Regularized Discriminant Analysis: A Large Dimensional Study

Xiaoke Yang, Khalil Elkhalil, Abla Kammoun, Tareq Y. Al-Naffouri and Mohamed-Slim Alouini
CEMSE Division, King Abdullah University of Science and Technology, Saudi Arabia
Emails: {xiaoke.yang, khalil.elkhalil, abla.kammoun, tareq.alnaffouri, slim.alouini}@kaust.edu.sa

Abstract—**This paper focuses on studying the performance of general regularized discriminant analysis (RDA) classifiers based on the Gaussian mixture model with different means and covariances. RDA offers a rich class of regularization options, covering as special cases the regularized linear discriminant analysis (RLDA) and the regularized quadratic discriminant analysis (RQDA) classifiers. Based on fundamental results from random matrix theory, we analyze RDA under the double asymptotic regime where the data dimension and the training size both increase in a proportional way. Under the double asymptotic regime and some mild assumptions, we show that the asymptotic classification error converges to a deterministic quantity that only depends on the data statistical parameters and dimensions. This result can be leveraged to select the optimal parameters that minimize the classification error, thus yielding the optimal classifier. Numerical results are provided to validate our theoretical findings on synthetic data showing high accuracy of our derivations.**

I. Introduction

Classification problems are widely studied in today's hot pattern recognition and machine learning fields [1]. Among them, we distinguish classification approaches based on discriminant analysis, which are extensively used in a large panel of applications, such as bioinformatics [2], finance [3] to name a few. Belonging to the larger set of model-based classification methods, their popularity owes to the fact that they rely on probabilistic foundations, making them optimal under the assumptions they have been built upon. Linear discriminant analysis (LDA) and Quadratic discriminant analysis (QDA) are the most typical cases of discriminant analysis. Both of them are based on the assumption that data from each class are drawn from a specific Gaussian distribution. The single difference lies in that LDA assumes the same covariance matrix for all classes, while QDA allows different covariance matrices. Under the Gaussian assumption and assuming the knowledge of the mean and covariance matrices for each class, LDA and QDA produce the classifiers that minimize the misclassification error rate. This high performance is, however, in practice not always guaranteed. The major reason for that is related to the fact that the means and covariance matrices associated with each class could not be perfectly acquired. They can instead be estimated based on available training sets for which the class labels are given. This results in performance losses which become all the more large when the sample size for each class is small with respect to their dimension, causing the estimated covariance matrices to be highly inaccurate and ill-conditioned. One approach to get around this issue, which dates back to the early works of Friedman [4], is to regularize the covariance matrix. The regularization artifice serves to improve the stability of these estimates by shrinking them towards the identity matrix. This can be accomplished by employing one regularization parameter in LDA or QDA, giving rise to what is known as regularized LDA (R-LDA) and regularized QDA (R-QDA). A better flexibility can be obtained by using two regularization parameters offering a better regulation of the weights associated with samples of each class. This approach, which seemingly can bring better performances, is known as regularized discriminant analysis (RDA).

A lot of attention has been devoted to analyzing the performances of R-LDA and R-QDA classifiers under several regimes. One frequently used regime, is the double asymptotic regime in which the number of samples and their dimensions grow large with the same pace. The interest of this regime lies in that it allows leveraging a large body of results from the theory of random matrices, opening up possibilities of accurate characterization of the asymptotic misclassification error rate. Towards this goal, the work in [5] studies the asymptotic misclassification error rate of the R-LDA, while the R-QDA has only recently been analyzed in [6]. To the best of the authors' knowledge, the RDA, which should offer better capabilities, has not been theoretically studied. The present work aims to fill this gap. Restricting our attention to binary classification, we analyse the asymptotic misclassification error rate for the RDA. In particular, we identify sufficient assumptions on the distance between covariance matrices and means of both classes ensuring non-trivial classification error rate. Under these assumptions, we show that the misclassification error rate can be approximated in the asymptotic regime by deterministic quantities that depend solely on the samples' size and their dimensions as well as the means and covariances of each class. The major interests behind this result are twofold. First, it allows to enlighten the impact of the intervening parameters on the classification performances. Second, it can be used in practice to properly tune the regularization parameters, so as to reap the full potential of RDA.

In a nutshell, the contributions are summarized as follows

- Under some mild assumptions, we show that the classification error rate approaches a non-trivial deterministic quantity that only depends on the classes' statistics and the problem's dimension.
- Based on the derived deterministic classification error, we

illustrate the importance of properly selecting the pair of parameters that minimizes the misclassification error rate.
- We validate the theoretical results using synthetic data and demonstrate the accuracy of our findings.

The rest of this paper is organized as follows: Section II introduces the RDA problem. Section III exposes the assumptions and the theoretical findings. The detailed proofs for the theorems can be found in the full version of this paper. Finally, we validate the accuracy of our derived results in section IV prior to concluding in section V.

II. Problem Statement

A. Notations

Throughout this paper, we use non-boldface lowercase letters to denote scalars, boldface lowercase letters to denote vectors and boldface uppercase letters to denote matrices. $\mathbf{I}_p$ denotes the p dimensional identity matrix. $\|.\|$ denotes the Euclidean norm for vectors and the spectral norm for matrices. $(.)^T$, $\operatorname{tr}(.)$ and $|.|$ respectively denote the transpose, the trace and the determinant of a matrix. For two functionals f and g, we say that $f = \mathcal{O}(g)$, if $\exists\, 0 < M < \infty$ such that $|f| \leq M|g|$. $\mathbb{P}(.)$ denotes the probability measure. $\xrightarrow{a.s.}$ denotes the almost sure convergence of random variables. $\Phi(x) = \int_{-\infty}^{x} \frac{\exp\left(-\frac{t^2}{2}\right)}{\sqrt{2\pi}} dt$ denotes the cumulative density function (CDF) of the standard normal distribution.

B. RDA classifier for Binary Classification

We consider the binary classification problem in which we aim to assign an observation $\mathbf{x} \in \mathbb{R}^p$ to the class $\mathcal{C}_i$, $i \in \{0,1\}$ which $\mathbf{x}$ most likely belongs to. We assume that observations vectors sampled from class $\mathcal{C}_i$, $i \in \{0,1\}$ follow a multivariate Gaussian distribution with mean $\boldsymbol{\mu}_i \in \mathbb{R}^p$ and covariance $\mathbf{\Sigma}_i \in \mathbb{R}^{p\times p}$. Denote the prior probability of $\mathbf{x}$ belonging to class $\mathcal{C}_i$ by π_i, $i \in \{0,1\}$. The Bayes classifier is thus the one that maximizes the posterior probability [7], which boils down to selecting the class that corresponds to the highest classification score $\delta_i^{RDA}(\mathbf{x})$, $i \in \{0,1\}$ where

$$\delta_i^{RDA}(\mathbf{x}) = -\frac{1}{2}\log|\mathbf{\Sigma}_i| - \frac{1}{2}\left(\mathbf{x}-\mu_i\right)^T \mathbf{\Sigma}_i^{-1}\left(\mathbf{x}-\mu_i\right) + \log\pi_i. \tag{1}$$

More formally, we assign $\mathbf{x}$ to class i^* where

$$i^* = \arg\max_{i\in\{0,1\}} \delta_i^{RDA}(\mathbf{x}). \tag{2}$$

The discriminant function $\delta_i^{RDA}(\mathbf{x})$ involves the exact statistics of each class, namely, their associated mean vectors and covariances. In practice, these parameters could not be perfectly known beforehand. They are estimated using available training data. We assume that n_i, $i \in \{0,1\}$ independent training samples are provided for class $\mathcal{C}_i$, respectively denoted by $\mathcal{S}_0 = \{\mathbf{x}_l \in \mathcal{C}_0\}_{l=1}^{n_0}$ and $\mathcal{S}_1 = \{\mathbf{x}_l \in \mathcal{C}_1\}_{l=n_0+1}^{n_0+n_1=n}$.

Based on this training data, we consider the following sample estimates of the mean and covariance matrices

$$\overline{\mathbf{x}}_i = \frac{1}{n_i}\sum_{l\in\mathcal{T}_i}\mathbf{x}_l, \quad i \in \{0,1\}.$$

$$\widehat{\mathbf{\Sigma}}_i = \frac{1}{n_i-1}\sum_{l\in\mathcal{T}_i}\left(\mathbf{x}_l-\overline{\mathbf{x}}_i\right)\left(\mathbf{x}_l-\overline{\mathbf{x}}_i\right)^T, \quad i \in \{0,1\}.$$

$$\widehat{\mathbf{\Sigma}} = \frac{n_0-1}{n-2}\widehat{\mathbf{\Sigma}}_0 + \frac{n_1-1}{n-2}\widehat{\mathbf{\Sigma}}_1,$$

where $\widehat{\mathbf{\Sigma}}$ is the pooled sample covariance. LDA and QDA classifiers are respectively obtained when $\widehat{\mathbf{\Sigma}}$ or $\widehat{\mathbf{\Sigma}}_i$ are used in place of $\mathbf{\Sigma}_i$, $i \in \{0,1\}$. They can be viewed as extreme cases of the RDA which consists in using the following regularized covariance matrix

$$\widehat{\mathbf{\Sigma}}_i(\lambda) = \frac{(1-\lambda)n_i\widehat{\mathbf{\Sigma}}_i + \lambda n\widehat{\mathbf{\Sigma}}}{(1-\lambda)n_i + \lambda n}, \;\; i \in \{0,1\},$$

where λ controls the shrinkage of the individual covariance of QDA toward the pooled covariance matrix in LDA. In this manner, they define a parametrized class of discriminant analysis classifiers ranging from LDA ($\lambda = 1$) to QDA ($\lambda = 0$). If the training size is much lower than the data dimension, $\widehat{\mathbf{\Sigma}}_i(\lambda)$ is singular. Singularity concerns will thus arise if it is directly plugged in $\delta_i^{RDA}(\mathbf{x})$ in place of $\mathbf{\Sigma}_i$. To overcome these issues, a common way is to use a second regularization parameter $\gamma \in [0,1]$ aiming to shrinking $\widehat{\mathbf{\Sigma}}_i(\lambda)$ towards the identity matrix. The regularized covariance matrix that is used in RDA takes thus the following form

$$\widehat{\mathbf{\Sigma}}_i(\lambda,\gamma) = \gamma\frac{(1-\lambda)n_i\widehat{\mathbf{\Sigma}}_i + \lambda n\widehat{\mathbf{\Sigma}}}{(1-\lambda)n_i + \lambda n} + (1-\gamma)\mathbf{I}_p.$$

Defining the following quantities

$$\mathbf{H}_0 = \left[(1-\gamma)\mathbf{I}_p + \alpha_0\widehat{\mathbf{\Sigma}}_0 + \beta_0\widehat{\mathbf{\Sigma}}_1\right]^{-1}.$$

$$\alpha_0 = \frac{n_0\gamma}{n_0+\lambda n_1}, \beta_0 = \frac{n_1\gamma\lambda}{n_0+\lambda n_1}.$$

$$\mathbf{H}_1 = \left[(1-\gamma)\mathbf{I}_p + \alpha_1\widehat{\mathbf{\Sigma}}_0 + \beta_1\widehat{\mathbf{\Sigma}}_1\right]^{-1}.$$

$$\alpha_1 = \frac{n_0\gamma\lambda}{n_1+\lambda n_0}, \beta_1 = \frac{n_1\gamma}{n_1+\lambda n_0},$$

the discriminant rule in (1) with plug-in estimators for RDA simplifies to

$$\widehat{\delta}_i^{RDA}(\mathbf{x}) = \frac{1}{2}\log|\mathbf{H}_i| - \frac{1}{2}\left(\mathbf{x}-\overline{\mathbf{x}}_i\right)^T\mathbf{H}_i\left(\mathbf{x}-\overline{\mathbf{x}}_i\right) + \log\pi_i. \tag{3}$$

Conditioning on the training samples $\mathcal{S}_i$, $i \in \{0,1\}$, the conditional misclassification error rate associated to class $\mathcal{C}_i$ is given by

$$\epsilon_i^{RDA} = \mathbb{P}\left[(-1)^i\,\widehat{\delta}_0^{RDA}(\mathbf{x}) < (-1)^i\,\widehat{\delta}_1^{RDA}(\mathbf{x})\,|\mathbf{x}\in\mathcal{C}_i\right]. \tag{4}$$

Thus, the total misclassification error rate writes as

$$\epsilon^{RDA} = \pi_0\epsilon_0^{RDA} + \pi_1\epsilon_1^{RDA}.$$

It is also important to note that using some basic manipulations, the conditional misclassification error rate can be written as

$$\epsilon_i^{RDA} = \mathbb{P}\left[\boldsymbol{\omega}^T \mathbf{B}_i \boldsymbol{\omega} + 2\boldsymbol{\omega}^T \mathbf{y}_i < \xi_i | \boldsymbol{\omega} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{I}_p\right)\right], \quad (5)$$

where

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

It entails from (5) that the misclassification error rate reduces to the cumulative distribution function (CDF) of bilinear forms of Gaussian random vectors, and as such cannot be derived in closed-form. However, as will be shown in the next section, an asymptotic evaluation of it can be obtained by utilizing the central limit theorem.

III. Main Results

In this part, we show that the classification error rate converges under some mild assumptions to a some deterministic quantity that depends on the means and covariances associated with each class. These assumptions are designed in such a way to avoid trivial misclassification error rates.

A. Technical Assumptions

The following assumptions are conceived in order to get non-trivial classification error rates. For $i \in \{0,1\}$, when n_i, $p \to \infty$, we make the following assumptions

Assumption *1 (Data scaling):* $\frac{n_i}{p} \to c \in \{0, \infty\}$ with $|n_0 - n_1| = o(1)$.

Assumption *2 (Mean scaling):* Let $\boldsymbol{\mu} \triangleq \boldsymbol{\mu}_0 - \boldsymbol{\mu}_1$, $\|\boldsymbol{\mu}\| = \|\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\|^2 = \mathcal{O}\left(\sqrt{p}\right)$.

Assumption *3 (Covariance scaling):* $\|\boldsymbol{\Sigma}_i\| = \mathcal{O}(1)$.

Assumption *4 (Covariance separation):* The matrix $\boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_1$ has exactly $\mathcal{O}\left(\sqrt{p}\right)$ eigenvalues of $\mathcal{O}(1)$ while the remaining ones decay at an order of $\mathcal{O}\left(1/\sqrt{p}\right)$.

Assumption 1 establishes the double asymptotic regime which implies that the number of samples are commensurable with their dimensions. As a byproduct of Assumption 1, we have $\pi_i \to \frac{1}{2}$ as $n, p \to \infty$, and $|\alpha_i - \beta_{1-i}| = o(1)$. This particularly implies that the regularization weight associated with $\boldsymbol{\Sigma}_i$ in $\mathbf{H}_0$ is approximately equal to that corresponding to $\boldsymbol{\Sigma}_{1-i}$ in $\mathbf{H}_1$. This is useful from a technical perspective to control the distance between $\mathbf{H}_1$ and $\mathbf{H}_0$. Assumption 2 states that the difference of Euclidean distance between the means should scales at the rate of $\mathcal{O}\left(\sqrt{p}\right)$ for RDA. As will be elaborated on later, this is the growth rate that allows RDA to leverage information about the means of both classes. Assumption 3 bounds the spectral norm of the covariance matrices and is of standard use in random matrix theory. Assumption 4 controls the distance between covariance matrices to avoid the situation of trivial misclassification error rates, which implies that $\frac{1}{\sqrt{p}} \operatorname{tr} \mathbf{A}(\boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_1) = \mathcal{O}(1)$ for any $\mathbf{A}$ with bounded spectral norm. Assumption 2 and Assumption 4 play important roles on controlling the distance between class means and class covariances so that the RDA classifier can present significative performance. The importance of these assumptions will be discussed later.

B. Central Limit Theorem(CLT)

Under Assumptions 1-4, using Lyapunov's CLT in [8], the work in [9] has proved that the bilinear form $\boldsymbol{\omega}^T \mathbf{B}_i \boldsymbol{\omega} + 2\boldsymbol{\omega}^T \mathbf{y}_i$ in the random vector $\boldsymbol{\omega}$ with $\boldsymbol{\omega} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{I}_p\right)$ follows a Gaussian distribution with mean $\operatorname{tr} \mathbf{B}_i$ and variance $2 \operatorname{tr} \mathbf{B}_i^2 + 4\mathbf{y}_i^T \mathbf{y}_i$. Based on this result, we thus prove that the condition classification error ϵ_i satisfies

Theorem 1: Assume $\lambda \neq 1$. Under assumptions 1-4, the conditional classification error in (4) satisfies

$$\epsilon_i^{RDA} - \Phi\left((-1)^i \frac{\xi_i - \operatorname{tr} \mathbf{B}_i}{\sqrt{2 \operatorname{tr} \mathbf{B}_i^2 + 4\mathbf{y}_i^T \mathbf{y}_i}}\right) \xrightarrow{a.s} 0. \quad (6)$$

R-LDA, which has been studied in [5] could not be directly derived from Theorem 1, since it is associated with a linear classifier, while RDA involves in general a quadratic form in its classification rule.

C. Deterministic Equivalent

In this part, we derive the asymptotic misclassification error rate. Prior to stating our results, we need to introduce the following notations which stems from the use of standard tools of random matrix theory. For $i \in \{0,1\}$, denote by δ_i the unique positive solution to the following fixed point equation

$$\delta_i = \frac{1}{n_i} \operatorname{tr} \boldsymbol{\Sigma}_i \left[(1-\gamma)\mathbf{I}_p + \frac{\alpha_i}{1+\alpha_i \delta_i}\boldsymbol{\Sigma}_0 + \frac{\beta_i}{1+\beta_i \delta_i}\boldsymbol{\Sigma}_1\right]^{-1}.$$

Define

$$\tilde{\delta}_i = \frac{\alpha_i}{1+\alpha_i \delta_i},$$

$$\mathbf{Q}_i = \left[(1-\gamma)\mathbf{I}_p + \frac{\alpha_i}{1+\alpha_i \delta_i}\boldsymbol{\Sigma}_0 + \frac{\beta_i}{1+\beta_i \delta_i}\boldsymbol{\Sigma}_1\right]^{-1},$$

and let

$$\phi = \frac{1}{n_1} \operatorname{tr} \boldsymbol{\Sigma}_1 \mathbf{Q}_1 \boldsymbol{\Sigma}_1 \mathbf{Q}_1.$$

With these notations at hand, we prove the following convergences

Proposition 1: Under Assumptions 1-4, we have

$$\frac{1}{\sqrt{p}}\xi_i - \bar{\xi}_i \xrightarrow{p} 0, \quad (7)$$

$$\frac{1}{\sqrt{p}} \operatorname{tr} \mathbf{B}_i - \overline{b_i} \xrightarrow{p} 0, \quad (8)$$

$$\frac{1}{p} \operatorname{tr} \mathbf{B}_i^2 - \overline{B_i} \xrightarrow{p} 0, \quad (9)$$

$$\frac{1}{p}\mathbf{y}_i^T \mathbf{y}_i \xrightarrow{p} 0, \quad (10)$$

where

$$\begin{aligned}\bar{\xi}_i \triangleq & \frac{1}{\sqrt{p}} \log \left(\frac{1+\alpha_0\delta_0}{1+\alpha_1\delta_1}\right)^{n_0} \left(\frac{1+\beta_0\delta_0}{1+\beta_1\delta_1}\right)^{n_1} + \frac{1}{\sqrt{p}} \log \frac{|\mathbf{Q}_1|}{|\mathbf{Q}_0|} \\ & + \frac{1}{\sqrt{p}} \left[\frac{\alpha_1\delta_1 n_0 - \alpha_0\delta_0 n_0}{(1+\alpha_1\delta_1)(1+\alpha_0\delta_0)} + \frac{\beta_1\delta_1 n_0 - \beta_0\delta_0 n_0}{(1+\beta_1\delta_1)(1+\beta_0\delta_0)}\right] \\ & + \frac{1}{\sqrt{p}} (-1)^{i+1} \boldsymbol{\mu}^T \mathbf{Q}_{1-i} \boldsymbol{\mu}. \\ \bar{b}_i = & \frac{1}{\sqrt{p}} \operatorname{tr} \boldsymbol{\Sigma}_i (\mathbf{Q}_1 - \mathbf{Q}_0). \\ \overline{B_i} \triangleq & \frac{1}{p} \frac{2n_1\phi}{1-\left(\tilde{\delta}_1^2+\tilde{\delta}_0^2\right)\phi} - \frac{1}{p}\frac{2n_1\phi}{1-2\tilde{\delta}_0\tilde{\delta}_1\phi}.\end{aligned}$$

The detailed proofs will be provided in the full version of this paper. Plugging these deterministic equivalents into the misclassification error rate in Theorem 1, we obtain

Theorem 2: Under assumptions 1-4, the following convergence holds for $i \in \{0, 1\}$

$$\epsilon_i^{RDA} - \Phi\left((-1)^i \frac{\bar{\xi}_i - \bar{b}_i}{\sqrt{2\overline{B_i}}}\right) \xrightarrow{p} 0.$$

Proof 1: It appears that $\bar{\xi}_i$ and $\bar{b}_i$ are going to blow up at the rate of $\mathcal{O}\left(\sqrt{p}\right)$ so that the classification error will converge to a trivial value since $\frac{1}{\sqrt{p}}\log|\mathbf{Q}_i|$ and $\frac{1}{\sqrt{p}}\operatorname{tr}\boldsymbol{\Sigma}_i Q_i$ are $\mathcal{O}(\sqrt{p})$. However, because of Assumption 4, the distance of class covariances are controlled, which results in $\frac{1}{\sqrt{p}}\log|\mathbf{Q}_1| - \frac{1}{\sqrt{p}}\log|\mathbf{Q}_0|$ being $\mathcal{O}(1)$ and $\frac{1}{\sqrt{p}}\operatorname{tr}\boldsymbol{\Sigma}_i(\mathbf{Q}_1 - \mathbf{Q}_0)$ being $\mathcal{O}(1)$. Take $\frac{1}{\sqrt{p}}\operatorname{tr}\boldsymbol{\Sigma}_i(\mathbf{Q}_1 - \mathbf{Q}_0)$ for example. We first use *resolvent identity* to expand $\bar{b}_i$ as follows:

$$\begin{aligned}\bar{b}_i &= \frac{1}{\sqrt{p}} \operatorname{tr} \boldsymbol{\Sigma}_i \mathbf{Q}_1 (\mathbf{Q}_0^{-1} - \mathbf{Q}_1^{-1}) \mathbf{Q}_0 \\ &= \frac{1}{\sqrt{p}} \left(\frac{\alpha_0}{1+\alpha_0\delta_0} - \frac{\alpha_1}{1+\alpha_1\delta_1}\right) \operatorname{tr} \mathbf{Q}_0 \boldsymbol{\Sigma}_i \mathbf{Q}_1 (\boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_1)\end{aligned}$$

Since $||Q_0\boldsymbol{\Sigma}_i Q_1||$ is bounded, from Assumption 4, we can derive that $\frac{1}{\sqrt{p}}(\frac{\alpha_0}{1+\alpha_0\delta_0} - \frac{\alpha_1}{1+\alpha_1\delta_1})\operatorname{tr}\mathbf{Q}_0\boldsymbol{\Sigma}_i\mathbf{Q}_1(\boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_1) = \mathcal{O}(1)$. Therefore, we complete the convergence proof for $\bar{b}_i$ under Assumption 4. It is also easy to prove that $\frac{1}{\sqrt{p}}(-1)^{i+1}\boldsymbol{\mu}^T\mathbf{Q}_{\mathbf{1-i}}\boldsymbol{\mu} = \mathcal{O}(1)$ using Assumption 2. In conclusion, these assumptions are carefully established to guarantee a non-trivial classification error.

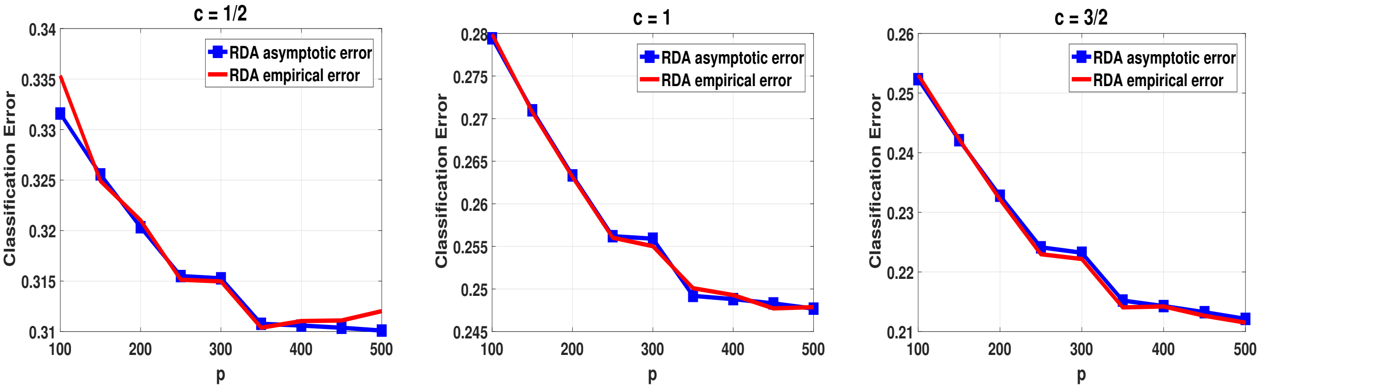
Theorem 2 reveals two important facts. First, the classification error rate can be characterized asymptotically by a deterministic quantity that depends solely on the parameters of the Gaussian model as well as the problem dimensions. The importance of this result lies in that it not only sheds light on the impact of these parameters but also it opens up possibilities of properly tuning the pair of γ and λ that correspond to the least asymptotic misclassification error rate.

Special cases:

1) As shown in [6], for the R-LDA to leverage the information about the mean classes, it suffices to have $\|\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\| = \mathcal{O}(1)$. In our case, when $\|\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\| = \mathcal{O}(1)$, the classification error of RDA can still converge to a non-trivial deterministic equivalence in which the contribution of the difference in means vanishes. This is because RDA suffers from a higher level of estimation noise, requiring the distance between the means $\|\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\|$ to scale as high as $O(\sqrt{p})$ so as to be leveraged by the classifier.
2) When $\|\boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_1\|_F = \mathcal{O}(1)$ and $\|\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\| = \mathcal{O}(1)$, it is easy to show that $\bar{b}_0 - \bar{b}_1 \to 0$. Therefore, we can show that $\epsilon_0^{RDA} = \phi(\omega)$ and $\epsilon_1^{RDA} = \phi(-\omega)$ where ω is some quantity that is the same for both classes. Finally, the classification error ϵ will converge to $\pi_0\epsilon_0^{RDA} + \pi_1\epsilon_1^{RDA} = 0.5\phi(\omega) + 0.5\phi(-\omega) = 0.5$. This indicates that when $\|\boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_1\|_F = O(1)$, the information about the covariance matrices is not exploited. In such circumstances, the use of R-LDA should be priorly considered.

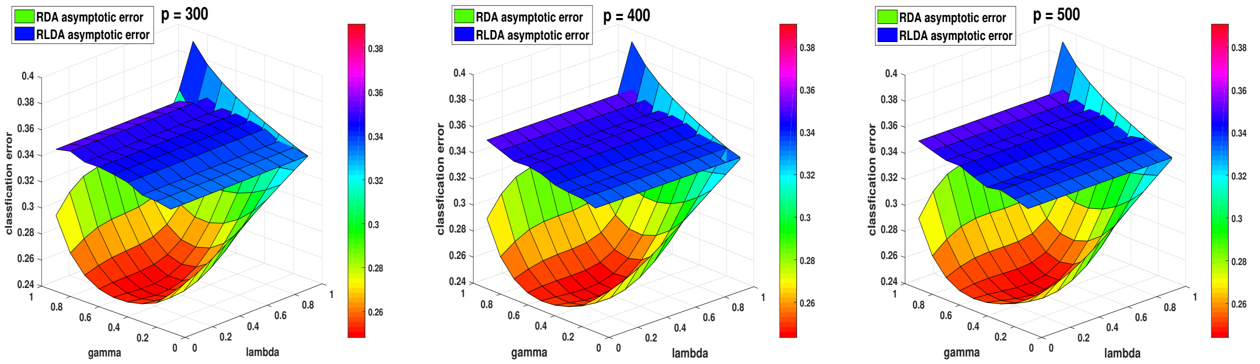
IV. Experiment

In this part, we provide numerical results to validate the accuracy of our theoretical findings. Define the following set of parameters for the Gaussian model: $[\boldsymbol{\Sigma}_0]_{i,j} = 0.6^{|i-j|}$, $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_0 + 2\mathbf{A}$, where $\mathbf{A} = \begin{bmatrix}\mathbf{I}_k & \mathbf{O}_{k\times(p-k)} \\ \mathbf{O}_{(p-k)\times k} & \mathbf{O}_{(p-k)\times(p-k)}\end{bmatrix}$ and $k = \lfloor\sqrt{p}\rfloor$. The statistical means are taken to be $\boldsymbol{\mu}_0 = \mathbf{1}_{p\times 1}$ and $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_0 + 2p^{-\frac{1}{4}}\mathbf{1}_{p\times 1}$. In the first experiment, we validate the correctness of our derived asymptotic error. We first estimate the means and covariances from the training data set and compute the empirical error by drawing $n_{test} = 2000$ samples from testing data set. We run 200 monte-carlo simulations for this process and take the average value for the final empirical error. Fixing $\gamma = 0.5$, $\lambda = 0.5$ and p varying from 100 to 500, Figure 1 shows the behavior of the classification error rate. We can see that the asymptotic error presents good agreement with the empirical error computed over the testing data. In a second experiment, taking $p \in \{300, 400, 500\}$ and $c = 1$, we want to verify what trends the classification error will present with the variation of both regularization parameters γ and λ. Since our derivations for RDA cannot generalize to R-LDA, we will compare the performance of R-LDA with RDA in terms of classification error separately. Finally, we get the 3D plots about RDA and R-LDA's variation trends of classification errors as shown in Figure 2. The blue curved planes are the misclassification error of R-LDA with variation of γ and the colorful surfaces are misclassification error of RDA with variation of γ and λ. It is clear from the plots that the minimum classification error is achieved in the most red region of the RDA surface below the R-LDA curved plane. Regularization parameter λ approaching the minimum classification error is neither 1 nor 0, which means that the optimal classifier minimizing the classification error is not one of the extreme cases, namely, neither R-LDA($\lambda = 1$) nor R-QDA ($\lambda = 0$). Instead, it lies somewhere between R-LDA and R-QDA, which aligns with our expectation that RDA offers better classification performance than R-LDA and R-QDA with proper regularizers selection.



(a) (b) (c)

Fig. 1. RDA classifier performance in terms of classification error with equal training, $n_0 = n_1$. The x axis is the number of the data dimension.



(a) (b) (c)

Fig. 2. Comparison of performance of RDA and R-LDA in terms of classification error with equal training, $n_0 = n_1$. The x axis is regularization parameter λ and y axis is regularization parameter γ for $p \in \{300, 400, 500\}$ and $c = 1$.

V. Conclusion

In this paper, we consider the double asymptotic regime where the sample size and data dimension increase comparatively in magnitude to study the performance of the RDA classifier. Under some mild assumptions controlling the distance between the class means and covariances, we show that the asymptotic classification error converges to a deterministic quantity relying merely on the data dimension and statistics of each class. This conclusion enables us to design an improved classifier by selecting a pair of regularization parameters that minimize the asymptotic classification error. We validate the accuracy of our theoretical findings using synthetic data which allows to see the advantage of using the RDA classifier as compared to its special cases R-LDA and R-QDA.

References

[1] G. McLachlan, *Discriminant analysis and statistical pattern recognition*. John Wiley & Sons, 2004, vol. 544.

[2] D. Bakirov, A. P. James, and A. Zollanvari, "An efficient method to estimate the optimum regularization parameter in RLDA," *Bioinformatics*, vol. 32, no. 22, pp. 3461–3468, 2016.

[3] P. Ou, "Prediction of Stock Market Index Movement by Ten Data Mining Techniques," *Modern Applied Science*, vol. 3, no. 12, Dec. 2009.

[4] J. Friedman, "Regularized discriminant analysis," *Journal of the American Statistical Association*, vol. 84, pp. 165–175, 1989.

[5] 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, 2015.

[6] K. Elkhalil, A. Kammoun, R. Couillet, T. Y. Al-Naffouri, and M. S. Alouini, "Asymptotic performance of regularized quadratic discriminant analysis based classifiers," in *IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP)*, Sept 2017, pp. 1–6.

[7] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*. Springer series in statistics New York, 2001, vol. 1.

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

[9] K. Elkhalil, A. Kammoun, R. Couillet, T. Y. Al-Naffouri, and M.-S. Alouini, "A Large Dimensional Study of Regularized Discriminant Analysis Classifiers," *ArXiv e-prints*, Nov. 2017.