Improved Design of Quadratic Discriminant Analysis Classifier in Unbalanced Settings

Thesis/Dissertation by

Amine Bejaoui

In Partial Fulfillment of the Requirements

For the Degree of

Masters of Science

King Abdullah University of Science and Technology

Thuwal, Kingdom of Saudi Arabia

March 2020

EXAMINATION COMMITTEE PAGE

The thesis of Amine Bejaoui is approved by the examination committee

Committee Chairperson: Mohamed-Slim Alouini
Committee Members: Raphaël Huser, Abla Kammoun

ABSTRACT

Improved Design of Quadratic Discriminant Analysis Classifier in Unbalanced Settings

Amine Bejaoui

The use of quadratic discriminant analysis (QDA) or its regularized version (R-QDA) for classification is often not recommended, due to its well-acknowledged high sensitivity to the estimation noise of the covariance matrix. This becomes all the more the case in unbalanced data settings for which it has been found that R-QDA becomes equivalent to the classifier that assigns all observations to the same class. In this paper, we propose an improved R-QDA that is based on the use of two regularization parameters and a modified bias, properly chosen to avoid inappropriate behaviors of R-QDA in unbalanced settings and to ensure the best possible classification performance. The design of the proposed classifier builds on a refined asymptotic analysis of its performance when the number of samples and that of features grow large simultaneously, which allows to cope efficiently with the high-dimensionality frequently met within the big data paradigm. The performance of the proposed classifier is assessed on both real and synthetic data sets and was shown to be much higher than what one would expect from a traditional R-QDA.

ACKNOWLEDGEMENTS

I would like to express my gratitude to my supervisor, Professor Mohamed-Slim Alouini for his continous support, encouragement and guidance.

I would like also to thank Doctor Abla Kammoun for her valuable help, guidance and critical evaluation throughout the course of this work. Her expertise in statistics and random matrix theory helped me to solve many complex problems that I faced during this thesis.

I would like also to thank Doctor Khalil Elkhalil from Duke University for his precious help and technical support during all the steps of this work.

I am also thankful to all my friends at KAUST for making this experience at this lovely university very enjoyable and exciting.

Finally, I am deeply grateful to my beloved parents, my sister and my brother for their unconditional love and support.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

QDA	Quadratic Discriminant Analysis
R-QDA	Regularized Quadratic Discriminant Analysis
RMT	Random Matrix theory
LDA	Linear Discriminant Analysis
LR	Logistic Regression
SVM	Support Vector Machine
DT	Decision Tree
RF	Random forest
KNN	K-Nearest Neighbors
NN	Neural Network

LIST OF SYMBOLS

p	Number of features
n	Number of samples
γ	Regularization parameter
$\mathcal{C}_i$	The class i to which belongs the observations
$\boldsymbol{\mu}_i$	True mean of the class $\mathcal{C}_i$
$\hat{\boldsymbol{\mu}}_i$	estimated mean of the class $\mathcal{C}_i$
$\boldsymbol{\Sigma}_i$	True covariance matrix of the class $\mathcal{C}_i$
$\widehat{\boldsymbol{\Sigma}}_i$	Estimated covariance matrix of the class $\mathcal{C}_i$
$\mathbf{H}_i(\gamma)$	Regularized estimator of the inverse of the covariance matrix $\widehat{\boldsymbol{\Sigma}}_i$ of the class $\mathcal{C}_i$

LIST OF FIGURES

Chapter 1

Introduction

1.1 Context and Motivation

Discriminant analysis encompasses a wide variety of techniques used for classification purposes. These techniques, commonly recognized among the class of model-based methods in the field of machine learning [1], rely merely on the fact that we assume a parametric model in which the outcome is described by a set of explanatory variables that follow a certain distribution. Among them, we particularly distinguish linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA) as the most representatives. LDA is often connected or confused with Fisher discriminant analysis (FDA) [2], a method of projecting the data into a subspace and turns out to coincide with LDA when the target subspace has two dimensions. Both LDA and QDA are obtained by maximizing the posterior probability under the assumption that observations follow normal distribution, with the single difference that LDA assumes common covariances across classes while QDA assumes the most general situation with classes possessing different means and covariances. If the data follow perfectly the normal distributions and the statistics are known in explicit form, QDA turns out to be the optimal classifier that achieves the lowest possible classification error rate [3]. It coincides with LDA when the covariances are equal but outperforms it when they are different. However, in practical scenarios, the use of QDA was not always shown to yield the expected performances. This is because the mean and covariance of each class, which are in general unknown, are estimated based on available training data

with perfectly known classes. The obtained estimates are then used as plug-in estimators in the classification rules associated with LDA and QDA. The estimation error of the class statistics causes a provably degradation of the performances which reaches very high levels when the number of samples is comparable or less than their sizes. In this latter situation, QDA and LDA, relying on computing the inverse of the covariance matrix could not be used. To overcome this issue, one technique consists in using a regularized estimate of the covariance matrix as a plug-in estimator of the covariance matrix giving the name to Regularized LDA (R-LDA) or Regularized QDA (R-QDA) to the associated classifiers. However, this solution does not allow for a significant reduction of the estimation noise. The situation is even worse for R-QDA, since the number of samples used to estimate the covariance matrix of each class is lower than that of LDA. This is probably the reason why LDA provided in many scenarios higher performances than QDA, although it might wrongly consider that the covariances across classes are equal.

A question of major theoretical and practical interest is to investigate to which extent the estimation noise of the covariance matrix impacts the performances of R-LDA and R-QDA. In this respect, the study of LDA and subsequently that of R-LDA have received a particular attention, dating back to the early works of Raudys [4], before being investigated again using recent advances of random matrix theory tools in a recent series of works [5, 6]. However, the theoretical analysis of QDA and R-QDA is more scarce and very often limited to specific situations in which the number of samples is higher than that of their dimensions [7], or under specific structures of the covariance matrices [8–10]. It was only recently that the work in [11] considered the analysis of R-QDA for general structures of the covariance matrices and identified the necessary asymptotic conditions under which QDA does not exhibit the trivial behavior by which it returns always the same class or randomly guesses it. Particularly, the work in [11] assumes balanced data across classes, because otherwise

R-QDA would tend to assign all observations to one class, thereby limiting the use of R-QDA in general settings.

This lies behind the main motivation of the present work. Based on a careful investigation of the asymptotic behavior of R-QDA under unbalanced settings in binary classification problems, we propose to amend the traditional R-QDA to cope with cases in which the proportions of training data from both classes are not equal. The new classifier is based on using two different regularization parameters instead of a common regularization parameter as well as an optimized bias properly chosen to minimize the misclassification error rates. Interestingly, we show that the proposed classifier not only outperforms R-LDA and R-QDA but also other state-of-the-art classification methods, opening promising avenues for the use of the proposed classifier in practical scenarios.

1.2 Notation

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. $(.)^T$,$(.)^H$, Tr and $|.|$ stands for the transpose, hermitian, the trace and the determinant of a matrix respectively. For two functions f and g, we say that $f = O(g)$, if $\exists 0 < M < \infty$ such that $|f| \leq Mg$. We say also that that $f = \Theta(g)$, if $\exists 0 < C_1 < C_2 < \infty$ such that $C_1 g \leq |f| \leq C_2 g$. $\mathcal{P}(.)$, $\xrightarrow{p}$ 0, and $\xrightarrow[M,K\to+\infty]{\text{a.s.}}$ respectively denote the probability measure, 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. $\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$.

Chapter 2

Random Matrix Theory

2.1 Introduction

In this chapter, we provide a quick overview about Random Matrix Theory and present some of the most important results that were established for Gram matrices in the random matrix theory that we are going to use in the next chapters.

2.2 Random Matrix Theory developments

Random Matrix Theory (RMT) is a powerful mathematical tool that focuses on characterizing spectral properties of large dimensional matrices made of random entries drawn from different probability distributions. It provides fascinating results that allow to transform random quantities into deterministic ones in the asymptotic regime where both the population size p and sample size n grow infinitely large while keeping a fixed ratio between these two dimensions. More importantly, RMT has the unique capacity to turn stochastically involved problems that do not have an exact mathematically intractable result into non-random ones which advocate for the great potential that emanates from this theory.

Primarily, the use of RMT can be traced back to the work of Wishart in 1928 with regards to his findings on matrices with Gaussian entries with finite dimensions. However, it is the results on nuclear physics brought by Wigner in 1967 that have unveiled the promising potential of RMT at that time. In fact, Wigner established asymptotic spectral characteristics of symmetric matrices, which are better known as

the Semicircular law. This has opened the door to the emergence of a wide range of other applications of RMT, such as wireless communication [12], signal processing [13], and financial mathematics [14].

Recently, RMT has received an unprecedented interest in the field of machine learning [15–17] for analyzing the performance of these popular machine learning algorithms. In fact, there has been a serious lack of understanding of machine learning algorithms over the years, even of the very basic ones. This is mainly due to the non-linear formulation of these algorithms added to the fact that the used dataset can be of any type (image, signal), which made these algorithms challenging to study. As a matter of fact, researchers were not able to provide relevant arguments on why a specific algorithm works very well on a specific dataset and performs poorly on another one, or to specify what is the best case to choose a specific algorithm rather than the other ones. So basically, the performance assessment of these algorithms is of big importance since it provides very useful insights on how to set their hyper-parameters. Thus far, the conventional methodology has been based on carrying out extensive simulations. However, this kind of methods presents two main drawbacks. Indeed, these simulations are often computationally expensive, and this becomes all the more the case when high-dimensional data are considered. Additionally, these simulations do not provide interpretable results in the sense that they do not provide insights into the impact of the statistics of the data (covariance, mean) on the performances. On the top of that, conducting a statistically rigorous analysis of these algorithms was not possible in the past due to the curse of dimensionality issues caused by the high dimensionality of the different variables, which resulted in an intractability of estimating a high dimensional quantity with precision. In contrast, thanks to RMT, the curse of dimensionality is no longer an issue, but it is considered as a blessing as described by Donoho [18]. This is mainly explained by the fact that RMT offers tools that allow us to control the fluctuations of random quantities in high dimensions and

allow for a proper estimation of the quantities of interest.

2.3 Resolvent and Gram matrices

Resolvent and Gram matrices [19] constitute a fundamental part in random matrix theory. They are defined as follows. Let $\mathbf{H} = (H_{m,k})$ be a $M \times K$ random matrix, then the matrices $\mathbf{H}\mathbf{H}^H$ and $\mathbf{Q}(\gamma) = (\gamma \mathbf{H}\mathbf{H}^H + \mathbf{I}_M)^{-1}$ denote respectively the Gram and resolvent matrices corresponding to $\mathbf{H}$. It is worth noting that it is essential to normalize the elements of $\mathbf{H}$ by $\frac{1}{K}$ in order to ensure that its norm is finite since the dimensions of these matrices can grow to infinity. As a result, the elements of $\mathbf{H}$ are often defined as:

$$H_{m,k} = \frac{\sigma_{m,k} W_{m,k}}{\sqrt{K}}$$

where $(W_{m,k})$ are i.i.d random variables with zero mean and variance 1, and $(\sigma_{m,k})$ is a bounded sequence of positive real numbers denoted as variance profile.

In summary, we can mainly identify three types of variance profiles of $(\sigma_{m,k})$ which are defined as follows

- Separable profile: The variance profile is said to be separable if there exist positive sequences d_m and $\tilde{d}_k$ such that:

$$\sigma_{m,k}^2 = d_m \tilde{d}_k$$

- Limit profile: The variance profile is said to be limit if there exists a continuous function $f(x, y)$ defined on $[0, 1]^2$ such that:

$$\sigma_{m,k}^2 = f\left(\tfrac{m}{M}, \tfrac{k}{K}\right)$$

- General profile: If the variance profile is not separable, it is said to be general.

Since the limiting spectral measure of Gram matrices exists only under special conditions, [20] established asymptotic equivalents for the asymptotic behavior of the spectral measure that are unique and valid for any variance profile.

Hereafter, we provide an important asymptotic equivalent result for separable variance profiles and present some of its properties.

2.3.1 Useful lemmas related to Gram matrices

Let $\mathbf{X} = [x_1, \ldots, x_N]^T$ an $p \times N$ matrix, where $\{x_i\}_{i=1}^N$ are drawn from a multivariate Gaussian distribution with mean $\mathbf{0}_{p\times 1}$ and covariance $\mathbf{\Sigma}_{p\times p}$. Let us assume also that p and n go to infinty while keeping a constant asymptotic ratio $\frac{p}{n} \longrightarrow c \in [0, 1]$. If $\|\mathbf{\Sigma}\| = \Theta(1)$, then we have the following convergence properties:

$$\frac{1}{p}\mathrm{Tr}\left(\mathbf{A}\mathbf{H}(t)\right) - \frac{1}{p}\mathrm{Tr}\left(\mathbf{A}\mathbf{T}(t)\right) \xrightarrow[M,K\to+\infty]{\text{a.s.}} 0, \qquad \text{for any matrix } \mathbf{A}$$

$$\mathbf{u}^T\mathbf{H}(t)\mathbf{v} - \mathbf{u}^T\mathbf{T}(t)\mathbf{v} \xrightarrow[M,K\to+\infty]{\text{a.s.}} 0, \qquad \forall \mathbf{u}, \mathbf{v} \text{ with finite norms}$$

where,

$$\mathbf{H}(\gamma) = \left(\gamma\frac{\mathbf{X}\mathbf{X}^\mathbf{T}}{N} + \mathbf{I}_p\right)^{-1}$$

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

$$\delta = \frac{1}{p}\mathrm{Tr}\left[\mathbf{\Sigma}\left(\frac{\gamma}{1+\gamma\delta}\mathbf{\Sigma} + \mathbf{I}_p\right)^{-1}\right]$$

It is worth noting that $\mathbf{T}(\gamma)$ is considered as the deterministic equivalent of $\mathbf{H}(\gamma)$. In fact, $\mathbf{T}(\gamma)$ does not depend on the observations $\{x_i\}_{i=1}^n$ and is only a function of the true covariance matrix $\mathbf{\Sigma}_{p\times p}$ of the multivariate Gaussian distribution.

The parameter δ is considered as a fixed point equation that can be solved iter-

atevely using the follwing algorithm.

Algorithm : Iterative algorithm for computing asymptotic equivalents.

Input : $\delta \leftarrow 0, \epsilon \leftarrow 10^{-3}$

Repeat:

1. $\delta^n \leftarrow \delta$
2. Compute $\mathbf{T}(\gamma) = \left(\frac{\gamma}{1+\gamma\delta}\mathbf{\Sigma} + \mathbf{I}_p\right)^{-1}$
3. Compute $\delta = \frac{1}{p}\mathrm{Tr}\left[\mathbf{\Sigma}\mathbf{T}(\gamma)\right]$

Until : $\frac{\delta - \delta^n}{\delta^n} < \epsilon$

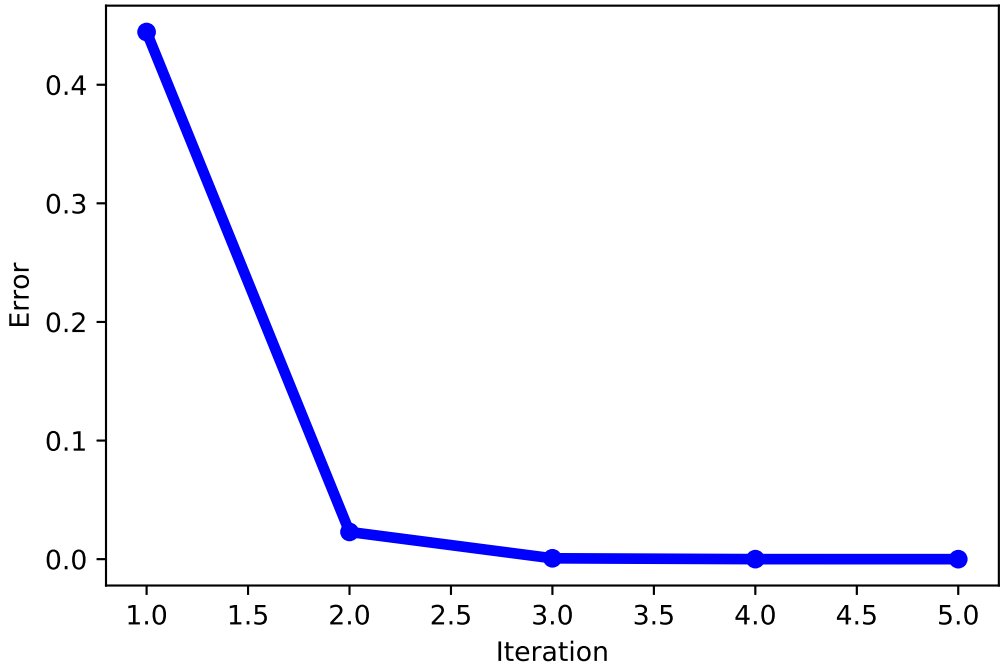


Figure 2.1: Number of iterations needed to converge.

2.4 Conclusion

In this chapter, we provided an introduction to RMT and we stated some important asymptotic results about Gram matrices that will come in handy in the next chapters.

Chapter 3

Quadratic discriminant analysis classifiers

3.1 Introduction

This chapter presents an introduction to the formulation of quadratic discriminant analysis classifier. It pinpoints exactly the problems related to the use of such classifier and paves the way for the formulation of a new improved version of a regularized quadratic discriminant analysis classifier that overcomes all these issues.

3.2 Regularized quadratic discriminant analysis

As aforementioned, R-QDA is equivalent to the classifier that assigns all observations to the same class when designed out of a set of unbalanced training data samples. Such a behavior has led the authors in [11] to consider the analysis of R-QDA only under a balanced training sample. In this section, we show that this behavior can be easily predicted through a close examination of the mean and variance of the classification rule associated with R-QDA. This constitutes an important step that will pave the way towards the improved R-QDA presented in the next section. But prior to that, we shall first review the traditional R-QDA for binary classification.

3.2.1 Regularized QDA for binary classification

For ease of presentation, we focus on binary classification problems where we have two distinct classes. We assume that the data follow a Gaussian mixture model, such that observations in class $\mathcal{C}_i$, $i \in \{0, 1\}$ are drawn from a multivariate Gaussian

distribution with mean $\boldsymbol{\mu}_i$ and covariance $\boldsymbol{\Sigma}_i$. More formally, we assume that

$$\mathbf{x} \in \mathcal{C}_i \quad \Leftrightarrow \quad \mathbf{x} = \boldsymbol{\mu}_i + \boldsymbol{\Sigma}_i^{1/2}\boldsymbol{z}, \quad \text{with} \quad \boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p) \tag{3.1}$$

$$\Leftrightarrow \quad P(\mathbf{x}|\mathbf{x} \in \mathcal{C}_i) = \frac{1}{\sqrt{(2\pi)^p |\boldsymbol{\Sigma}_i|}} \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1}(\mathbf{x} - \boldsymbol{\mu}_i)\right] \tag{3.2}$$

Let π_i, $i = 0, 1$, denote the prior probability that $\mathbf{x}$ belongs to class $\mathcal{C}_i$. By using the Bayes rule, we get:

$$P(\mathbf{x} \in \mathcal{C}_i|\mathbf{x}) = \frac{P(\mathbf{x}|\mathbf{x} \in \mathcal{C}_i)\pi_i}{P(\mathbf{x})}$$

We are intrested in finding the class $\mathcal{C}_i$, to which the observation $\mathbf{x}$ is more likely to belong. In other words, we are intrested in the sign of the following quantity

$$\log\left[\frac{P(\mathbf{x} \in \mathcal{C}_0|\mathbf{x})}{P(\mathbf{x} \in \mathcal{C}_1|\mathbf{x})}\right] = \log\left[\frac{P(\mathbf{x}|\mathbf{x} \in \mathcal{C}_0)\pi_0}{P(\mathbf{x}|\mathbf{x} \in \mathcal{C}_1)\pi_1}\right]$$

Thus, we end up with the following classification rule associated with the QDA classifier which is given by

$$\begin{aligned} W^{QDA}(\mathbf{x}) = &-\frac{1}{2}\log\frac{|\boldsymbol{\Sigma}_0|}{|\boldsymbol{\Sigma}_1|} - \frac{1}{2}\mathbf{x}^T\left(\boldsymbol{\Sigma}_0^{-1} - \boldsymbol{\Sigma}_1^{-1}\right)\mathbf{x} + \mathbf{x}^T\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 - \mathbf{x}^T\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 \\ &- \frac{1}{2}\boldsymbol{\mu}_0^T\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \frac{1}{2}\boldsymbol{\mu}_1^T\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 - \log\frac{\pi_1}{\pi_0} \end{aligned} \tag{3.3}$$

which is used to classify the observations based on the following rule:

$$\begin{cases} \mathbf{x} \in \mathcal{C}_0, & \text{if} \quad W^{QDA} > 0, \\ \mathbf{x} \in \mathcal{C}_1, & \text{otherwise.} \end{cases} \tag{3.4}$$

As seen from (3.3), the classification rule of QDA involves the true parameters of the Gaussian distribution, namely the means and covariances associated with each class. In practice, these parameters are not known. One approach to solve this issue

is to estimate them using the available training data. The obtained estimates are then used as plug-in estimators in (3.3). In particular, consider the case in which $n_i, i \in \{0,1\}$, training observations for each class $\mathcal{C}_i, i \in \{0,1\}$, are available and denote by $\mathcal{T}_0 = \{\mathrm{x}_l \in \mathcal{C}_0\}_{l=1}^{n_0}$ and $\mathcal{T}_1 = \{\mathrm{x}_l \in \mathcal{C}_1\}_{l=n_0+1}^{n_0+n_1=n}$ their respective samples. The sample estimates of the mean and covariances of each class are then given by:

$$\begin{aligned}
\hat{\boldsymbol{\mu}}_i &= \frac{1}{n_i} \sum_{l \in \mathcal{T}_i} \mathbf{x}_l, \quad i \in \{0,1\}, \\
\widehat{\boldsymbol{\Sigma}}_i &= \frac{1}{n_i - 1} \sum_{l \in \mathcal{T}_i} \left(\mathbf{x}_l - \hat{\boldsymbol{\mu}}_i\right) \left(\mathbf{x}_l - \hat{\boldsymbol{\mu}}_i\right)^T, \quad i \in \{0,1\}.
\end{aligned}$$

In case the number of samples n_0 or n_1 is less than the number of features p, the use of the sample covariance matrix as plug-in estimator is not permitted since the inverse $\boldsymbol{\Sigma}_i^{-1}$ could not be defined. A popular approach to circumvent this issue is to consider a regularized estimator of the inverse of the covariance matrix given by

$$\mathbf{H}_i(\gamma) = \left(\mathbf{I}_p + \gamma \widehat{\boldsymbol{\Sigma}}_i\right)^{-1} \quad , i \in \{0,1\} \tag{3.5}$$

where γ is a regularization parameter, which serves to shrink the sample covariance matrix towards identity. Replacing $\boldsymbol{\Sigma}_i^{-1}$ by $\mathbf{H}_i(\gamma)$ yields the following classification rule

$$\begin{aligned}
\widehat{W}^{R-QDA}(\mathbf{x}) = \frac{1}{2} \log \frac{|\mathbf{H}_0(\gamma)|}{|\mathbf{H}_1(\gamma)|} &- \frac{1}{2} \left(\mathbf{x} - \hat{\boldsymbol{\mu}}_0\right)^T \mathbf{H}_0(\gamma) \left(\mathbf{x} - \hat{\boldsymbol{\mu}}_0\right) \\
&+ \frac{1}{2} \left(\mathbf{x} - \hat{\boldsymbol{\mu}}_1\right)^T \mathbf{H}_1(\gamma) \left(\mathbf{x} - \hat{\boldsymbol{\mu}}_1\right) - \log \frac{\pi_1}{\pi_0}.
\end{aligned} \tag{3.6}$$

The R-QDA classifier assigns wrongly observation $\mathbf{x}$ if $\widehat{W}^{R-QDA}(\mathbf{x}) < 0$ when $\mathbf{x} \in \mathcal{C}_0$ or if $\widehat{W}^{R-QDA}(\mathbf{x}) > 0$ when $\mathbf{x} \in \mathcal{C}_1$. Conditioning on the training sample $\mathcal{T}_i, i \in \{0,1\}$,

the classification error associated with class $\mathcal{C}_i$, is thus given by

$$\epsilon_i^{R-QDA} = \mathbb{P}\left[(-1)^i \widehat{W}^{R-QDA}(\mathbf{x}) < 0 | \mathbf{x} \in \mathcal{C}_i, \mathcal{T}_0, \mathcal{T}_1\right] \tag{3.7}$$

which gives the following expression for the total misclassification error probability

$$\epsilon^{R-QDA} = \pi_0 \epsilon_0^{R-QDA} + \pi_1 \epsilon_1^{R-QDA}. \tag{3.8}$$

3.2.2 Identification of the problems of the R-QDA classifier in unbalanced data settings

In this section, we unveil several issues pertaining to the use of the classification rule (3.3) of R-QDA in high-dimensional settings. These issues can be revealed through a careful investigation of the asymptotic distribution of the classification rule associated with R-QDA. We first recall that the classification rule associated with R-QDA is a quadratic function of the Gaussian test observation $\mathbf{x}$ and as such behaves like a Gaussian distribution with a certain mean and variance as long as the Lyapunov conditions are met [21]. To get direct insights into how the R-QDA behaves, we assume that there is asymptotically no error in assuming that $\frac{1}{\sqrt{p}}\widehat{W}^{R-QDA}(\mathbf{x})$ when $\mathbf{x}$ belongs to class $\mathcal{C}_i$ behaves like a Gaussian distribution with mean $\overline{S}_i = E_{\mathbf{x}}(\frac{1}{\sqrt{p}}\widehat{W}^{R-QDA}(\mathbf{x}))$ and variance $\overline{V}_i = \mathrm{var}(\frac{1}{\sqrt{p}}\widehat{W}^{R-QDA}(\mathbf{x}))$ where here the expected value and variances are taken with respect to the distribution of the testing observation $\mathbf{x} \in \mathcal{C}_i$, and the scaling factor $\frac{1}{\sqrt{p}}$ is used to produce fluctuations of order $O(1)$. For the R-QDA to lead to appropriate behavior (including perfect classification error rate), the means $\overline{S}_i$ should be of opposite signs (namely $\overline{S}_0 > 0$ and $\overline{S}_1 < 0$) and at least of order $O(1)$ while the variances $\overline{V}_i$ be $O(1)$. This latter condition on the variance is already ensured provided that spectral norms of the covariances is bounded and the difference between mean vectors have a norm at most $O(p^{\frac{1}{4}})$. Under these assumptions, and

taking the expectation over the testing observations, $\overline{S}_i$ and $\overline{V}_i$ satisfy:

$$\overline{S}_i = \frac{1}{2\sqrt{p}} \log \frac{|\mathbf{H}_0(\gamma)|}{|\mathbf{H}_1(\gamma)|} - \frac{1}{2\sqrt{p}} \left(\boldsymbol{\mu}_i - \hat{\boldsymbol{\mu}}_0\right)^T \mathbf{H}_0(\gamma) \left(\boldsymbol{\mu}_i - \hat{\boldsymbol{\mu}}_0\right) + \frac{1}{2\sqrt{p}} \left(\boldsymbol{\mu}_i - \hat{\boldsymbol{\mu}}_1\right)^T \mathbf{H}_1(\gamma) \left(\boldsymbol{\mu}_i - \hat{\boldsymbol{\mu}}_1\right) - \frac{1}{2\sqrt{p}} \mathrm{Tr}\big[\boldsymbol{\Sigma}_i \mathbf{H}_0(\gamma)\big] + \frac{1}{2\sqrt{p}} \mathrm{Tr}\big[\boldsymbol{\Sigma}_i \mathbf{H}_1(\gamma)\big] - \frac{1}{\sqrt{p}} \log \frac{\pi_1}{\pi_0}. \tag{3.9}$$

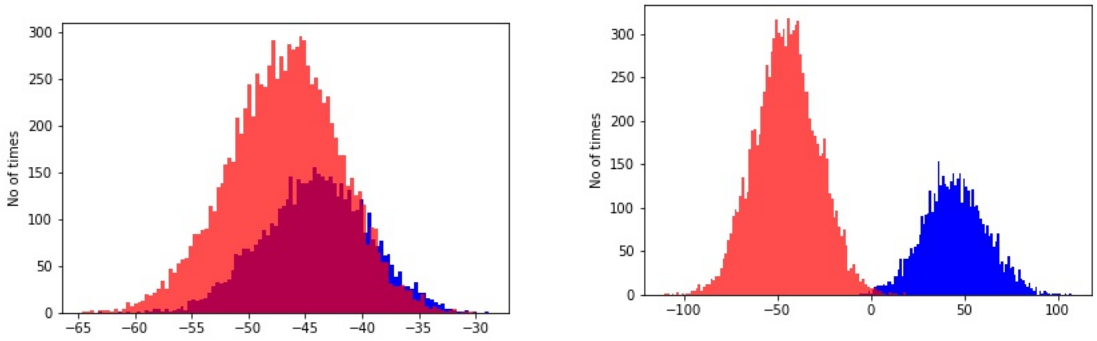
$$\overline{V}_i = O(1). \tag{3.10}$$

It can be easily seen that under the assumption that $\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0\| = O(p^{\frac{1}{4}})$, and the spectral norms of $\boldsymbol{\Sigma}_i, i = \{0, 1\}$ are bounded uniformly in p, the means $\overline{S}_i$ are asymptotically approximated as:

$$\overline{S}_i = \frac{1}{2\sqrt{p}} \log \frac{|\mathbf{H}_0(\gamma)|}{|\mathbf{H}_1(\gamma)|} - \frac{1}{2\sqrt{p}} \mathrm{Tr}\big[\boldsymbol{\Sigma}_i \mathbf{H}_0(\gamma)\big] + \frac{1}{2\sqrt{p}} \mathrm{Tr}\big[\boldsymbol{\Sigma}_i \mathbf{H}_1(\gamma)\big] + O(1) \tag{3.11}$$

Several important remarks are in order regarding (3.11). First, we note that the prior probabilities π_1 and π_0 do not play asymptotically any role in the classification, since the term $\frac{1}{2\sqrt{p}} \log \frac{\pi_1}{\pi_0}$ tends to zero. Second, one can easily see that if the distance between the covariances is such that $\frac{1}{\sqrt{p}} \mathrm{Tr}\big[\boldsymbol{\Sigma}_1 \mathbf{H}_0\big] - \frac{1}{\sqrt{p}} \mathrm{Tr}\big[\boldsymbol{\Sigma}_0 \mathbf{H}_0\big] = O(1)$ and $\frac{1}{\sqrt{p}} \mathrm{Tr}\big[\boldsymbol{\Sigma}_1 \mathbf{H}_1\big] - \frac{1}{\sqrt{p}} \mathrm{Tr}\big[\boldsymbol{\Sigma}_0 \mathbf{H}_1\big] = O(1)$ which occurs for instance when $\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_0$ has at most rank $\sqrt{p}$ [11], the means $\overline{S}_i$ are given by:

$$\overline{S}_i = \frac{1}{2\sqrt{p}} \log \frac{|\mathbf{H}_0(\gamma)|}{|\mathbf{H}_1(\gamma)|} - \frac{1}{2\sqrt{p}} \mathrm{Tr}\big[\boldsymbol{\Sigma}_1 \mathbf{H}_0(\gamma)\big] + \frac{1}{2\sqrt{p}} \mathrm{Tr}\big[\boldsymbol{\Sigma}_1 \mathbf{H}_1(\gamma)\big] + O(1).$$



(a) Regularized covariance estimate

(b) Perfect knowledge of the covariance matrices

Figure 3.1: Histogram of the classification rule for the case with regularized covariance estimate where $\gamma_0 = 10$ and the case with perfect knowledge of the covariance matrices. We consider $p = 1000$ features with unbalanced training size where $n_0 = 500, n_1 = 1000$, $[\mathbf{\Sigma}_0] = 10\mathbf{I}_p, \mathbf{\Sigma}_1 = \mathbf{\Sigma}_0, \boldsymbol{\mu}_0 = \mathbf{0}_{p\times 1}$ and $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_0 + \frac{3}{\sqrt{p}}\mathbf{1}_{p\times 1}$. The testing set is of size 5000 and 10000 samples for the first and second class respectively.

It appears thus that the direct use of R-QDA poses two main issues. The first one concerns the bias term, the contribution of which in $\overline{S}_1$ and $\overline{S}_0$ is asymptotically independent of the mean vectors and the prior probabilities. This makes R-QDA perform classification only on the basis of the covariance matrix. It is thus important to modify the bias term. The second issue is that unlike the balanced case for which $\overline{S}_1$ and $\overline{S}_0$ were shown $O(1)$ when there are exactly $\Theta(\sqrt{p})$ of eigenvalues with order $\Theta(1)$ [11], $\overline{S}_1$ and $\overline{S}_0$ are up to order $O(\sqrt{p})$ which is the same for both classes. This can be clearly illustrated through Figure 3.1 which displays the histogram associated with the classification rule of R-QDA and that of QDA with perfect knowledge of the statistics. As can be seen, the use of R-QDA does not allow discrimination between both classes since the means of the classification rule under class $\mathcal{C}_0$ or class $\mathcal{C}_1$ at the highest order is the same. Based on Random Matrix Theory results, we can prove that such a behavior is caused by the use of the same regularization parameterc γ for both $\mathbf{H}_0(\gamma)$ and $\mathbf{H}_1(\gamma)$. In light of these observations, we propose to replace the

classification rule of R-QDA by the following rule:

$$\widehat{W}^{R-QDA^{\text{imp}}}(\mathbf{x}) = \frac{-\theta}{2}\sqrt{p} - \frac{1}{2}\left(\mathbf{x} - \hat{\boldsymbol{\mu}}_0\right)^T \mathbf{H}_0(\gamma_0)\left(\mathbf{x} - \hat{\boldsymbol{\mu}}_0\right) \quad + \frac{1}{2}\left(\mathbf{x} - \hat{\boldsymbol{\mu}}_1\right)^T \mathbf{H}_1(\gamma_1)\left(\mathbf{x} - \hat{\boldsymbol{\mu}}_1\right). \tag{3.12}$$

where 1) $\gamma_0 \geq 0$ and $\gamma_1 \geq 0$ are two regularization parameters for each class carefully devised so that the means $\mathbb{E}_{\mathbf{x}}\left[\widehat{W}^{R-QDA^{\text{imp}}}(\mathbf{x})\right]$ when $\mathbf{x} \in \mathcal{C}_0$ or $\mathcal{C}_1$ are $O(1)$ and reflect the class under consideration and 2) θ is a bias term that will be set to the value that minimizes the asymptotic classification error rate.

3.3 Conclusion

In this chapter, we have presented an overview about quadratic discriminant analysis and identified the problems related to the use of the R-QDA classifier in unbalanced data settings, mainly the problem of the bias and the problem of perfect misclassification error rate characterized by a total disregard of the class with respect to the other class. In the light of these observations, an improved version of the R-QDA based on two regularization parameters and a bias has been proposed to overcome all these issues.

Chapter 4

Design of the improved R-QDA classifier

In this chapter, we study the design of our improved R-QDA classifier and provides guidelines on how to choose its optimal parameters. Specifically, in Section (4.1), we present the general framework and classification settings for this problem and state the assumptions on the data that we used. Section (4.2) establishes an asymptotic expression for the error of misclassification that will pave the way for determining the optimal parameters for our improved RQDA classifier in Section (4.3). Finally, Section (4.3.3) presents consistent estimates of these parameters that depend solely on the training samples.

4.1 General framework

In this section, we propose an improved design of the R-QDA classifier that fixes the aforementioned issues met in unbalanced settings. The design will be based on an asymptotic analysis of the statistics in (3.12) under the following asymptotic regime, which was also considered in [11]:

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

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

Assumption. 3 (Covariance scaling). $\|\boldsymbol{\Sigma}_i\| = \Theta(1)$, $i = 0, 1$.

Assumption. 4. Matrix $\boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_1$ has $\Theta(\sqrt{p})$ eigenvalues of order $\Theta(1)$. The remaining eigenvalues are of order $\left(\frac{1}{\sqrt{p}}\right)$.

Assumptions 1 and 3 are standard and are often used to describe a growth regime

in which the number of features scales comparably with that of samples and the spectral norm of both covariance matrices remain bounded. Assumption 2 defines the smallest distance between the mean vectors so that they are used to discriminate between both classes, while Assumption 4, introduced in [11] is used to ensure that the difference between covariances has a contribution that is of the same order of magnitude as that of the difference between the mean vectors. In other words, it allows us to ensure that for any matrix A of finite spectral norm, we have:

$$\frac{1}{\sqrt{p}}\mathrm{Tr}\big[\mathbf{A}\left(\boldsymbol{\Sigma}_0-\boldsymbol{\Sigma}_1\right)\big]=O(1).$$

4.2 Asymptotic estimate of the misclassification error

Under the asymptotic regime specified by Assumptions 1–4 and along the same lines as in [11], we analyze the classification error rate of the proposed classifier with the classification rule (3.12). In order to do so, we start by providing a simplified expression of the R-QDA classification error defined in (3.7) which can be written as:

$$\epsilon_i^{R-QDA}=\mathbb{P}\left[\boldsymbol{Z}^T\mathbf{B}_i\boldsymbol{Z}+2\boldsymbol{Z}^T\mathbf{r}_i<\xi_i|\boldsymbol{Z}\sim\mathcal{N}\left(\mathbf{0},\mathbf{I}_p\right),\mathcal{T}_0,\mathcal{T}_1\right] \tag{4.1}$$

where,

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

Using Proposition 1 from [11], we can say that:

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

. This result allows us to express the classification error rate of the class $\mathcal{C}_i$ in closed form as a function of a cumulative distribution function (CDF) of a normal distribution which paves the way down for the derivation of a deterministic approximation of the R-QDA classification error that depends only on the true parameters associated with each class.

Before presenting the corresponding result, we shall first introduce the following notation which defines deterministic objects that naturally appears when using Random Matrix Theory results.

For $i = 0, 1$, let δ_i be the unique positive solution to the following fixed point equation:

$$\delta_i = \frac{1}{n_i}\mathrm{Tr}\left[\mathbf{\Sigma}_i\left(\mathbf{I}_p + \frac{\gamma_i}{1+\gamma_i\delta_i}\mathbf{\Sigma}_i\right)^{-1}\right] \tag{4.2}$$

The existence and uniqueness of δ_i follows from standard results in Random Matrix Theory [22]. For $i = 0, 1$, we also define matrices $\mathbf{T}_i$ as:

$$\mathbf{T}_i = \left(\mathbf{I}_p + \frac{\gamma_i}{1+\gamma_i\delta_i}\mathbf{\Sigma}_i\right)^{-1}, \tag{4.3}$$

and the scalars ϕ_i and $\tilde{\phi}_i$ as:

$$\phi_i = \frac{1}{n_i}\mathrm{Tr}\left[\mathbf{\Sigma}_i^2\mathbf{T}_i^2\right], \quad \tilde{\phi}_i = \frac{1}{(1+\gamma_i\delta_i)^2}. \tag{4.4}$$

With this notation at hand, we are now in position to state our first asymptotic result:

Theorem 1. *Under Assumption 1–4, and assuming that the regularization parameters γ_0 and γ_1 are $\Theta(1)$, the classification error rate associated with class $\mathcal{C}_i$ satisfies:*

$$\epsilon_i^{R-QDA} - \Phi\left((-1)^i\frac{\overline{\xi}_i - \overline{b_i}}{\sqrt{2\overline{B_i} + 4\overline{r}_i}}\right) \xrightarrow{p} 0, \tag{4.5}$$

where

$$\overline{\xi}_i \triangleq \frac{1}{\sqrt{p}}\left[(-1)^{i+1}\boldsymbol{\mu}^T\mathbf{T}_{1-i}\boldsymbol{\mu}\right] + \theta, \quad \text{with} \quad \boldsymbol{\mu} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_0, \tag{4.6}$$

$$\overline{b}_i = \frac{1}{\sqrt{p}}\mathrm{Tr}\boldsymbol{\Sigma}_i\left(\mathbf{T}_1 - \mathbf{T}_0\right) \tag{4.7}$$

$$\overline{B_i} = \frac{\phi_i}{1-\gamma_i^2\phi_i\tilde{\phi}_i}\frac{n_i}{p} + \frac{1}{p}\mathrm{Tr}\left[\boldsymbol{\Sigma}_i^2\mathbf{T}_{1-i}^2\right] + \frac{n_i}{p}\frac{\gamma_{1-i}^2\tilde{\phi}_{1-i}}{1-\gamma_{1-i}^2\phi_{1-i}\tilde{\phi}_{1-i}}\left(\frac{1}{n_i}\mathrm{Tr}\left[\boldsymbol{\Sigma}_i\boldsymbol{\Sigma}_{1-1}\mathbf{T}_{1-i}^2\right]\right)^2 - \frac{2}{p}\mathrm{Tr}\left[\boldsymbol{\Sigma}_i\mathbf{T}_1\boldsymbol{\Sigma}_i\mathbf{T}_0\right] \tag{4.8}$$

$$\overline{r}_i = \frac{\frac{1}{p}\boldsymbol{\mu}^T\boldsymbol{\Sigma}_{1-i}\mathbf{T}_{1-i}^2\boldsymbol{\mu}}{1-\gamma_{1-i}^2\phi_{1-i}\tilde{\phi_{1-i}}}. \tag{4.9}$$

Proof. The proof follows along the same lines in [11] and is as such omitted. □

Remark: Under Assumption 4, it can be shown that $\overline{B_i}$ can be asymptotically simplified to

$$\overline{B_i} \triangleq \frac{2n_i}{p}\frac{\gamma_i^2\tilde{\phi}_i\phi_i^2}{1-\gamma_i^2\phi_i\tilde{\phi}_i} + \Theta(\frac{1}{\sqrt{p}}).$$

Moreover, the term $\overline{r}_i$ is $O(\frac{1}{\sqrt{p}})$ and as such converges to zero as p, n grow to infinity. However, in our simulations, we chose to work with the non-simplified expressions for $\overline{B}_i$ and to keep the term $\overline{r}_i$, since we observed that in doing so a better accuracy is obtained in finite-dimensional simulations.

4.3 Selection of optimal parameters for our proposed R-QDA classification algorithm

4.3.1 Optimal choice of the regularization parameters

The result of Theorem 1 allows us to provide guidelines on how to choose γ_0 and γ_1 and the optimal bias θ. As discussed before, the design should require the mean of the classification rule to be $\Theta(1)$ and to reflect the class under consideration. This

mean is represented in the asymptotic expression of the classification error rate by the quantity $\overline{\xi}_i - \overline{b}_i$ which, at first sight, is $\Theta(\sqrt{p})$ as $\overline{b}_i = \Theta(\sqrt{p})$ and $\overline{\xi}_i = \Theta(1)$. Moreover, the class of the testing observation is not reflected in $\overline{b}_i$ since under Assumption 3–4, in case $\overline{b}_i = O(\sqrt{p})$, $\overline{b}_i = \frac{1}{\sqrt{p}}\mathrm{Tr}\boldsymbol{\Sigma}_1(\mathbf{T}_1 - \mathbf{T}_0) + \Theta(1)$. To solve this issue, we need to design γ_1 and γ_0 such that $\overline{b}_i$ is $\Theta(1)$ or equivalently,

$$\frac{1}{p}\mathrm{Tr}\big[\boldsymbol{\Sigma}_1(\mathbf{T}_1 - \mathbf{T}_0)\big] = \Theta(\frac{1}{\sqrt{p}}) \tag{4.10}$$

so that $\overline{b}_0$ becomes different from $\overline{b}_1$ at its highest order. To this end, we prove that it suffices to select the regularization parameter associated with the class with the largest number of samples as:

Theorem 2. *Under assumption 1-4, and assume that $n_1 > n_0$, if*

$$\gamma_1 = \frac{\gamma_0}{1 - \left(\frac{1}{n_1} - \frac{1}{n_0}\right)\gamma_0 \mathrm{Tr}\big[\boldsymbol{\Sigma}_0\mathbf{T}_0\big]}, \tag{4.11}$$

where γ_0 is fixed to a given constant then $\overline{b}_i = O(1)$.

Proof. See appendix B. □

It is worth mentioning that in the balanced case, plugging $n_0 = n_1$ into (4.11) yields $\gamma_1 = \gamma_0$. It is thus not necessary to use different regularization parameters when the classes are balanced.

4.3.2 Optimal choice of the bias term

With this choice of the regularization parameters being set, the optimal bias can be chosen so that the asymptotic classification error rate given by

$$\overline{\epsilon} = \pi_0\Phi\left(-\frac{\overline{\xi}_0 - \overline{b}_0}{\sqrt{2\overline{B}_0}}\right) + \pi_1\Phi\left(-\frac{\overline{\xi}_1 - \overline{b}_1}{\sqrt{2\overline{B}_1}}\right)$$

is minimized.

Theorem 3. *The optimal bias that allows to minimize the asymptotic classification error rate is given by:*

$$\theta^* = \frac{\beta_1 - \beta_0}{2} - \frac{2\alpha^2}{\beta_1 + \beta_0} \log(\frac{\pi_1}{\pi_0}), \tag{4.12}$$

where

$$\begin{cases} \beta_0 = \frac{1}{\sqrt{p}} \left[-\boldsymbol{\mu}^T \mathbf{T}_1 \boldsymbol{\mu} \right] - \frac{1}{\sqrt{p}} \operatorname{tr} \boldsymbol{\Sigma}_0 \left(\mathbf{T}_1 - \mathbf{T}_0 \right); \\ \beta_1 = \frac{1}{\sqrt{p}} \left[-\boldsymbol{\mu}^T \mathbf{T}_0 \boldsymbol{\mu} \right] + \frac{1}{\sqrt{p}} \operatorname{tr} \boldsymbol{\Sigma}_1 \left(\mathbf{T}_1 - \mathbf{T}_0 \right); \\ \alpha = \sqrt{2\overline{B_0}}. \end{cases}$$

Proof. See Appendix C. □

Before proceeding further, it is important to note that thanks to the careful choice of the regularization parameters γ_0 and γ_1 provided in Theorem 2, the term $\frac{1}{\sqrt{p}} \mathrm{Tr}\left[\boldsymbol{\Sigma}_i \left(\mathbf{T}_1 - \mathbf{T}_0 \right) \right]$ is $\Theta(1)$ for i $\in \{0, 1\}$, Additionally, it can be shown easily that the term $\frac{1}{\sqrt{p}} \left[-\boldsymbol{\mu}^T \mathbf{T}_i \boldsymbol{\mu} \right]$ is of order $\Theta(1)$. As a result, both β_0 and β_1 are $\Theta(1)$.
On another note, it is worth mentioning that even in the case of balanced classes $n_0 = n_1$, characterized by $\gamma_1 = \gamma_0$ as proved in Theorem 2, the optimal bias is different from the one used in R-QDA. As such, the proposed design improves on the traditional R-QDA studied in [11] in the balanced case by optimally adapting the bias term to the case where the covariance matrix are not known.

4.3.3 General consistent estimates of the optimal parameters of our proposed R-QDA classifier

Theorem 2 and Theorem 3 provided in Section 4.3.2 can be used to obtain an optimized design of the proposed R-QDA classifier. As can be seen, the improved classifier employs only one regularization parameter associated with the class that presents the smallest number of training samples. Assume $\mathcal{C}_0$ is such a class. The regularization parameter associated with the other class cannot be arbitrarily chosen and should be set as (4.11), while the bias is selected according to (4.12). However, pursuing this design is not possible in practice due to the dependence of (4.11) and (4.12) on the true covariance matrices. To solve this issue, we propose in the following theorem a consistent estimator to estimate quantities arising in (4.11) and (4.12) that depend only on the training samples.

Theorem 4. *Assume $n_1 > n_0$ and let γ_0 be the regularization parameter associated with class $\mathcal{C}_0$. Let $\hat{\delta}_0$ be given by:*

$$\hat{\delta}_0 = \frac{1}{\gamma_0} \frac{\frac{p}{n_0} - \frac{1}{n_0}\mathrm{Tr}\big[\mathbf{H}_0(\gamma_0)\big]}{1 - \frac{p}{n_0} + \frac{1}{n_0}\mathrm{Tr}\big[\mathbf{H}_0(\gamma_0)\big]}$$

and define $\hat{\gamma}_1$ as:

$$\hat{\gamma}_1 = \frac{\gamma_0}{1 - \gamma_0\left(\frac{n_0}{n_1}\hat{\delta}_0 - \hat{\delta}_0\right)} \tag{4.13}$$

Then,

$$\hat{\gamma}_1 - \gamma_1 \xrightarrow[M,K\to+\infty]{\text{a.s.}} 0$$

where γ_1 is given in (4.11). *Define $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\alpha}$ as:*

$$\begin{cases} \hat{\beta}_0 = -\frac{1}{\sqrt{p}} \left(\hat{\boldsymbol{\mu}}_0 - \hat{\boldsymbol{\mu}}_1\right)^T \mathbf{H}_1(\hat{\gamma}_1) \left(\hat{\boldsymbol{\mu}}_0 - \hat{\boldsymbol{\mu}}_1\right) - \frac{1}{\sqrt{p}} \operatorname{tr}\left(\hat{\boldsymbol{\Sigma}}_0 \mathbf{H}_1(\hat{\gamma}_1)\right) + \frac{n_0}{\sqrt{p}} \hat{\delta}_0; \\ \hat{\beta}_1 = -\frac{1}{\sqrt{p}} \left(\hat{\boldsymbol{\mu}}_0 - \hat{\boldsymbol{\mu}}_1\right)^T \mathbf{H}_0(\gamma_0) \left(\hat{\boldsymbol{\mu}}_0 - \hat{\boldsymbol{\mu}}_1\right) - \frac{1}{\sqrt{p}} \operatorname{tr}\left(\hat{\boldsymbol{\Sigma}}_1 \mathbf{H}_0(\gamma_0)\right) + \frac{n_1}{\sqrt{p}} \hat{\delta}_1; \\ \hat{\alpha} = \sqrt{2\hat{B}_0}. \end{cases} \tag{4.14}$$

where $\hat{B}_0$ writes as:

$$\begin{aligned} \hat{B}_0 = & \left(1 + \gamma_0 \widehat{\delta}_0\right)^4 \frac{1}{p} \operatorname{Tr}\left[\widehat{\boldsymbol{\Sigma}}_0 \mathbf{H}_0(\gamma_0) \widehat{\boldsymbol{\Sigma}}_0 \mathbf{H}_0(\gamma_0)\right] \\ & - \frac{n_0}{p} \widehat{\delta}_0^2 \left(1 + \gamma_0 \widehat{\delta}_0\right)^2 + \frac{1}{p} \operatorname{Tr}\left[\widehat{\boldsymbol{\Sigma}}_0 \mathbf{H}_1(\hat{\gamma}_1) \widehat{\boldsymbol{\Sigma}}_0 \mathbf{H}_1(\hat{\gamma}_1)\right] \\ & - \frac{n_0}{p} \left(\frac{1}{n_0} \operatorname{Tr}\left[\widehat{\boldsymbol{\Sigma}}_0 \mathbf{H}_1(\hat{\gamma}_1)\right]\right)^2 \\ & - 2 \left(1 + \gamma_0 \hat{\delta}_0\right)^2 \frac{1}{p} \operatorname{Tr}\left[\widehat{\boldsymbol{\Sigma}}_0 \mathbf{H}_0(\gamma_0) \widehat{\boldsymbol{\Sigma}}_0 \mathbf{H}_1(\hat{\gamma}_1)\right] \\ & + \widehat{\delta}_0 \left(1 + \gamma_0 \widehat{\delta}_0\right) \frac{2}{p} \operatorname{Tr}\left[\widehat{\boldsymbol{\Sigma}}_0 \mathbf{H}_1(\hat{\gamma}_1)\right] \end{aligned} \tag{4.15}$$

Let $\hat{\theta}^\star$ be given by:

$$\hat{\theta}^\star = \frac{\hat{\beta}_1 - \hat{\beta}_0}{2} - \frac{2\hat{\alpha}^2}{\hat{\beta}_1 + \hat{\beta}_0} \log(\frac{\pi_1}{\pi_0}) \tag{4.16}$$

Then,

$$\widehat{\theta^\star} - \theta^\star \xrightarrow[M,K \to +\infty]{\text{a.s.}} 0,$$

where $\theta^\star$ is given in (4.12).

Proof. See Appendix D. □

It is worth mentioning that unlike γ_0, the regularization parameter γ_1 is random. It does not satisfy equality (4.11), but ensures (4.10) with high probability. Its use as a replacement of γ_1 would lead asymptotically to the same results as the improved classifier using γ_1.

With these consistent estimators at hand, we are now in position to present the

improved design of the R-QDA classifier:

Input: Assuming $n_1 \geq n_0$, let γ_0 the regularization parameter associated with class $\mathcal{C}_0$, $\mathcal{T}_0 = \{\mathbf{x}_l\}_{l=1}^{n_0}$ training samples in $\mathcal{C}_0$ and $\mathcal{T}_1 = \{\mathbf{x}_l\}_{l=n_0+1}^{n=n_0+n_1}$.

Output: Estimation of the parameters γ_1 and $\theta^\star$ to be plugged in (3.12).

1 Compute $\hat{\gamma}_1$ as in (4.13);

2 Compute $\hat{\theta}$ as in (4.16);

3 Return $\hat{\theta}$ and $\hat{\gamma}_1$ that will be plugged in the classification rule (3.12).

Algorithm 1: Improved design of the R-QDA classifier

4.4 Conclusion

In this chapter, we have conducted an asymptotic analysis of our proposed R-QDA classifier that allowed us to derive the optimal parameters for this classifier. Namely, this offered guidelines on how to properly select the regularization terms and the bias for this classifier. Then, we presented an algorithm that summarizes the design of our improved classifier using consistent estimates of these parameters that depend only on the training samples.

Chapter 5

Numerical Results

5.1 Introduction

In this chapter, we derive a general consistent estimator of the testing error of our proposed R-QDA classifier. Then, we validate its performance on both synthetic and real data sets.

5.2 General consistent estimator of the testing error

The improved design described in Algorithm 1 in Section 4.3.3 depends on the regularization parameter γ_0 associated with the class with the smallest number of training samples. One possible way to adjust this parameter is to resort to a traditional cross-validation approach which consists in estimating using a set of testing data the classification error rate for a set of candidate values for the regularization parameter γ_0. Such an approach is however computationally expensive and could not be used to test a large number of candidate values for γ_0. Additionally, it can lead to unreliable estimators characterized by a high variance especially in high dimensional data settings where data is scarce in which this issue becomes more striking. This is explained by the curse of dimensionality where the number of samples needed to support the result grows exponentially with the dimensionality [23] [24].

As an alternative we propose rather to build a consistent estimator of the classification error rate based on results from Random Matrix Theory. The main advantage of this approach is that it allows not only to have a reliable estimator that approx-

imates the empirical classification error with high precision, but it allows to build a consistent estimator of the error that is only dependent on the training data which is quick to compute. In other words, such an estimator represents a reliable indicator of the performance of the classifier and allows us to tune the classifier parameters. This is the objective of the following theorem:

Theorem 5. *Under Assumptions 1–4, a consistent estimator of the misclassification error rate associated with class $\mathcal{C}_i$ is given by:*

$$\hat{\epsilon}_i = \Phi\left((-1)^i \frac{\hat{\xi}_i - \hat{b}_i}{\sqrt{2\hat{B}_i + 4\hat{r}_i}}\right)$$

where $\hat{B}_0$ is given in (4.15) *and*

$$\begin{aligned}
\hat{\xi}_i &= \hat{\theta}^\star - \frac{1}{\sqrt{p}}\left(\hat{\boldsymbol{\mu}}_0 - \hat{\boldsymbol{\mu}}_1\right)^T \mathbf{H}_{1-i}(\gamma_i)\left(\hat{\boldsymbol{\mu}}_0 - \hat{\boldsymbol{\mu}}_1\right),\ i \in \{0,1\} \\
\hat{\delta}_i &= \frac{1}{\gamma_i}\frac{\left[\frac{p}{n_i} - \frac{1}{n_i}\mathrm{Tr}\left(\mathbf{H}_i(\gamma_i)\right)\right]}{1 - \frac{p}{n_i} + \frac{1}{n_i}\mathrm{Tr}\left[\mathbf{H}_i(\gamma_i)\right]}, \quad i \in \{0,1\} \\
\hat{b}_i &= \frac{(-1)^i}{\sqrt{p}}\mathrm{Tr}\left[\hat{\boldsymbol{\Sigma}}_i \mathbf{H}_{1-i}(\hat{\gamma}_{1-i})\right] + \frac{(-1)^{i+1} n_i}{\sqrt{p}}\hat{\delta}_i,\ i \in \{0,1\} \\
\hat{B}_1 &= \left(1 + \gamma_1\hat{\delta}_1\right)^4 \frac{1}{p}\mathrm{Tr}\left[\hat{\boldsymbol{\Sigma}}_1 \mathbf{H}_1(\hat{\gamma}_1)\hat{\boldsymbol{\Sigma}}_1 \mathbf{H}_1(\hat{\gamma}_1)\right] - \frac{n_1}{p}\hat{\delta}_1^2\left(1 + \gamma_1\hat{\delta}_1\right)^2 + \frac{1}{p}\mathrm{Tr}\left[\hat{\boldsymbol{\Sigma}}_1 \mathbf{H}_0(\gamma_0)\hat{\boldsymbol{\Sigma}}_1 \mathbf{H}_0(\gamma_0)\right] \\
&\quad - \frac{n_1}{p}\left(\frac{1}{n_1}\mathrm{Tr}\left[\hat{\boldsymbol{\Sigma}}_1 \mathbf{H}_0(\gamma_0)\right]\right)^2 - 2\left(1 + \gamma_1\hat{\delta}_1\right)^2 \frac{1}{p}\mathrm{Tr}\left[\hat{\boldsymbol{\Sigma}}_1 \mathbf{H}_1 \hat{\boldsymbol{\Sigma}}_1 \mathbf{H}_0(\gamma_0)\right] \\
&\quad + \hat{\delta}_1\left(1 + \gamma_1\hat{\delta}_1\right)\frac{2}{p}\mathrm{Tr}\left[\hat{\boldsymbol{\Sigma}}_1 \mathbf{H}_0(\gamma_0)\right] \\
\hat{r}_i &= \frac{1}{p}(\hat{\boldsymbol{\mu}}_0 - \hat{\boldsymbol{\mu}}_1)^T \mathbf{H}_{1-i}(\hat{\gamma}_{1-i})\hat{\boldsymbol{\Sigma}}_i \mathbf{H}_{1-i}(\hat{\gamma}_{1-i})(\hat{\boldsymbol{\mu}}_0 - \hat{\boldsymbol{\mu}}_1)
\end{aligned}$$

in the sense that:

$$\widehat{\epsilon}_i - \epsilon_i^{R-QDA} \xrightarrow[M,K\to+\infty]{\text{a.s.}} 0$$

***Proof**: The proof of this theorem can be derived from the results established in The-*

orem 2 in [11] and as such is omitted.

5.3 Numerical results

5.3.1 Validation with synthetic data

In this section, we assess the performance of our improved R-QDA classifier and compare it with the standard QDA classifier in the case of unbalanced data. To this end, we start by generating synthetic data for both classes that are compliant with the different assumptions used thoughout this work in order to validate our theoretical results.

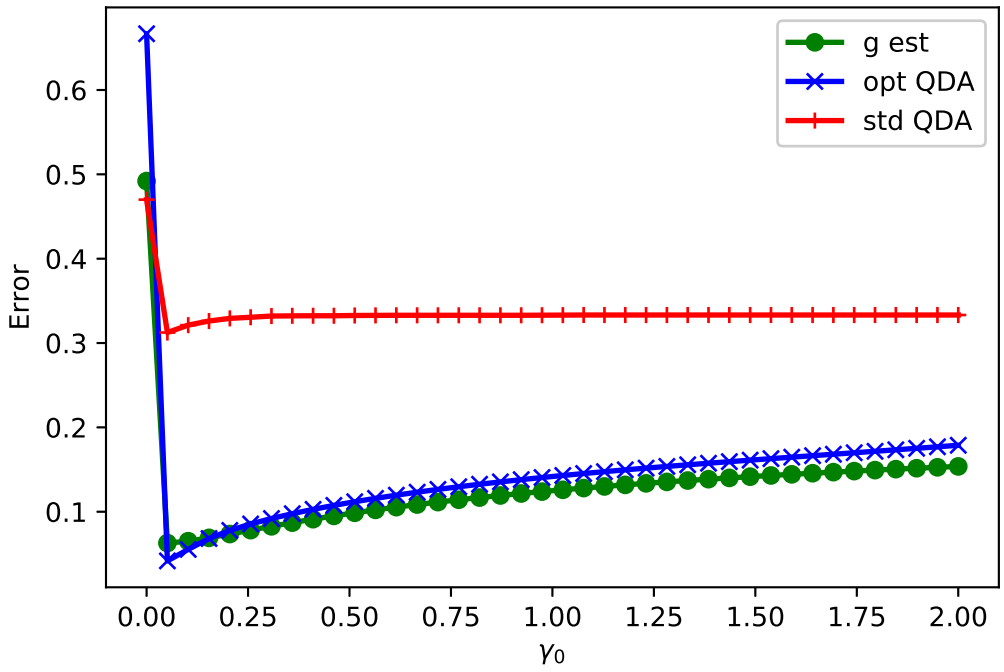


Figure 5.1: Average misclassification error rate versus the regularization parameter γ_0 using the G-estimator. We consider $p = 1000$ features with unbalanced training size where $n_0 = 2n_1$, $[\mathbf{\Sigma}_0] = 4\mathbf{I}_p$,$\mathbf{\Sigma}_1 = \mathbf{\Sigma}_0 + 3\mathbf{Q}_p\mathbf{D}_p\mathbf{Q}_p^T$, $\mathbf{Q}_p \in \mathcal{O}_n(R)$, $\mathbf{D}_p = diag\left[\mathbf{1}_{\sqrt{p}}, \mathbf{0}_{(p-\sqrt{p})}\right]$ and $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_0 + \frac{3}{\sqrt{p}}\mathbf{1}_{p\times 1}$.

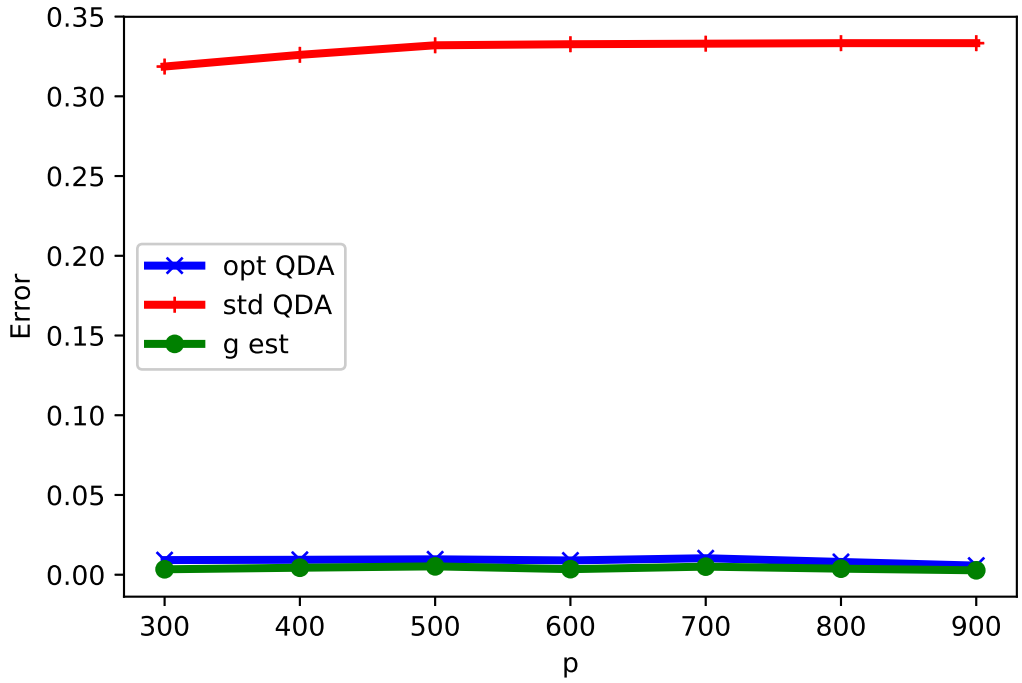


Figure 5.2: Average misclassification error rate versus the dimension p. We consider $\gamma_0 = 1$ with unbalanced training size where $n_0 = 2n_1$, $[\mathbf{\Sigma}_0] = 4\mathbf{I}_p$,$\mathbf{\Sigma}_1 = \mathbf{\Sigma}_0 + 3\mathbf{Q}_p\mathbf{D}_p\mathbf{Q}_p^T$, $\mathbf{Q}_p \in \mathcal{O}_n(R)$, $\mathbf{D}_p = diag\left[\mathbf{1}_{\sqrt{p}}, \mathbf{0}_{(p-\sqrt{p})}\right]$ and $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_0 + \frac{3}{\sqrt{p}}\mathbf{1}_{p\times 1}$.

In Figure 5.1 and Figure 5.2, we plot the classification error rate of the improved classifier and the traditional R-QDA classifier with respect to the regularization parameter γ_0 and the features' dimension p, respectively. As can be seen, we note that the standard R-QDA has a classification error rate that converges to the prior of the most dominant class, which reveals that as expected, it tends to assign all observations to the same class, which in this case coincides with the class that presents the highest number of training samples. On the opposite, the proposed R-QDA classifier presents a much higher performance, making it more suitable to cope with unbalanced settings. We finally note that the consistent estimator based on the results of Theorem 5 is accurate and as such can be used to properly adjust the regularization parameter γ_0.

5.3.2 Experiment with real data

In this section, we test the performance of the proposed R-QDA classifier on the public USPS dataset of handwritten digits [25] and the EEG dataset. The USPS dataset is composed of 42000 labeled digit images, and each image has $p = 784$ features represented by 28×28 pixels. The EEG dataset is composed of 5 classes that contain 4097 observations, and each observation has $p = 178$ features. We consider the classification of two classes from each dataset composed of n_0 and n_1 samples. Based on the results of Theorem 5, we tune the regularization factor γ_0 to the value that minimizes the consistent estimate of the misclassification error rate. The values of θ and $\hat{\gamma}_1$ are then computed based in (4.13) and (4.16). Fig. 5.3 and Fig.5.4 compares the performance of the proposed classifier with other state-of-the-art classification algorithms using cross-validation for different proportions of n_0 and n_1. As seen, our classifier, termed in the figure RQDA^{imp}, not only outperforms the standard QDA but also other existing classification algorithms. This suggests that the use of different regularization across classes in the QDA classification rule along with an adequate tune of the bias makes the QDA classifier more robust to the estimation noise of the covariance matrices in unbalanced settings.

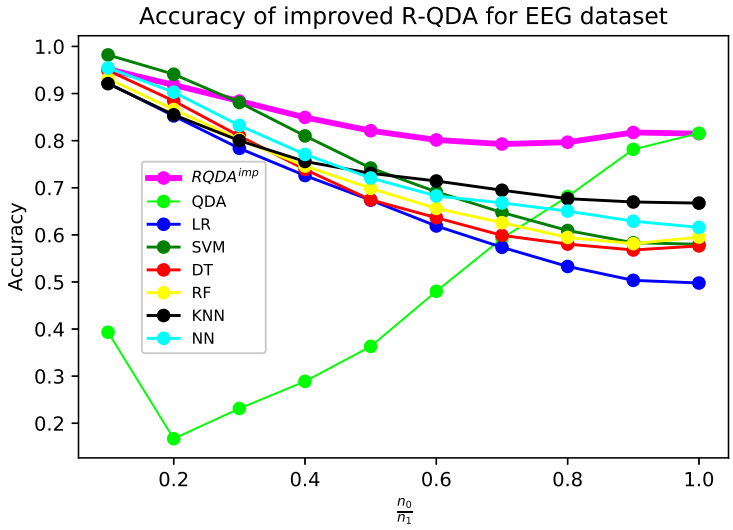


Figure 5.3: Comparaison between the performance of the our improved RQDA classifier with respect to other machine learning algorithms on the EEG dataset.

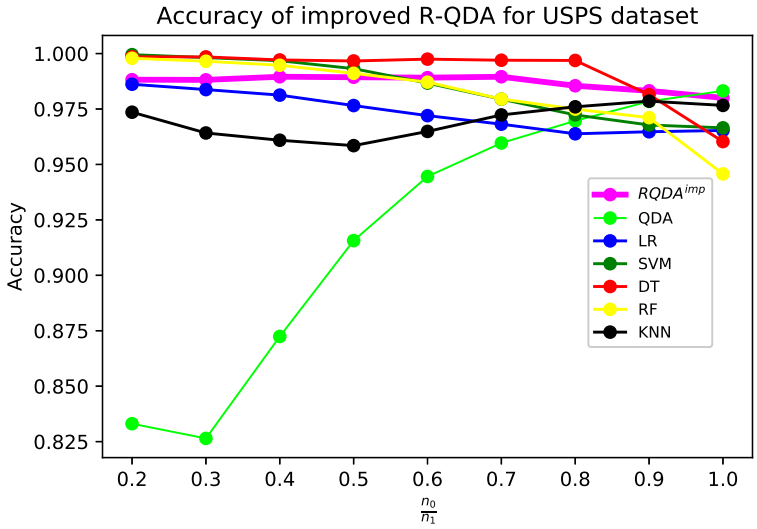


Figure 5.4: Comparaison between the performance of the our improved RQDA classifier with respect to other machine learning algorithms on the USPS dataset.

5.4 Conclusion

In this chapter, we have formulated a consistent estimator for the misclassification error. Then, we presented simulations on both synthetic data sets that accentuate the effectiveness of our proposed classifier with respect to other machine learning

algorithms.

Chapter 6

Conclusion

A common belief holds that the use of R-QDA leads in general to lower classification performances than many other existing classification methods, even though it is a classifier derived from the maximum likelihood principle under a general Gaussian mixture model. As a matter of fact, contrary to the other existing classifiers, the main issues of the R-QDA lies in its high sensitivity to the estimation noise of the parameters of the Gaussian mixture model. Through a careful investigation of the classification rule of R-QDA, we prove that in case of unbalanced training data, the estimation noise leads the R-QDA to assign all the observations to the same class, which explains its inefficiency to classify data under such settings. In this work, we propose to modify the design of R-QDA so that it becomes more resilient to the estimation noise. Particularly, we propose to use two regularization parameters (one for each class) as well as a carefully designed bias to optimize the classification performance. Our design, which leverages advanced results from Random Matrix Theory, clearly shows that there is room for improvement of basic classification methods based on the use of advanced statistical tools.

APPENDICES

A Useful Lemmas

A.1 Useful Lemmas

In this appendix, we present some technical results that are derived from random matrix theory with regards to the asymptotic behaviour of large random matrices. Throughout this section, Let $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_K]$ be a $M \times K$ standard complex Gaussian matrix. Let $t > 0$ and $\mathbf{D} = \mathrm{diag}\,(\alpha_1, \cdots, \alpha_K)$. The resolvent matrix of $\mathbf{X}\mathbf{D}\mathbf{X}^{\mathrm{H}}$ is denoted as:

$$\mathbf{Q}(t) = \left(\frac{t}{K} \sum_{i=1}^{K} \alpha_i \mathbf{x}_i \mathbf{x}_i^{\mathrm{H}} + \mathbf{I}_M \right)^{-1} = \left(\frac{t}{K} \mathbf{X}\mathbf{D}\mathbf{X}^{\mathrm{H}} + \mathbf{I}_M \right)^{-1}. \tag{A.1}$$

Lemma 6 (Convergence of quadratic forms)**.** *Let* $\mathbf{x} = [x_1, \ldots, x_M]^T \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_M)$. *Let* $\mathbf{A}$ *be an* $M \times M$ *random matrix with a bounded spectral norm such as* $\mathbf{A}$ *is independent of* $\mathbf{x}_M$*, Then, we have*

$$\frac{1}{M} \mathbf{x}^H \mathbf{A} \mathbf{x} - \frac{1}{M} \mathrm{tr}(\mathbf{A}) \xrightarrow[M \to +\infty]{\text{a.s.}} 0.$$

The following Lemma provides results allowing to approximate random quantities involving the resolvent matrix when their dimensions grow simultaneously large:

Lemma 7. *Let* $\delta(t)$ *be the unique positive solution to the following equation:*

$$\delta(t) = \frac{M}{K \left(1 + \frac{t}{K} \sum_{i=1}^{K} \frac{\alpha_i}{1 + t\delta(t)\alpha_i} \right)}. \tag{A.2}$$

Consider the asymptotic regime in which M and K grow to infinity with:

$$0 < \liminf \frac{M}{K} < \limsup \frac{M}{K} < \infty.$$

Then, we have the following convergance for a closed bonded interval $[a, b]$ in $[0, \infty)$:

$$\sup_{t\in[a,b]} \left| \frac{1}{K} \operatorname{tr} \mathbf{Q}(t) - \delta(t) \right| \xrightarrow[M,K\to+\infty]{\text{a.s.}} 0.$$

Also, if $\mathbf{y}_1, \cdots, \mathbf{y}_K$ denotes standard complex Gaussian vectors independent from $\mathbf{x}_1, \cdots, \mathbf{x}_K$, we get:

$$\max_{1\le j\le K} \sup_{t\in[a,b]} \left| \mathbf{y}_j^H \mathbf{Q}(t) \mathbf{y}_j - \delta(t) \right| \xrightarrow[M,K\to+\infty]{\text{a.s.}} 0.$$

B Appendix B

B.1 Proof of Theorem 2

As discussed in the paper, the design of the regularization parameters γ_0 and γ_1 should ensure that:

$$\frac{1}{\sqrt{p}}\mathrm{Tr}\big[\boldsymbol{\Sigma}_i\left(\mathbf{T}_1-\mathbf{T}_0\right)\big]=O(1) \tag{B.1}$$

where $\mathbf{T}_i=\left(\mathbf{I}+\gamma_i\widetilde{\delta}_i\boldsymbol{\Sigma}_i\right)^{-1}$, with $\widetilde{\delta}_i=\frac{1}{1+\gamma_i\delta_i}$. Using the relation $\mathbf{A}^{-1}-\mathbf{B}^{-1}=\mathbf{A}^{-1}(\mathbf{B}-\mathbf{A})\mathbf{B}^{-1}$ for any two square matrices $\mathbf{A}$ and $\mathbf{B}$, (B.1) boils down to:

$$\frac{1}{\sqrt{p}}\mathrm{Tr}\left[\boldsymbol{\Sigma}_i\mathbf{T}_1\left(\gamma_0\tilde{\delta}_0\boldsymbol{\Sigma}_0-\gamma_1\tilde{\delta}_1\boldsymbol{\Sigma}_1\right)\mathbf{T}_0\right]=O(1)$$

or equivalently:

$$\frac{\gamma_0\tilde{\delta}_0}{\sqrt{p}}\mathrm{Tr}\big[\boldsymbol{\Sigma}_i\mathbf{T}_1\left(\boldsymbol{\Sigma}_0-\boldsymbol{\Sigma}_1\right)\mathbf{T}_0\big]+\frac{\gamma_0\tilde{\delta}_0-\gamma_1\tilde{\delta}_1}{\sqrt{p}}\mathrm{Tr}\big[\boldsymbol{\Sigma}_i\mathbf{T}_1\boldsymbol{\Sigma}_1\mathbf{T}_0\big]=O(1)$$

Using Assumption 4, it can be readily seen that the first term $\frac{\gamma_0\tilde{\delta}_0}{\sqrt{p}}\mathrm{Tr}\big[\boldsymbol{\Sigma}_i\mathbf{T}_1\left(\boldsymbol{\Sigma}_0-\boldsymbol{\Sigma}_1\right)\mathbf{T}_0\big]=\Theta(1)$. To satisfy (B.1), we thus only need to design γ_0 and γ_1 such that:

$$\gamma_0\tilde{\delta}_0-\gamma_1\tilde{\delta}_1=O(1/\sqrt{p})$$

or equivalently:

$$\gamma_0+\frac{\gamma_0\gamma_1}{n_1}\mathrm{Tr}\big[\boldsymbol{\Sigma}_1\mathbf{T}_1\big]-\gamma_1-\frac{\gamma_0\gamma_1}{n_0}\mathrm{Tr}\big[\boldsymbol{\Sigma}_0\mathbf{T}_0\big]=\Theta(1/\sqrt{p})$$

Under Assumption 4,

$$\frac{1}{n_0}\mathrm{Tr}\big[\mathbf{\Sigma}_0\mathbf{T}_0\big] = \frac{1}{n_0}\mathrm{Tr}\big[\mathbf{\Sigma}_1\mathbf{T}_1\big] + O(\frac{1}{\sqrt{p}})$$

which proves that in choosing γ_1 given by:

$$\gamma_1 = \frac{\gamma_0}{1 - \left(\frac{1}{n_1} - \frac{1}{n_0}\right)\gamma_0\mathrm{Tr}\big[\mathbf{\Sigma}_0\mathbf{T}_0\big]}$$

the condition (B.1) becomes satisfied.

C Appendix C

C.1 Proof of Theorem 3

The choice of the regularization parameters γ_0 and γ_1 allows to ensure that:

$$\overline{B}_0 = \overline{B}_1 + O(\frac{1}{\sqrt{p}})$$

As a result, the expression of the asymptotic equivalents for the classification error rate of both classes defined in (4.5) for $i \in \{0, 1\}$ can be reduced to:

$$\epsilon_i^{R-QDA} - \Phi\left((-1)^i \frac{\overline{\xi}_i - \overline{b_i}}{\sqrt{2\overline{B}_0}}\right) \xrightarrow{p} 0 \tag{C.1}$$

Then, the total classification error can be written as:

$$\epsilon^{R-QDA} = \pi_0 \Phi\left(\frac{\beta_0 + \theta}{\alpha}\right) + \pi_1 \Phi\left(\frac{\beta_1 - \theta}{\alpha}\right)$$

$$\text{where} \begin{cases} \beta_0 = \frac{1}{\sqrt{p}}\left[-\boldsymbol{\mu}^T \mathbf{T}_1 \boldsymbol{\mu}\right] - \frac{1}{\sqrt{p}} \mathrm{Tr}\left[\boldsymbol{\Sigma}_0 \left(\mathbf{T}_1 - \mathbf{T}_0\right)\right] \\ \beta_1 = \frac{1}{\sqrt{p}}\left[-\boldsymbol{\mu}^T \mathbf{T}_0 \boldsymbol{\mu}\right] + \frac{1}{\sqrt{p}} \mathrm{Tr}\left[\boldsymbol{\Sigma}_1 \left(\mathbf{T}_1 - \mathbf{T}_0\right)\right] \\ \alpha = \sqrt{2\overline{B}_0} \end{cases}$$

Taking the derivative of this expression with respect to θ and setting it to zero, the optimal bias $\theta^\star$ should satisfy:

$$\frac{\pi_0}{\pi_1} e^{(\frac{\beta_1 - \theta^\star}{2\alpha})^2 - (\frac{\beta_0 + \theta^\star}{2\alpha})^2} = 1$$

Applying the logarithmic function on both sides, we obtain:

$$\log(\frac{\pi_0}{\pi_1}) + \left(\frac{\beta_1 - \theta^\star}{2\alpha}\right)^2 - \left(\frac{\beta_0 + \theta^\star}{2\alpha}\right)^2 = 0$$

thus leading to

$$\theta^* = \frac{\beta_1 - \beta_0}{2} - \frac{2\alpha^2}{\beta_1 + \beta_0} \log(\frac{\pi_1}{\pi_0})$$

D Appendix D

D.1 Proof Theorem 4

In Theorem 4, we provide a consistent estimator for the regularization parameter γ_1 that satisfies (4.10) with high probability and a consistent estimator for the optimal bias $\theta^\star$.

D.1.1 Consistent estimator for γ_1

We start by proving that $\gamma_1 - \hat{\gamma}_1 \xrightarrow[M,K\to+\infty]{\text{a.s.}} 0$. To this end, we need to provide a consistent estimator for $(\frac{1}{n_1} - \frac{1}{n_0})\text{Tr}\big[\mathbf{\Sigma}_0\mathbf{T}_0\big]$. We start by noticing that:

$$(\frac{1}{n_1} - \frac{1}{n_0})\text{Tr}\big[\mathbf{\Sigma}_0\mathbf{T}_0\big] = (\frac{n_0}{n_1} - 1)\delta_0$$

A consistent estimator for δ_0 has been provided in [11] and is given by:

$$\hat{\delta}_0 = \frac{1}{\gamma_0}\frac{\frac{p}{n_0} - \frac{1}{n_0}\text{Tr}\big[\mathbf{H}_0(\gamma_0)\big]}{1 - \frac{p}{n_0} + \frac{1}{n_0}\text{Tr}\big[\mathbf{H}_0(\gamma_0)\big]}$$

and as such a consistent estimator for γ_1 in (4.11) is given by:

$$\hat{\gamma}_1 = \frac{\gamma_0}{1 - \gamma_0(\frac{n_0}{n_1}\hat{\delta}_0 - \hat{\delta}_0)}$$

Note that the replacement of γ_1 by $\hat{\gamma}_1$ still ensures condition (B.1) since from standard results of random matrix theory $\hat{\delta}_0 - \delta_0 = O(\frac{1}{p})$ with high probability.

D.1.2 Consistent estimator for $\theta^\star$

Recall that

$$\theta^\star = \frac{\beta_1 - \beta_0}{2} - \frac{2\alpha^2}{\beta_1 + \beta_0}\log(\frac{\pi_1}{\pi_0})$$

To provide a consistent estimator for $\theta^\star$, it is thus required to provide that of β_0, β_1 and α. Since $\alpha = \sqrt{2\overline{B_0}}$ and $\hat{B}_0 - \overline{B}_0 \overset{\text{a.s.}}{\to} 0$, we thus have: $\hat{\alpha} - \alpha \overset{\text{a.s.}}{\to} 0$ where $\hat{\alpha} = \sqrt{2\hat{B}_0}$. As for β_i, $i = 0, 1$, it can be written as:

$$\begin{aligned}\beta_i &= -\frac{1}{\sqrt{p}}\boldsymbol{\mu}^T\mathbf{T}_{1-i}\boldsymbol{\mu} + \frac{1}{\sqrt{p}}\mathrm{Tr}\big[\boldsymbol{\Sigma}_i\mathbf{T}_i\big] - \frac{1}{\sqrt{p}}\mathrm{Tr}\big[\boldsymbol{\Sigma}_i\mathbf{T}_{1-i}\big] \\ &= -\frac{1}{\sqrt{p}}\boldsymbol{\mu}^T\mathbf{T}_{1-i}\boldsymbol{\mu} - \frac{1}{\sqrt{p}}\mathrm{Tr}\big[\boldsymbol{\Sigma}_i\mathbf{T}_{1-i}\big] + \frac{n_i}{\sqrt{p}}\delta_i\end{aligned}$$

Due to the independence of $\boldsymbol{\Sigma}_i$ from $\mathbf{H}_{1-i}$ and of $\hat{\mu}_1$ and $\hat{\mu}_0$ and $\mathbf{H}_i$, $i = 0, 1$, we have:

$$\frac{1}{\sqrt{p}}\mathrm{Tr}\big[\hat{\boldsymbol{\Sigma}}_i\mathbf{H}_{1-i}\big] - \frac{1}{\sqrt{p}}\mathrm{Tr}\big[\boldsymbol{\Sigma}_i\mathbf{T}_{1-i}\big] \xrightarrow[M,K\to+\infty]{\text{a.s.}} 0$$

and

$$\frac{1}{\sqrt{p}}(\hat{\boldsymbol{\mu}}_0 - \hat{\boldsymbol{\mu}}_1)\mathbf{H}_{1-i}(\hat{\boldsymbol{\mu}}_0 - \hat{\boldsymbol{\mu}}_1) - \frac{1}{\sqrt{p}}(\hat{\boldsymbol{\mu}}_0 - \hat{\boldsymbol{\mu}}_1)\mathbf{T}_{1-i}(\hat{\boldsymbol{\mu}}_0 - \hat{\boldsymbol{\mu}}_1) \xrightarrow[M,K\to+\infty]{\text{a.s.}} 0.$$

Papers Submitted

• A. Bejaoui, K. Elkhalil A. Kammoun, M. S. Alouini, T. Alnaffouri "Improved Quadratic Discriminant Analysis in unbalanced settings", *Submitted to ICML* , 2019.

• A. Bejaoui, K-H. Park, M. S. Alouini "A QoS-Oriented Trajectory Optimization in Swarming Unmanned-Aerial-Vehicles Communications", *Accepted in IEEE WCL*, 2019.

REFERENCES

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

[2] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of eugenics*, vol. 7, no. 2, pp. 179–188, 1936.

[3] T. H. J. Friedman and R. Tibshirani, *The Elements of Statistical Learning*, 2009.

[4] S. Raudys, "On determining training sample size of a linear classifier," *Computing Systems*, vol. 28, pp. 79–87, 1967, in Russian.

[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, June 2015.

[6] C. Wang and B. Jiang, "On the dimension effect of regularized linear discriminant analysis," *Electronic Journal of Statistics*, vol. 12, pp. 2709–2742, 2018.

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

[8] Y. Cheng, "Asymptotic probabilities of misclassification of two discriminant functions in cases of high dimensional data," *Statistics & Probability Letters*, vol. 67, pp. 9–17, 03 2004.

[9] Q. Li and J. Shao, "Sparse quadratic discriminant analysis for high dimensional data," *Statistica Sinica*, vol. 25, 04 2015.

[10] B. Jiang, X. Wang, and C. Leng, "Quda: A direct approach for sparse quadratic discriminant analysis," *Journal of Machine Learning Research*, vol. 19, 09 2015.

[11] K. Elkhalil, A. Kammoun, R. Couillet, T. Y. Al-Naffouri, and M.-S. Alouini, "A large dimensional study of regularized discriminant analysis classifiers," vol. abs/1711.00382, 2017. [Online]. Available: https://arxiv.org/abs/1711.00382

[12] R. Couillet and M. Debbah, *Random Matrix Methods for Wireless Communications*. New York, NY, USA: Cambridge University Press, 2011.

[13] R. Couillet and M. Debbah, "Signal processing in large systems: A new paradigm," *IEEE Signal Processing Magazine*, vol. 30, no. 1, pp. 24–39, Jan 2013.

[14] R. Couillet and M. McKay, "Large dimensional analysis and optimization of robust shrinkage covariance matrix estimators," *Journal of Multivariate Analysis*, vol. 131, 01 2014.

[15] G. Wainrib and J. Touboul, "Topological and dynamical complexity of random neural networks," *Physical Review Letters*, vol. 110, 10 2012.

[16] T. J. Hastie, A. Montanari, S. Rosset, and R. J. Tibshirani, "Surprises in high-dimensional ridgeless least squares interpolation," *ArXiv*, vol. abs/1903.08560, 2019.

[17] C. Thrampoulidis, E. Abbasi, and B. Hassibi, "The lasso with non-linear measurements is equivalent to one with linear measurements," 06 2015.

[18] D. L. Donoho, "High-dimensional data analysis: The curses and blessings of dimensionality," in *AMS CONFERENCE ON MATH CHALLENGES OF THE 21ST CENTURY*, 2000.

[19] R. Couillet, M. Debbah, and J. W. Silverstein, "A deterministic equivalent for the analysis of correlated mimo multiple access channels," *IEEE Transactions on Information Theory*, vol. 57, no. 6, pp. 3493–3514, June 2011.

[20] W. Hachem, P. Loubaton, and J. Najim, "Deterministic equivalents for certain functionals of large random matrices." *Ann. Appl. Probab*, vol. 17, pp. 875–930, 2007.

[21] P. Billingsley, *Probability and Measure*. Wiley, 1995, 3rd edition.

[22] 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.

[23] P. Domingos, "A few useful things to know about machine learning," *Commun. ACM*, vol. 55, no. 10, pp. 78–87, Oct. 2012. [Online]. Available: http://doi.acm.org/10.1145/2347736.2347755

[24] E. Dougherty, C. Sima, J. Hua, B. Hanczar, and U. Braga-Neto, "Performance of error estimators for classification," *Current Bioinformatics*, vol. 5, pp. 53–67, 03 2010.

[25] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov 1998.