{1}{\sqrt{p}} \log|\mathbf{T}_1| - \frac{1}{\sqrt{p}} \log|\mathbf{T}_0|$

[3]Mathematical details treating the existence and uniqueness of δ_i can be found in [14].

and $\frac{1}{\sqrt{p}}(n_1\delta_1 - n_0\delta_0)$ are of order $O(1)$. In particular, the following is proven in Appendix D

Proposition. 2. *Under Assumption 5-8 The deterministic quantities $\overline{\xi}_i$ and $\overline{b}_i$ are uniformly bounded when p grows to infinity.*

Proof. The proof is deferred to Appendix D □

Remark. 2. *The results of Theorem 2 along with proposition 2 show that the classification error converges to a non-trivial deterministic quantity that depends only on the statistical means and covariances within each class. The major importance of this result is that it allows to find a good choice of the regularization γ as the value that minimizes the asymptotic classification error. While it seems to be elusive for such value to possess a closed-form expression, it can be numerically approximated by using a simple one-dimensional line search algorithm.*

Remark. 3. *Using Assumption 8, it can be shown that $\overline{B}_i$ can asymptotically simplified to*

$$\overline{B}_i = \frac{1}{c}\frac{\phi^2\tilde{\phi}}{1-\gamma^2\phi\tilde{\phi}} + o(1). \tag{33}$$

where $\phi = \phi_0$ or $\phi = \phi_1$. The above relation comes from the fact that, up to an error of order $o(1)$, matrices $\mathbf{\Sigma}_1$ or $\mathbf{\Sigma}_0$ can be used interchangeably in ϕ_0 or ϕ_1 and in the terms involved in $\overline{B}_i$. This, in particular, implies that $\overline{B}_0$ and $\overline{B}_1$ are the same up to a vanishing error. It is noteworthy to see that the same artifice could not work for the terms $\overline{\xi}_i$ and $\overline{b}_i$ because the normalization, being with $\frac{1}{\sqrt{p}}$, is not sufficient to provide vanishing terms. We should also mention that, although (33) *takes a simpler form, we chose to work in the simulations and when computing the consistent estimates of $\overline{B}_i$ with the expression* (32) *since we found that it provides the highest accuracy.*

4) Some Special cases: a) It is important to note that we could have considered $\|\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\| = O(1)$. In this case, the classification error rate would still converge to a non trivial limit but would not asymptotically depend on the difference $\|\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\|$. This is because in this case, the difference in covariance matrices dominate that of the means and as such represent the discriminant metric that asymptotically matters.
b) Another interesting case to highlight is the one in which $\|\mathbf{\Sigma}_0 - \mathbf{\Sigma}_1\| = o\left(p^{-\frac{1}{2}}\right)$. From Theorem 2 and using (33), it is easy to show that the total classification error converges as

$$\epsilon^{R-QDA} - \Phi\left(-\frac{\boldsymbol{\mu}^T\mathbf{T}\boldsymbol{\mu}}{2\sqrt{p}}\sqrt{\frac{c(1-\gamma^2\phi\tilde{\phi})}{\gamma^2\phi^2\tilde{\phi}}}\right) \to_{prob.} 0, \tag{34}$$

where ϕ, $\tilde{\phi}$ and $\mathbf{T}$ have respectively the same definitions as ϕ_i, $\tilde{\phi}_i$ and $\mathbf{T}_i$ upon dropping the class index i, since quantities associated with class 0 or class 1 can be used interchangeably in the asymptotic regime. It is easy to see that in this case if $\|\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\|^2$ scales slower than $O\left(\sqrt{p}\right)$, classification is asymptotically impossible. This must be contrasted with the results of R-LDA, which provides non-vanishing misclassification rates for $\|\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\| = O(1)$. This means that in this particular setting, R-QDA is asymptotically beaten by R-LDA which achieves perfect classification.
c) When $\|\mathbf{\Sigma}_0 - \mathbf{\Sigma}_1\|_F = O(1)$ occurring for instance when $\|\mathbf{\Sigma}_0 - \mathbf{\Sigma}_1\|_1 = O(p^{-\frac{1}{2}})$ or $\mathbf{\Sigma}_0 - \mathbf{\Sigma}_1$ is of finite rank, and $\|\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\|^2 = O(1)$, then $\overline{b}_i \to b$ where b does not depend on i and as such the misclassification error probability associated with both classes converge respectively to $1-\eta$ and η with η some probability depending solely on the statistics. The total misclassification error associated with R-QDA converges to 0.5.
d) When $n_1 - n_0 \to \infty$, quantities $\overline{\xi}_i$ and $\overline{b}_i$ grow unboundedly as the dimension increases. This unveils that asymptotically, the discriminant score of R-QDA will keep the same sign for all observations. The classifier would thus return the same class regardless of the observation under consideration.

The above remarks should help to draw some hints on when R-LDA or R-QDA should be used. Particularly, if the Frobenius norm of $\mathbf{\Sigma}_0 - \mathbf{\Sigma}_1$ is $O(1)$, using the information on the difference between the class covariance matrices is not recommended. We should rather rely on using the information on the difference between the classes' means, or in other words favoring the use of R-LDA against R-QDA.

IV. General Consistent Estimator of the Testing Error

In the machine learning field, evaluating the performances of algorithms is a crucial step that not only serves to ensure their efficacy but also to properly set the parameters involved in the design thereof, a process known in the machine learning parlance as model selection. The traditional way to evaluate performances consists in devoting a part of the training data to the design of the underlying method whereas performances are tested on the remaining data called testing data, treated as unseen data since they do not intervene in the design step. Among the many existing computational methods that are built on these ideas are the cross-validation [23], [24] and the bootstrap [25], [26] techniques. Despite being widely used in the machine learning community, these methods have the drawback of being computationally expensive and most importantly of relying on mere computations, which does not lead to gain a better understanding of the performances of the underlying algorithm. As far as LDA and QDA classifiers are considered, the results of the previous section allow to gain a deeper understanding of the classification performances with respect to the covariances and means associated with both classes. However, as these results are expressed in terms of the unknown covariances and means, they could not be relied upon to assess the classification performances. In this section, we address this question and provide consistent estimators of the classification performances for both R-LDA and R-QDA classifiers that approximate in probability their asymptotic expressions.

A. R-LDA

The following theorem provides the expression of the class-conditional true error estimator ϵ_i^{R-LDA}, for $i \in \{0,1\}$.

Theorem. 3. *Under Assumptions 1-4, denote*

$$\widehat{\epsilon}_i^{R-LDA} = \Phi\left(\frac{(-1)^{i+1} G\left(\hat{\boldsymbol{\mu}}_i, \hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\mu}}_1, \mathbf{H}\right) + \widehat{\theta}_i + (-1)^i \log\frac{\pi_1}{\pi_0}}{\widehat{\psi}_i \sqrt{D\left(\hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\mu}}_1, \mathbf{H}, \widehat{\boldsymbol{\Sigma}}_i\right)}}\right) \tag{35}$$

where

$$\widehat{\theta}_i = \frac{\frac{1}{n_i}\operatorname{tr}\widehat{\boldsymbol{\Sigma}}_i\mathbf{H}}{1-\frac{\gamma}{n-2}\operatorname{tr}\widehat{\boldsymbol{\Sigma}}_i\mathbf{H}}, \tag{36}$$

$$\widehat{\psi}_i = \frac{1}{1-\frac{\gamma}{n-2}\operatorname{tr}\widehat{\boldsymbol{\Sigma}}_i\mathbf{H}}. \tag{37}$$

Then,

$$\epsilon_i^{R-LDA} - \widehat{\epsilon}_i^{R-LDA} \to_{a.s.} 0.$$

Proof. The proof is postponed to Appendix E. □

Remark. 4. *From Theorem 3, it is easy to recover the general consistent estimator of the conditional classification error constructed in [8]. In particular, in the case where* $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}$*, we have the following*

$$\frac{1}{n_i}\operatorname{tr}\widehat{\boldsymbol{\Sigma}}\mathbf{H} = \frac{1}{\gamma}\left(\frac{p}{n_i} - \frac{1}{n_i}\operatorname{tr}\mathbf{H}\right).$$

Thus, upon dropping the class index i, $\widehat{\theta}$ *is equivalent to* $\widehat{\delta}$ *used in [8].*

B. R-QDA

Based on the deterministic equivalent of the conditional classification error derived in Theorem 2 , we construct a general consistent estimator of ϵ_i^{R-QDA} denoted by $\widehat{\epsilon}_i^{R-QDA}$. The general consistent estimator of the R-QDA misclassification error is given by the following Theorem.

Theorem. 4. *Under Assumptions 5-8, define*

$$\widehat{\epsilon}_i^{R-QDA} = \Phi\left((-1)^i \frac{\widehat{\xi}_i - \widehat{b}_i}{\sqrt{2\widehat{B}_i}}\right), \tag{38}$$

Then,

$$\widehat{\epsilon}_i^{R-QDA} - \epsilon_i^{R-QDA} \to_{prob.} 0.$$

where

$$\widehat{\xi}_i = -\frac{1}{\sqrt{p}}\log\frac{|\mathbf{H}_0|}{|\mathbf{H}_1|} + \frac{(-1)^{i+1}}{\sqrt{p}}\left(\hat{\boldsymbol{\mu}}_0 - \hat{\boldsymbol{\mu}}_1\right)^T \mathbf{H}_{1-i}\left(\hat{\boldsymbol{\mu}}_0 - \hat{\boldsymbol{\mu}}_1\right).$$

$$\widehat{\delta}_i = \frac{1}{\gamma}\frac{\frac{p}{n_i} - \frac{1}{n_i}\operatorname{tr}\mathbf{H}_i}{1 - \frac{p}{n_i} + \frac{1}{n_i}\operatorname{tr}\mathbf{H}_i}.$$

$$\widehat{b}_i = \frac{(-1)^i}{\sqrt{p}}\operatorname{tr}\widehat{\boldsymbol{\Sigma}}_i\mathbf{H}_{1-i} + \frac{(-1)^{i+1}n_i}{\sqrt{p}}\widehat{\delta}_i.$$

$$\widehat{B}_i = \left(1+\gamma\widehat{\delta}_i\right)^4 \frac{1}{p}\operatorname{tr}\widehat{\boldsymbol{\Sigma}}_i\mathbf{H}_i\widehat{\boldsymbol{\Sigma}}_i\mathbf{H}_i - \frac{n_i}{p}\widehat{\delta}_i^2\left(1+\gamma\widehat{\delta}_i\right)^2$$
$$+ \frac{1}{p}\operatorname{tr}\widehat{\boldsymbol{\Sigma}}_i\mathbf{H}_{1-i}\widehat{\boldsymbol{\Sigma}}_i\mathbf{H}_{1-i} - \frac{n_i}{p}\left(\frac{1}{n_i}\operatorname{tr}\widehat{\boldsymbol{\Sigma}}_i\mathbf{H}_{1-i}\right)^2$$
$$- 2\left(1+\gamma\widehat{\delta}_i\right)^2\frac{1}{p}\operatorname{tr}\widehat{\boldsymbol{\Sigma}}_i\mathbf{H}_i\widehat{\boldsymbol{\Sigma}}_i\mathbf{H}_{1-i} + \widehat{\delta}_i\left(1+\gamma\widehat{\delta}_i\right)\frac{2}{p}\operatorname{tr}\widehat{\boldsymbol{\Sigma}}_i\mathbf{H}_{1-i}$$

Proof. See Appendix F. □

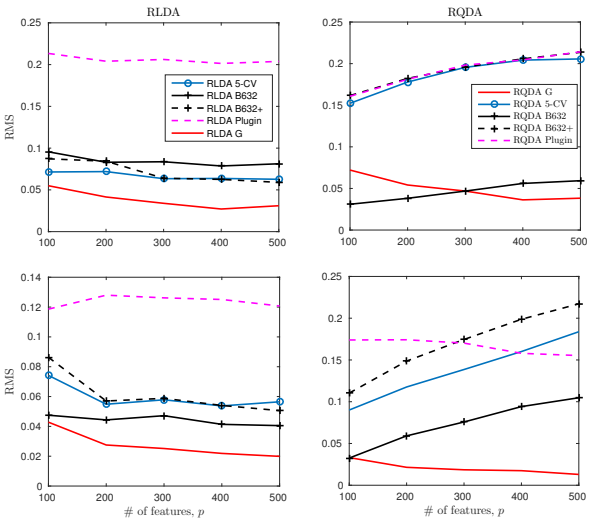

Fig. 1. RMS performance of the proposed general consistent estimators (RLDA G and RQDA G) compared with the benchmark estimation techniques. We consider equal training size ($n_0 = n_1$), $\gamma = 1$ and $[\boldsymbol{\Sigma}_0]_{i,j} = 0.6^{|i-j|}$, $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_0 + 3\mathbf{S}_p$, $\boldsymbol{\mu}_0 = \left[1, \mathbf{0}_{1\times(p-1)}\right]^T$ and $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_0 + \frac{0.8}{\sqrt{p}}\mathbf{1}_{p\times 1}$. The first row treats the case where $n_0 = p/2$ whereas the second row treats the case $n_0 = p$. The testing error is evaluated over a testing set of size 1000 samples for both classes and averaged over 1000 realizations.

C. Validation with synthetic data

Unless otherwise stated, we model the distance between the covariance matrices as $\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_0 = \alpha\mathbf{S}_p$ for some bounded $\alpha \in \mathbb{R}$ where $\mathbf{S}_p = \begin{bmatrix}\mathbf{I}_k & \mathbf{0}_{k\times(p-k)} \\ \mathbf{0}_{(p-k)\times k} & \mathbf{0}_{(p-k)\times(p-k)}\end{bmatrix}$ with $k = \lfloor\sqrt{p}\rfloor$. We validate the results of Theorems 3 and 4 by examining the accuracy of the proposed general consistent estimators in terms of the RMS defined as follows[4]

$$\operatorname{RMS}\left(\widehat{\epsilon}\right) = \sqrt{\operatorname{Bias}\left(\widehat{\epsilon}\right)^2 + \operatorname{var}\left(\widehat{\epsilon} - \epsilon\right)}, \tag{39}$$

where

$$\operatorname{Bias}\left(\widehat{\epsilon}\right) = \mathbb{E}\left[\widehat{\epsilon} - \epsilon\right]. \tag{40}$$

We also compare the proposed general consistent estimator (that we denote by the G-estimator) for both R-LDA and R-QDA with the following benchmark estimation techniques fully described in [27]

- 5-fold cross-validation with 5 repetitions (5-CV).
- 0.632 bootstrap (B632).
- 0.632+ bootstrap (B632+).
- Plugin estimator consisting of replacing the statistics in the deterministic equivalents by their corresponding sample estimates.

In Figures 1 and 2, we observe that the naive plugin estimator has the worst RMS performance for both classifiers in most cases. This is simply explained by the fact that when p and n_i have the same order of magnitude, the sample estimates are inaccurate which leads to a medicocre RMS performance.

[4]Since both synthetic and real data are of finite dimensions, we kept vanishing parts of the estimator when implementing the proposed consistent estimators in our simulations. This was shown to yield a better accuracy of our results.

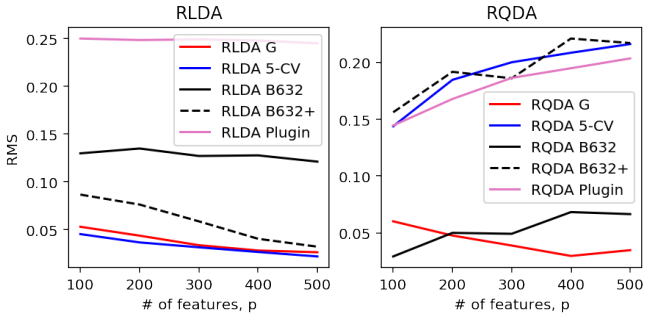

Fig. 2. RMS performance of the proposed general consistent estimators (RLDA G and RQDA G) compared with the benchmark estimation techniques. We consider $n_0 = p/2$, $n_1 = n_0 + \lfloor \sqrt{p} \rfloor$, $\gamma = 1$ and $[\mathbf{\Sigma}_0]_{i,j} = 0.6^{|i-j|}$, $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_0 + 3\mathbf{S}_p$, $\boldsymbol{\mu}_0 = \left[1, \mathbf{0}_{1\times(p-1)}\right]^T$ and $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_0 + \frac{0.8}{\sqrt{p}}\mathbf{1}_p$.

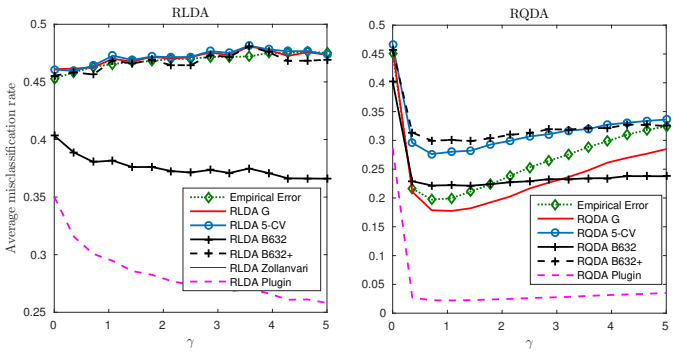

Fig. 3. Average misclassification rate versus the regularization parameter γ. We consider $p = 100$ features with equal training size ($n_0 = n_1 = p$), $[\mathbf{\Sigma}_0]_{i,j} = 0.6^{|i-j|}$, $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_0 + 3\mathbf{S}_p$, $\boldsymbol{\mu}_0 = \left[1, \mathbf{0}_{1\times(p-1)}\right]^T$ and $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_0 + \frac{0.8}{\sqrt{p}}\mathbf{1}_{p\times 1}$. The testing error is evaluated over a testing set of size 1000 samples for both classes and averaged over 1000 realizations.

On another front, it is clear for both settings ($n_i = p/2$ and $n_i = p$) that the proposed G-estimator achieves a suitable RMS performance beating 5-fold cross validation and the bootstrap. In Figure 3, we examine the performance of the different error estimators against the regularization parameter. As shown in the Figure 3, R-LDA is less vulnerable to the choice of γ as compared to R-QDA where the choice of γ tends to have a higher influence on the performance. Also, for both classifiers, the proposed G-estimator is able to track the empirical error and thus permits to predict the optimal regularizer with high accuracy.

V. Experiments with real data

In this section, we examine the performance of the proposed G estimator on the public USPS dataset of handwritten digits [28]. The dataset consists of 7291 training samples of 16×16 grayscale images ($p = 256$ features) and 2007 testing images http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/multiclass.html#usps [5]. First, we examine the RMS performance of the different error estimators on the data for different values of the training size and for different class labels. The RMS is determined by averaging the error over a number of training sets randomly selected from the total training dataset. As shown in Figure 4, the proposed G-estimator gives a good RMS performance especially for R-QDA where it can actually outperform state-of-the art estimators such as cross validation and Bootstrap. Moreover, it is clear that the plugin estimator has a higher RMS performance for most of the considered scenarios.

[5]All the results of this paper can be reproduced using our Julia codes available in https://github.com/KhalilElkhalil/Large-Dimensional-Discriminant-Analysis-Classifiers-with-Random-Matrix-Theory

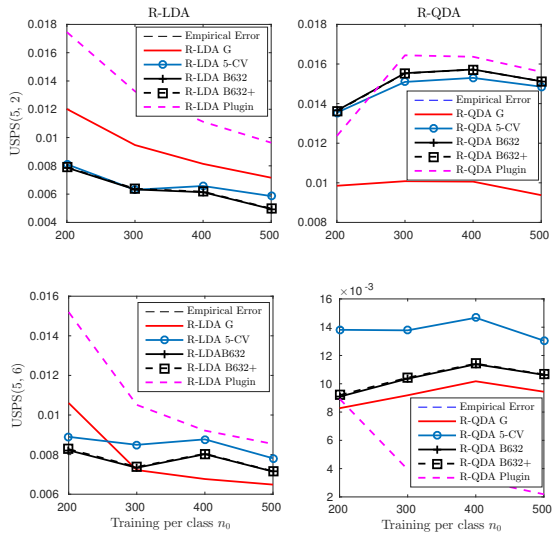

Fig. 4. RMS performance of the proposed general consistent estimators (R-LDA G and R-QDA G) compared with the benchmark estimation techniques. We consider equal training size ($n_0 = n_1$) and $\gamma = 1$. The first row gives the performance for the USPS data with digits (5, 2) whereas the second row considers the digits (5, 6).

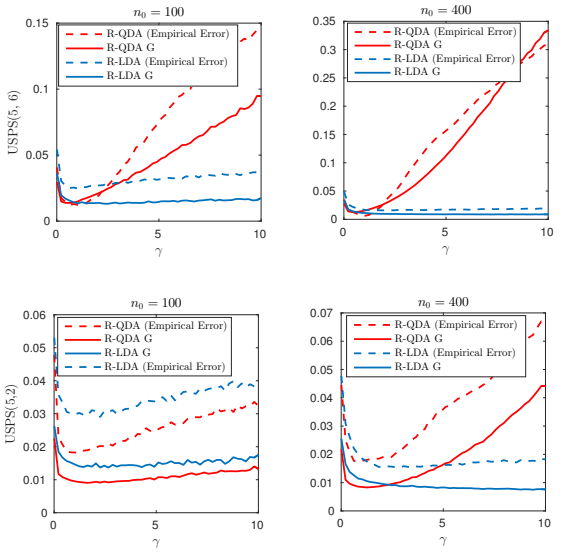

Fig. 5. Average misclassification rate versus the regularization parameter γ of the USPS dataset for different instances of digits and assuming equal training size ($n_0 = n_1$). The solid red line refers to the performance of the proposed G-estimator whereas the dotted black line refers to the empirical error computed using the testing data.

TABLE I
ESTIMATES OF THE OPTIMAL REGULARIZER USING THE TWO-STEP OPTIMIZATION METHOD WITH THEIR CORRESPONDING TESTING ERROR.

	$\widehat{\gamma}^{R-LDA}$	$\epsilon_{testing}^{R-LDA}$	$\widehat{\gamma}^{R-QDA}$	$\epsilon_{testing}^{R-QDA}$
USPS(5, 2), $n_0 = 100$	3.54	0.0307	0.896	0.0111
USPS(5, 2), $n_0 = 400$	20.98	0.0251	1.049	0.0223
USPS(5, 6), $n_0 = 100$	4.753	0.0272	0.567	0.0272
USPS(5, 6), $n_0 = 400$	12.1572	0.01818	0.562	0.009

Now, we turn our attention to finding the *optimal* γ that results in the minimum testing error. Since the construction of the G-estimator is heavily based on the Gaussian assumption of the data, picking the regularizer that minimizes the estimated error using the G-estimator will not necessarily minimize the error computed on the testing data for USPS. One straightforward approach is to compute the testing error for all possible value of γ in the range $(0, \infty)$, then pick the regularizer resulting in the minimum error[6]. Obviously, this approach is far from being practical and is simply unfeasible. Motivated by this issue, we propose a two-stage optimization explained as follows.

A. Two-stage optimization

Although real data are far from being Gaussian, the proposed G-estimator can be used to have a glimpse on the optimal regularizer. More specifically, we can use the G-estimator to determine the interval in which the optimal regularizer is likely to belong, then we perform cross validation (or testing if we have enough testing data) for multiple values of γ inside that interval and finally pick the value that results on the minimum cross-validation error (or testing error). As seen in Figure 5, both the R-LDA and the R-QDA G-estimators are able to mimic the real behavior of the testing error when γ varies for both situations when $n_0 < p$ and $n_0 > p$. Similarly to synthetic data, Figure 5 also shows how R-QDA is vulnerable to the choice of γ which justifies the need to find a *good* regularization parameter γ. In Table I, we provide numerical values for the output of the two-step optimization using a confidence interval $\left(\left(\widehat{\gamma}_G - \frac{2}{\sqrt{p}}\right)^+, \widehat{\gamma}_G + \frac{2}{\sqrt{p}}\right)$ [7] with a uniform grid of 50 points where $\widehat{\gamma}_G$ is a minimizer of the G-estimator built based on the Gaussian assumption.

VI. CONCLUDING REMARKS

In this work, we carried out a performance analysis of the asymptotic misclassification rate for R-LDA and R-QDA based classifiers in the regime where the dimension of the training data and their number grow large with the same pace. By leveraging results from random matrix theory, we identify the growth rate regimes in which R-LDA and R-QDA result in non trivial mis-classification rates. These latter are characterized in the asymptotic regime by closed-form expressions reflecting the impact of the means and covariances of each class on the classification performance. Several insights are drawn from our results, which can guide the practitioners to choose the best classifier according to the setting into consideration. Particularly, we highlight that R-LDA achieves perfect classification rates when the difference in the mean vectors is higher than $O(p^\alpha)$ for $\alpha > 0$. The R-QDA, on the other hand, results in perfect classification when the number of significant eigenvalues of the difference between both covariance matrices scales larger than $O(\sqrt{p})$ or the difference in means is higher than $O(\sqrt{p})$. Such findings reveal a fundamental difference in the way the information about the classes means and covariances are leveraged by both methods. Unlike the R-LDA which tends to leverage only the information about the means, the R-QDA exploits both discriminative statistics, but requires a higher order in the mean difference so that it is reflected in the classification performances.

This work can be extended to study more sophisticated classifiers such as kernel Fisher discriminant analysis [29] or kernel discriminant analysis [30] where data is mapped into a feature space through a non-linear kernel prior to the application of the classifier. The extension is though not trivial since the non-linear mapping of data makes it difficult to examine the misclassification probability in closed-form.

[6] Usually, we perform cross validation or Bootstrap to have an estimate of the error from the training set, but since we have enough testing data we rely on the testing error for the USPS dataset.

[7] $x^+ = \max(x, 0)$, for $x \in \mathbb{R}$.

REFERENCES

[1] K. Elkhalil, A. Kammoun, R. Couillet, T. Y. Al-Naffouri, and M. S. Alouini, "Asymptotic Performance of Regularized Quadratic Discriminant Analysis based Classifiers," in *2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP)*, Sept 2017, pp. 1–6.

[2] P. Devijver and J. Kittler, *Pattern Recognition: A Statistical Approach*. Prentice Hall, 1982.

[3] D. G. S. Richard O. Duda, Peter E. Hart, *Pattern Classification*. Wiley, 2000.

[4] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[5] J. Friedman, T. Hastie, and R. Tibshirani, *The Elements of Statistical Learning*. Springer, 2009.

[6] A. Ghodsi, "Dimensionality Reduction A Short Tutorial," 2005.

[7] I. Fodor, "A survey of dimension reduction techniques," Tech. Rep., 2002.

[8] A. Zollanvari and E. R. Dougherty, "Generalized Consistent Error Estimator of Linear Discriminant Analysis," *IEEE Transactions on Signal Processing*, vol. 63, no. 11, pp. 2804–2814, June 2015.

[9] H. R. McFarland and D. S. P. Richards, "Exact Misclassification Probabilities for Plug-In Normal Quadratic Discriminant Functions," *Journal of Multivariate Analysis*, vol. 82, p. 299–330, 2002.

[10] J. Shao, Y. Wang, X. Deng, and S. Wang, "Sparse linear discriminant analysis by thresholding for high dimensional data," *The Annals of Statistics*, vol. 39, no. 2, pp. 1241–1265, 2011. [Online]. Available: http://www.jstor.org/stable/29783672

[11] Q. Li and J. Shao, "Sparse Quadratic Discriminant Analysis For High Dimensional Data," *Statistica Sinica*, vol. 25, pp. 457–473, 2015.

[12] V. Girko, *Statistical Analysis of Observations of Increasing Dimension*. Springer, 1995.

[13] J. W. Silverstein and Z. D. Bai, "On the Empirical Distribution of Eigenvalues of a Class of Large Dimensional Random Matrices," *Journal of Multivariate Analysis*, vol. 54, pp. 175–192, May 2002.

[14] W. Hachem, O. Khorunzhiy, P. Loubaton, J. Najim, and L. Pastur, "A New Approach for Mutual Information Analysis of Large Dimensional Multi-Antenna Channels," *IEEE Transactions on Information Theory*, vol. 54, no. 9, pp. 3987–4004, Sept 2008.

[15] R. Couillet and M. Debbah, *Random Matrix Methods for Wireless Communications*. Cambridge University Press, 2011.

[16] F. Benaych-Georges and R. Couillet, "Spectral Analysis of the Gram Matrix of Mixture Models," *ESAIM: Probability and Statistics*, vol. 20, pp. 217–237, 2016.

[17] v. Raudys and D. M. Young, "Results in statistical discriminant analysis: A review of the former soviet union literature," *J. Multivar. Anal.*, vol. 89, no. 1, pp. 1–35, Apr. 2004. [Online]. Available: http://dx.doi.org/10.1016/S0047-259X(02)00021-0

[18] H. Watanabe, M. Hyodo, T. Seo, and T. Pavlenko, "Asymptotic properties of the misclassification rates for Euclidean Distance Discriminant rule in high-dimensional data," *Journal of Multivariate Analysis*, vol. 78, pp. 234–244, 2015.

[19] G. McLachlan, *Discriminant Analysis and Statistical Pattern Recognition*. Wiley, 2009.

[20] U. M. Braga-Neto, A. Zollanvari, and E. R. Dougherty, "Cross-Validation under Separate Sampling: Strong Bias and How to Correct it ," *Bioinformatics*, vol. 30, no. 23, p. 3349–3355, 2014.

[21] P. W. Wahl and R. A. Kronmal, "Discriminant Functions when Covariances are Unequal and Sample Sizes are Moderate," *Biometrics*, vol. 33, no. 20, pp. 479–484, 1977.

[22] P. Billingsley, *Probability and measure*. John Wiley & Sons, 2008.

[23] S. Geisser, "Discrimination, Allocatory and Separatory, Linear Aspects," *in Classification and Clustering, ed. J. Van Ryzin, New York: Academic Press*, pp. 301–330, 1977.

[24] P. A. Lachenbruch, *Discriminant Analysis*. New York: Hafner Press, 1975.

[25] B. Efron, "Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation," *Journal of the American Statistical Association*, vol. 78, pp. 316–331, 1983.

[26] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. CRC press, 1994.

[27] E. R. Dougherty, C. S. Hua, B. Hanczar, and U. M. Braga-Neto., "Performance of Error Estimators for Classification," *Current Bioinformatics*, vol. 5, pp. 53–67, 2010.

[28] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-Based Learning Applied To Document Recognition," *Proceedings of the IEEE*, vol. 86, no. 11, p. 2278–2324, 1998.

[29] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K.-R. Muller, "FISHER DISCRIMINANT ANALYSIS WITH KERNELS," *Neural Networks for Signal Processing: Proceedings of the 1999 IEEE Signal Processing Society Workshop*, 1999.

[30] C. H. PARK and H. PARK, "NONLINEAR DISCRIMINANT ANALYSIS USING KERNEL FUNCTIONS AND THE GENERALIZED SINGULAR VALUE DECOMPOSITION," *SIAM J. MATRIX ANAL. APPL.*, vol. 27, no. 1, 2005.

[31] R. J. Serfling, *Approximations Theorems of Mathematical Statistics*. John Wiley & Sons, 2002.

[32] H. D. Lee, "On Some Matrix Inequalities," *Korean J. Math.*, no. 4, p. 565–571, 2008.

APPENDIX A
PROOF OF THEOREM 1

A. Notations

Through this appendix, the following notations are used. For $i \in \{0,1\}$, we let $\mathbf{X}_i \in \mathbb{R}^{p\times n_i}$ the matrix of n_i observations associated with class i. Thus, there exists $\mathbf{Y}_i = \mathbf{\Sigma}_i^{1/2}\mathbf{Z}_i$ such that $\mathbf{X}_i = \mathbf{\Sigma}_i^{\frac{1}{2}}\mathbf{Z}_i + \boldsymbol{\mu}_i\mathbf{1}_{n_i}^T$ where $\mathbf{1}_{n_i} \in \mathbb{R}^{n_i}$ is the vector of all ones, and $\mathbf{Z}_i = [\mathbf{z}_{i,1}, \cdots, \mathbf{z}_{i,n_i}] \in \mathbb{R}^{p\times n_i}$ where $\mathbf{z}_{i,j}$ are independent random vectors with standard multivariate Gaussian distribution. We define the following resolvent matrix:

$$\mathbf{Q}(z) = \left(\frac{\mathbf{Y}_0\mathbf{Y}_0^T}{p} + \frac{\mathbf{Y}_1\mathbf{Y}_1^T}{p} - z\mathbf{I}_p\right)^{-1} \tag{41}$$

the behavior of which has been extensively studied in [16, Proposition 5]. Particularly, it was shown that under assumptions 1, 2 and 3, $\mathbf{Q}(z)$ is equivalent to a deterministic matrix $\bar{\mathbf{Q}}(z)$ or $\mathbf{Q}(z) \leftrightarrow \bar{\mathbf{Q}}(z)$ in the sense that

$$\begin{aligned} &\frac{1}{p}\operatorname{tr}\mathbf{M}\left(\mathbf{Q}(z) - \bar{\mathbf{Q}}(z)\right) \to_{prob.} 0. \\ &\mathbf{u}^T\left(\mathbf{Q}(z) - \bar{\mathbf{Q}}(z)\right)\mathbf{v} \to_{prob.} 0, \end{aligned} \tag{42}$$

for all deterministic matrices $\mathbf{M}$ of bounded spectral norms and all deterministic vectors $\mathbf{u}$ and $\mathbf{v}$ of bounded euclidean norms. Moreover, it has been shown in [16, Proposition 6] that

$$\mathbf{Q}(z)\,\mathbf{\Sigma}_i\mathbf{Q}(z) \leftrightarrow \widetilde{\mathbf{Q}}_i(z), \text{ for } i \in \{0,1\}. \tag{43}$$

where $\widetilde{\mathbf{Q}}_i$ is given in (18). Based on the results of (42) and (43), we successively prove (22) and (23).

B. Proof of (22)

With the aforementioned notations at hand, it is easy to show that $\widehat{\mathbf{\Sigma}}_i$ can be expressed as

$$\widehat{\mathbf{\Sigma}}_i = \frac{1}{n_i - 1}\left(\mathbf{Y}_i\mathbf{Y}_i^T - \mathbf{Y}_i\frac{\mathbf{1}_{n_i}\mathbf{1}_{n_i}^T}{n_i}\mathbf{Y}_i^T\right).$$

Let $\frac{\mathbf{1}_i\mathbf{1}_i^T}{n_i} = \mathbf{O}_i\mathbf{E}_i\mathbf{O}_i^T$, be the eigenvalue decomposition of $\frac{\mathbf{1}_i\mathbf{1}_i^T}{n_i}$ where $\mathbf{E}_i = \operatorname{diag}\left(\left[1, \mathbf{0}_{(n_i-1)\times 1}\right]\right)$ and $\mathbf{O}_i$ is a $n_i \times n_i$ orthogonal matrix with first column $\frac{1}{\sqrt{n_i}}\mathbf{1}_{n_i}$. Let $\tilde{\mathbf{Y}}_i = \mathbf{Y}_i\mathbf{O}_i$. Hence

$$\begin{aligned} \widehat{\mathbf{\Sigma}}_i &= \frac{1}{n_i - 1}\mathbf{Y}_i\mathbf{O}_i\mathbf{O}_i^T\mathbf{Y}_i^T - \frac{1}{n_i - 1}\mathbf{Y}_i\mathbf{O}_i\mathbf{E}_i\mathbf{O}_i^T\mathbf{Y}_i^T \\ &= \frac{1}{n_i - 1}\widetilde{\mathbf{Y}}_i\widetilde{\mathbf{Y}}_i^T - \frac{1}{n_i - 1}\widetilde{\mathbf{y}}_{i,1}\widetilde{\mathbf{y}}_{i,1}^T, \end{aligned} \tag{44}$$

where $\widetilde{\mathbf{y}}_{i,1}$ being the first column of $\widetilde{\mathbf{Y}}_i$. Since the Gaussian distribution is invariant to multiplication by a unitary matrix, $\tilde{\mathbf{Y}}_i$ has the same distribution as $\mathbf{Y}_i$. and as such matrix $\mathbf{H}$ can be expressed as:

$$\mathbf{H} = \left[\mathbf{I}_p + \frac{\gamma}{n-2}\overline{\mathbf{Y}}_0\overline{\mathbf{Y}}_0^T + \frac{\gamma}{n-2}\overline{\mathbf{Y}}_1\overline{\mathbf{Y}}_1^T\right]^{-1}, \tag{45}$$

where $\overline{\mathbf{Y}}_0$ and $\overline{\mathbf{Y}}_1$ are obtained by respectively removing the first column of $\widetilde{\mathbf{Y}}_0$ and $\widetilde{\mathbf{Y}}_1$. Then, the following relation holds for $z = -\frac{n}{p\gamma}$

$$\mathbf{H} = -z\mathbf{Q}(z) + O_{\|.\|}(p^{-1}),$$

where $O_{\|.\|}(p^{-1})$ refers to a matrix whose spectral norm is $O(p^{-1})$. We are now ready to handle the term $G(\boldsymbol{\mu}_i, \hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\mu}}_1, \mathbf{H})$. To start, we first express it as

$$\begin{aligned} G(\boldsymbol{\mu}_i, \hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\mu}}_1, \mathbf{H}) = &\left(\frac{(-1)^i}{2}\boldsymbol{\mu}^T - \frac{1}{2n_0}\mathbf{1}^T\mathbf{Y}_0 - \frac{1}{2n_1}\mathbf{1}_{n_1}^T\mathbf{Y}_1\right) \\ &\times \mathbf{H}\left(\frac{1}{n_0}\mathbf{Y}_0\mathbf{1}_{n_0} - \frac{1}{n_1}\mathbf{Y}_1\mathbf{1}_{n_1} + \boldsymbol{\mu}\right). \end{aligned}$$

and thus can be expanded as

$$
\begin{aligned}
G\left(\boldsymbol{\mu}_i, \hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\mu}}_1, \mathbf{H}\right) &= \frac{(-1)^i}{2}\boldsymbol{\mu}^T\mathbf{H}\boldsymbol{\mu} + \frac{(-1)^i}{2n_0}\boldsymbol{\mu}^T\mathbf{H}\mathbf{Y}_0^T\mathbf{1}_{n_0} \\
&+ \frac{(-1)^{i+1}}{2n_1}\boldsymbol{\mu}^T\mathbf{H}\mathbf{Y}_1^T\mathbf{1}_{n_1} - \frac{1}{2n_0}\boldsymbol{\mu}^T\mathbf{H}\mathbf{Y}_0\mathbf{1}_{n_0} \\
&- \frac{1}{2n_0^2}\mathbf{1}_{n_0}^T\mathbf{Y}_0^T\mathbf{H}\mathbf{Y}_0\mathbf{1}_{n_0} + \frac{1}{2n_0n_1}\mathbf{1}_{n_0}^T\mathbf{Y}_0^T\mathbf{H}\mathbf{Y}_1\mathbf{1}_{n_1} \\
&- \frac{1}{2n_1}\boldsymbol{\mu}^T\mathbf{H}\mathbf{Y}_1\mathbf{1}_{n_1} - \frac{1}{2n_0n_1}\mathbf{1}_{n_0}^T\mathbf{Y}_0^T\mathbf{H}\mathbf{Y}_1\mathbf{1}_{n_1} \\
&+ \frac{1}{2n_1^2}\mathbf{1}_{n_1}^T\mathbf{Y}_1^T\mathbf{H}\mathbf{Y}_1\mathbf{1}_{n_1}.
\end{aligned}
$$

It follows from (45) that $\tilde{\mathbf{y}}_{0,1} = \frac{1}{\sqrt{n_0}}\mathbf{Y}_0\mathbf{1}$ and $\tilde{\mathbf{y}}_{1,1} = \frac{1}{\sqrt{n_1}}\mathbf{Y}_1\mathbf{1}$ are independent of $\mathbf{H}$. The following convergence holds thus true

$$
\begin{aligned}
&\frac{1}{n_0}\boldsymbol{\mu}^T\mathbf{H}\mathbf{Y}_0\mathbf{1}_{n_0} \to_{a.s.} 0. \\
&\frac{1}{n_1}\boldsymbol{\mu}^T\mathbf{H}\mathbf{Y}_1\mathbf{1}_{n_1} \to_{a.s.} 0. \\
&\frac{1}{n_0n_1}\mathbf{1}_{n_0}^T\mathbf{Y}_0^T\mathbf{H}\mathbf{Y}_1\mathbf{1}_{n_1} \to_{a.s.} 0.
\end{aligned}
$$

On the other hand, we have

$$
\boldsymbol{\mu}^T\mathbf{H}\boldsymbol{\mu} = -z\boldsymbol{\mu}^T\mathbf{Q}\left(z\right)\boldsymbol{\mu} + o(1). \tag{46}
$$

and thus from (42)

$$
\boldsymbol{\mu}^T\mathbf{H}\boldsymbol{\mu} + z\boldsymbol{\mu}^T\bar{\mathbf{Q}}\left(z\right)\boldsymbol{\mu} \to_{prob.} 0. \tag{47}
$$

Moreover,

$$
\frac{1}{n_i^2}\mathbf{1}_{n_i}^T\mathbf{Y}_i^T\mathbf{H}\mathbf{Y}_i\mathbf{1}_{n_i} = \frac{1}{n_i}\widetilde{\mathbf{y}}_{i,1}^T\mathbf{H}\widetilde{\mathbf{y}}_{i,1}
$$

Again, from the independence of $\widetilde{\mathbf{y}}_i$ and $\mathbf{H}$ and the application of the trace Lemma [15, Theorem 3.7] it follows that:

$$
\frac{1}{n_i^2}\mathbf{1}_{n_i}\mathbf{Y}_i^T\mathbf{H}\mathbf{Y}_i\mathbf{1}_{n_i} - \frac{1}{n_i}\operatorname{tr}\boldsymbol{\Sigma}_i\mathbf{H} \to_{a.s.} 0.
$$

which gives using (42),

$$
\frac{1}{n_i^2}\mathbf{1}_{n_i}\mathbf{Y}_i^T\mathbf{H}\mathbf{Y}_i\mathbf{1}_{n_i} + \frac{z}{n_i}\operatorname{tr}\boldsymbol{\Sigma}_i\bar{\mathbf{Q}}\left(z\right) \to_{prob.} 0.
$$

This completes the proof of (22).

C. Proof of (23)

Using the notations employed in the proof of (22), $D(\overline{\mathbf{x}}_0, \overline{\mathbf{x}}_1, \mathbf{H}, \boldsymbol{\Sigma}_i)$ can be expressed as:

$$
\begin{aligned}
D(\hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\mu}}_1, \mathbf{H}, \boldsymbol{\Sigma}_i) = &\left(\boldsymbol{\mu}^T + \mathbf{1}_{n_0}^T\frac{\mathbf{Y}_0}{n_0} - \mathbf{1}_{n_1}^T\frac{\mathbf{Y}_1}{n_1}\right)\mathbf{H}\boldsymbol{\Sigma}_i\mathbf{H} \\
&\times\left(\boldsymbol{\mu} + \mathbf{Y}_0\frac{\mathbf{1}_{n_0}}{n_0} - \mathbf{Y}_1\frac{\mathbf{1}_{n_1}}{n_1}\right).
\end{aligned} \tag{48}
$$

As in the proof (22), from the independence of $\frac{1}{n_1}\mathbf{Y}_1\mathbf{1}$ and $\frac{1}{n_0}\mathbf{Y}_0\mathbf{1}$ of $\mathbf{H}$, it is easy to see that the cross-products in (48) will converge to zero almost surely. We thus have:

$$
\begin{aligned}
D\left(\hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\mu}}_1, \mathbf{H}, \boldsymbol{\Sigma}_i\right) = &\boldsymbol{\mu}^T\mathbf{H}\boldsymbol{\Sigma}_i\mathbf{H}\boldsymbol{\mu} + \frac{1}{n_0^2}\mathbf{1}_{n_0}^T\mathbf{Y}_0^T\mathbf{H}\boldsymbol{\Sigma}_i\mathbf{H}\mathbf{Y}_0\mathbf{1}_{n_0} \\
&+ \frac{1}{n_1^2}\mathbf{1}_{n_1}^T\mathbf{Y}_1^T\mathbf{H}\boldsymbol{\Sigma}_i\mathbf{H}\mathbf{Y}_1\mathbf{1}_{n_1}.
\end{aligned}
$$

Finally, we use (43) to obtain

$$
\begin{aligned}
&\boldsymbol{\mu}^T\mathbf{H}\boldsymbol{\Sigma}_i\mathbf{H}\boldsymbol{\mu} - z^2\boldsymbol{\mu}^T\widetilde{\mathbf{Q}}_i\left(z\right)\boldsymbol{\mu} \to_{prob.} 0 \\
&\frac{1}{n_j^2}\mathbf{1}_{n_j}^T\mathbf{Y}_j^T\mathbf{H}\boldsymbol{\Sigma}_i\mathbf{H}\mathbf{Y}_j\mathbf{1}_{n_j} - \frac{z^2}{n_j}\operatorname{tr}\boldsymbol{\Sigma}_j\widetilde{\mathbf{Q}}_i\left(z\right) \to_{prob.} 0,\ j = 1 - i.
\end{aligned}
$$

which completes the proof of (23).

Appendix B
Proof of Proposition 1

To reduce the amount of notations, we drop the class subscript i. In all the proof $\mathbf{B}$ plays the role of $\mathbf{B}_i$ and $\mathbf{y}$ plays the role of $\mathbf{y}_i$, for $i \in \{0, 1\}$. To begin with, let $\mathbf{B} = \mathbf{U}_b\mathbb{B}\mathbf{U}_b^T$ be the eigenvalue decomposition of $\mathbf{B}$, so that $\boldsymbol{z}^T\mathbf{B}\boldsymbol{z} + 2\boldsymbol{z}^T\mathbf{r}$ has the same distribution as:

$$
g\left(\mathbf{z}\right) \triangleq \sum_{j=1}^{p}\left(\alpha_j z_j^2 + 2z_j\tilde{r}_j\right),
$$

where $\tilde{\mathbf{y}} = \mathbf{U}_b\mathbf{r}$, α_i diagonal elements of $\mathbb{B}$ and z_j and $\tilde{r}_j$ are respectively the jth entries of $\boldsymbol{\omega}$ and $\tilde{\mathbf{y}}$. Let $\boldsymbol{\Psi} = [\mathbf{X}_0, \mathbf{X}_1]$ be the observations associated with class 0 and 1. Then, conditioning on $\boldsymbol{\Psi}$, $g\left(\boldsymbol{z}\right)$ is the sum of independent but not identically distributed r.v's. $q_j = \alpha_j z_j^2 + 2z_j\tilde{r}_j$. To prove the CLT, we resort to the Lyapunov CLT Theorem, [22, Theorem 27.3]. We first calculate the mean and the variance of r_j conditioned on $\boldsymbol{\Psi}$

$$
\begin{aligned}
\mathbb{E}\left[q_j|\boldsymbol{\Psi}\right] &= \alpha_j \\
\operatorname{var}\left[q_j|\boldsymbol{\Psi}\right] &= \sigma_j^2 = 2\alpha_j^2 + 4\tilde{r}_j^2.
\end{aligned}
$$

Define the total variance s_p^2 as

$$
s_p^2 = \sum_{j=1}^{p}\sigma_j^2 = 2\operatorname{tr}\mathbf{B}^2 + 4\tilde{\mathbf{r}}^T\tilde{\mathbf{r}}. \tag{49}
$$

To prove the CLT, it suffices to check the Lyapunov's condition. Under the setting of Proposition 1,

$$
\begin{aligned}
&\lim_{p\to\infty}\frac{1}{s_p^4}\sum_{j=1}^{p}\mathbb{E}\left[\left|q_j - \alpha_j\right|^4||\boldsymbol{\Psi}\right] \\
&= \lim_{p\to\infty}\frac{\sum_{j=1}^{p}60\alpha_j^4 + 240\alpha_j^2\tilde{r}_j^2 + 48\tilde{r}_j^4}{\left(2\operatorname{tr}\mathbf{B}^2 + 4\tilde{\mathbf{r}}^T\tilde{\mathbf{r}}\right)^2} \\
&\leq \lim_{p\to\infty}\frac{60/p^2\operatorname{tr}\mathbf{B}^2 + 240/p^2\operatorname{tr}\mathbf{B}^2\|\tilde{\mathbf{r}}\|_2^2 + 48/p^2\|\tilde{\mathbf{r}}\|_2^4}{\left(2/p\operatorname{tr}\mathbf{B}^2 + 4/p\|\tilde{\mathbf{r}}\|_2^2\right)^2}.
\end{aligned}
$$

Appendix C
Proof of Theorem 2

The proof consists in showing the following convergences

$$
\frac{1}{\sqrt{p}}\xi_i - \overline{\xi}_i \to_{prob.} 0. \tag{50}
$$

$$
\frac{1}{\sqrt{p}}\operatorname{tr}\mathbf{B}_i - \overline{b_i} \to_{a.s.} 0. \tag{51}
$$

$$
\frac{1}{p}\operatorname{tr}\mathbf{B}_i^2 - \overline{B_i} \to_{a.s.} 0. \tag{52}
$$

$$
\frac{1}{p}\mathbf{r}_i^T\mathbf{r}_i \to_{a.s.} 0. \tag{53}
$$

and establishing that the condition in 1 holds with probability 1. We will prove sequentially equations (50)-(53).

A. Proof of (50)

Using the simplified expression of $\widehat{\mathbf{\Sigma}}_i$ in (44), we can write

$$\mathbf{H}_i = \left(\mathbf{I}_p + \frac{\gamma}{n_i - 1} \mathbf{Y}_i \mathbf{Y}_i^T - \frac{\gamma}{n_i - 1} \mathbf{y}_i \mathbf{y}_i^T \right)^{-1}.$$

Recall that $\frac{1}{\sqrt{p}} \xi_i$ writes as

$$\begin{aligned} \frac{1}{\sqrt{p}} \xi_i = & -\frac{1}{\sqrt{p}} \log \frac{|\mathbf{H}_0|}{|\mathbf{H}_1|} + \frac{1}{\sqrt{p}} (\boldsymbol{\mu}_i - \hat{\boldsymbol{\mu}}_0)^T \mathbf{H}_0 (\boldsymbol{\mu}_i - \hat{\boldsymbol{\mu}}_0) \\ & - \frac{1}{\sqrt{p}} (\boldsymbol{\mu}_i - \hat{\boldsymbol{\mu}}_1)^T \mathbf{H}_1 (\boldsymbol{\mu}_i - \hat{\boldsymbol{\mu}}_1) + \frac{2}{\sqrt{p}} \log \frac{\pi_1}{\pi_0}. \end{aligned}$$

Under Assumption 7, Matrix $\mathbf{H}_i$ follows the model in [14]. According to [14, Theorem 1],

$$\frac{1}{p} \log |\mathbf{H}_i| - \frac{1}{p} \left(\log |\mathbf{T}_i| - n_i \log (1 + \gamma \delta_i) + \gamma \frac{n_i \delta_i}{1 + \gamma \delta_i} \right) \to_{a.s.} 0. \tag{54}$$

The convergence holds with rate $O(p^{-1})$ hence,

$$\frac{1}{\sqrt{p}} \log |\mathbf{H}_i| - \frac{1}{\sqrt{p}} \left(\log |\mathbf{T}_i| - n_i \log (1 + \gamma \delta_i) + \gamma \frac{n_i \delta_i}{1 + \gamma \delta_i} \right) \to_{prob.} 0.$$

and

$$\begin{aligned} & \frac{1}{\sqrt{p}} \left((\boldsymbol{\mu}_i - \hat{\boldsymbol{\mu}}_0)^T \mathbf{H}_0 (\boldsymbol{\mu}_i - \hat{\boldsymbol{\mu}}_0) - (\boldsymbol{\mu}_i - \hat{\boldsymbol{\mu}}_1)^T \mathbf{H}_1 (\boldsymbol{\mu}_i - \hat{\boldsymbol{\mu}}_1) \right) \\ & - \frac{(-1)^{i+1}}{\sqrt{p}} \boldsymbol{\mu}^T \mathbf{T}_{1-i} \boldsymbol{\mu} \to_{prob.} 0. \end{aligned}$$

B. Proof of (51)

From [14], we know that

$$\frac{1}{p} \operatorname{tr} \mathbf{\Sigma}_i \mathbf{H}_i - \frac{1}{p} \operatorname{tr} \mathbf{\Sigma}_i \mathbf{T}_i \to_{a.s.} 0 \tag{55}$$

where the above convergence holds with rate $O(p^{-1})$.

Thus,

$$\frac{1}{\sqrt{p}} \operatorname{tr} \mathbf{B}_i - \frac{1}{\sqrt{p}} \operatorname{tr} \mathbf{\Sigma}_i (\mathbf{T}_1 - \mathbf{T}_0) \to_{prob.} 0.$$

C. Proof of (52)

To prove (52), we need the following lemma, the proof of which is omitted since it follows from the techniques established in [14]

Lemma. 1. *Let* $\mathbf{A}$ *be a matrix with uniformly bounded spectral norm. Then, for* $i \in \{0, 1\}$ *the following convergence holds true*

$$\begin{aligned} & \frac{1}{n_i} \operatorname{tr} \mathbf{A} \mathbf{H}_i \mathbf{A} \mathbf{H}_i - \left[\frac{1}{n_i} \operatorname{tr} \mathbf{T}_i^2 \mathbf{A}^2 + \frac{\gamma^2 \tilde{\phi}_i}{1 - \gamma^2 \phi_i \tilde{\phi}_i} \left(\frac{1}{n_i} \operatorname{tr} \mathbf{A} \mathbf{\Sigma}_i \mathbf{T}_i^2 \right)^2 \right] \\ & \to_{a.s.} 0. \end{aligned} \tag{56}$$

With the above Lemma at hand, we are now ready to handle $\frac{1}{p} \operatorname{tr} \mathbf{B}_i^2$.

$$\begin{aligned} \frac{1}{p} \operatorname{tr} \mathbf{B}_i^2 &= \frac{1}{p} \operatorname{tr} \mathbf{\Sigma}_i (\mathbf{H}_1 - \mathbf{H}_0) \mathbf{\Sigma}_i (\mathbf{H}_1 - \mathbf{H}_0) \\ &= \frac{1}{p} \operatorname{tr} \mathbf{\Sigma}_i \mathbf{H}_1 \mathbf{\Sigma}_i \mathbf{H}_1 + \frac{1}{p} \operatorname{tr} \mathbf{\Sigma}_i \mathbf{H}_0 \mathbf{\Sigma}_i \mathbf{H}_0 - \frac{2}{p} \operatorname{tr} \mathbf{\Sigma}_i \mathbf{H}_0 \mathbf{\Sigma}_i \mathbf{H}_1. \end{aligned}$$

By product of Lemma 1, we can easily get

$$\frac{1}{p} \operatorname{tr} \mathbf{\Sigma}_i \mathbf{H}_i \mathbf{\Sigma}_i \mathbf{H}_i - \frac{c \phi_i}{1 - \gamma^2 \phi_i \tilde{\phi}_i} \to_{a.s.} 0 \tag{57}$$

$$\begin{aligned} & \frac{1}{p} \operatorname{tr} \mathbf{\Sigma}_i \mathbf{H}_{1-i} \mathbf{\Sigma}_i \mathbf{H}_{1-i} \\ & - \left[\frac{1}{p} \operatorname{tr} \mathbf{\Sigma}_i^2 \mathbf{T}_{1-i}^2 + \frac{\gamma^2 c \tilde{\phi}_{1-i} \left(\frac{1}{n_i} \operatorname{tr} \mathbf{\Sigma}_i \mathbf{\Sigma}_{1-i} \mathbf{T}_{1-i}^2 \right)^2}{1 - \gamma^2 \phi_{1-i} \tilde{\phi}_{1-i}} \right] \\ & \to_{a.s.} 0. \end{aligned} \tag{58}$$

Finally, it is straightforward to obtain

$$\frac{1}{p} \operatorname{tr} \mathbf{\Sigma}_i \mathbf{H}_1 \mathbf{\Sigma}_i \mathbf{H}_0 - \frac{1}{p} \operatorname{tr} \mathbf{\Sigma}_i \mathbf{T}_1 \mathbf{\Sigma}_i \mathbf{T}_0 \to_{a.s.} 0. \tag{59}$$

This completes the proof of (52).

D. Proof of (53)

Let $i \in \{0, 1\}$. Then, one can see that:

$$\frac{1}{p} \mathbf{r}_i^T \mathbf{r}_i - \frac{1}{p} \boldsymbol{\mu}^T \mathbf{H}_{1-i} \mathbf{\Sigma}_i \mathbf{H}_{1-i} \boldsymbol{\mu} \to_{prob.} 0, \tag{60}$$

where by Assumptions 6 and 7

$$\frac{1}{p} \boldsymbol{\mu}^T \mathbf{H}_{1-i} \mathbf{\Sigma}_i \mathbf{H}_{1-i} \boldsymbol{\mu} = O\left(\frac{1}{\sqrt{p}} \right).$$

Finally, by applying the continous mapping theorem [31], we complete the proof of Theorem 2.

Now, to conclude we need to check that the condition in Proposition 1 holds with probability 1. This can be easily seen by replacing in (27), $\frac{1}{p} \operatorname{tr} \mathbf{B}^2$ by its deterministic equivalent and noting that it has order $O(1)$.

Appendix D
Proof of Proposition 2

To prove Proposition 2, it suffices to show

$$\frac{1}{\sqrt{p}} \left(\frac{n_0 \delta_0}{1 + \gamma \delta_0} - \frac{n_1 \delta_1}{1 + \gamma \delta_1} \right) = O(1) \tag{61}$$

$$\frac{1}{\sqrt{p}} \left(n_1 \log (1 + \gamma \delta_1) - n_0 \log (1 + \gamma \delta_0) \right) = O(1) \tag{62}$$

$$\frac{1}{\sqrt{p}} \left(\log |\mathbf{T}_0| - \log |\mathbf{T}_1| \right) = O(1) \tag{63}$$

Relying on Assumption 5, we can assume $n_0 = n_1 = \frac{n}{2}$. To begin, we first note that

$$\frac{\delta_0}{1 + \gamma \delta_0} - \frac{\delta_1}{1 + \gamma \delta_1} = \frac{1}{(1 + \gamma \delta_0)(1 + \gamma \delta_1)} (\delta_0 - \delta_1).$$

On the other hand,

$$\begin{aligned} \delta_0 - \delta_1 &= \frac{2}{n} \operatorname{tr} \mathbf{\Sigma}_0 \mathbf{T}_0 - \frac{2}{n} \operatorname{tr} \mathbf{\Sigma}_1 \mathbf{T}_1 \\ &= \frac{2}{n} \operatorname{tr} \mathbf{\Sigma}_0 (\mathbf{T}_0 - \mathbf{T}_1) + \frac{2}{n} \operatorname{tr} (\mathbf{\Sigma}_0 - \mathbf{\Sigma}_1) \mathbf{T}_1. \end{aligned}$$

Recall that for invertible square matrices $\mathbf{A}$ and $\mathbf{B}$, we have the *resolvent identity* given by

$$\mathbf{A}^{-1} - \mathbf{B}^{-1} = \mathbf{A}^{-1} (\mathbf{B} - \mathbf{A}) \mathbf{B}^{-1}. \tag{64}$$

Thus,

$$\begin{aligned}
&\delta_0-\delta_1\\
&=\frac{2}{n}\operatorname{tr}\boldsymbol{\Sigma}_0\mathbf{T}_0\left(\mathbf{T}_1^{-1}-\mathbf{T}_0^{-1}\right)\mathbf{T}_1+\frac{2}{n}\operatorname{tr}\left(\boldsymbol{\Sigma}_0-\boldsymbol{\Sigma}_1\right)\mathbf{T}_1\\
&=\frac{2}{n}\operatorname{tr}\boldsymbol{\Sigma}_0\mathbf{T}_0\left(\frac{\gamma}{1+\gamma\delta_1}\boldsymbol{\Sigma}_1-\frac{\gamma}{1+\gamma\delta_0}\boldsymbol{\Sigma}_0\right)\mathbf{T}_1+\frac{2}{n}\operatorname{tr}\left(\boldsymbol{\Sigma}_0-\boldsymbol{\Sigma}_1\right)\mathbf{T}_1\\
&=\frac{2\gamma}{n}\operatorname{tr}\boldsymbol{\Sigma}_0\mathbf{T}_0\left(\frac{\boldsymbol{\Sigma}_1}{1+\gamma\delta_1}-\frac{\boldsymbol{\Sigma}_1}{1+\gamma\delta_0}+\frac{\boldsymbol{\Sigma}_1}{1+\gamma\delta_0}-\frac{\boldsymbol{\Sigma}_0}{1+\gamma\delta_0}\right)\mathbf{T}_1\\
&+\frac{2}{n}\operatorname{tr}\left(\boldsymbol{\Sigma}_0-\boldsymbol{\Sigma}_1\right)\mathbf{T}_1\\
&=\frac{\gamma^2\left(\delta_0-\delta_1\right)}{\left(1+\gamma\delta_0\right)\left(1+\gamma\delta_1\right)}\frac{2}{n}\operatorname{tr}\boldsymbol{\Sigma}_0\mathbf{T}_0\boldsymbol{\Sigma}_1\mathbf{T}_1\\
&-\frac{\gamma}{1+\gamma\delta_0}\frac{2}{n}\operatorname{tr}\boldsymbol{\Sigma}_0\mathbf{T}_0\left(\boldsymbol{\Sigma}_0-\boldsymbol{\Sigma}_1\right)\mathbf{T}_1+\frac{2}{n}\operatorname{tr}\left(\boldsymbol{\Sigma}_0-\boldsymbol{\Sigma}_1\right)\mathbf{T}_1.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\left(\delta_0-\delta_1\right)\left[1-\frac{2\gamma^2}{\left(1+\gamma\delta_0\right)\left(1+\gamma\delta_1\right)}\frac{1}{n}\operatorname{tr}\boldsymbol{\Sigma}_0\mathbf{T}_0\boldsymbol{\Sigma}_1\mathbf{T}_1\right]\\
&=-\frac{\gamma}{1+\gamma\delta_0}\frac{2}{n}\operatorname{tr}\boldsymbol{\Sigma}_0\mathbf{T}_0\left(\boldsymbol{\Sigma}_0-\boldsymbol{\Sigma}_1\right)\mathbf{T}_1+\frac{2}{n}\operatorname{tr}\left(\boldsymbol{\Sigma}_0-\boldsymbol{\Sigma}_1\right)\mathbf{T}_1.
\end{aligned}$$

or equivalently

$$\begin{aligned}
\delta_0-\delta_1=&\left[1-\frac{2\gamma^2}{\left(1+\gamma\delta_0\right)\left(1+\gamma\delta_1\right)}\frac{1}{n}\operatorname{tr}\boldsymbol{\Sigma}_0\mathbf{T}_0\boldsymbol{\Sigma}_1\mathbf{T}_1\right]^{-1}\\
&\times\left[-\frac{2\gamma}{1+\gamma\delta_0}\frac{1}{n}\operatorname{tr}\boldsymbol{\Sigma}_0\mathbf{T}_0\left(\boldsymbol{\Sigma}_0-\boldsymbol{\Sigma}_1\right)\mathbf{T}_1+\frac{2}{n}\operatorname{tr}\left(\boldsymbol{\Sigma}_0-\boldsymbol{\Sigma}_1\right)\mathbf{T}_1\right].
\end{aligned}$$

All in all, we have

$$\begin{aligned}
&\frac{\delta_0}{1+\gamma\delta_0}-\frac{\delta_1}{1+\gamma\delta_1}\\
&=\frac{1}{\left(1+\gamma\delta_0\right)\left(1+\gamma\delta_1\right)}\left[1-\frac{2\gamma^2}{\left(1+\gamma\delta_0\right)\left(1+\gamma\delta_1\right)}\frac{1}{n}\operatorname{tr}\boldsymbol{\Sigma}_0\mathbf{T}_0\boldsymbol{\Sigma}_1\mathbf{T}_1\right]^{-1}\\
&\times\left[-\frac{\gamma}{1+\gamma\delta_0}\frac{2}{n}\operatorname{tr}\boldsymbol{\Sigma}_0\mathbf{T}_0\left(\boldsymbol{\Sigma}_0-\boldsymbol{\Sigma}_1\right)\mathbf{T}_1+\frac{1}{n}\operatorname{tr}\left(\boldsymbol{\Sigma}_0-\boldsymbol{\Sigma}_1\right)\mathbf{T}_1\right]
\end{aligned}\tag{65}$$

To guarantee that the left hand side of (65) does not blow up, we shall prove that

$$\liminf_p\left(1-\frac{2\gamma^2}{\left(1+\gamma\delta_0\right)\left(1+\gamma\delta_1\right)}\frac{1}{n}\operatorname{tr}\boldsymbol{\Sigma}_0\mathbf{T}_0\boldsymbol{\Sigma}_1\mathbf{T}_1\right)>0.\tag{66}$$

or equivalently

$$\limsup_p\frac{2\gamma^2}{\left(1+\gamma\delta_0\right)\left(1+\gamma\delta_1\right)}\frac{1}{n}\operatorname{tr}\boldsymbol{\Sigma}_0\mathbf{T}_0\boldsymbol{\Sigma}_1\mathbf{T}_1<1.\tag{67}$$

For that, recall that for a symmetric matrix $\mathbf{A}$ and non-negative definite matrix $\mathbf{B}$ [32], we have

$$\operatorname{tr}\mathbf{AB}\leq\|\mathbf{A}\|\operatorname{tr}\mathbf{B}.\tag{68}$$

Thus,

$$\begin{aligned}
&\frac{\gamma^2}{\left(1+\gamma\delta_0\right)\left(1+\gamma\delta_1\right)}\frac{1}{n}\operatorname{tr}\boldsymbol{\Sigma}_0\mathbf{T}_0\boldsymbol{\Sigma}_1\mathbf{T}_1\\
&\leq\frac{\gamma^2\|\boldsymbol{\Sigma}_0\mathbf{T}_0\|}{\left(1+\gamma\delta_0\right)\left(1+\gamma\delta_1\right)}\underbrace{\frac{1}{n}\operatorname{tr}\boldsymbol{\Sigma}_1\mathbf{T}_1}_{\delta_1/2}=\underbrace{\frac{\gamma\delta_1/2}{1+\gamma\delta_1}}_{<1/2}\frac{\gamma\|\boldsymbol{\Sigma}_0\mathbf{T}_0\|}{1+\gamma\delta_0}\\
&<\frac{\gamma/2\|\boldsymbol{\Sigma}_0\mathbf{T}_0\|}{1+\gamma\delta_0}.
\end{aligned}$$

Since $\boldsymbol{\Sigma}_0$ and $\mathbf{T}_0$ share the same eigenvectors, there exists a λ an eigenvalue of $\boldsymbol{\Sigma}_0$ such that

$$\|\boldsymbol{\Sigma}_0\mathbf{T}_0\|=\frac{\lambda}{1+\frac{\gamma\lambda}{1+\gamma\delta_0}}.$$

Thus,

$$\begin{aligned}
&\frac{2\gamma^2}{\left(1+\gamma\delta_0\right)\left(1+\gamma\delta_1\right)}\frac{1}{n}\operatorname{tr}\boldsymbol{\Sigma}_0\mathbf{T}_0\boldsymbol{\Sigma}_1\mathbf{T}_1<\frac{\gamma\|\boldsymbol{\Sigma}_0\mathbf{T}_0\|}{1+\gamma\delta_0}\\
&=\frac{\frac{\gamma\lambda}{1+\frac{\gamma\lambda}{1+\gamma\delta_0}}}{1+\gamma\delta_0}=\frac{\frac{\gamma\lambda}{1+\gamma\delta_0}}{1+\frac{\gamma\lambda}{1+\gamma\delta_0}}<1.
\end{aligned}$$

Thus, (67) holds. Using Assumptions 5 and 7 and by (67)

$$\begin{aligned}
&\frac{1}{\left(1+\gamma\delta_0\right)\left(1+\gamma\delta_1\right)}\\
&\times\left(1-\frac{2\gamma^2}{\left(1+\gamma\delta_0\right)\left(1+\gamma\delta_1\right)}\frac{1}{n}\operatorname{tr}\boldsymbol{\Sigma}_0\mathbf{T}_0\boldsymbol{\Sigma}_1\mathbf{T}_1\right)^{-1}=O\left(1\right),
\end{aligned}$$

which implies using Assumption 8 that

$$\delta_0-\delta_1=O\left(\frac{1}{\sqrt{p}}\right).\tag{69}$$

This gives the claim of (61). For (62), and using the inequality $\log\left(x\right)\leq x-1$, for $x>0$ we can show that

$$\begin{aligned}
\left|\log\left(1+\gamma\delta_1\right)-\log\left(1+\gamma\delta_0\right)\right|&=\left|\log\tfrac{1+\gamma\delta_1}{1+\gamma\delta_0}\right|\\
&=\left|\log\left(1+\gamma\tfrac{\delta_1-\delta_0}{1+\gamma\delta_0}\right)\right|\leq\frac{\gamma\left|\delta_0-\delta_1\right|}{1+\gamma\min\left(\delta_0,\delta_1\right)}.
\end{aligned}$$

Following the result of (61), (62) also holds.
As for (63), it suffices to notice that:

$$\begin{aligned}
&\frac{1}{\sqrt{p}}\log\det\mathbf{T}_0\mathbf{T}_1^{-1}\\
&=\frac{1}{\sqrt{p}}\log\det\left(\mathbf{I}_p+\mathbf{T}_0^{\frac{1}{2}}\left(\gamma\tilde{\delta_1}\boldsymbol{\Sigma}_1-\gamma\tilde{\delta_0}\boldsymbol{\Sigma}_0\right)\mathbf{T}_0^{\frac{1}{2}}\right)\\
&=\frac{1}{\sqrt{p}}\log\det\left[\mathbf{I}_p+\gamma\tilde{\delta_1}\mathbf{T}_0(\boldsymbol{\Sigma}_1-\boldsymbol{\Sigma}_0)+\frac{\tilde{\delta}_1-\tilde{\delta}_0}{\tilde{\delta}_0}\left(\mathbf{I}_p-\mathbf{T}_0\right)\right]
\end{aligned}$$

Define $\boldsymbol{\Phi}$ the matrix that has the same eigenvectors as $\boldsymbol{\Sigma}_1-\boldsymbol{\Sigma}_0$ with eigenvalues $\phi_i=\left|\lambda_i\left(\boldsymbol{\Sigma}_1-\boldsymbol{\Sigma}_0\right)\right|$, $i=1,\cdots,p$. Then, since $\mathbf{T}_0\preceq\mathbf{I}_p$, we have the following

$$\begin{aligned}
&\left|\frac{1}{\sqrt{p}}\log\det\mathbf{T}_0\mathbf{T}_1^{-1}\right|\\
&\leq\frac{1}{\sqrt{p}}\log\det\left[\mathbf{I}_p+\gamma\tilde{\delta_1}\mathbf{T}_0^{\frac{1}{2}}\boldsymbol{\Phi}\mathbf{T}_0^{\frac{1}{2}}+\frac{\left|\tilde{\delta}_1-\tilde{\delta}_0\right|}{\tilde{\delta}_0}\mathbf{I}_p\right].
\end{aligned}$$

Given that $0 \preceq \mathbf{\Phi}$, we have

$$\frac{1}{\sqrt{p}} \log\det\left[\mathbf{I}_p + \gamma\tilde{\delta}_1 \mathbf{T}_0^{\frac{1}{2}} \mathbf{\Phi} \mathbf{T}_0^{\frac{1}{2}} + \frac{\left|\tilde{\delta}_1 - \tilde{\delta}_0\right|}{\tilde{\delta}_0}\mathbf{I}_p\right]$$
$$\leq \frac{1}{\sqrt{p}} \operatorname{tr}\left[\gamma\tilde{\delta}_1 \mathbf{\Phi}\mathbf{T}_0 + \frac{\left|\tilde{\delta}_1 - \tilde{\delta}_0\right|}{\tilde{\delta}_0}\mathbf{I}_p\right]$$

By Assumption 8, $\frac{1}{\sqrt{p}} \operatorname{tr}\left[\gamma\tilde{\delta}_1 \mathbf{\Phi}\mathbf{T}_0 + \frac{\left|\tilde{\delta}_1 - \tilde{\delta}_0\right|}{\tilde{\delta}_0}\mathbf{I}_p\right] = O(1)$. This completes the proof.

Appendix E
Proof of Theorem 3

We start the proof by showing the following

$$G(\hat{\boldsymbol{\mu}}_i, \hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\mu}}_1, \mathbf{H}) + (-1)^{i+1}\widehat{\theta}_i - G(\boldsymbol{\mu}_i, \hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\mu}}_1, \mathbf{H}) \to_{a.s.} 0.$$

To this end, note that

$$G(\hat{\boldsymbol{\mu}}_i, \hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\mu}}_1, \mathbf{H}) = G(\boldsymbol{\mu}_i, \hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\mu}}_1, \mathbf{H}) + \frac{\mathbf{1}_{n_i}^T}{n_i}\mathbf{Y}_i\mathbf{H}(\hat{\boldsymbol{\mu}}_0 - \hat{\boldsymbol{\mu}}_1).$$

Using the same arguments used to prove (22), we can easily show that

$$G(\hat{\boldsymbol{\mu}}_i, \hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\mu}}_1, \mathbf{H}) = G(\boldsymbol{\mu}_i, \hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\mu}}_1, \mathbf{H}) + \frac{(-1)^i}{n_i}\operatorname{tr}\mathbf{\Sigma}_i\mathbf{H} + o(1)$$

It remains now to show that

$$\widehat{\theta}_i - \frac{1}{n_i}\operatorname{tr}\mathbf{\Sigma}_i\mathbf{H} \to_{a.s.} 0. \tag{70}$$

To this end, we examine the convergence of the quantity $\frac{1}{n_i}\operatorname{tr}\widehat{\mathbf{\Sigma}}_i\mathbf{H}$. Recall from (44) that

$$\widehat{\mathbf{\Sigma}}_i = \frac{1}{n_i - 1}\widetilde{\mathbf{Y}}_i\widetilde{\mathbf{Y}}_i^T - \frac{1}{n_i - 1}\widetilde{\mathbf{y}}_{i,1}\widetilde{\mathbf{y}}_{i,1}^T$$
$$= \frac{1}{n_i - 1}\sum_{j=1}^{n_i}\widetilde{\mathbf{y}}_{i,j}\widetilde{\mathbf{y}}_{i,j}^T - \frac{1}{n_i - 1}\widetilde{\mathbf{y}}_{i,1}\widetilde{\mathbf{y}}_{i,1}^T$$

Let $\mathbf{H}_{[j]} = \left(\gamma\widehat{\mathbf{\Sigma}} - \frac{\gamma}{n-2}\widetilde{\mathbf{y}}_{i,j}\widetilde{\mathbf{y}}_{i,j}^T + \mathbf{I}_p\right)^{-1}$ Thus,

$$\frac{1}{n_i}\operatorname{tr}\widehat{\mathbf{\Sigma}}_i\mathbf{H} = \frac{1}{n_i}\sum_{j=2}^{n_i}\frac{\frac{1}{n_i-1}\widetilde{\mathbf{y}}_{i,j}^T\mathbf{H}_{[j]}\widetilde{\mathbf{y}}_{i,j}}{1 + \frac{\gamma}{n-2}\widetilde{\mathbf{y}}_{i,j}^T\mathbf{H}_{[j]}\widetilde{\mathbf{y}}_{i,j}}$$

Thanks to the independence between $\mathbf{H}_j$ and $\widetilde{\mathbf{y}}_{i,j}$ and by simple application of the trace Lemma [15], we have

$$\frac{1}{n_i}\operatorname{tr}\widehat{\mathbf{\Sigma}}_i\mathbf{H} - \frac{1}{n_i}\operatorname{tr}\mathbf{\Sigma}_i\mathbf{H}\frac{1}{1 + \frac{\gamma}{n-2}\operatorname{tr}\mathbf{\Sigma}_i\mathbf{H}} \to_{a.s.} 0.$$

By simple manipulations, we have the convergence in (70).

Now, using the same tricks consisting in using the inversion Lemma along with the trace Lemma, we obtain

$$\widehat{\psi}_i^2 D\left(\hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\mu}}_1, \mathbf{H}, \widehat{\mathbf{\Sigma}}_i\right) - D(\hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\mu}}_1, \mathbf{H}, \mathbf{\Sigma}_i) \to_{a.s.} 0.$$

Appendix F
Proof of Theorem 4

The proof consists in proving the following convergences

$$\widehat{\xi}_i - \frac{1}{\sqrt{p}}\xi_i \to_{prob.} 0. \tag{71}$$

$$\widehat{b}_i - \frac{1}{\sqrt{p}}\operatorname{tr}\mathbf{B}_i \to_{prob.} 0. \tag{72}$$

$$\widehat{B}_i - \frac{1}{p}\operatorname{tr}\mathbf{B}_i^2 \to_{prob.} 0. \tag{73}$$

The proof of (71) is straightforward and relies on the following facts

$$\frac{1}{\sqrt{p}}(\boldsymbol{\mu}_i - \hat{\boldsymbol{\mu}}_i)^T\mathbf{H}_i(\boldsymbol{\mu}_i - \hat{\boldsymbol{\mu}}_i) \to_{prob.} 0.$$

$$\frac{1}{\sqrt{p}}\left(\bar{\hat{\boldsymbol{\mu}}}_0 - \hat{\boldsymbol{\mu}}_1\right)^T\mathbf{H}_{1-i}(\hat{\boldsymbol{\mu}}_0 - \hat{\boldsymbol{\mu}}_1)$$
$$- \frac{1}{\sqrt{p}}(\boldsymbol{\mu}_i - \hat{\boldsymbol{\mu}}_{1-i})^T\mathbf{H}_{1-i}(\boldsymbol{\mu}_i - \hat{\boldsymbol{\mu}}_{1-i}) \to_{prob.} 0.$$

The proof of (72) relies on the fact that $\widehat{\delta}_i$ is a consistent estimator of δ_i as shown in [8], the variance of which can be shown to be of order $O(p^{-2})$. Thus,

$$\frac{1}{\sqrt{p}}\operatorname{tr}\mathbf{\Sigma}_i\mathbf{T}_i - \frac{n_i}{\sqrt{p}}\widehat{\delta}_i \to_{prob.} 0.$$

Also, we have, for $i \in \{0, 1\}$,

$$\frac{1}{\sqrt{p}}\operatorname{tr}\widehat{\mathbf{\Sigma}}_i\mathbf{H}_{1-i} - \frac{1}{\sqrt{p}}\operatorname{tr}\mathbf{\Sigma}_i\mathbf{H}_{1-i} \to_{prob.} 0.$$

which gives the convergence in (72).

A. Proof of (73)

The proof of (73) is a bit more involved than those of (71) and (72) as we will show in the following. The proof mainly relies on the application of the inversion lemma followed by the trace lemma [15]. Recall that

$$\frac{1}{p}\operatorname{tr}\mathbf{B}_i^2 = \frac{1}{p}\operatorname{tr}\mathbf{\Sigma}_i\mathbf{H}_1\mathbf{\Sigma}_i\mathbf{H}_1 - \frac{2}{p}\operatorname{tr}\mathbf{\Sigma}_i\mathbf{H}_0\mathbf{\Sigma}_i\mathbf{H}_1$$
$$+ \frac{1}{p}\operatorname{tr}\mathbf{\Sigma}_i\mathbf{H}_0\mathbf{\Sigma}_i\mathbf{H}_0.$$

Without loss of generality, we can assume $i = 1$, the other case follows naturally. We start by handling the term $\frac{1}{p}\operatorname{tr}\mathbf{\Sigma}_1\mathbf{H}_1\mathbf{\Sigma}_1\mathbf{H}_1$. The common method here is to replace $\mathbf{\Sigma}_1$ by its sample estimate $\widehat{\mathbf{\Sigma}}_1$, then compute the limit of the obtained expression and perform the necessary corrections to obtain the estimate of interest. In fact, we have

$$\frac{1}{p}\operatorname{tr}\widehat{\mathbf{\Sigma}}_1\mathbf{H}_1\widehat{\mathbf{\Sigma}}_1\mathbf{H}_1 = \frac{1}{p}\sum_{j=1}^{n_1-1}\sum_{k=1}^{n_1-1}\frac{1}{p}\operatorname{tr}\frac{\widetilde{\mathbf{y}}_{1,j}\widetilde{\mathbf{y}}_{1,j}^T}{n_1-1}\mathbf{H}_1\frac{\widetilde{\mathbf{y}}_{1,k}\widetilde{\mathbf{y}}_{1,k}^T}{n_1-1}\mathbf{H}_1$$
$$= \frac{1}{p}\sum_{j=1}^{n_1-1}\sum_{k\neq j}\operatorname{tr}\frac{\widetilde{\mathbf{y}}_{1,j}\widetilde{\mathbf{y}}_{1,j}^T}{n_1-1}\mathbf{H}_1\frac{\widetilde{\mathbf{y}}_{1,k}\widetilde{\mathbf{y}}_{1,k}^T}{n_1-1}\mathbf{H}_1$$
$$+ \frac{1}{p}\sum_{j=1}^{n_1-1}\operatorname{tr}\frac{\widetilde{\mathbf{y}}_{1,j}\widetilde{\mathbf{y}}_{1,j}^T}{n_1-1}\mathbf{H}_1\frac{\widetilde{\mathbf{y}}_{1,j}\widetilde{\mathbf{y}}_{1,j}^T}{n_1-1}\mathbf{H}_1$$

Using the inversion lemma, we handle the first term in the previous equation as follows

$$\frac{1}{p}\sum_{j=2}^{n_1-1}\sum_{k\neq j}\operatorname{tr}\frac{\widetilde{\mathbf{y}}_{1,j}\widetilde{\mathbf{y}}_{1,j}^T}{n_1-1}\mathbf{H}_1\frac{\widetilde{\mathbf{y}}_{1,k}\widetilde{\mathbf{y}}_{1,k}^T}{n_1-1}\mathbf{H}_1$$
$$=\frac{1}{p}\sum_{j=2}^{n_1-1}\sum_{k\neq j}\frac{\left(\frac{1}{n_i-1}\widetilde{\mathbf{y}}_{1,j}^T\mathbf{H}_{1,j,k}\widetilde{\mathbf{y}}_{1,k}\right)^2}{(1+\frac{\gamma}{n_1-1}\widetilde{\mathbf{y}}_{1,j}^T\mathbf{H}_{1,j}\widetilde{\mathbf{y}}_{1,j})^2(1+\frac{\gamma}{n_i}\widetilde{\mathbf{y}}_{1,k}\mathbf{H}_{1,j,k}\widetilde{\mathbf{y}}_{1,k})^2}$$

where

$$\mathbf{H}_{1,j}=\left(\mathbf{I}_p+\gamma\widehat{\boldsymbol{\Sigma}}_i-\frac{\gamma}{n_1-1}\widetilde{\mathbf{y}}_{1,j}\widetilde{\mathbf{y}}_{1,j}^T\right)^{-1}$$

and

$$\mathbf{H}_{1,j,k}=\left(\mathbf{I}_p+\gamma\widehat{\boldsymbol{\Sigma}}_i-\frac{\gamma}{n_1-1}\widetilde{\mathbf{y}}_{1,j}\widetilde{\mathbf{y}}_{1,j}^T-\frac{\gamma}{n_1-1}\widetilde{\mathbf{y}}_{1,k}\widetilde{\mathbf{y}}_{1,k}^T\right)^{-1}.$$

We now refer to the use of the trace lemma to replace the denominator by its deterministic equivalents, thus we get

$$\frac{1}{p}\sum_{j=2}^{n_1-1}\sum_{k\neq j}\operatorname{tr}\frac{\widetilde{\mathbf{y}}_{1,j}\widetilde{\mathbf{y}}_{1,j}^T}{n_1-1}\mathbf{H}_1\frac{\widetilde{\mathbf{y}}_{1,k}\widetilde{\mathbf{y}}_{1,k}^T}{n_1-1}\mathbf{H}_1$$
$$=\frac{n_i}{p}\frac{\left(\frac{1}{n_1-1}\operatorname{tr}\boldsymbol{\Sigma}\mathbf{H}_1\boldsymbol{\Sigma}\mathbf{H}_1\right)}{(1+\frac{\gamma}{n_1-1}\operatorname{tr}\boldsymbol{\Sigma}_1\mathbf{H}_1)^4}+o(1).$$

Using similar steps, the second term can be approximated as follows

$$\frac{1}{p}\sum_{j=1}^{n_1-1}\operatorname{tr}\frac{\widetilde{\mathbf{y}}_{1,j}\widetilde{\mathbf{y}}_{1,j}^T}{n_1-1}\mathbf{H}_1\frac{\widetilde{\mathbf{y}}_{1,j}\widetilde{\mathbf{y}}_{1,j}^T}{n_1-1}\mathbf{H}_1=\frac{n_i}{p}\frac{\left(\frac{1}{n_i-1}\operatorname{tr}\boldsymbol{\Sigma}_1\mathbf{H}_1\right)^2}{(1+\frac{\gamma}{n_1-1}\operatorname{tr}\boldsymbol{\Sigma}_1\mathbf{H}_1)^2}+o(1).$$

We thus obtain

$$\frac{1}{p}\operatorname{tr}\widehat{\boldsymbol{\Sigma}}_1\mathbf{H}_1\widehat{\boldsymbol{\Sigma}}_1\mathbf{H}_1=\frac{n_i}{p}\frac{\left(\frac{1}{n_i-1}\operatorname{tr}\boldsymbol{\Sigma}_1\mathbf{H}_1\boldsymbol{\Sigma}_1\mathbf{H}_1\right)}{(1+\frac{\gamma}{n_i-1}\operatorname{tr}\boldsymbol{\Sigma}_1\mathbf{H}_1)^4}$$
$$+\frac{n_i}{p}\frac{\left(\frac{1}{n_i-1}\operatorname{tr}\boldsymbol{\Sigma}_1\mathbf{H}_1\right)^2}{(1+\frac{\gamma}{n_1-1}\operatorname{tr}\boldsymbol{\Sigma}_1\mathbf{H}_1)^2}+o(1).$$

We will now handle the term $\frac{1}{p}\operatorname{tr}\boldsymbol{\Sigma}_1\mathbf{H}_0\boldsymbol{\Sigma}_1\mathbf{H}_0$. Again, we start by replacing $\boldsymbol{\Sigma}_1$ by $\widehat{\boldsymbol{\Sigma}}_1$. In doing so, we obtain:

$$\frac{1}{p}\operatorname{tr}\widehat{\boldsymbol{\Sigma}}_1\mathbf{H}_0\widehat{\boldsymbol{\Sigma}}_1\mathbf{H}_0=\frac{1}{p}\sum_{j=2}^{n_1-1}\sum_{k=2}^{n_1-1}\frac{\widetilde{\mathbf{y}}_{1,j}\widetilde{\mathbf{y}}_{1,j}^T}{n_1-1}\mathbf{H}_0\frac{\widetilde{\mathbf{y}}_{1,k}\widetilde{\mathbf{y}}_{1,k}}{n_1-1}\mathbf{H}_0$$
$$=\frac{1}{p}\sum_{j=2}^{n_1-1}\left(\frac{\widetilde{\mathbf{y}}_{1,j}^T\mathbf{H}_0\widetilde{\mathbf{y}}_{1,j}}{n_1-1}\right)^2+\frac{1}{p}\sum_{j=2}\sum_{k\neq j}\frac{\widetilde{\mathbf{y}}_{1,j}\widetilde{\mathbf{y}}_{1,j}^T}{n_1-1}\mathbf{H}_0\frac{\widetilde{\mathbf{y}}_{1,k}\widetilde{\mathbf{y}}_{1,k}}{n_1-1}\mathbf{H}_0$$
$$=\frac{n_i}{p}\left(\frac{1}{n_i-1}\operatorname{tr}\boldsymbol{\Sigma}_1\mathbf{H}_0\right)^2+\frac{1}{p}\operatorname{tr}\boldsymbol{\Sigma}_1\mathbf{H}_0\boldsymbol{\Sigma}_1\mathbf{H}_0+o(1)$$

It remains now to handle the term $\frac{1}{p}\operatorname{tr}\boldsymbol{\Sigma}_1\mathbf{H}_1\boldsymbol{\Sigma}_1\mathbf{H}_0$. Using the same reasoning, we have:

$$\frac{1}{p}\operatorname{tr}\widehat{\boldsymbol{\Sigma}}_1\mathbf{H}_1\widehat{\boldsymbol{\Sigma}}_1\mathbf{H}_0=\frac{1}{p}\sum_{j=2}^{n_1-1}\sum_{k=2}^{n_1-1}\frac{\widetilde{\mathbf{y}}_{1,j}\widetilde{\mathbf{y}}_{1,j}^T}{n_1-1}\mathbf{H}_1\frac{\widetilde{\mathbf{y}}_{1,k}\widetilde{\mathbf{y}}_{1,k}^T}{n_1-1}\mathbf{H}_0$$
$$=\frac{1}{p}\sum_{j=2}^{n_1-1}\frac{1}{(n_i-1)^2}\widetilde{\mathbf{y}}_{1,j}^T\mathbf{H}_1\widetilde{\mathbf{y}}_{1,j}\widetilde{\mathbf{y}}_{1,j}^T\mathbf{H}_0\widetilde{\mathbf{y}}_{1,j}$$
$$+\frac{1}{p}\sum_{j=2}^{n_1-1}\sum_{k\neq j}\frac{1}{(n_1-1)^2}\operatorname{tr}\widetilde{\mathbf{y}}_{1,j}\widetilde{\mathbf{y}}_{1,j}^T\mathbf{H}_1\widetilde{\mathbf{y}}_{1,k}\widetilde{\mathbf{y}}_{1,k}^T\mathbf{H}_0$$

Using the inversion Lemma along with the trace Lemma, we ultimately find:

$$\frac{1}{p}\operatorname{tr}\widehat{\boldsymbol{\Sigma}}_1\mathbf{H}_1\widehat{\boldsymbol{\Sigma}}_1\mathbf{H}_0=\frac{n_i}{p}\frac{1}{n_i}\operatorname{tr}\boldsymbol{\Sigma}_1\mathbf{H}_0\frac{\frac{1}{n_i-1}\operatorname{tr}\boldsymbol{\Sigma}_1\mathbf{H}_1}{1+\frac{\gamma}{n_1-1}\operatorname{tr}\boldsymbol{\Sigma}_1\mathbf{H}_1}$$
$$+\frac{\frac{1}{p}\operatorname{tr}\boldsymbol{\Sigma}_1\mathbf{H}_1\boldsymbol{\Sigma}_1\mathbf{H}_0}{(1+\frac{\gamma}{n_1-1}\operatorname{tr}\boldsymbol{\Sigma}_1\mathbf{H}_1)^2}+o(1).$$

Now, we will put things together. We have the following

$$\frac{1}{p}\operatorname{tr}\boldsymbol{\Sigma}_1\mathbf{H}_1\boldsymbol{\Sigma}_1\mathbf{H}_1=(1+\frac{\gamma}{n_1-1}\operatorname{tr}\boldsymbol{\Sigma}_1\mathbf{H}_1)^4\frac{1}{p}\operatorname{tr}\widehat{\boldsymbol{\Sigma}}_1\mathbf{H}_1\widehat{\boldsymbol{\Sigma}}_1\mathbf{H}_1$$
$$-\frac{n_1}{p}\left(\frac{1}{n_1}\operatorname{tr}\boldsymbol{\Sigma}_1\mathbf{H}_1\right)^2\left(1+\frac{\gamma}{n_1-1}\operatorname{tr}\boldsymbol{\Sigma}_1\mathbf{H}_1\right)^2+o(1).$$

$$\frac{1}{p}\operatorname{tr}\boldsymbol{\Sigma}_1\mathbf{H}_0\boldsymbol{\Sigma}_1\mathbf{H}_0=\frac{1}{p}\operatorname{tr}\widehat{\boldsymbol{\Sigma}}_1\mathbf{H}_0\widehat{\boldsymbol{\Sigma}}_1\mathbf{H}_0-\frac{n_1}{p}\left(\frac{1}{n_1}\operatorname{tr}\boldsymbol{\Sigma}_1\mathbf{H}_0\right)^2$$
$$+o(1).$$

and

$$\frac{1}{p}\operatorname{tr}\boldsymbol{\Sigma}_1\mathbf{H}_1\boldsymbol{\Sigma}_1\mathbf{H}_0=(1+\frac{\gamma}{n_1-1}\operatorname{tr}\boldsymbol{\Sigma}_1\mathbf{H}_1)^2\frac{1}{p}\widehat{\boldsymbol{\Sigma}}_1\mathbf{H}_1\widehat{\boldsymbol{\Sigma}}_1\mathbf{H}_0$$
$$-\frac{n_1}{p}\frac{1}{n_1}\operatorname{tr}\boldsymbol{\Sigma}_1\mathbf{H}_0\frac{1}{n_1}\boldsymbol{\Sigma}_1\mathbf{H}_1(1+\frac{\gamma}{n_1-1}\operatorname{tr}\boldsymbol{\Sigma}_1\mathbf{H}_1)+o(1).$$

A consistent estimator of $\frac{1}{p}\operatorname{tr}\mathbf{B}_1^2$ is thus given by

$$\frac{1}{p}\operatorname{tr}\mathbf{B}_1^2=\left(1+\frac{\gamma}{n_1-1}\operatorname{tr}\boldsymbol{\Sigma}_1\mathbf{H}_1\right)^4\frac{1}{p}\operatorname{tr}\widehat{\boldsymbol{\Sigma}}_1\mathbf{H}_1\widehat{\boldsymbol{\Sigma}}_1\mathbf{H}_1$$
$$-\frac{n_1}{p}\left(\frac{1}{n_1}\operatorname{tr}\boldsymbol{\Sigma}_1\mathbf{H}_1\right)^2\left(1+\frac{\gamma}{n_1-1}\operatorname{tr}\boldsymbol{\Sigma}_1\mathbf{H}_1\right)^2$$
$$+\frac{1}{p}\operatorname{tr}\widehat{\boldsymbol{\Sigma}}_1\mathbf{H}_0\widehat{\boldsymbol{\Sigma}}_1\mathbf{H}_0-\frac{n_1}{p}\left(\frac{1}{n_1}\operatorname{tr}\boldsymbol{\Sigma}_1\mathbf{H}_0\right)^2$$
$$-2\left(1+\frac{\gamma}{n_1-1}\operatorname{tr}\boldsymbol{\Sigma}_1\mathbf{H}_1\right)^2\frac{1}{p}\operatorname{tr}\widehat{\boldsymbol{\Sigma}}_1\mathbf{H}_1\widehat{\boldsymbol{\Sigma}}_1\mathbf{H}_0$$
$$+\frac{2n_1}{p}\frac{1}{n_1}\operatorname{tr}\boldsymbol{\Sigma}_1\mathbf{H}_0\frac{1}{n_1}\operatorname{tr}\boldsymbol{\Sigma}_1\mathbf{H}_1\left(1+\frac{\gamma}{n_1-1}\operatorname{tr}\boldsymbol{\Sigma}_1\mathbf{H}_1\right).$$

We replace $\frac{1}{n_1}\operatorname{tr}\boldsymbol{\Sigma}_1\mathbf{H}_1$ and $\frac{1}{n_1}\operatorname{tr}\boldsymbol{\Sigma}_1\mathbf{H}_0$ by their respective consistent estimates $\widehat{\delta}_1$ and $\frac{1}{n_1}\operatorname{tr}\widehat{\boldsymbol{\Sigma}}_1\mathbf{H}_0$ to get the consistent estimate for $\frac{1}{p}\operatorname{tr}\mathbf{B}_1^2$. By this, we achieve the proof of the theorem.