2017 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, SEPT. 25–28, 2017, TOKYO, JAPAN

ASYMPTOTIC PERFORMANCE OF REGULARIZED QUADRATIC DISCRIMINANT ANALYSIS BASED CLASSIFIERS

Khalil Elkhalil[*], *Abla Kammoun*[*], *Romain Couillet*[†], *Tareq Y. Al-Naffouri*[*] *and Mohamed-Slim Alouini*[*]

[*] CEMSE Division, King Abdullah University of Science and Technology, Saudi Arabia
[†] CNRS-Centrale Supélec-Université Paris-Sud, France

ABSTRACT

This paper carries out a large dimensional analysis of the standard regularized quadratic discriminant analysis (QDA) classifier designed on the assumption that data arise from a Gaussian mixture model. The analysis relies on fundamental results from random matrix theory (RMT) when both the number of features and the cardinality of the training data within each class grow large at the same pace. Under some mild assumptions, we show that the asymptotic classification error converges to a deterministic quantity that depends only on the covariances and means associated with each class as well as the problem dimensions. Such a result permits a better understanding of the performance of regularized QDA and can be used to determine the optimal regularization parameter that minimizes the misclassification error probability. Despite being valid only for Gaussian data, our theoretical findings are shown to yield a high accuracy in predicting the performances achieved with real data sets drawn from popular real data bases, thereby making an interesting connection between theory and practice.

Index Terms— QDA, classification, machine learning, deterministic equivalent, random matrix theory.

1. INTRODUCTION

1.1. Overview of Discriminant Analysis for Classification

Discriminant analysis is part of a larger class of classification methods commonly known in the machine learning community as model-based classification methods [1–3]. These methods rely on the assumption that the input data follow a certain distribution. A classifier is then designed so as to minimize a certain classification metric [2]. Linear and Quadratic discriminant analysis (LDA and QDA), merely relying on the assumption of the data following Gaussian distribution, are among the most popular representatives [4]. Both methods are designed to assign for a given input data the class that presents the highest posterior probability. Their major unique difference is that LDA presumes equal covariance matrices for both classes but different means whereas QDA assumes different covariances and means across classes. By construction, they both require the knowledge of the Gaussian parameters for each class. This can be performed by estimating these parameters from the available training points using maximum likelihood estimation, a way that should be effective if the number of training samples is sufficiently high. However, when the number of training samples is small compared to their dimensions, maximum likelihood covariance matrix estimates can be poorly conditioned, leading to high misclassification error rates. One popular approach to solve the ill-posed estimation consists in regularizing the covariance estimation [5]. It has led to the emergence of regularized versions of discriminant analysis, termed as regularized LDA (R-LDA) and regularized QDA (R-QDA). In this paper, the focus is on regularized QDA.

1.2. Previous works

A large body of research has been conducted to analyze the performance of discriminant analysis classifiers. One approach, carried out under the assumption of exact dimensions and hinging on properties of the Wishart distribution, has been pursued in [6] to derive the exact misclassification error rate of the QDA. Such an analysis was limited to the case in which the training sample size for each class is greater than the number of features. Moreover, it cannot be easily generalized to handle regularized discriminant analysis. A second asymptotic approach has arisen in several recent works, leading to concurrent results about the misclassification error rates associated with discriminant analysis classifiers. Particularly, based on sparsity assumptions on the mean and covariance matrices, sparse variants of LDA and QDA has been proposed in [7] and [8] and analyzed under the asymptotic regime in which the number of features p is much larger than the number of the training samples n. A different possible regime is the one in which n and p grow large with the same pace, often termed as the double asymptotic regime. The major advantage of this regime is that it lends itself to the use of results from random matrix theory. Most importantly, the work of Raudys in [9] which permits to derive the asymptotic misclassification error in the double asymptotic regime under the assumption of equal covariance matrices. It has also recently been considered in

978-1-5090-6341-3/17/$31.00

the analaysis of the regularized LDA [10], but to the best of the authors' knowledge has not been considered for the most general case in which the covariances across classes are different, henceforth, calling for the use of QDA based classifiers.

1.3. Contributions

The present work aims to provide a comprehensive understanding of the performance of regularized QDA under the asymptotic regime in which the number of training samples with each class grow large with the number of features. Under some mild assumptions controlling the distance between class covariances and means, we show that the classification error converges to a non-trivial deterministic quantity that only depends on the Gaussian distribution parameters of each class and the problem dimensions. Although real data are far from being Gaussian, our asymptotic approach has been shown to yield good accuracy when applied to real data.

- Under some mild assumptions, building on fundamental results from random matrix theory [11] we establish the convergence of the classification error to a deterministic error that reveals the mathematical connection between the classification error and the statistical parameters associated with each class.
- We leverage this result to propose a more efficient design of the regularized QDA classifier by selecting the regularization parameter that minimizes the asymptotic classification error.
- We validate our theoretical findings using both synthetic data and real data drawn from available data bases. We particularly illustrate the good accuracy of our results for both settings.

In the remainder of this paper, we give an overview of regularized QDA for binary classification in Section 2. The main results are presented in Section 3 while all proofs are available in an extended version of this paper. We validate our analysis in Section 4 and conclude the paper in Section 5.

Notations:

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

2. QDA CLASSIFIER FOR BINARY CLASSIFICATION

We consider the problem of classifying a multivariate observation $\mathbf{x} \in \mathbb{R}^{p\times 1}$ to one of two classes under the assumption that $\mathbf{x}$ belongs to class $\mathcal{C}_i$, $i = 0, 1$, if and only if

$$\mathbf{x} = \boldsymbol{\mu}_i + \boldsymbol{\Sigma}_i^{\frac{1}{2}}\boldsymbol{\omega},$$

with $\boldsymbol{\omega} \sim \mathcal{N}(\mathbf{0}_{p\times 1}, \mathbf{I}_p)$, $\boldsymbol{\Sigma}_i \in \mathbb{R}^{p\times p}$ and $\boldsymbol{\mu}_i$ are the covariance and the mean vector associated with class i. Let $\pi_i, i = 0, 1$ denote the prior probabability that $\mathbf{x}$ belongs to class $\mathcal{C}_i$. Based on these assumptions, the Bayes rule classifier is the one that assigns $\mathbf{x}$ to the class that presents the highest posterior probability, $\mathbb{P}(\mathbf{x} \in \mathcal{C}_i|\mathbf{x})$. This amounts to selecting the class which achieves the highest value of the following classification score

$$W_i^{QDA}(\mathbf{x}) = -\frac{1}{2}\log|\boldsymbol{\Sigma}_i| - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) + \log\pi_i. \tag{1}$$

In particular, the classification rule is given by

$$\begin{cases}\mathbf{x}\in\mathcal{C}_0 & \text{if } W_0^{QDA}(\mathbf{x}) > W_1^{QDA}(\mathbf{x}) \\ \mathbf{x}\in\mathcal{C}_1 & \text{otherwise.}\end{cases} \tag{2}$$

The resulting classification approach produces quadratic class boundaries, giving the name quadratic discriminant analysis (QDA) to the corresponding classifier. As a Bayes classifier, it is associated with the lowest possible expected misclassification error rate if the data follow the assumed Gaussian mixture model. However, in practice the class parameters $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ are not known. To solve this issue, a set of independent training data with known class labels is used to estimate the covariance matrix $\boldsymbol{\Sigma}_i$ and the mean vector $\boldsymbol{\mu}_i$ associated with each class. Such estimates are used as a plug-in estimators in the discriminant analysis cost (1). In particular, let n_i be the number of available samples in class $\mathcal{C}_i$ and denote by $\mathcal{T}_0 = \{\mathbf{x}_l \in \mathcal{C}_0\}_{l=1}^{n_0}$ and $\mathcal{T}_1 = \{\mathbf{x}_l \in \mathcal{C}_1\}_{l=n_0+1}^{n_0+n_1}$ the corresponding samples. Denote by $\overline{\mathbf{x}}_i$ and $\widehat{\boldsymbol{\Sigma}}_i$ the empirical estimates of the mean vector and covariance matrix associated with class $\mathcal{C}_i$

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

Then, the empirical discriminant analysis score becomes

$$\widehat{W}_i^{QDA}(\mathbf{x}) = -\frac{1}{2}\log|\widehat{\boldsymbol{\Sigma}}_i| - \frac{1}{2}(\mathbf{x}-\overline{\mathbf{x}}_i)^T \widehat{\boldsymbol{\Sigma}}_i^{-1}(\mathbf{x}-\overline{\mathbf{x}}_i) + \log\pi_i, \tag{3}$$

The empirical QDA formulated in (3) requires all covariance estimates $\widehat{\boldsymbol{\Sigma}}_i$ to be non-singular. However, in some practical scenarios, $\widehat{\boldsymbol{\Sigma}}_i$ might be ill-conditioned if not singular, a situation arising in particular when the number of samples n_i is lower than the number of features. To get around this issue, regularized estimators shrinking the sample covariance estimate to identity have often been proposed [5]. In this paper, we consider the following regularized estimate of the inverse covariance matrix

$$\mathbf{H}_i = \left(\mathbf{I}_p + \gamma\widehat{\boldsymbol{\Sigma}}_i\right)^{-1}, \tag{4}$$

where $\gamma > 0$ is a regularizer. The regularized discriminant analysis is thus obtained by replacing $\boldsymbol{\Sigma}_i^{-1}$ by $\mathbf{H}_i$, thus yielding

$$\widehat{W}_i^{RQDA} = \frac{1}{2}\log|\mathbf{H}_i| - \frac{1}{2}(\mathbf{x}-\overline{\mathbf{x}}_i)^T\mathbf{H}_i(\mathbf{x}-\overline{\mathbf{x}}_i) + \log\pi_i.$$

Conditioned on the training samples $\mathcal{T}_i$, $i\in\{0,1\}$, the classification error contributed by class $\mathcal{C}_i$ is given by

$$\epsilon_i = \mathbb{P}\left[(-1)^i\widehat{W}_0^{RQDA}(\mathbf{x}) < (-1)^i\widehat{W}_1^{RQDA}(\mathbf{x})\,|\mathbf{x}\in\mathcal{C}_i\right], \tag{5}$$

which yields the following total mis-classification error probability

$$\epsilon = \pi_0\epsilon_0 + \pi_1\epsilon_1. \tag{6}$$

On the other hand, the conditional classification error in (5) can easily be shown to write as

$$\epsilon_i = \mathbb{P}\left[\boldsymbol{\omega}^T\mathbf{B}_i\boldsymbol{\omega} + 2\boldsymbol{\omega}^T\mathbf{y}_i < \xi_i\right], \tag{7}$$

where $\boldsymbol{\omega}\sim\mathcal{N}(\mathbf{0}_p,\mathbf{I}_p)$,

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

It thus amounts to computing the cumulative distribution function (CDF) of quadratic forms of Gaussian random vectors, and hence cannot be derived in closed form in general. However, it can be still approximated by considering asymptotic regimes that allow to exploit results about central limit theorem involving quadratic forms, as will be shown in the next section. This is in striking difference with LDA classifiers, for which the conditional probability coincides with that of a Gaussian random variable (since $\mathbf{B}_i = 0$).

3. STATEMENT OF THE MAIN RESULTS

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

3.1. Technical Assumptions

We consider the following double asymptotic regime in which n_i, p grow to ∞ for $i\in\{0,1\}$ and the following assumptions are met

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

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

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

Assumption 4. $\limsup\frac{1}{\sqrt{p}}\operatorname{tr}\mathbf{A}(\boldsymbol{\Sigma}_0-\boldsymbol{\Sigma}_1) = \mathcal{O}(1)$*, for all* $\mathbf{A}\in\mathbb{R}^{p\times p}$ *satisfying* $\|\mathbf{A}\| = \mathcal{O}(1)$.

The first assumption states that the number of features and that of training samples are commensurable. This is of standard use within the framework of random matrix theory and allows to obtain closed-form approximations of the mis-classification error probabilities. Assumption 1 implies also that $\pi_i\to\frac{1}{2}$ for $i\in\{0,1\}$. The second assumption governs the distance between the two classes in terms of the Euclidean distance of the difference between the means. This is mandatory in order to avoid asymptotic perfect classification. A similar assumption is required to control the distance between the covariance matrices. Particularly, the spectral norm of the covariance matrices are required to be bounded while their difference should satisfy $\frac{1}{\sqrt{p}}\operatorname{tr}\mathbf{A}(\boldsymbol{\Sigma}_0-\boldsymbol{\Sigma}_1) = \mathcal{O}(1)$. This latter condition is met for instance when at most $\lceil\sqrt{p}\rceil$ eigenvalues of $\boldsymbol{\Sigma}_0-\boldsymbol{\Sigma}_1$ are $\mathcal{O}(1)$ while the remaining are $\mathcal{O}(p^{-\alpha})$, for $\alpha\geq\frac{1}{2}$. It allows, together with the fact that $\pi_i\to\frac{1}{2}$ for $i\{0,1\}$, terms involving the difference $\mathbf{H}_1-\mathbf{H}_0$ to decrease at a rate of $\mathcal{O}\left(p^{-\frac{1}{2}}\right)$.

3.2. Central Limit Theorem (CLT)

Under Assumptions 1-4, a central limit theorem (CLT) on the random variable $\boldsymbol{\omega}^T\mathbf{B}_i\boldsymbol{\omega}+2\boldsymbol{\omega}^T\mathbf{y}_i$ when $\boldsymbol{\omega}\sim\mathcal{N}(\mathbf{0}_{p\times1},\mathbf{I}_p)$ is established. This result is essential to evaluate the asymptotic approximation of the mis-classification rate and is stated as follows

Proposition 1 (CLT)**.** *Under assumptions 1-4, the following convergence holds*

$$\frac{\boldsymbol{\omega}^T \mathbf{B}_i \boldsymbol{\omega} + 2\boldsymbol{\omega}^T \mathbf{y}_i - \operatorname{tr} \mathbf{B}_i}{\sqrt{2 \operatorname{tr} \mathbf{B}_i^2 + 4 \mathbf{y}_i^T \mathbf{y}_i}} \xrightarrow{d} \mathcal{N}(0,1). \tag{8}$$

Proof. The proof is mainly based on the application of the Lyapunov's CLT for the sum of independent but non identically distributed random variables [12]. □

As a by-product of the above Proposition, we obtain the following expression for the conditional classification error ϵ_i

Corollary 1. *Under the setting of Proposition 1 , the conditional classification error in (5) satisfies*

$$\epsilon_i - \Phi\left((-1)^i \frac{\xi_i - \operatorname{tr} \mathbf{B}_i}{\sqrt{2 \operatorname{tr} \mathbf{B}_i^2 + 4 \mathbf{y}_i^T \mathbf{y}_i}} \right) \to 0. \tag{9}$$

3.3. Deterministic Equivalent

With the term above at hand, we are now in position to derive the deterministic equivalent for the conditional classification error. First, we shall introduce the following notations, which stems from standard results from random matrix theory. We define for $i \in \{0,1\}$, δ_i as the solution of the following fixed point equation [1]

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

Define $\mathbf{T}_i$ as

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

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

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

Let $\boldsymbol{\mu} = \boldsymbol{\mu}_0 - \boldsymbol{\mu}_1$, and set $\overline{\xi}_i$, $\overline{b}_i$ and $\overline{B}_i$ to

$$\begin{aligned}
\overline{\xi}_i &\triangleq \frac{1}{\sqrt{p}} \left[-\log \frac{|\mathbf{T}_0|}{|\mathbf{T}_1|} + \log \frac{(1+\gamma\delta_0)^{n_0}}{(1+\gamma\delta_1)^{n_1}} \right. \\
&\quad \left. + \gamma \left(\frac{n_1 \delta_1}{1+\gamma\delta_1} - \frac{n_0 \delta_0}{1+\gamma\delta_0} \right) + (-1)^{i+1} \boldsymbol{\mu}^T \mathbf{T}_{1-i} \boldsymbol{\mu} \right]. \\
\overline{b}_i &= \frac{1}{\sqrt{p}} \operatorname{tr} \boldsymbol{\Sigma}_i (\mathbf{T}_1 - \mathbf{T}_0). \\
\overline{B_i} &\triangleq c \left[\frac{\phi_0}{1-\gamma^2\phi_0\tilde{\phi}_0} + \frac{\phi_1}{1-\gamma^2\phi_1\tilde{\phi}_1} \right] - \frac{2}{p} \operatorname{tr} \boldsymbol{\Sigma}_i \mathbf{T}_1 \boldsymbol{\Sigma}_i \mathbf{T}_0.
\end{aligned}$$

[1] Mathematical details treating the existence and uniqueness of δ_i can be found in [11].

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

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

Proof. The proof relies on showing that $\overline{B}_i$, $\overline{b}_i$ $\overline{\xi}_i$ are respectively the limits in probability of $\frac{1}{p} \operatorname{tr} \mathbf{B}_i^2$, $\frac{1}{\sqrt{p}} \operatorname{tr} \mathbf{B}_i$ and $\frac{1}{\sqrt{p}} \xi_i$. Further details can be found in Appendix A of the extended version. □

Theorem 1 shows that the mis-classification error converges to a non-trivial deterministic quantity that depends only on the statistical means and covariances within each class. The major importance of this result is that it can be used to determine the regularization γ that minimizes the asymptotic classification error. While it seems to be elusive for such value to possess a closed-form expression, it can be numerically obtained by using a simple one-dimensional line search algorithm.

Special cases

1) It is important to note that we could have considered $\|\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\| = O(1)$. In this case, the classification error rate would still converge to a non trivial limit but would not asymptotically depend on the difference $\|\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\|$. This is because in this case, the difference in covariance matrices dominate that of the means and as such represent the discriminant metric that asymptotically matters.
2) Another interesting case to highlight is the one in which $\|\boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_1\| = O(p^{-\frac{1}{2}-\alpha})$, $\alpha > 0$. From Theorem 1 and using some basic manipulations, it is easy to show that the total classification error converges to

$$\epsilon - \Phi\left(-\frac{\boldsymbol{\mu}^T \mathbf{T} \boldsymbol{\mu}}{2\sqrt{p}} \sqrt{\frac{1-\gamma^2\phi\tilde{\phi}}{c\gamma^2\phi^2\tilde{\phi}}} \right) \xrightarrow{p} 0, \tag{10}$$

where ϕ, $\tilde{\phi}$ and $\mathbf{T}$ have respectively the same definitions as ϕ_i, $\tilde{\phi}_i$ and $\mathbf{T}_i$ upon dropping the class index i, since quantities associated with class 0 or class 1 can be used interchangably in the asymptotic regime. It is easy to see that in this case if $\|\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\|^2$ scales slower than $\mathcal{O}(\sqrt{p})$, classification is asymptotically impossible. This must be contrasted with the results of LDA [10], which provides non-vanishing misclassification rates for $\|\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\| = O(1)$.
3) When $\|\boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_1\|_F = O(1)$ occurring for instance when $\|\boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_1\|_2 = O(p^{-\frac{1}{2}})$ or $\boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_1$ is of finite rank, and $\|\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\|^2 = O(1)$, we can prove that the misclassification error probability associated with each class converges respectively to $1-\eta$ and η where η is some probability depending solely on the statistics. Hence, the total mis-classification error probability associated with regularized QDA converges to

$$\epsilon \to 0.5.$$

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

4. EXPERIMENTS

In this section, we carry out simulations to validate our results for synthetic and real data.

4.1. Synthetic Data

For synthetic data, we choose the following models for $\boldsymbol{\mu}_i$ and $\mathbf{\Sigma}_i$: $\{\mathbf{\Sigma}_0\}_{i,j} = 0.6^{|i-j|}$, $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_0 + 2\begin{bmatrix} \mathbf{I}_k & \mathbf{0}_{k\times(p-k)} \\ \mathbf{0}_{(p-k)\times k} & \mathbf{0}_{(p-k)\times(p-k)} \end{bmatrix}$, $k = \lfloor \sqrt{p} \rfloor$, $\boldsymbol{\mu}_0 = \left[1, \mathbf{0}_{(p-1)\times 1}\right]$, $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_0 + p^{-\frac{1}{4}}\mathbf{1}_{p\times 1}$. We validate the theoretical results of the previous section by evaluating the empirical misclassification rate of the regularized QDA classifier over a testing set of $n_{test} = 1000$ samples from $\mathcal{C}_0$ and $\mathcal{C}_1$. The statistics $\bar{x}_i$ and $\widehat{\mathbf{\Sigma}}_i$ are estimated using training sets $\mathcal{T}_i$ independent of the testing set, with cardinalities n_i respectively for $i \in \{0, 1\}$. This process is averaged over 50 Monte Carlo realizations. In a first experiment, we fix $\gamma = 1$ and quantify the classification error as a function of the number of features p ranging from 100 to 500. The results of this experiment are shown in the first row of Figure 1 for different values of $c \in \{\frac{1}{2}, 1, \frac{3}{2}\}$. For all values of c, it is clear that the asymptotic error presents a good match with the real error computed over the given testing set.

In a second experiment, we fix $p = 300$ features and examine the behavior of the classification error with respect to the regularization parameter γ. As seen, there exists an optimal γ denoted by γ^* that gives the lowest classification error for R-QDA. The theoretical asymptotic error can be then used to determine the optimal regularization parameter. It is worth mentioning that as the ratio $\frac{p}{n_i}$ increases, the optimal value of the regularization parameter γ becomes closer to zero. This can be explained by the fact that in this case, the empirical covariance matrix becomes ill-conditioned, and hence, putting more weight on the bias (identity matrix in this case) should yield better performance.

4.2. Real Data

The Gaussianity assumption of the training set and testing set has been extensively used in our derivation. However, in practice, real data are not Gaussian. In this section, we assess how accurate are our results when applied on real data. Surprisingly, when applied to the real data sets namely the MNIST

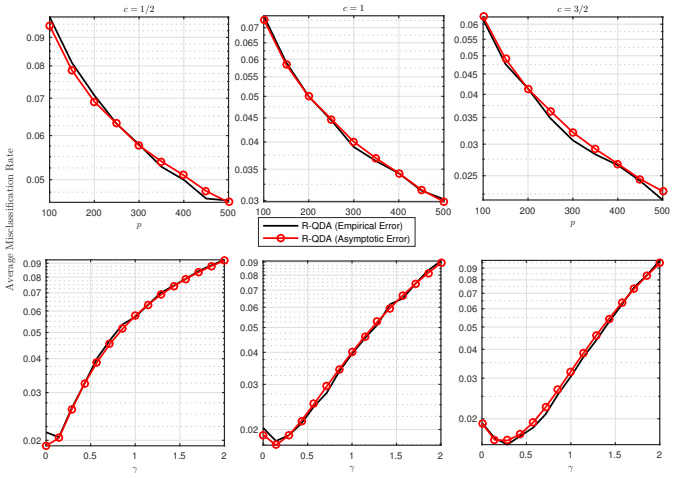

Fig. 1: Performance in terms of the testing classification error of the regularized QDA classifier with equal training, $n_0 = n_1$. The x-axis in the first row is the number of features p for $\gamma = 1$ while in the second row is the regularization parameter γ for $p = 300$ features.

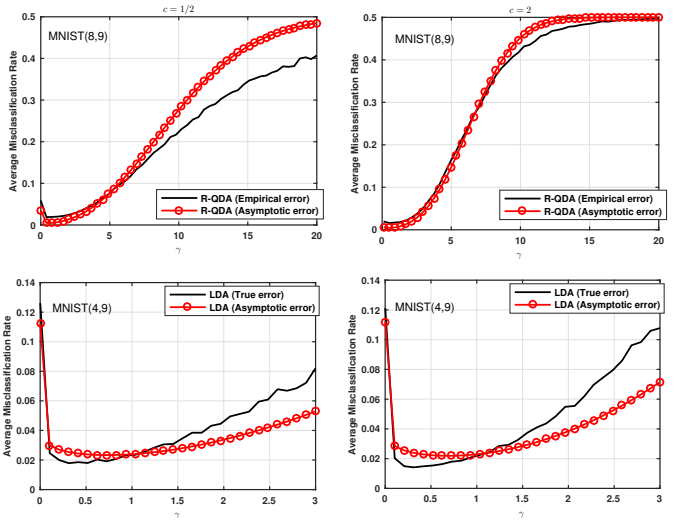

Fig. 2: Performance of the R-QDA classifier applied to both data sets given by Table 1 in terms of the average classification error for different values of the ratio c. The first row gives the performance of the R-QDA classifier when applied to MNIST data set with digits 8 and 9 whereas the second row gives its performance when applied to digits 4 and 9. The x-axis is the regularization parameter γ. In both data sets, we consider equal training, i.e. $n_0 = n_1$.

Table 1: Data sets Description where both statistics $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$, $i \in \{0, 1\}$ are estimated using the total samples available in the data sets.

	MNIST (8,9)	MNIST (4,9)
p	784	784
N_0	5851	5842
N_1	5949	5949
$\frac{1}{\sqrt{p}} \left\|\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\right\|^2$	4.6707	1.9687
$\frac{1}{\sqrt{p}} \operatorname{tr}\left(\boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_1\right)$	2.2406	0.7345

data set [13] (class 0 is given by instances of the digit 8 or 4 and class 1 is given by instances of the digit 9). Our derivations are found to mimic the real behavior of the classification error with a reasonable discrepancy. In Table 1, we summarize both data sets parameters in terms of the number of features p, the total number of samples N_0 of class 0 and the total number of samples of class 1 N_1 along with their associated distances in means and covariance matrices. The training for each class i is performed using n_i samples randomly selected from the available N_i samples. The testing is then performed by randomly selecting n_{test} samples from the remaining $N_i - n_i$ samples. This process is repeated for 100 times, over which the mis-classification error is averaged. Surprisingly as shown in Figure 2, the deterministic equivalent of the classification error computed based on the empirical means and the covariance matrices provides a good way to approximate the real mis-classification error for both data sets. More importantly, the deterministic equivalent is able to track the regularization parameter γ that would minimize the average classification error as shown in Figure 2.

5. CONCLUSION

This paper studies the asymptotic mis-classification error rate of the regularized QDA classifier. It is shown that under the regime in which the dimension of the training feature vectors and their numbers in each class grow large at the same pace, the mis-classification error converges to a deterministic quantity that depends solely on the problem dimensions and the statistical parameters in each class. By setting the regularization parameter to the value that minimizes the asymptotic misclassification rate, such a result should set the stage for a better design of the regularized QDA. This becomes all the more of a major practical importance, given that a good accuracy of our derivations is shown for synthetic and real data.

6. REFERENCES

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

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

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

[4] S. Srivastava, M. R. Gupta, and B. A. Frigyik, "Bayesian quadratic discriminant analysis," *The Journal of Machine Learning Research*, vol. 8, pp. 1277–1305, 2007.

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

[6] H. R. McFarland and D. S. P. Richards, "Exact Misclassification Probabilities for Plug-In Normal Quadratic Discriminant Functions," *Journal of Multivariate Analysis*, vol. 82, p. 299330, 2002.

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

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

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

[10] 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.

[11] 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.

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

[13] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-Based Learning Applied To Document Recognition," *Proceedings of the IEEE*, vol. 86, no. 11, p. 22782324, 1998.