IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 3 July 2024, accepted 28 July 2024, date of publication 2 August 2024, date of current version 28 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3437246

RESEARCH ARTICLE

An Improved Binary Quadratic Discriminant Analysis Classifier by Using Robust Regularization

ALAM ZAIB[1], SHAHID KHATTAK[1], GHULAM MUJTABA[1], SHAHID KHAN[1], AND AMAL AL-RASHEED[2]

[1]Department of Electrical and Computer Engineering, COMSATS University Islamabad, Abbottabad Campus, Abbottabad 22060, Pakistan
[2]Department of Information Systems, College of Computer and Information Sciences, Princess Nourah Bint Abdulrahman University, P.O. Box 84428, Riyadh 11671, Saudi Arabia

Corresponding author: Amal Al-Rasheed (aaalrasheed@pnu.edu.sa)

This work was supported by the Princess Nourah Bint Abdulrahman University Researchers Supporting, Riyadh, Saudi Arabia, under Project PNURSP2024R235.

ABSTRACT In many real classification problems where a limited number of training samples is available, the linear classifiers based on discriminant analysis are unable to deliver accurate results. Moreover, the testing and/or the training data can be erroneous due to noise contamination which further degrades their performance. Regularization techniques become imperative to deal with these problems. However, the existing regularization techniques, to some extent, mainly focus on data scarcity issues but completely ignore the noisy nature of the testing and/or the training data. We propose a novel regularized quadratic discriminant analysis (R-QDA) classifier which addresses both issues simultaneously. The procedure involves a reformulation of the discriminant function of the conventional QDA classifier into least square problems and then solving them by using regularized least squares (Reg-LS) based on ℓ_2-norm. In contrast to existing R-QDA techniques, the proposed R-QDA classifier employs two regularization parameters pertaining to each class, which can be independently selected by various robust techniques. Numerical results demonstrate the effectiveness of the proposed method over classical R-QDA methods, especially in high noise regimes.

INDEX TERMS Discriminant analysis, LDA, QDA, regularization, data classification.

I. INTRODUCTION

Classification is a supervised learning approach in machine learning, in which a computer program learns from the input training data set and then uses this learning to classify new unseen observations [1]. Classification problems may be categorized as binary-class or multi-class problems. In the former, the data set belongs to only two classes while in the latter it may belong to more than two classes. The statistical classifier aims to use object characteristics to declare which class it belongs to. Speech and handwriting recognition, biometric verification, and documents classification are some of the common examples of classification problems.

Classifiers based on discriminant analysis (DA) are used numerously in various classification problems and applications [2], [3]. DA aims to categorize the objects to one of the predefined classes by thresholding a discriminant function of the data. Two distinct types of DA are linear discriminant analysis (LDA) and quadratic discriminant analysis(QDA) which are distinguished from one another by linear and quadratic decision boundaries respectively [4]. Linear classification classifies the data by making decisions based on the value of a linear combination of features. It is a simple and computationally attractive approach that works better when the data is linearly separable. Due to

The associate editor coordinating the review of this manuscript and approving it for publication was Dost Muhammad Khan.

the simplicity of LDA, it has been used effectively in various classification and face recognition problems [5], [6], [7]. Some standard assumptions associated with LDA are: (i) Data from each class has the same mean vector and the same covariance matrix (ii) Data are multivariate normally distributed and (iii) Data entities exhibit independence. Classification problems that are dealt with the QDA are those where the data pairs are discriminated by boundaries modeled by quadratic functions [8]. Decision boundaries are in the form of curves in this case. Gaussian assumption is maintained in case of the QDA as well, but covariance matrices are considered different for each class. If LDA is used in case of different class covariances, it may result in high variability. Although both approaches can handle static data very well, QDA is more suitable for complex decision boundaries.

DA classifiers produce least misclassification error under Gaussian/Normal assumption if the mean vector and the covariance matrix of each class are known exactly. However, these parameters are not available in practice and must be estimated. Generally, an independent set of data along with known labels, called training data, is used to estimate these class parameters that are used in discriminant functions. In many practical problems, the number of data samples is less than the number of features which is often referred to as data scarcity. Moreover, the performance of the DA-based classifiers deteriorates significantly when the test data is also contaminated with noise that is not observed during the training stage. The data scarcity problem results in ill-conditioning or even non-invertibility of covariance matrices as the number of features is usually large. Different techniques can be considered to address these issues. One way is to use dimensionality reduction to retain the most important features from the data and discard the ones which are not important from a classification perspective [1], [3]. However, it is often challenging to decide which features should be retained and to what extent the dimensions be reduced. Also, the dimensionality reduction results in some loss of information that cannot be recovered. Although the issue of robustness of QDA has been addressed using various techniques in recent works [9], [10], an alternative and finest way is to employ regularization techniques which give rise to regularized discriminant analysis (RDA) classifiers [11], [12] and they are the main focus of this paper.

In literature, different regularized versions of LDA and QDA, namely R-LDA and R-QDA respectively, have been proposed [13], [14]. The basic premise behind these methods is to replace the covariance matrices' estimates with their ridge estimates [15], [16], thereby stabilizing the inverse of covariance matrices. The performance of the resulting R-LDA and R-QDA classifiers heavily depends on the value of the ridge or regularization parameter and, therefore, it must be chosen appropriately. Different techniques have been proposed in the literature for finding the best value of the regularization parameter, such as, generalized cross-validation [17], L-curve [18], quasi-optimal [19] etc. These methods use grid search to find the best value of the regularization parameter, which has its drawbacks as the grid size and grid interval are not always well defined. More recently, the authors in [20] and [21] have proposed a bounded perturbation regularization (BPR) and a constrained perturbation regularization approach (COPRA), where the regularization parameter is found by minimizing the mean squared error (MSE). These methods have been shown to perform faster than the grid search methods as the regularization parameter is obtained by solving non-linear equations via Newton's method [22]. In another line of research, [23], [24] and [25] developed R-LDA classifiers by using asymptotic analysis of the probability of misclassification based on random matrix theory. In [26] the results of R-LDA were extended to the case of R-QDA as well. The regularization parameter in these methods is found by the grid search technique where the asymptotic misclassification error rate has to be computed over a fine grid of predefined parameter values. Moreover, these methods are strongly built on Gaussian assumptions of the data as well as on a few specific assumptions on covariance matrices. These assumptions do not apply equally well to the real data (that might not be Gaussian) and to the case of generic covariance matrices. Furthermore, all the existing R-LDA and R-QDA methods do not address the noise contamination problem, and therefore, suffer from performance degradation. This has been the motivation for the present work.

In this paper, we present an improved binary R-QDA classifier based on a robust regularization approach by reformulating the score function of conventional R-QDA as a regularized least square (Reg-LS) problem. The resulting Reg-LS problem can be solved by using any of the robust regularization techniques as mentioned above.[1] The proposed method takes care of the ill-conditioning of the covariance matrices due to data scarcity as well as the perturbations in the training and/or the test data due to noise contamination. The proposed method has the following distinctive features:

- Two regularization parameters are calculated corresponding to each class based on both the training and the test data which can be tuned independently to cope with the perturbations in the test data. This is to be contrasted with existing approaches which utilize a single regularization operation based solely on the training data. This feature makes the proposed approach more robust to noise that is unobserved in the training data but occurs in the test data.
- The regularization parameter selection approach is agnostic to the underlying distribution of the data contrary to existing works [23], [24], [25], [26] which strongly rely on the Gaussian assumption.

A. NOTATIONS

Throughout this paper, we used non-bold letters to denote scalars (e.g., W), boldface lowercase letters to denote column

[1] The authors have recently developed a similar technique for the LDA classifier [27], where the COPRA gave the best performance.

vectors (e.g., $\mathbf{x}$), and boldface uppercase letters to denote matrices (e.g., $\mathbf{H}$). The notation $\mathbf{I}_p$ denotes an identity matrix of dimension p, and $\mathbf{0}_{p_1 \times p_2}$ represents a $p_1 \times p_2$ matrix with all zero elements. We use tr(.) and $(.)^{\mathrm{T}}$ to denote the matrix trace and matrix/vector transpose operations, respectively. The notation $\hat{x}$ indicates an estimate of the variable x. The set of real numbers is denoted by $\mathbb{R}$ and the l_2 norm of a vector is denoted by $\|.\|_2$. The probability density function and the statistical expectation of a random variable x are denoted by $\mathrm{P}(x)$ and $\mathbb{E}(x)$, respectively. The symbol $\approx$ stands for "approximately equivalent to," while $:=$ means "defined to be equal to". Finally, "s.t." is an abbreviation for "subject to."

The rest of the paper is organized as follows. Section II gives an overview of binary discriminant analysis based classification, in Section III we develop the proposed R-QDA classifier, Section IV describes various techniques to find the regularization parameter and also summarizes the R-QDA algorithm, simulation results are presented in Section V and finally, Section VI is conclusion.

II. DISCRIMINANT ANALYSIS BASED CLASSIFICATION

In this paper, we consider a binary classification problem based on the discriminant rule that assigns an input data vector to one of the two classes it most likely belongs to. The classifier is designed based on n available training data samples with known class labels. We consider Bayesian discriminant rule and assume that observations from each class $\mathcal{C}_i, i \in \{0, 1\}$ are independent and sampled from a multivariate Gaussian distribution with mean $\boldsymbol{\mu}_i \in \mathbb{R}^{p\times 1}$ and non-negative covariance matrix $\Sigma_i \in \mathbb{R}^{p\times p}$. More specifically, a multivariate data vector $\mathbf{x} \in \mathbb{R}^{p\times 1}$ is assigned to the class $\mathcal{C}_i$, if

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

For different covariance matrices Σ_0 and Σ_1, the score function of QDA classifier reads as [1]:

$$\begin{aligned} W^{\mathrm{QDA}}(\mathbf{x}) = &-\frac{1}{2}\log\frac{|\Sigma_0|}{|\Sigma_1|} - \frac{1}{2}\mathbf{x}^{\mathrm{T}}\left(\Sigma_0^{-1} - \Sigma_1^{-1}\right)\mathbf{x} \\ &+ \mathbf{x}^{\mathrm{T}}\Sigma_0^{-1}\boldsymbol{\mu}_0 - \mathbf{x}^{\mathrm{T}}\Sigma_1^{-1}\boldsymbol{\mu}_1 - \frac{1}{2}\boldsymbol{\mu}_0^{\mathrm{T}}\Sigma_0^{-1}\boldsymbol{\mu}_0 \\ &+ \frac{1}{2}\boldsymbol{\mu}_1^{\mathrm{T}}\Sigma_1^{-1}\boldsymbol{\mu}_1 - \log\frac{\pi_1}{\pi_0}, \end{aligned} \tag{2}$$

where π_i represents prior probability of class i. The class assignment rule for $\mathbf{x}$ is given by,

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

Applying the assumption of equal class covariance matrices, i.e., $\Sigma_0 = \Sigma_1$, the QDA reduces to LDA having the score function,

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

Note that $W^{\mathrm{LDA}}(\mathbf{x})$ is a linear function of data, and hence called linear discriminant function. The corresponding decision rule in this case is the same as given in (3) with $W^{\mathrm{QDA}}(\mathbf{x})$ replaced by $W^{\mathrm{LDA}}(\mathbf{x})$. As clear from (2) and (4), the computation of discriminant functions requires the knowledge of class statistics in the form of class mean vectors and class covariance matrices. Since these class statistics are rarely available in practice, they must be estimated from the training data with known labels. Therefore, we assume that $n_i, i \in \{0, 1\}$ independent training samples $\mathcal{T}_0 = \{\mathbf{x}_l \in \mathcal{C}_0\}_{l=0}^{n_0}$ and $\mathcal{T}_1 = \{\mathbf{x}_l \in \mathcal{C}_1\}_{l=n_0+1}^{n_0+n_1=n}$ are available to estimate the class statistics. In particular, the sample estimates of the mean vector and covariance matrix of each class i are given as follows:

$$\begin{aligned} \hat{\boldsymbol{\mu}}_i &= \frac{1}{n_i}\sum_{l\in\mathcal{T}_i}\mathbf{x}_l, \quad i \in \{0, 1\} \\ \hat{\Sigma}_i &= \frac{1}{n_i - 1}\sum_{l\in\mathcal{T}_i}(\mathbf{x}_l - \hat{\boldsymbol{\mu}}_i)(\mathbf{x}_l - \hat{\boldsymbol{\mu}}_i)^{\mathrm{T}}, \quad i \in \{0, 1\}. \end{aligned} \tag{5}$$

A major source of error in the above formulation is the inversion of the covariance matrix $\hat{\Sigma}_i$. In many practical setups where n is comparable to p, $\hat{\Sigma}$ becomes ill-conditioned, or even singular. To get around this issue, the inverse of covariance matrix $\hat{\Sigma}_i^{-1}$ is often replaced with a ridge estimator $\mathbf{H}_i = (\mathbf{I}_p + \gamma\hat{\Sigma}_i)^{-1}$, where $\gamma \in \mathbb{R}^+$ is the regularization parameter and $\mathbf{I}_p$ is the identity matrix of dimension p. The resulting classifier is referred to as regularized QDA (R-QDA). Hence, the discriminant function of R-QDA based on sample estimates takes the form,

$$\begin{aligned} &W^{RQDA}(\mathbf{x}) \\ &= \frac{1}{2}\log\frac{|\mathbf{H}_0|}{|\mathbf{H}_1|} - \frac{1}{2}\mathbf{x}^{\mathrm{T}}\left(\mathbf{H}_0 - \mathbf{H}_1\right)\mathbf{x} + \mathbf{x}^{\mathrm{T}}\mathbf{H}_0\hat{\boldsymbol{\mu}}_0 \\ &\quad - \mathbf{x}^{\mathrm{T}}\mathbf{H}_1\hat{\boldsymbol{\mu}}_1 - \frac{1}{2}\hat{\boldsymbol{\mu}}_0^{\mathrm{T}}\mathbf{H}_0\hat{\boldsymbol{\mu}}_0 + \frac{1}{2}\hat{\boldsymbol{\mu}}_1^{\mathrm{T}}\mathbf{H}_1\hat{\boldsymbol{\mu}}_1 - \log\frac{\pi_1}{\pi_0} \end{aligned} \tag{6}$$

As opposed to the above strategy, we employ a different form of regularization to that in (6). In the proposed regularized QDA classifier, we apply two separate regularization operations for each class, which help in accounting for singularities of the covariance matrix of each class and providing robustness against error contributions that are present in the noisy test data.

III. THE PROPOSED R-QDA CLASSIFIER

It is clear from above that simply replacing the covariance matrix inverse $\hat{\Sigma}_i^{-1}$ by ridge estimator $\mathbf{H}$ only caters to singularity of covariance matrices when $n_i < p$. As such, it does not address the noise perturbation in training and/or the testing data. Further, the regularization parameter γ in (6) is usually computed using only the training data. Hence existing methods are more prone to error perturbations in the training and/or the testing data, especially when the error statistics of the test data deviate from those of the training data.

To address these issues, we reformulate the QDA score function in (2) as follows. Using the sample estimates of class mean vectors and covariance matrices, (2) can be re-written as,

$$\hat{W}^{QDA}(\mathbf{x}) = -\frac{1}{2}\log\frac{|\hat{\Sigma}_0|}{|\hat{\Sigma}_1|} - \frac{1}{2}\left(\mathbf{x}-\hat{\boldsymbol{\mu}}_0\right)^{\mathrm{T}}\hat{\Sigma}_0^{-1}\left(\mathbf{x}-\hat{\boldsymbol{\mu}}_0\right) + \frac{1}{2}\left(\mathbf{x}-\hat{\boldsymbol{\mu}}_1\right)^{\mathrm{T}}\hat{\Sigma}_1^{-1}\left(\mathbf{x}-\hat{\boldsymbol{\mu}}_1\right) - \log\frac{\pi_1}{\pi_0} \tag{7}$$

In case that $n_i < p$, the sample mean and covariance estimates in (7) are not accurate hence, the discriminant function will be subsequently modified using the regularization parameters that take care of these problems. Note that the two quadratic terms $\left(\mathbf{x}-\hat{\boldsymbol{\mu}}_i\right)^{\mathrm{T}}\Sigma_i^{-1}\left(\mathbf{x}-\hat{\boldsymbol{\mu}}_i\right)$ for $i \in \{0, 1\}$ appearing in (7) can be expressed as the inner product of two vectors as follows:

$$\left(\mathbf{x}-\hat{\boldsymbol{\mu}}_i\right)^{\mathrm{T}}\hat{\Sigma}_i^{-1}\left(\mathbf{x}-\hat{\boldsymbol{\mu}}_i\right) = \left(\mathbf{x}-\hat{\boldsymbol{\mu}}_i\right)^{\mathrm{T}}\hat{\Sigma}_i^{-\frac{1}{2}}\underbrace{\hat{\Sigma}_i^{-\frac{1}{2}}\left(\mathbf{x}-\hat{\boldsymbol{\mu}}_i\right)}_{\mathbf{z}_i} = \mathbf{z}_i^{\mathrm{T}}\mathbf{z}_i, \quad i \in \{0, 1\}$$

where,

$$\mathbf{z}_i := \hat{\Sigma}_i^{-\frac{1}{2}}(\mathbf{x}-\hat{\boldsymbol{\mu}}_i), \quad i \in \{0, 1\}. \tag{8}$$

By using (8) in (7) we get,

$$\hat{W}^{\mathrm{QDA}}(\mathbf{x}) = -\frac{1}{2}\log\frac{|\hat{\Sigma}_0|}{|\hat{\Sigma}_1|} - \frac{1}{2}\mathbf{z}_0^{\mathrm{T}}\mathbf{z}_0 + \frac{1}{2}\mathbf{z}_1^{\mathrm{T}}\mathbf{z}_1 - \log\frac{\pi_0}{\pi_1}. \tag{9}$$

To avoid the singularity issue associated with the covariance matrices, we define $\mathbf{x}_i := \mathbf{x}-\hat{\boldsymbol{\mu}}_i$, so that (8) is equivalent to a set of linear equations:

$$\mathbf{x}_i = \hat{\Sigma}_i^{\frac{1}{2}}\mathbf{z}_i, \quad i \in \{0, 1\}. \tag{10}$$

Now, considering the error perturbations in training and/or testing data due to unknown noise, (10) can be modeled as:

$$\mathbf{x}_i = \hat{\Sigma}_i^{\frac{1}{2}}\mathbf{z}_i + \mathbf{v}_i, \quad i \in \{0, 1\}. \tag{11}$$

where $\mathbf{v}_i$ is the additive noise term for the class i. Note that the noise term includes the error perturbations in training and/or testing data (through $\mathbf{x}_i$), the estimation noise due to insufficient training as well as the modeling inaccuracies. To simplify our derivations we assume that, the noise vector $\mathbf{v}_i$ has zero mean and an unknown covariance matrix $\sigma_v^2\mathbf{I}_p$, the unknown random vector $\mathbf{z}_i$ is also assumed as zero mean with an unknown positive semi-definite diagonal covariance matrix and the vectors $\mathbf{v}_i$ and $\mathbf{z}_i$ are mutually independent. In Section V, we will see that these simplifying assumptions still work for different classification examples.

Focusing on (11), different regularization methods, commonly called ridge regression or Tikhonov regularization [16], [28], [29], can be applied to obtain a stable estimate of $\mathbf{z}_i$ in the presence of noise and the singularity of the covariance matrix. This estimate can be expressed in a closed form as [28]

$$\hat{\mathbf{z}}_i = (\hat{\Sigma}_i + \gamma_i\mathbf{I}_p)^{-1}\hat{\Sigma}_i^{\frac{1}{2}}\mathbf{x}_i, \quad i \in \{0, 1\}, \tag{12}$$

where γ_i is the regularization parameter associated with class i. Let $\hat{\Sigma}_i = \mathbf{U}_i\mathbf{D}_i^2\mathbf{U}_i^{\mathrm{T}}$ be the eigenvalue decomposition (EVD)[2] of the covariance matrix $\hat{\Sigma}_i$, where $\mathbf{U}_i$ is the matrix of eigenvectors satisfying orthonormality property $\mathbf{U}_i\mathbf{U}_i^{\mathrm{T}} = \mathbf{U}_i^{\mathrm{T}}\mathbf{U}_i = \mathbf{I}_p$ and $\mathbf{D}_i^2$ is the diagonal matrix consisting of the eigenvalues of $\hat{\Sigma}_i$. Then, invoking the EVD of covariance matrix in (12), the vector estimate $\hat{\mathbf{z}}_i$ can be expressed as,

$$\hat{\mathbf{z}}_i = \mathbf{U}_i(\mathbf{D}_i^2 + \gamma_i\mathbf{I}_p)^{-1}\mathbf{D}_i\mathbf{U}_i^{\mathrm{T}}\mathbf{x}_i, \quad i \in \{0, 1\}. \tag{13}$$

Now, by replacing $\mathbf{z}_i$ in (9) with their estimates $\hat{\mathbf{z}}_i$ given in (13) and simplifying the resulting expression, the modified form of R-QDA score function is obtained as,

$$\begin{aligned}
&\hat{W}^{\mathrm{R}QDA}(\mathbf{x}) \\
&= -\frac{1}{2}\log\frac{|\hat{\Sigma}_0|}{|\hat{\Sigma}_1|} - \frac{1}{2}\hat{\mathbf{z}}_0^{\mathrm{T}}\hat{\mathbf{z}}_0 + \frac{1}{2}\hat{\mathbf{z}}_1^{\mathrm{T}}\hat{\mathbf{z}}_1 - \log\frac{\pi_0}{\pi_1} \\
&= -\frac{1}{2}\log\frac{|\hat{\Sigma}_0|}{|\hat{\Sigma}_1|} - \frac{1}{2}\mathbf{x}_0^{\mathrm{T}}\mathbf{U}_0\mathbf{D}_0^{\mathrm{T}}(\mathbf{D}_0^2 + \gamma_0\mathbf{I}_p)^{-2}\mathbf{D}_0\mathbf{U}_0^{\mathrm{T}}\mathbf{x}_0 \\
&\quad + \frac{1}{2}\mathbf{x}_1^{\mathrm{T}}\mathbf{U}_1\mathbf{D}_1^{\mathrm{T}}(\mathbf{D}_1^2 + \gamma_1\mathbf{I}_p)^{-2}\mathbf{D}_1\mathbf{U}_1^{\mathrm{T}}\mathbf{x}_1 - \log\frac{\pi_0}{\pi_1}
\end{aligned} \tag{14}$$

Since $\mathbf{D}_i$ is diagonal matrix and so is the matrix $(\mathbf{D}_i^2+\gamma_i\mathbf{I}_p)^{-2}$, we can combine $\mathbf{D}_i$'s on either side of parenthesis in (14) to get $(\mathbf{D}_i^2+\gamma_i\mathbf{I}_p)^{-2}\mathbf{D}_i^2$. Finally, by approximating $\mathbf{D}_i^2$ with $(\mathbf{D}_i^2+\gamma_i\mathbf{I}_p)$ i.e., by adding the regularization term, we get

$$\hat{W}^{\mathrm{R}QDA}(\mathbf{x}) = \frac{1}{2}\log\frac{|\mathbf{H}_0|}{|\mathbf{H}_1|} - \frac{1}{2}\mathbf{x}_0^{\mathrm{T}}\mathbf{H}_0\mathbf{x}_0 + \frac{1}{2}\mathbf{x}_1^{\mathrm{T}}\mathbf{H}_1\mathbf{x}_1 - \log\frac{\pi_0}{\pi_1} \tag{15}$$

where, the matrix $\mathbf{H}_i$, which essentially represents the regularized estimate of the inverse of covariance matrix of the class $\mathcal{C}_i$, is defined as

$$\mathbf{H}_i = \hat{\Sigma}_i^{-1} \approx \mathbf{U}_i\left(\mathbf{D}_i^2 + \gamma_i\mathbf{I}_p\right)^{-1}\mathbf{U}_i^{\mathrm{T}}, \quad i \in \{0, 1\}. \tag{16}$$

From above we observe that the proposed method employs $\mathbf{H}_i$ defined in (16) with two independent regularization parameters pertaining to each class. Further, these parameters are found based on both the training and the testing data, as can be seen from (11). In contrast, the classical methods employ $\mathbf{H}_i = (\mathbf{I}_p + \gamma\hat{\Sigma}_i)^{-1}$, as the regularized estimate of the inverse of the covariance matrix, with common regularization parameter γ for both classes. Further, the common parameter γ is computed based only on the training data. Also, from the noisy data model (11), it is clear that the proposed R-QDA classifier not only takes care of the singularity of the class covariance matrix but also the noise perturbations in testing and/or training data. Therefore, the proposed classifier is

[2] The rationale behind using EVD is to partition the significant and insignificant eigenvalues of the covariance matrices so that the inversion of these matrices is stable.

robust as it is less prone to errors, and is expected to yield a solution with better stability than the the existing methods.

Now, it only remains to set the values of the regularization parameters $\gamma_i, i \in \{0, 1\}$ pertaining to each class. In the following section, we present different robust methods to compute these regularization parameters for the Reg-LS solution in (12).

IV. REGULARIZATION PARAMETER SELECTION

There exist several methods for finding the regularization parameter γ, such as the L-curve [18], the GCV [17], the quasi-optimal method [19], the BPR method [20] and COPRA [21]. These methods use different criteria which results in different values of the regularization parameter (see [30]). In practice, the performance of each of these methods may vary significantly depending on the data distribution or the problem at hand. The BPR and COPRA have shown immense success in signal estimation and beamforming [20], image restoration [21] and LDA classification [27]. They also have the advantage of being robust against data distributions and the fastest runtime.

In this section, we give a brief account of these methods and see how they can be applied to find two regularization parameters for the proposed R-QDA classifier developed in Section III. Towards this end, we first re-state the linear model of (11) in a more generalized form as follows:

$$\mathbf{y} = \mathbf{A}\mathbf{z} + \mathbf{v}, \tag{17}$$

where matrix $\mathbf{A}$ plays the role of the square-root of the sample covariance matrix i.e., $\mathbf{A} := \hat{\Sigma}_i^{\frac{1}{2}}$ and $\mathbf{y} = \mathbf{x} - \hat{\boldsymbol{\mu}}_i$ is the observed data vector contaminated with the noise vector $\mathbf{v}$. The Reg-LS solution of (17) based on Tikohonov regularization [28], is given by:

$$\hat{\mathbf{z}}_\gamma = (\mathbf{A}^{\mathrm{T}}\mathbf{A} + \gamma \mathbf{I}_p)^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{y}, \tag{18}$$

where the subscript γ used with $\hat{\mathbf{z}}$ explicitly shows the γ dependency of the estimate $\hat{\mathbf{z}}$. Note that, in the absence of any regularization i.e., $\gamma = 0$, the Reg-LS solution converges to the ordinary least squares (OLS) solution,

$$\hat{\mathbf{z}}^{\mathrm{LS}} = \left(\mathbf{A}^{\mathrm{T}}\mathbf{A}\right)^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{y}. \tag{19}$$

The major drawback of the OLS solution is the sensitivity to noise perturbations. Further, OLS is not feasible when the transformation matrix $\mathbf{A}$ is singular, hence it is not applicable when $n < p$. In this paper, we use Reg-LS solution (18), where the regularization parameter is selected by using the aforementioned techniques which are briefly discussed below.

A. GENERALIZED CROSS VALIDATION (GCV)

GCV is a popular approach used for selecting the regularization parameter. It involves selecting the parameter value γ that minimizes the GCV function [17]

$$G(\gamma) = \frac{\|\mathbf{A}\hat{\mathbf{z}}_\gamma - \mathbf{y}\|_2^2}{\mathrm{trace}(\mathbf{I}_p - \mathbf{A}\mathbf{A}_\gamma^{\#})} \tag{20}$$

where trace(.) is the matrix trace operator and $\mathbf{A}_\gamma^{\#}$ denotes the regularized pseudo-inverse of $\mathbf{A}$ defined as, $\mathbf{A}^{\#} := \mathbf{A}(\mathbf{A}^{\mathrm{T}}\mathbf{A} + \gamma\mathbf{I}_p)^{-1}\mathbf{A}^{\mathrm{T}}$. The GCV function is evaluated several times for different γ values which are sequentially selected from a predefined interval. The desired solution is obtained for the value of γ that minimizes the GCV function. Practically, this grid search approach seems quite attractive if the matrix $\mathbf{A}$ is small enough and its singular value decomposition (SVD) can be computed rapidly. In this case, the GCV function can be easily calculated several times. However, in case of ill-conditioning, the matrix $\mathbf{A}$ is supposed to be large enough and the GCV is a computationally expensive approach.

B. L-CURVE

L-curve is a robust technique used in finding the optimal regularization parameter. Like other techniques, the L-curve is also a trade-off method. L-curve criterion (LCC) is proposed by Hansen and O'Leary [18] and is used in many practical applications. L-curve is the logarithmic plot of the solution norm $\|\hat{\mathbf{z}}_\gamma\|_2$ versus residual norm $\|\mathbf{A}\hat{\mathbf{z}}_\gamma - \mathbf{y}\|_2$ along with the regularization parameter γ. The use of log scale makes the plot insensitive to scaling of $\mathbf{A}$ and $\mathbf{z}$ and hence makes it a robust approach. L-curve mainly consists of two parts - the flat part and the steep part. The regularization and the perturbation error dominate in both of these parts respectively, whereas the optimal value of γ lies near the corner of the curve.

As L-curve is a trade-off method, if too much damping is imposed or an equivalently large value of γ is used, it can lead to higher residual error. The same is the case with smaller γ values, which may result in higher data errors. Hence, the method requires one to define an approximate range for the optimal value of γ in advance. In contrast to other methods like GCV, L-curve provides a robust estimation where the GCV sometimes is not possible.

C. QUASI-OPTIMAL

Quasi-optimal criteria is a robust technique for deciding the regularization parameter γ. It performs much better in many practical scenarios than the L-curve and the GCV methods. In quasi-optimal method, we consider the regularized solution based on Thikonov regularization is given in (18). In Quasi-optimal criteria, we choose $\gamma > 0$ such that [19]

$$\left\| \frac{d\hat{\mathbf{z}}_\gamma}{d\gamma} \right\| \rightarrow \min_\gamma. \tag{21}$$

The equation above declares the robustness of the method as it does not rely on the solution knowledge and the noise level.

D. BOUNDED PERTURBATION REGULARIZATION (BPR)

The BPR is a recently developed technique by Tarig Ballal et al. that has been shown to outperform GCV, L-curve, and quasi-optimal methods in certain applications [20]. The basic idea behind BPR is to introduce an artificial perturbation in the linear model to improve the singular-value structure of the model matrix $\mathbf{A}$. To adapt BPR for the proposed R-QDA classifier, we consider replacing the matrix $\mathbf{A}$ in (17) by its perturbed version to get,

$$\mathbf{y} \approx (\mathbf{A} + \Delta)\,\mathbf{z} + \mathbf{v}, \tag{22}$$

where Δ is an unknown perturbation matrix which is norm bounded by a positive number λ, i.e., $\|\Delta\|_2 \leq \lambda$, where $\|\Delta\|_2$ represents the spectral norm of Δ. The perturbation Δ can be thought of as an error in the model due to the noisy nature of $\mathbf{A}$, which is the case for (17). The vector $\mathbf{z}$ is estimated by minimizing the worst-case residual error,

$$\min_{\hat{\mathbf{z}}} \max_{\Delta} ||\mathbf{y} - (\mathbf{A} + \Delta)\,\hat{\mathbf{z}}||_2, \;\; \text{s.t. } ||\Delta||_2 \leq \lambda. \tag{23}$$

The min-max problem (23) can be converted to a minimization problem whose solution is given by (18) with the constraint [21], [31], [32],

$$\gamma||\hat{\mathbf{z}}||_2 = \lambda||\mathbf{y} - \mathbf{A}\hat{\mathbf{z}}||_2. \tag{24}$$

We observe that the solution of (23) depends on the bound λ and not on the structure of the perturbation matrix Δ. Also, both λ and $\mathbf{z}$ are unknown. However, we can substitute (18) and the EVD: $\mathbf{A} = \mathbf{U}\mathbf{D}^2\mathbf{U}^{\mathrm{T}}$ in (24) and manipulate to obtain,

$$\lambda^2 = \frac{\operatorname{trace}\left(\left(\mathbf{D}^2 + \gamma\mathbf{I}_p\right)^{-2}\mathbf{U}^{\mathrm{T}}\,\mathbb{E}\left(\mathbf{y}\mathbf{y}^{\mathrm{T}}\right)\,\mathbf{U}\right)}{\operatorname{trace}\left(\mathbf{D}^2\left(\mathbf{D}^2 + \gamma\mathbf{I}_p\right)^{-2}\mathbf{U}^{\mathrm{T}}\,\mathbb{E}\left(\mathbf{y}\mathbf{y}^{\mathrm{T}}\right)\,\mathbf{U}\right)}. \tag{25}$$

Here, we have replaced $\mathbf{y}\mathbf{y}^{\mathrm{T}}$ with its expected value to get the optimal value of bound λ averaged over many realizations of $\mathbf{y}$. From (17), we get $\mathbb{E}\left(\mathbf{y}\mathbf{y}^{\mathrm{T}}\right) = \mathbf{U}\mathbf{D}\mathbf{U}^{\mathrm{T}}\Sigma_{\mathbf{z}}\mathbf{U}\mathbf{D}\mathbf{U}^{\mathrm{T}} + \sigma_v^2\mathbf{I}_p$. An interesting property of BPR is that it results in a regularization parameter that minimizes the MSE criterion, which, in our case, is given by,

$$\begin{aligned}\text{MSE} &\triangleq \operatorname{trace}\left(\mathbb{E}\left((\mathbf{z}-\hat{\mathbf{z}})(\mathbf{z}-\hat{\mathbf{z}})^{\mathrm{T}}\right)\right) = \sigma_v^2\operatorname{trace}\Big(\mathbf{D}^2\times \\ &\left(\mathbf{D}^2+\gamma\mathbf{I}_p\right)^{-2}\Big) + \gamma^2\operatorname{trace}\left(\left(\mathbf{D}^2+\gamma\mathbf{I}_p\right)^{-2}\mathbf{U}^{\mathrm{T}}\Sigma_{\mathbf{z}}\mathbf{U}\right),\end{aligned} \tag{26}$$

The second equality in (26) follows by substituting (18) and using the eigenvalue decomposition (EVD) of $\mathbf{A}$. By differentiating the MSE, the value of γ that minimizes the MSE can be obtained as follows:

$$\frac{\partial\,(\text{MSE})}{\partial\gamma} = 0 \implies \gamma \approx \frac{n\sigma_v^2}{\operatorname{trace}\left(\Sigma_{\mathbf{z}}\right)}. \tag{27}$$

To make the derivation short, the equations (24), (25) and (27) can be combined and manipulated to absorb the unknown parameters, leading to the BPR equation,

$$\begin{aligned} f(\gamma) = \operatorname{trace}\left(\left(\mathbf{D}^2 + \gamma\mathbf{I}_p\right)^{-1}\right)\operatorname{trace}\left(\left(\mathbf{D}^2 + \gamma\mathbf{I}_p\right)^{-1}\mathbf{d}\mathbf{d}^{\mathrm{T}}\right) \\ - p\;\operatorname{trace}\left(\left(\mathbf{D}^2 + \gamma\mathbf{I}_p\right)^{-2}\mathbf{d}\mathbf{d}^{\mathrm{T}}\right) = 0. \end{aligned} \tag{28}$$

where $\mathbf{d} := \mathbf{U}^{\mathrm{T}}\mathbf{y}$. The nonlinear BPR equation (28) depends only on γ, which can be solved by Newton's method [22] to get an optimal value of the regularization parameter.

E. CONSTRAINED PERTURBATION REGULARIZATION (COPRA)

The COPRA approach is an extension of BPR and hinges on the same basic principle as the BPR [21]. Therefore, the derivation of the COPRA algorithm takes similar steps to those in BPR except for the EVD of the model matrix $\mathbf{A}$. By exploiting the fact that model matrix $\mathbf{A}$ is ill-conditioned, COPRA may yield a more robust solution than BPR in some applications, e.g. see [21], [27]. In fact, due to the ill-conditioning of $\mathbf{A}$, some of its eigenvalues are likely very close, or even equal, to zero. Therefore, the EVD of $\mathbf{A}$ can be written in the block matrices form as,

$$\mathbf{A} = [\mathbf{U}_1\;\;\mathbf{U}_2]\begin{bmatrix}\mathbf{D}_1^2 & \mathbf{0}_{p_1\times p_2}\\ \mathbf{0}_{p_2\times p_1} & \mathbf{D}_2^2\end{bmatrix}\begin{bmatrix}\mathbf{U}_1^{\mathrm{T}}\\ \mathbf{U}_2^{\mathrm{T}}\end{bmatrix} \simeq \mathbf{U}_1\mathbf{D}_1^2\mathbf{U}_1^{\mathrm{T}}, \tag{29}$$

where $\mathbf{D}_1$ and $\mathbf{D}_2$ are diagonal matrices containing the p_1 most significant and $p_2 = p - p_1$ least significant eigenvalues, respectively. A threshold value can be set to find the point of this partitioning as recommended in [21]. However, a simple and intuitive rule is used here to determine the value of p_1 as the smaller value of p (the number of features) and n (the number of training samples), i.e., $p_1 = \min(n, p)$. The main purpose of (29) is to improve numerical stability by removing extremely small eigenvalues. In the case, where all eigenvalues are significant, there is no need for partitioning of EVD. In this case, the COPRA solution would converge to the BPR. Hence, COPRA can be thought of as a more generalized form of BPR. By following similar steps as in BPR, it can be shown that the optimal regularization parameter γ that minimizes the MSE satisfies the COPRA equation (30), as shown at the bottom of the next page. Equation (30), which is nonlinear in γ, can be solved by using Newton's method [22] to obtain the optimal value of γ. The iterations should be initialized from a positive initial guess close to zero to avoid missing the positive root, as explained in [21].

F. SUMMARY OF THE PROPOSED R-QDA ALGORITHM

The main steps involved in the proposed R-QDA algorithm based on the robust regularization are summarized as follows:

1) *Estimate the class mean vectors $\hat{\boldsymbol{\mu}}_i$ and covariance matrices $\hat{\Sigma}_i$ based on the training data by using (5).*
2) *Compute the EVD of $\hat{\Sigma}_i$ to determine the matrices $\mathbf{D}_i$ and $\mathbf{U}_i$, for $i = 0, 1$.*
3) *For the given test sample $\mathbf{x}$, set $\mathbf{y} = \mathbf{x}_i$ and $\mathbf{A} = \hat{\Sigma}_i^{1/2}$, for $i = 0, 1$ in (17), and determine the regularization parameters γ_i, $i = 0, 1$ using regularization techniques discussed in Section IV.*
4) *Compute the matrices $\mathbf{H}_i$ for $i = 0, 1$, by using (16), and the R-QDA score function $\hat{W}^{RQDA}$ given in (15).*
5) *Classify the given test sample $\mathbf{x}$ according to the rule: $\mathbf{x} \in \mathcal{C}_0$ if $\hat{W}^{RQDA} > 0$, and $\mathbf{x} \in \mathcal{C}_1$, otherwise.*

Remark: Sec.IV develops the selection of one general parameter for the model (17). However, for selecting the two

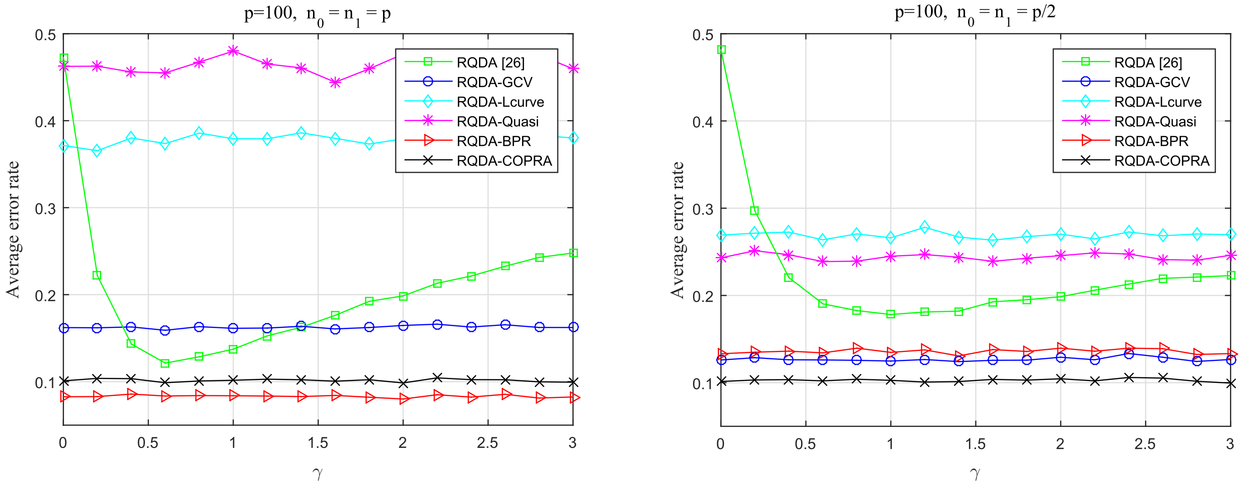

FIGURE 1. Average misclassification error rate versus γ used in benchmark paper [26]. The optimal value of γ can be read as 0.6 that minimizes the testing error. The results are based on 1000 test samples independently generated from the two classes with statistics given in (31) and (32). The error rate is averaged over 100 Monte Carlo trials.

parameters γ_i, $i = 0, 1$ for each class in step 3, the model (17) is adapted to (11) by replacing $\mathbf{y}$ with $\mathbf{x}_i$, and matrix $\mathbf{A}$ with $\hat{\Sigma}_i^{1/2}$, for $i = 0, 1$.

It is also emphasized here that the proposed R-QDA algorithm uses only the statistics from the training data set (step 1). The computations in steps 3 and 5 are solely based on the given test sample and not on the test data or the noise statistics. In fact, the test data or the noise statistics are assumed completely unknown to the classifier.

V. RESULTS AND DISCUSSIONS

In this section, we demonstrate the classification performance of the proposed R-QDA classifier against the benchmark technique of [26]. For our classifier, we consider robust regularization techniques discussed in Section IV i.e., GCV, L-curve, quasi-optimal, BPR, and COPRA, for selecting the two regularization parameters.

A. SYNTHETIC DATA

For simulations with synthetic data, the Gaussian data model is considered with class statistics given as,

$$\boldsymbol{\mu}_0 = [1, \mathbf{0}_{1\times(p-1)}]^{\mathrm{T}}, \quad \boldsymbol{\mu}_1 = \boldsymbol{\mu}_0 + \frac{0.8}{\sqrt{p}}\mathbf{1}_{p\times 1} \tag{31}$$

$$[\Sigma_0]_{i,j} = 0.9^{|i-j|},\ i, j = 1, 2, .., p, \quad \Sigma_1 = \Sigma_0 + 3\mathbf{S}_p, \tag{32}$$

where,

$$\mathbf{S}_p = \begin{bmatrix} \mathbf{I}_k & \mathbf{0}_{k\times p-k} \\ \mathbf{0}_{p-k\times k} & \mathbf{0}_{p-k\times p-k} \end{bmatrix}, \quad k = \lceil\sqrt{p}\rceil.$$

We would like to mention that a similar set of class means and covariance matrices fulfilling the desired assumptions was considered in the benchmark paper [26], and therefore, these statistics represent the best-case scenario for the benchmark paper [26]. For training the classifiers, a training set of size n_i for the class C_i is generated in each simulation trial. Without loss of generality, we set $n_0 = n_1$. For computing the classification error rate, a testing data set comprising 500 samples is generated independently from each class during each trial, and simulation results are averaged over 100 Monte Carlo trials.

Fig. 1, shows the average misclassification error rate against parameter γ used in benchmark paper [26]. We set $p = 100$ and consider two different training scenarios $n_0 = n_1 = p$ and $n_0 = n_1 = p$. The results show that the choice of γ strongly influences the error rate and the best choice minimizing the testing error for [26] is $\gamma = 0.6$. Nevertheless, the proposed RQDA algorithm with BPR and COPRA regularization techniques outperforms the benchmark method over the considered range of γ values. The GCV also performs reasonably well when $p > n$, while other regularization techniques, L-curve and Quasi-optimal,

$$\begin{aligned} h(\gamma) = {} & \operatorname{trace}\left(\mathbf{D}^2\left(\mathbf{D}^2+\gamma\mathbf{I}_p\right)^{-2}\mathbf{d}\mathbf{d}^{\mathrm{T}}\right)\operatorname{trace}\left(\left(\mathbf{D}_1^2+\gamma\mathbf{I}_{p_1}\right)^{-2}\left(\frac{p}{p_1}\mathbf{D}_1^2+\gamma\mathbf{I}_{p_1}\right)\right) \\ & + \frac{(p-p_1)}{\gamma}\operatorname{trace}\left(\mathbf{D}^2\left(\mathbf{D}^2+\gamma\mathbf{I}_p\right)^{-2}\mathbf{d}\mathbf{d}^{\mathrm{T}}\right) \\ & - \operatorname{trace}\left(\left(\mathbf{D}^2+\gamma\mathbf{I}_p\right)^{-2}\mathbf{d}\mathbf{d}^{\mathrm{T}}\right)\operatorname{trace}\left(\mathbf{D}_1^2\left(\mathbf{D}_1^2+\gamma\mathbf{I}_{p_1}\right)^{-2}\left(\frac{p}{p_1}\mathbf{D}_1^2+\gamma\mathbf{I}_{p_1}\right)\right) = 0 \end{aligned} \tag{30}$$

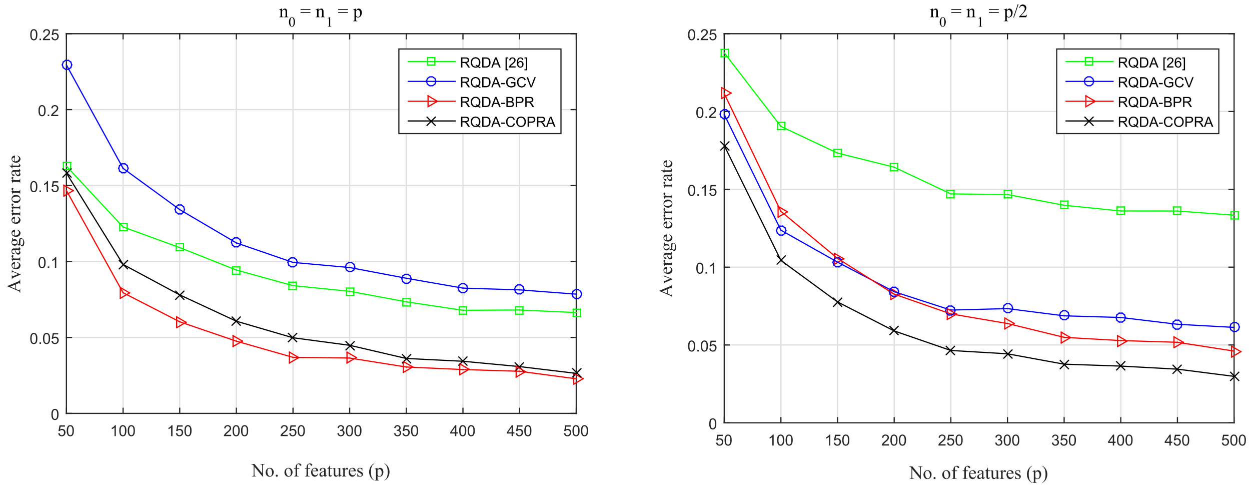

FIGURE 2. Average misclassification error rate vs p for two different training scenarios, $n_0 = n_1 = p$ and $n_0 = n_1 = p/2$. The parameter γ is optimally tuned for the benchmark method [26]. The results are based on 1000 test samples and averaged over 100 Monte Carlo trials.

are not well suited to the proposed RQDA classifier. For this reason, they are excluded from the remaining results.

Next, in Fig. 2 we present the classification performance against different values of features p when the training size is $n_0 = n_1 = p$ and $n_0 = n_1 = p/2$. The parameter γ for the benchmark method [26] is optimally tuned to $\gamma = 0.6$. It is clear from the results that the proposed RQDA method with COPRA and BPR outperforms the benchmark RQDA technique over all the values of p. Also observe that the performance of proposed techniques is more pronounced when $n < p$, compared to the case when $n = p$. For $n < p$, the estimated class statistics are more noisy in addition to unobserved noise in the test data, which is catered for by the two regularization parameters used in the proposed RQDA classifier.

We also study the classification performances against varying numbers of training samples n, for a fixed value of $p = 100$. Again, the parameter γ for the benchmark technique [26] was chosen to minimize the expected testing error for the considered training set of 1000 sample. The results presented in Fig. 3 demonstrate the better classification performance for the proposed RQDA algorithm with COPRA, BPR, and GCV regularization methods as compared to the benchmark technique [26].

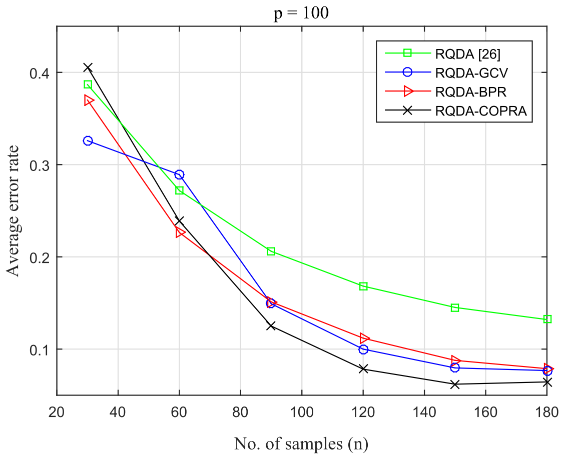

FIGURE 3. Average misclassification error rate vs number of training samples $n = n_0 + n_1$, where we have used $n_0 = n_1$. The value of p is fixed to 100. The results are generated with 1000 test samples and averaged over 100 Monte Carlo trials.

FIGURE 4. A sample of handwritten digits from MNIST dataset.

B. REAL DATA

To validate the performance of proposed R-QDA classifier with real data, we consider the MNIST dataset which consists of 20×20 gray-scale images of handwritten digits between 0 and 9, and is publicly available. A sample of this dataset is shown in Fig. 4. For binary classification, we only use selected images of most confusing digits (1,7) and (7,9). To test classification performance, we randomly selected equal number of training samples i.e., $n_0 = n_1$ and 500 test samples from each class. Both training and testing images are vectorized to form the data samples of dimensionality $p = 400$. The optimal value of single regularization parameter for the benchmark method [26] was set to $\gamma = 0.65$, and determined empirically as explained in Sec. V-A. The results, averaged over 500 monte carlo trials are presented in Fig. 5, show the classification error rate against the number of training samples $n = n_0 + n_1$. It is evident that in each case, the proposed regularized algorithm with BPR

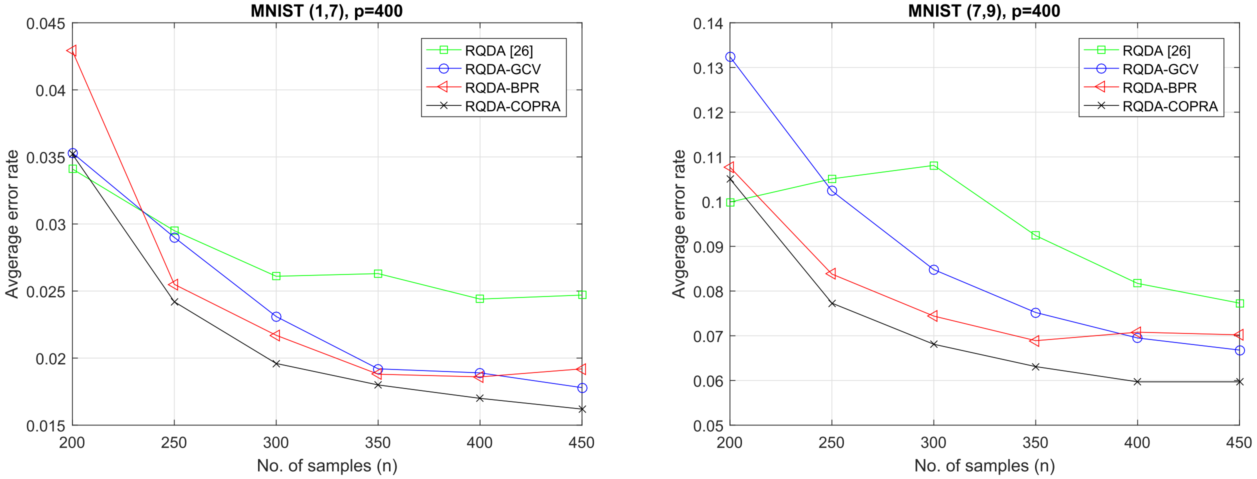

FIGURE 5. **Average misclassification error rate vs number of training samples $n = n_0 + n_1$ with $n_0 = n_1$ for the selected pair of digits from MNIST dataset. The results are generated with 1000 test samples and averaged over 500 Monte Carlo trials.**

and COPRA regularization techniques is more convincing than other methods. The results also validate that the dual regularization approach proposed in this paper is better than the benchmark method relying on single regularization parameter.

VI. CONCLUSION

We have presented a novel R-QDA classifier for binary classification of data that employs two regularization parameters pertaining to each class based on both training and testing data. In the proposed R-QDA approach, the discriminant function of the conventional R-QDA classifier is modified in such a way that it is resilient against the noise in training and /or testing data. The effectiveness of our approach is validated by experiments with Gaussian distributed data as well as real images of handwritten digits from MNIST dataset. The results demonstrate the robustness of the proposed approach. The proposed binary R-QDA classifier with COPRA and BPR gave the best overall performance compared to other regularization techniques. The latter shows more robustness when the dimensionality of data is large compared to observations.

ACKNOWLEDGMENT

The authors would like to acknowledge the mentorship provided by Dr. Tarig Ballal and Dr. Tareq Al-Naffouri at KAUST, Saudi Arabia, and the efforts of Sana Khurshid in shaping certain aspects for this work during the M.S. studies at COMSATS, Abbottabad.

DECLARATION

The authors declare that they have no conflicts of interest.

REFERENCES

[1] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. Cham, Switzerland: Springer, 2001.

[2] G. J. McLachlan, *Discriminant Analysis and Statistical Pattern Recognition*. Hoboken, NJ, USA: Wiley, 2005.

[3] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Cham, Switzerland: Springer, 2006.

[4] B. Ghojogh and M. Crowley, "Linear and quadratic discriminant analysis: Tutorial," *Int. J. Appl. Pattern Recognit.*, vol. 3, no. 2, p. 145, 2016.

[5] S. Kim, E. R. Dougherty, I. Shmulevich, K. R. Hess, S. R. Hamilton, J. M. Trent, G. N. Fuller, and W. Zhang, "Identification of combination gene sets for glioma classification," *Mol. Cancer Therapeutics*, vol. 1, no. 13, pp. 1229–1236, Nov. 2002.

[6] D. Huang, Y. Quan, M. He, and B. Zhou, "Comparison of linear discriminant analysis methods for the classification of cancer based on gene expression data," *J. Experim. Clin. Cancer Res.*, vol. 28, no. 1, pp. 1–8, Dec. 2009.

[7] D. L. Swets and J. J. Weng, "Using discriminant eigenfeatures for image retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, no. 8, pp. 831–836, Aug. 1996.

[8] A. Tharwat, "Linear vs. Quadratic discriminant analysis classifier: A tutorial," *Int. J. Appl. Pattern Recognit.*, vol. 3, no. 2, p. 145, 2016, doi: 10.1504/ijapr.2016.079050.

[9] O. K. Sajana and T. A. Sajesh, "Robust quadratic discriminant analysis using sn covariance," *Commun. Statist. Simul. Comput.*, vol. 52, no. 3, pp. 735–744, Mar. 2023, doi: 10.1080/03610918.2020.1868512.

[10] P. Houdouin, A. Wang, M. Jonckheere, and F. Pascal, "Robust classification with flexible discriminant analysis in heterogeneous data," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 5717–5721.

[11] J. H. Friedman, "Regularized discriminant analysis," *J. Amer. Stat. Assoc.*, vol. 84, no. 405, pp. 165–175, 1989.

[12] J. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos, "Regularized discriminant analysis for the small sample size problem in face recognition," *Pattern Recognit. Lett.*, vol. 24, no. 16, pp. 3079–3087, Dec. 2003.

[13] A. Aries, Z. Nashed, and V. Morozov, *Methods for Solving Incorrectly Posed Problems*. Cham, Switzerland: Springer, 2012.

[14] J. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos, "Regularization studies of linear discriminant analysis in small sample size scenarios with application to face recognition," *Pattern Recognit. Lett.*, vol. 26, no. 2, pp. 181–191, Jan. 2005.

[15] Z. Zhang, G. Dai, C. Xu, and M. I. Jordan, "Regularized discriminant analysis, ridge regression and beyond," *J. Mach. Learn. Res.*, vol. 11, pp. 2199–2228, Oct. 2010.

[16] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 42, no. 1, p. 80, Feb. 2000.

[17] M. A. Lukas, "Robust generalized cross-validation for choosing the regularization parameter," *Inverse Problems*, vol. 22, no. 5, pp. 1883–1902, Oct. 2006.

[18] P. C. Hansen and D. P. O'Leary, "The use of the L-curve in the regularization of discrete ill-posed problems," *SIAM J. Sci. Comput.*, vol. 14, no. 6, pp. 1487–1503, Nov. 1993, doi: 10.1137/0914086.

[19] F. Bauer and M. Reiß, "Regularization independent of the noise level: An analysis of quasi-optimality," *Inverse Problems*, vol. 24, no. 5, Oct. 2008, Art. no. 055009.

[20] T. Ballal, M. A. Suliman, and T. Y. Al-Naffouri, "Bounded perturbation regularization for linear least squares estimation," *IEEE Access*, vol. 5, pp. 27551–27562, 2017.

[21] M. A. Suliman, T. Ballal, and T. Y. Al-Naffouri, "Perturbation-based regularization for signal estimation in linear discrete ill-posed problems," *Signal Process.*, vol. 152, pp. 35–46, Nov. 2018.

[22] C. Zarowski, *An Introduction to Numerical Analysis for Electrical and Computer Engineers*. Hoboken, NJ, USA: Wiley, 2004.

[23] A. Zollanvari and E. R. Dougherty, "Generalized consistent error estimator of linear discriminant analysis," *IEEE Trans. Signal Process.*, vol. 63, no. 11, pp. 2804–2814, Jun. 2015.

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

[25] H. Sifaou, A. Kammoun, and M.-S. Alouini, "Improved LDA classifier based on spiked models," in *Proc. IEEE 19th Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Jun. 2018, pp. 1–5.

[26] K. Elkhalil, A. Kammoun, R. Couillet, T. Y. Al-Naffouri, and M.-S. Alouini, "Asymptotic performance of regularized quadratic discriminant analysis based classifiers," in *Proc. IEEE 27th Int. Workshop Mach. Learn. Signal Process. (MLSP)*, Sep. 2017, pp. 1–6.

[27] A. Zaib, T. Ballal, S. Khattak, and T. Y. Al-Naffouri, "A doubly regularized linear discriminant analysis classifier with automatic parameter selection," *IEEE Access*, vol. 9, pp. 51343–51354, 2021.

[28] A. N. Tikhonov, "Solution of incorrectly formulated problems and the regularization method," *Sov. Math. Dokl.*, vol. 4, pp. 1035–1038, Aug. 1963.

[29] P. C. Hansen, *Discrete Inverse Problems: Insight and Algorithms*. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 2010.

[30] F. Bauer and M. A. Lukas, "Comparingparameter choice methods for regularization of ill-posed problems," *Math. Comput. Simul.*, vol. 81, no. 9, pp. 1795–1841, May 2011, doi: 10.1016/j.matcom.2011.01.016.

[31] S. Chandrasekaran, G. H. Golub, M. Gu, and A. H. Sayed, "Parameter estimation in the presence of bounded data uncertainties," *SIAM J. Matrix Anal. Appl.*, vol. 19, no. 1, pp. 235–252, Jan. 1998, doi: 10.1137/s0895479896301674.

[32] T. Ballal and T. Y. Al-Naffouri, "Improved linear least squares estimation using bounded data uncertainty," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 3427–3431.

ALAM ZAIB received the B.Sc. degree (Hons.) in electrical engineering from the University of Engineering and Technology (UET), Peshawar, Pakistan, in 2002, the joint M.Sc. degree in electrical engineering and information technology from Universitat Politècnica de Catalunya (UPC), Barcelona, Spain, and Karlsruhe Institute of Technology (KIT), Germany, in 2009, and the Ph.D. degree in electrical engineering from King Fahd University of Petroleum and Minerals (KFUPM), Dhahran, Saudi Arabia, in 2015. He was an Erasmus Mundus Scholar at MERIT Master Program, from 2007 to 2009. Currently, he is an Associate Professor with the Department of Electrical and Computer Engineering, COMSATS University Islamabad, Abbottabad Campus. His research interests include statistical signal processing, wireless communications, and applications of machine learning and artificial neural networks in antenna arrays and wireless communications.

SHAHID KHATTAK received the B.Sc. degree from the University of Engineering and Technology (UET), Peshawar, Pakistan, in 1993, the M.S.E.E. degree from Purdue University, USA, in 1997, and the Ph.D. degree from Technische Universität Dresden, Germany, in 2008. Currently, he is a Professor with the Department of ECE, COMSATS University Islamabad, Abbottabad Campus, Abbottabad. His research interests include wireless communications and signal processing.

GHULAM MUJTABA received the B.Sc. degree in computer systems engineering from GIKIEST, Pakistan, in 2003, and the Postgraduate Diploma and Ph.D. degrees in electrical engineering from the High-Speed Networks Laboratory, Loughborough University, in 2011. Currently, he is an Assistant Professor with the Department of Electrical and Computer Engineering, COMSATS University Islamabad, Abbottabad Campus. His research interests include network security and machine learning.

SHAHID KHAN was born in Landikotal, Pakistan, in 1986. He received the B.S. degree in telecommunication engineering from the University of Engineering and Technology, Peshawar, Pakistan, the M.S. degree in satellite navigation and related applications from Politecnico de Torino, Italy, in 2011, and the Ph.D. degree from the University of Lorraine, France, in 2021. He was a Visiting Fellow at the 5G Innovation Center, University of Surrey. Currently, he is working as an Assistant Professor with the Department of Electrical and Computer Engineering, COMSATS University Islamabad, Abbottabad Campus. His research interests include development of circularly polarized phased array DRAs for satellite application, implantable antennas, and reconfigurable dielectric resonator antenna for different wireless applications.

AMAL AL-RASHEED received the Ph.D. degree in information systems from King Saud University, in 2017. She is currently an Associate Professor with the Department of Information Systems, College of Computer and Information Sciences, Princess Nourah Bint Abdulrahman University (PNU), Riyadh, Saudi Arabia. She has been involved in many projects related to learning technologies, cyber security, and virtual reality. Her contributions in research projects in academia led to the publication of papers in many journals and conferences. Her research interests include education, knowledge management, data mining, data analytics, cyber security, and natural language processing. In 2017, she was awarded the Research Excellence Award, by PNU, for her publications during performing Ph.D.