IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received March 4, 2021, accepted March 19, 2021, date of publication March 24, 2021, date of current version April 8, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3068611

A Doubly Regularized Linear Discriminant Analysis Classifier With Automatic Parameter Selection

ALAM ZAIB[1,*], TARIG BALLAL[2,*], (Member, IEEE), SHAHID KHATTAK[1], AND TAREQ Y. AL-NAFFOURI[2], (Senior Member, IEEE)

[1]Department of Electrical and Computer Engineering, COMSATS University Islamabad (CUI), Abbottabad 22060, Pakistan
[2]Electrical and Computer Engineering, CEMSE, King Abdullah University of Science and Technology (KAUST), Thuwal 23955, Saudi Arabia

Corresponding author: Tarig Ballal (tarig.ahmed@kaust.edu.sa)

[*]Alam Zaib and Tarig Ballal contributed equally to this work and are co-first authors.

This work was supported by the King Abdullah University of Science and Technology (KAUST) Office of Sponsored Research (OSR) under Award OSR-CRG2019-4041.

ABSTRACT Linear discriminant analysis (LDA) based classifiers tend to falter in many practical settings where the training data size is smaller than, or comparable to, the number of features. As a remedy, different regularized LDA (RLDA) methods have been proposed. These methods may still perform poorly depending on the size and quality of the available training data. In particular, the test data deviation from the training data model, for example, due to noise contamination, can cause severe performance degradation. Moreover, these methods commit further to the Gaussian assumption (upon which LDA is established) to tune their regularization parameters, which may compromise accuracy when dealing with real data. To address these issues, we propose a doubly regularized LDA classifier that we denote as R2LDA. In the proposed R2LDA approach, the RLDA score function is converted into an inner product of two vectors. By substituting the expressions of the regularized estimators of these vectors, we obtain the R2LDA score function that involves two regularization parameters. To set the values of these parameters, we adopt three existing regularization techniques; the constrained perturbation regularization approach (COPRA), the bounded perturbation regularization (BPR) algorithm, and the generalized cross-validation (GCV) method. These methods are used to tune the regularization parameters based on linear estimation models, with the sample covariance matrix's square root being the linear operator. Results obtained from both synthetic and real data demonstrate the consistency and effectiveness of the proposed R2LDA approach, especially in scenarios involving test data contaminated with noise that is not observed during the training phase.

INDEX TERMS Linear discriminant analysis, LDA, RLDA, regularization, covariance matrix estimation, classification algorithms.

I. INTRODUCTION

The idea of linear discriminant analysis (LDA) was originally conceived by Fisher [1] and is based on the assumption that the data follows a Gaussian distribution with a common class covariance matrix. Owing to its simplicity, LDA has been successfully applied to various classification and recognition tasks such as detection [2], speech recognition [3], cancer genomics [4], [5] and face recognition [6] to mention a few. In addition, LDA is a classical tool for feature extraction [7].

The performance of LDA-based classifiers depends heavily on accurate estimation of the class statistics, namely, the sample covariance matrix and class mean vectors. These statistics can be estimated with fairly high accuracy when the number of available samples is large compared to the data dimensionality. In practical high-dimensional data settings, the challenge is to cope with a limited number of available samples. In this case, the sample covariance estimates become highly perturbed and ill-conditioned resulting in severe performance degradation. To alleviate this problem, the sample covariance matrix is replaced with a regularized or ridge covariance matrix [8], giving the name regularized LDA (RLDA). The values of the regularization parameters

The associate editor coordinating the review of this manuscript and approving it for publication was Aysegul Ucar.

ultimately dictate the performance of RLDA classifiers. Hence, it is essential to judiciously tune the regularization parameters' values to reap the full benefit of the regularization process. Towards this end, various regularization techniques have been proposed. For example, cross-validation [9] has been one of the classical techniques for estimating the ridge parameter as evidenced in [5], [10]–[13].

An optimal regularization method that minimizes the asymptotic classification error is derived in [14], [15]. The method is based on recent results from random matrix theory. In [16], [17], the method of [14], [15] is extended to a more general class of discriminant analysis based classifiers, with LDA obtained as a special case. In [18], [19], improved RLDA classifiers are proposed, with the required parameters given in closed forms. These classifiers are designed for spiked-model covariance structures. Nevertheless, the authors demonstrate their usefulness when the data is generated from other (non-spiked) models.

In all the above-mentioned RLDA approaches, a regularization parameter is tuned based only on the data available in the training phase. Such a regularization parameter may produce satisfactory results when the test data follows the exact model of the training data. In some practical situations, it occurs that the test data deviates from the training data model. For example, the training data and the test data might represent measurements obtained from non-identical devices. In such a case, the value of the regularization parameter computed during the training phase may no longer be adequate, let alone be optimal. Consequently, the above-mentioned approaches' performance might deteriorate significantly. Moreover, these methods use the Gaussian assumption of the underlying data distribution for finding the value of the regularization parameter. This assumption may not hold in practical settings, e.g., with real data. Even though the Gaussian assumption is essential in deriving the basic LDA, excessive reliance on the assumption may eventually compromise the RLDA classifier's performance. To tackle these issues, we propose a new approach to regularized LDA classification. Focusing on binary classification, this paper develops a doubly regularized LDA (R2LDA) classifier by expressing the LDA score function as an inner product of two vectors that are linearly related to the mean vectors and the data covariance matrix. Regularized estimators are used to obtain the values of the two vectors and the value of the score function. The regularization parameter used in the estimation of one of the two vectors is tuned based on the current sample of the test data, hence providing robustness against any irregularities in the test data.

We summarize our main innovations and the most prominent features of the proposed R2LDA approach as follows:

(a) We deviate from the classical covariance matrix estimation approach to RLDA, where the focus is to obtain a regularized linear estimator of the data covariance matrix. Instead, we reformulate the problem as a vector estimation problem. We apply regularization to estimate two vector quantities. This implicitly results in a regularized *nonlinear* estimator of the data covariance matrix.

(b) R2LDA is designed not only to cope with the insufficiency of the training data but also with perturbations in the test data that are not observed during training. This is achieved by adjusting two regularization parameters independently; one is computed based only on the training data, and another is dynamically tuned to the test data sample. This is to be contrasted with existing approaches that compute their regularization parameters based solely on the training data.

(c) We automate the regularization parameter selection process based on existing methods that are well suited to the task. We theoretically motivate the main approaches adopted to tune the regularization parameters.

(d) The regularization parameter selection approach is agnostic to the underlying distribution of the data contrary to [15], [16], [18], which rely on the Gaussian assumption. Even though the Gaussian assumption is embedded in LDA, further commitment to Gaussianity in the regularization parameter tuning process might impede classification performance, especially with real data.

A. NOTATIONS

Throughout this paper, we use non-bold letters to denote scalars (e.g., W), boldface lowercase letters to denote column 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 $(.)^T$ 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 $P(x)$ 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 remainder of this paper is organized as follows. In Section II, we present a concise overview of regularized LDA classification. In Section III, we present our proposed R2LDA approach, along with three regularization parameter selection methods. Performance evaluation of the proposed approach and comparisons with existing techniques are presented in Section IV. We close this paper by making a concluding remark in Section IV.

II. RLDA CLASSIFICATION

We consider the binary classification problem of assigning a multivariate observation vector $\mathbf{x} \in \mathbb{R}^{p\times 1}$ to one of two classes $\mathcal{C}_i, i = 0, 1$. Let π_i be the prior probability that $\mathbf{x}$ belongs to the class $\mathcal{C}_i$, and assume that the class conditional densities $P\left(\mathbf{x}|\mathbf{x} \in \mathcal{C}_i\right), i = 0, 1,$ are Gaussian with mean vectors $\mathbf{m}_i \in \mathbb{R}^{p\times 1}$ and positive semidefinite covariance matrices $\Sigma_i \in \mathbb{R}^{p\times p}$.

LDA employs the Bayesian discriminant rule, which assigns $\mathbf{x}$ to the class with the maximum posterior probability. Let $\mathcal{S}_0 = \{\mathbf{x}_l\}_{l=0}^{n_0}$ and $\mathcal{S}_1 = \{\mathbf{x}_l\}_{l=n_0+1}^{n_0+n_1}$ represent the available training samples pertaining to the two classes, where n_i is the number of training samples for class $\mathcal{C}_i$ and $n = n_0 + n_1$ is the total number of training samples. The LDA score function reads [20]

$$W^{LDA}(\mathbf{x}) = \left(\mathbf{x} - \frac{\hat{\mathbf{m}}_0 + \hat{\mathbf{m}}_1}{2}\right)^{\mathrm{T}} \hat{\Sigma}^{-1} \left(\hat{\mathbf{m}}_0 - \hat{\mathbf{m}}_1\right). \quad (1)$$

The unbiased mean vector estimates $\hat{\mathbf{m}}_i$, and the pooled sample covariance matrix $\hat{\Sigma}$, are computed according to

$$\hat{\mathbf{m}}_i = \frac{1}{n_i}\sum_{l\in\mathcal{S}_i} \mathbf{x}_l, \quad \hat{\Sigma} = \frac{(n_0-1)\hat{\Sigma}_0 + (n_1-1)\hat{\Sigma}_1}{n_0+n_1+1}, \quad (2)$$

where the sample covariance matrices $\hat{\Sigma}_i$ are computed using

$$\hat{\Sigma}_i = \frac{1}{n_i - 1}\sum_{l\in\mathcal{S}_i} (\mathbf{x}_l - \hat{\mathbf{m}}_i)(\mathbf{x}_l - \hat{\mathbf{m}}_i)^{\mathrm{T}}. \quad (3)$$

The class assignment rule for $\mathbf{x}$ is as follows:

$$\mathbf{x} \in \begin{cases} \mathcal{C}_0, & \text{if } W(\mathbf{x}) > \log(\pi_1/\pi_0); \\ \mathcal{C}_1, & \text{otherwise.} \end{cases} \quad (4)$$

A major source of error in the above formulation is the inversion of the sample covariance matrix $\hat{\Sigma}$. In many practical setups where n is comparable to p, $\hat{\Sigma}$ becomes ill-conditioned, or even singular. To circumvent this issue, $\hat{\Sigma}^{-1}$ in (1) is replaced with a regularized estimator. Typically, $\mathbf{H} = (\mathbf{I}_p + \gamma\hat{\Sigma})^{-1}$ is used, where $\gamma \in \mathbb{R}^+$ is a regularization parameter and $\mathbf{I}_p$ is the identity matrix of dimension p. This replacement results in the RLDA score function [14], [15]

$$W^{RLDA}(\mathbf{x}) = \left(\mathbf{x} - \frac{\hat{\mathbf{m}}_0 + \hat{\mathbf{m}}_1}{2}\right)^{\mathrm{T}} \mathbf{H}\left(\hat{\mathbf{m}}_0 - \hat{\mathbf{m}}_1\right). \quad (5)$$

In this work, we apply a different regularization form to (1). In the proposed regularized LDA classifier, we employ two separate regularization operations to account for the deficiency in the training data. The proposed approach also improves the classifier's robustness to error contributions that are present only in the test data.

III. THE PROPOSED R2LDA CLASSIFICATION APPROACH

Many existing RLDA techniques are based on (5), with $\mathbf{H}$ estimated by selecting the regularization parameter γ using only the training data. This makes these techniques vulnerable to errors in the test data. To address this issue, we express the LDA score function (1) as

$$W^{LDA}(\mathbf{x}) = (\mathbf{x}')^{\mathrm{T}}\hat{\Sigma}^{-\frac{1}{2}}\hat{\Sigma}^{-\frac{1}{2}}\hat{\mathbf{m}}^{-} = \mathbf{z}^{\mathrm{T}}\mathbf{b}, \quad (6)$$

where $\mathbf{x}' := \mathbf{x} - \frac{1}{2}\hat{\mathbf{m}}^{+}$, $\hat{\mathbf{m}}^{+} := \hat{\mathbf{m}}_0 + \hat{\mathbf{m}}_1$, $\hat{\mathbf{m}}^{-} := \hat{\mathbf{m}}_0 - \hat{\mathbf{m}}_1$, $\mathbf{z} := \hat{\Sigma}^{-\frac{1}{2}}\mathbf{x}'$, and $\mathbf{b} := \hat{\Sigma}^{-\frac{1}{2}}\hat{\mathbf{m}}^{-}$. Based on the last two definitions, our proposed R2LDA method aims to obtain regularized estimates of $\mathbf{z}$ and $\mathbf{b}$ to improve the computation of the score function (6). To this end, we utilize the linear models

$$\mathbf{x}' = \hat{\Sigma}^{\frac{1}{2}}\mathbf{z} + \mathbf{v}_x, \quad (7)$$

$$\hat{\mathbf{m}}^{-} = \hat{\Sigma}^{\frac{1}{2}}\mathbf{b} + \mathbf{v}_m, \quad (8)$$

where $\mathbf{v}_x$ and $\mathbf{v}_m$ are additive noise vectors. These noise vectors can be interpreted as the contribution of the errors in estimating the mean vectors. In addition, $\mathbf{v}_x$ can also be used to absorb any noise contributions that occur in the test data vector $\mathbf{x}$. Each of (7) and (8) can be represented by the linear model

$$\mathbf{y} = \hat{\Sigma}^{\frac{1}{2}}\mathbf{c} + \mathbf{v}, \quad (9)$$

where (7) or (8) can be obtained by setting $\{\mathbf{y} = \mathbf{x}', \mathbf{c} = \mathbf{z}, \mathbf{v} = \mathbf{v}_x\}$, or $\{\mathbf{y} = \hat{\mathbf{m}}^{-}, \mathbf{c} = \mathbf{b}, \mathbf{v} = \mathbf{v}_m\}$, respectively.

Focusing on (9), regularization methods, commonly named ridge regression or Tikhonov regularization [21]–[23], can be applied to obtain a stabilized estimate of $\mathbf{c}$. This estimate can be expressed in a closed form as [24]

$$\hat{\mathbf{c}} = (\hat{\Sigma} + \gamma\mathbf{I}_p)^{-1}\hat{\Sigma}^{\frac{1}{2}}\mathbf{y}. \quad (10)$$

Based on (10), we can estimate $\mathbf{z}$ and $\mathbf{b}$ and substitute the results in (6) to obtain the R2LDA score function in the form

$$\begin{aligned} W^{R2LDA}(\mathbf{x}) &= \hat{\mathbf{z}}^{\mathrm{T}}\hat{\mathbf{b}} \\ &= (\mathbf{x}')^{\mathrm{T}}\mathbf{U}\mathbf{D}^2\left(\mathbf{D}^2 + \gamma_z\mathbf{I}_p\right)^{-1} \\ &\quad \times \left(\mathbf{D}^2 + \gamma_b\mathbf{I}_p\right)^{-1}\mathbf{U}^{\mathrm{T}}\hat{\mathbf{m}}^{-}, \end{aligned} \quad (11)$$

where $\gamma_z \in \mathbb{R}^+$ and $\gamma_b \in \mathbb{R}^+$ are the regularization parameters associated with the linear models (7) and (8), respectively. The second equality in (11) follows directly from substituting (in (10)) the eigenvalue decomposition (EVD) $\hat{\Sigma} = \mathbf{U}\mathbf{D}^2\mathbf{U}^{\mathrm{T}}$, where $\mathbf{U}$ is the matrix of eigenvectors and $\mathbf{D}^2$ is the diagonal matrix of eigenvalues of $\hat{\Sigma}$.

Now, it only remains to set the values of the regularization parameters γ_z and γ_b, which will be discussed in the following subsections.

Remark 1: Compared to the conventional RLDA score function (5), the new formulation (11) involves two regularization operations. Note that the estimation of the class mean vectors $\mathbf{m}_i$ results in perturbations in both $\hat{\mathbf{m}}^{-}$ and $\mathbf{x}'$. Besides, $\mathbf{x}'$ also has errors coming from the test data. By carrying out two independent estimations to obtain regularized estimates of $\mathbf{z}$ and $\mathbf{b}$ (see (6)), we can optimize the choice of two different regularization parameters to cope with the different perturbations in $\mathbf{x}'$ and $\hat{\mathbf{m}}^{-}$. This is a key advantage of the proposed R2LDA method over the classical RLDA based on (5) that employs a single regularization operation based only on the training data.

A. REGULARIZATION PARAMETER SELECTION

Several methods have been proposed in the literature for selecting the regularization parameter γ required in (10), e.g., [25]–[28], to mention a few. These methods are based on different criteria, which results in different regularization parameter values (see [29]).

In this work, we pursue three regularization methods; the constrained perturbation regularization approach (COPRA) [30], bounded perturbation regularization (BPR) [31], and the generalized cross-validation (GCV) [26]. The choice of COPRA and BPR is motivated by the fact that these algorithms are designed to optimize the *mean squared error* of a vector estimation. Also, these two methods are based on a very relevant model to the setup under consideration. As will be shown subsequently, BPR is a special case of COPRA. On the other hand, cross-validation, a method based on a totally different concept compared to BPR and COPRA, is a widely adopted heuristic technique that has shown immense success in machine-learning applications.

Next, we provide details on the three selected regularization methods and how they can be combined with R2LDA.

B. THE CONSTRAINED PERTURBATION REGULARIZATION ALGORITHM (COPRA)

To simplify the derivations, we make the following assumptions on the model (9):

1) The noise vector $\mathbf{v}$ has zero mean and an unknown covariance matrix $\sigma_v^2 \mathbf{I}_p$.
2) The unknown random vector $\mathbf{c}$ is zero mean with an unknown positive semidefinite diagonal covariance matrix $\Sigma_{\mathbf{cc}}$.
3) The vectors $\mathbf{v}$ and $\mathbf{c}$ are mutually independent.

COPRA is based on the principle of introducing an artificial perturbation in a linear model to improve the singular-value structure of the resulting model matrix. For the linear model in (9), $\hat{\Sigma}^{\frac{1}{2}}$ is replaced by a perturbed version to obtain the model

$$\mathbf{y} \approx \left(\hat{\Sigma}^{\frac{1}{2}} + \Delta\right) \mathbf{c} + \mathbf{v}, \tag{12}$$

where $\Delta \in \mathbb{R}^{p \times p}$ is an unknown perturbation matrix which is norm bounded by a positive quantity λ, i.e., $\|\Delta\|_2 \leq \lambda$. The original method in [30] utilizes the perturbation Δ to stabilize the estimation of $\mathbf{c}$ based on the model (9). However, in this specific application, Δ can be viewed as a genuine *uncertainty* in the model due to the noisy nature of $\hat{\Sigma}^{\frac{1}{2}}$. In other words, (12) is the natural model for our vector estimation problem. These two different interpretations of Δ in (12) yield identical estimators of the vector $\mathbf{c}$ (i.e., the same value of the regularization parameter in (10)). This makes COPRA an excellent candidate for computing the regularization parameters for R2LDA.

To obtain an estimate of $\mathbf{c}$, we consider the minimization of the worst-case residual error. Namely, we pursue the following optimization:

$$\min_{\hat{\mathbf{c}}} \max_{\Delta} \left\| \mathbf{y} - \left(\hat{\Sigma}^{\frac{1}{2}} + \Delta\right) \hat{\mathbf{c}} \right\|_2, \quad \text{s.t. } \|\Delta\|_2 \leq \lambda. \tag{13}$$

Interestingly, as shown in [30], [32], [33], the min-max problem (13) can be converted to a minimization problem whose solution is given by (10), with the additional constraint

$$\gamma \|\hat{\mathbf{c}}\|_2 = \lambda \left\| \mathbf{y} - \hat{\Sigma}^{\frac{1}{2}} \hat{\mathbf{c}} \right\|_2. \tag{14}$$

Based on (14), we observe that the solution of (13) depends on the bound λ (in addition to the other system parameters) and is agnostic to the structure of the perturbation matrix Δ.

Now, we can substitute (10) and the EVD of $\hat{\Sigma}$ in (14) and manipulate to obtain

$$\lambda^2 = \frac{\operatorname{tr}\left(\left(\mathbf{D}^2 + \gamma \mathbf{I}_p\right)^{-2} \mathbf{U}^T \mathbf{y}\mathbf{y}^T \mathbf{U}\right)}{\operatorname{tr}\left(\mathbf{D}^2 \left(\mathbf{D}^2 + \gamma \mathbf{I}_p\right)^{-2} \mathbf{U}^T \mathbf{y}\mathbf{y}^T \mathbf{U}\right)}. \tag{15}$$

where tr(.) is the matrix trace operation. Since λ in (15) is stochastic in nature (due to the involvement of $\mathbf{y}$), we consider a value of λ that would represent the average case. To this end, we replace $\mathbf{y}\mathbf{y}^T$ with its expected value $\mathbb{E}(\mathbf{y}\mathbf{y}^T)$, which can be written based on (9) in the following form:

$$\mathbb{E}(\mathbf{y}\mathbf{y}^T) = \mathbf{U}\mathbf{D}\mathbf{U}^T \Sigma_{\mathbf{cc}} \mathbf{U}\mathbf{D}\mathbf{U}^T + \sigma_v^2 \mathbf{I}_p. \tag{16}$$

Owing to the ill-conditioning of $\hat{\Sigma}$, it is likely that some of its eigenvalues are very close, or even equal, to zero. Therefore, the EVD of $\hat{\Sigma}$ can be written in the form

$$\hat{\Sigma} = [\mathbf{U}_1 \;\; \mathbf{U}_2] \begin{bmatrix} \mathbf{D}_1^2 & \mathbf{0}_{p_1 \times p_2} \\ \mathbf{0}_{p_2 \times p_1} & \mathbf{D}_2^2 \end{bmatrix} \begin{bmatrix} \mathbf{U}_1^T \\ \mathbf{U}_2^T \end{bmatrix} \simeq \mathbf{U}_1 \mathbf{D}_1^2 \mathbf{U}_1^T, \tag{17}$$

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 based approach to find the point of this partitioning is recommended in [30]. 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 (17) is to improve numerical stability by removing extremely small eigenvalues.

Now, we substitute (16) and (17) in (15) and manipulate to obtain (18) (as shown at the bottom of the next page). Next, we proceed to eliminate σ_v and $\Sigma_{\mathbf{cc}}$ from (18) by using the *mean squared error* (MSE) as a performance criterion. The MSE of the RLS estimator (10) can be written as [24]

$$\begin{aligned} \text{MSE} = \operatorname{tr}\left(\mathbb{E}\left((\mathbf{c} - \hat{\mathbf{c}})(\mathbf{c} - \hat{\mathbf{c}})^T\right)\right) = \sigma_v^2 \operatorname{tr}\left(\mathbf{D}^2 \left(\mathbf{D}^2 + \gamma \mathbf{I}_p\right)^{-2}\right) \\ + \gamma^2 \operatorname{tr}\left(\left(\mathbf{D}^2 + \gamma \mathbf{I}_p\right)^{-2} \mathbf{U}^T \Sigma_{\mathbf{cc}} \mathbf{U}\right). \end{aligned} \tag{19}$$

By differentiating (19), the regularization parameter γ that minimizes the MSE can be obtained using

$$\frac{\partial\, (\text{MSE})}{\partial\, \gamma} = 0 \implies \gamma = \frac{p\, \sigma_v^2}{\operatorname{tr}\left(\Sigma_{\mathbf{cc}}\right)}. \tag{20}$$

By substituting (20) in (18), we obtain (21), as shown at the bottom of the page, which shows a bound λ that does not depend on the statistics of $\mathbf{c}$ or those of the noise. Note that the derivations of (16) and (18) require Assumptions 1–3 to be satisfied–otherwise, these results will hold only in an approximation way.

Ultimately, by using (21), we can eliminate λ from (15) to obtain (22), where $\mathbf{d} := \mathbf{U}^{\mathrm{T}}\mathbf{y}$. Equation (22), as shown at the bottom of the page, which is nonlinear in γ, can be solved by using Newton's method [34] 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 [30].

C. BOUNDED PERTURBATION REGULARIZATION (BPR)

Similar to COPRA, the BPR approach is also based on the model (12) [31]. The derivation of the BPR algorithm takes similar steps to those of COPRA except for the eigenvalue matrix partitioning step (17), which is omitted. In fact, the BPR algorithm can be obtained by setting $p_1 = p$ and manipulating (22), which results in

$$\operatorname{tr}\left(\left(\mathbf{D}^2+\gamma\mathbf{I}_p\right)^{-1}\right)\operatorname{tr}\left(\left(\mathbf{D}^2+\gamma\mathbf{I}_p\right)^{-1}\mathbf{d}\mathbf{d}^{\mathrm{T}}\right) \\ -p\operatorname{tr}\left(\left(\mathbf{D}^2+\gamma\mathbf{I}_p\right)^{-2}\mathbf{d}\mathbf{d}^{\mathrm{T}}\right)=0. \qquad (23)$$

The above nonlinear equation can be solved using Newton's method to obtain the regularization parameter pertaining to the BPR algorithm.

D. THE GENERALIZED CROSS-VALIDATION (GCV) METHOD

One may consider using the GCV for automating the regularization parameter selection for R2LDA. In contrast to COPRA and BPR, GCV hinges on a different philosophy and is based on minimizing the GCV function [26]:

$$G(\gamma)=\frac{\left\|\hat{\Sigma}^{\frac{1}{2}}(\hat{\Sigma}+\gamma\mathbf{I}_p)^{-1}\hat{\Sigma}^{\frac{1}{2}}\mathbf{y}-\mathbf{y}\right\|_2^2}{\left(\operatorname{tr}\left(\mathbf{I}_p-\hat{\Sigma}^{\frac{1}{2}}(\hat{\Sigma}+\gamma\mathbf{I}_p)^{-1}\hat{\Sigma}^{\frac{1}{2}}\right)\right)^2}, \qquad (24)$$

which can be manipulated to the form

$$G(\gamma)=\frac{\left\|\mathbf{D}^2(\mathbf{D}^2+\gamma\mathbf{I}_p)^{-1}\mathbf{d}-\mathbf{d}\right\|_2^2}{\left(p-\operatorname{tr}\left(\mathbf{D}^2(\mathbf{D}^2+\gamma\mathbf{I}_p)^{-1}\right)\right)^2}. \qquad (25)$$

The GCV approach can be thought of as an approximation of leave-one-out cross-validation (the reader can refer to [26], chapter 4). To compute the regularization parameter using the GCV, a line search that evaluates $G(\gamma)$ over a suitably chosen γ interval is carried out. To set up the interval, we apply the technique described in [35].

E. SUMMARY OF THE PROPOSED R2LDA APPROACH

The main steps involved in the proposed R2LDA approach are summarized as follows:

1) Estimate the class statistics $\hat{\mathbf{m}}_i$, $\hat{\Sigma}_i$ and $\hat{\Sigma}$ from the training data by using (2) and (3).
2) Compute $\hat{\mathbf{m}}^+$, $\hat{\mathbf{m}}^-$ and the EVD of $\hat{\Sigma}$.
3) Set $\mathbf{y} = \hat{\mathbf{m}}^-$ in the model (9) and obtain γ_b using the chosen regularization parameter selection method.
4) For a given test sample, compute $\mathbf{x}'$.
5) Set $\mathbf{y} = \mathbf{x}'$ in the model (9) and obtain γ_z using the chosen regularization parameter selection method.
6) Compute the R2LDA score function using (11), and assign the test sample to a class according to (4).

In Step 3 and Step 5, we apply any of the three regularization parameter selection methods discussed in the previous subsections (COPRA, BPR or GCV). Henceforth, the resulting classification algorithm will be referred to as COPRA-R2LDA, BPR-R2LDA, or GCV-R2LDA, depending on the regularization parameter selection method used.

IV. PERFORMANCE EVALUATION

We demonstrate the performance of the proposed R2LDA classifiers with different regularization parameter selection techniques against the RLDA classifiers of the asymptotic error estimator (Asym-RLDA) [15] and the optimal-intercept-improved RLDA (OII-RLDA) [19]. We consider both synthetic and real data for performance evaluation. The codes used to generate the results are available online[1].

$$\lambda^2\left(\operatorname{tr}\left(\left(\mathbf{D}_1^2+\gamma\mathbf{I}_{p_1}\right)^{-2}\left(\mathbf{D}_1^2+\frac{p_1\sigma_v^2}{\operatorname{tr}(\Sigma_{\mathbf{cc}})}\mathbf{I}_{p_1}\right)\right)+\frac{(p-p_1)p_1\sigma_v^2}{\gamma^2\operatorname{tr}(\Sigma_{\mathbf{cc}})}\right)\simeq\operatorname{tr}\left(\mathbf{D}_1^2\left(\mathbf{D}_1^2+\gamma\mathbf{I}_{p_1}\right)^{-2}\left(\mathbf{D}_1^2+\frac{p_1\sigma_v^2}{\operatorname{tr}(\Sigma_{\mathbf{cc}})}\mathbf{I}_{p_1}\right)\right) \qquad (18)$$

$$\lambda^2\left(\operatorname{tr}\left(\left(\mathbf{D}_1^2+\gamma\mathbf{I}_{p_1}\right)^{-2}\left(\frac{p}{p_1}\mathbf{D}_1^2+\gamma\mathbf{I}_{p_1}\right)\right)+\frac{(p-p_1)}{\gamma}\right)\simeq\operatorname{tr}\left(\mathbf{D}_1^2\left(\mathbf{D}_1^2+\gamma\mathbf{I}_{p_1}\right)^{-2}\left(\frac{p}{p_1}\mathbf{D}_1^2+\gamma\mathbf{I}_{p_1}\right)\right) \qquad (21)$$

$$\operatorname{tr}\left(\mathbf{D}^2\left(\mathbf{D}^2+\gamma\mathbf{I}_p\right)^{-2}\mathbf{d}\mathbf{d}^{\mathrm{T}}\right)\operatorname{tr}\left(\left(\mathbf{D}_1^2+\gamma\mathbf{I}_{p_1}\right)^{-2}\left(\frac{p}{p_1}\mathbf{D}_1^2+\gamma\mathbf{I}_{p_1}\right)\right)+\frac{(p-p_1)}{\gamma}\operatorname{tr}\left(\mathbf{D}^2\left(\mathbf{D}^2+\gamma\mathbf{I}_p\right)^{-2}\mathbf{d}\mathbf{d}^{\mathrm{T}}\right) \\ -\operatorname{tr}\left(\left(\mathbf{D}^2+\gamma\mathbf{I}_p\right)^{-2}\mathbf{d}\mathbf{d}^{\mathrm{T}}\right)\operatorname{tr}\left(\mathbf{D}_1^2\left(\mathbf{D}_1^2+\gamma\mathbf{I}_{p_1}\right)^{-2}\left(\frac{p}{p_1}\mathbf{D}_1^2+\gamma\mathbf{I}_{p_1}\right)\right)=0 \qquad (22)$$

[1] https://github.com/tBallal/R2LDA

We use the average percentage classification error as the performance metric. This section also discusses the computational complexity of various algorithms.

A. DATASETS DESCRIPTION

Synthetic Data: The synthetic data is generated based on a Gaussian data model with dimension $p = 100$. The class covariance matrix Σ_0 is generated with diagonal elements equal to 1 and off-diagonal elements equal to 0.1, while the other class covariance matrix is generated as $\Sigma_1 = \Sigma_0 + \mathbf{I}$. As for the model mean vectors, we set $\mathbf{m}_1 = -\mathbf{m}_0$, where $\mathbf{m}_0 = [a, a, \ldots, a]^{\mathrm{T}}$. The parameter a is chosen according to the between-class Mahalanobis distance, δ, defined according to $\delta^2 = (\mathbf{m}_0 - \mathbf{m}_1)^{\mathrm{T}} \Sigma^{-1} (\mathbf{m}_0 - \mathbf{m}_1)$ [15]. We use $\delta^2 = 9$. A training set $\mathcal{S}_i$ of size n_i is generated independently in each training trial, where $n_0 = n_1$. For the test data, we generate an independent set of samples for each class.

Real Data: We use (i) the MNIST dataset that consists of 20 × 20 gray-scale images of handwritten digits [36], (ii) the phonemes dataset considered in [37], and (iii) the sonar classification dataset [38]. These datasets are available for download from the UCI Machine Learning Repository [2].

The MNIST images are vectorized to result in data of dimensionality $p = 400$. For binary classification, selected pairs of images are used.

The phonemes dataset is based on log-periodogram (of length $p = 256$) of digitized speech frames extracted from the TIMIT database (TIMIT Acoustic-Phonetic Continuous Speech Corpus, NTIS, U.S. Department of Commerce) [37], which is widely used in speech recognition. The phonemes are transcribed as: (1) "sh"as in "she", (2) "dcl"as in "dark", (3) "iy"as the vowel in "she", (4) "aa"as the vowel in "dark", and (5) "ao"as the first vowel in "water". For binary classification, selected pairs of phonemes are formed from the above five phonemes.

The sonar dataset consists of 208 examples, each with 60 attributes representing sonar returns from a metal cylinder (class 0) or a rough cylindrical rock (class 1).

B. EXPERIMENTS DESCRIPTION

For both the synthetic and real datasets, 500 training trials were carried out, each followed by a number between 50 and 500 test trials, depending on the size of the available of data from the dataset. Each training or test trial is based on a randomly generated/selected data. As a pre-processing step, all datasets are translated to the interval $[-1, 1]$ to facilitate comparison of results across different datasets.

For all datasets, we test the case where zero-mean Gaussian noise with standard deviation σ is added *only* to the test data. For each dataset, we test σ values that allow us to observe reasonable performance variability (some datasets are more resilient to noise than others). The statistical properties of this noise are not known to the proposed R2LDA classifier, nor are they known to any of the benchmark methods.

C. DIMENSIONALITY REDUCTION

In scenarios involving high-dimensional data and a limited number of observations, one can reduce the dimensionality of the data by extracting a small set of the most significant features present in the data. While there are myriad of feature reduction/selection methods available [39], we apply the simple t-test and use the p-values of each feature as a criterion for feature selection. In our experiments, we apply dimensionality reduction to the MNIST dataset by selecting the top 12.5% features based on the p-values. This exercise aims to investigate the behavior of the proposed classifiers in setups with reduced dimensionality.

D. RESULTS DISCUSSION

Figs. 1–5 plot the percentage classification errors versus the training data size (n) for different datasets under different test data noise levels. Fig.1 presents the results for the (synthetic) Gaussian data, while Fig.2, Fig.3 and Fig.4 show the results for the MNIST, phonemes and sonar datasets, respectively. On the other hand, Fig. 5 depicts results for an example from the MNIST dataset with reduced dimensionality. The MNIST results are based on the image/digit pairs (1,7), (5,8), and (7,9), while the phonemes dataset results use the phoneme combinations (1,2), (1,3), (1,5), and (4,5). From the results in Figs.1–5, we observe the following:

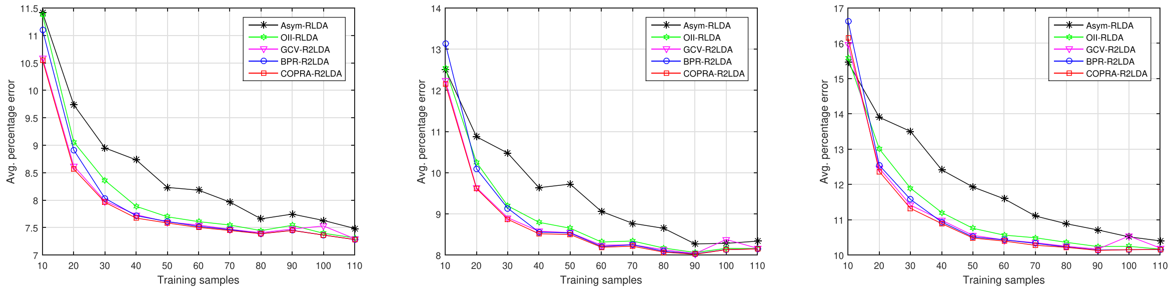
- On average, the R2LDA methods outperform the RLDA methods.
- The R2LDA methods remain more consistent and stable than the RLDA methods as the noise level in the test data increases. This is more visible in real datasets that deviate from Gaussianity.
- Amongst the R2LDA classifiers, COPRA-R2LDA and BPR-R2LDA appear to be slightly more consistent than GCV-R2LDA. GCV-R2LDA seems to occasionally falter, as in Fig.2(a), Fig. 2(d) and Fig.2(g).
- For the MNIST dataset with reduced dimensionality, the R2LDA methods preserve their superiority over the RLDA counterparts, especially in noisy conditions. This is evident from Fig. 5, where the top 50 features are selected out of 400 features present in the MNIST data.

E. COMPUTATIONAL COMPLEXITY

We consider the computational complexity of the proposed algorithms when classifying a test dataset of size k. Let l_{COPRA} and l_{BPR} be the maximum number of iterations required for the COPRA and BPR algorithms to converge. Also, let g_{GCV} and g_{Asym} be the number of grid points used in the search processes of the GCV and Asym methods, respectively. The worst-case time complexities of the proposed algorithms (including all the steps listed in Subsection III-E) and the benchmark methods are given in Table 1 using the big-O notation.

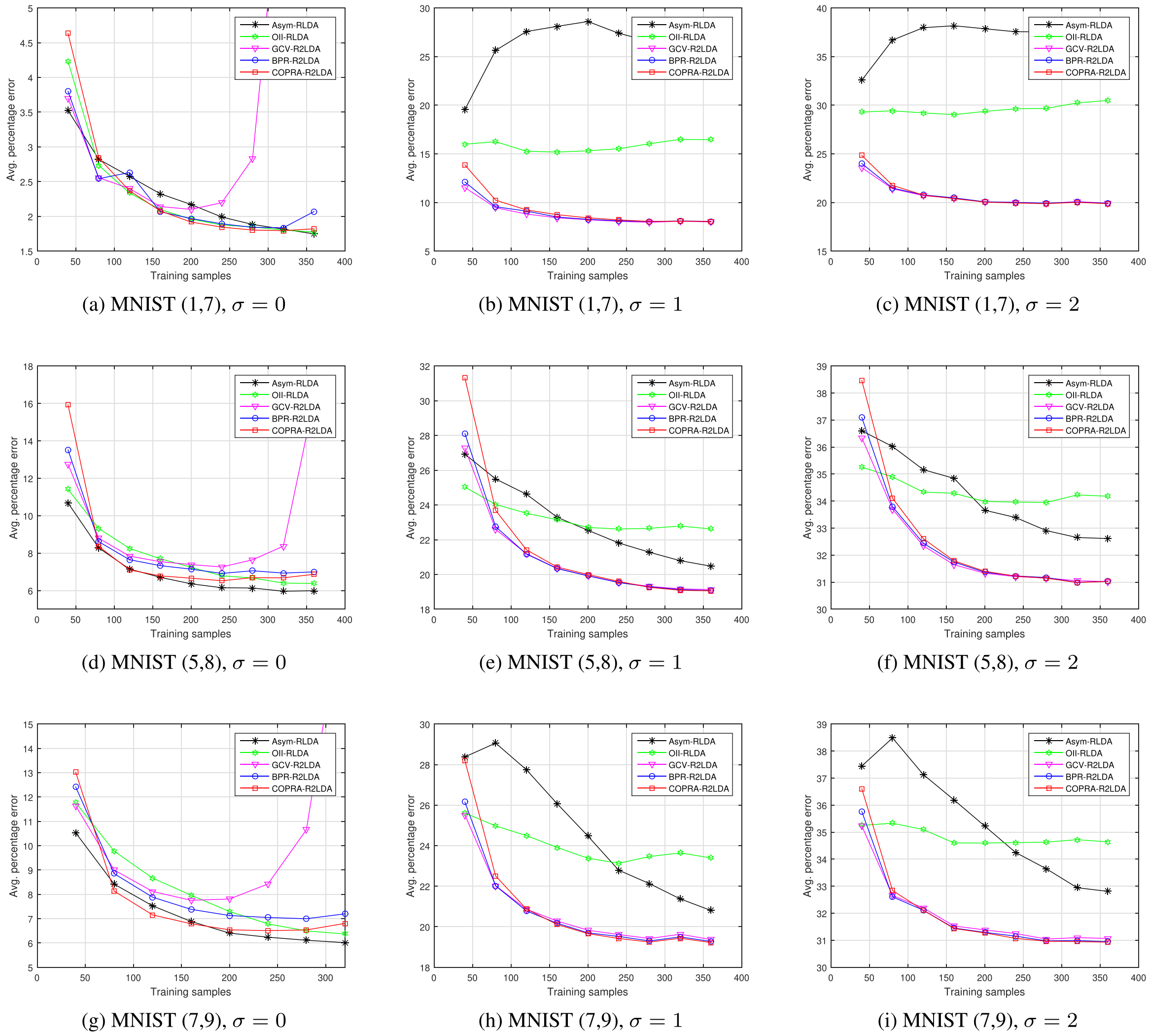
Note that all the five complexity expressions listed in Table 1 feature the terms np^2 and p^3. These two terms

[2] https://archive.ics.uci.edu/ml/datasets



(a) Gaussian, $\sigma = 0$ (b) Gaussian, $\sigma = 0.1$ (c) Gaussian, $\sigma = 0.2$

FIGURE 1. **Gaussian data misclassification rates versus training data size for different test data noise levels.**



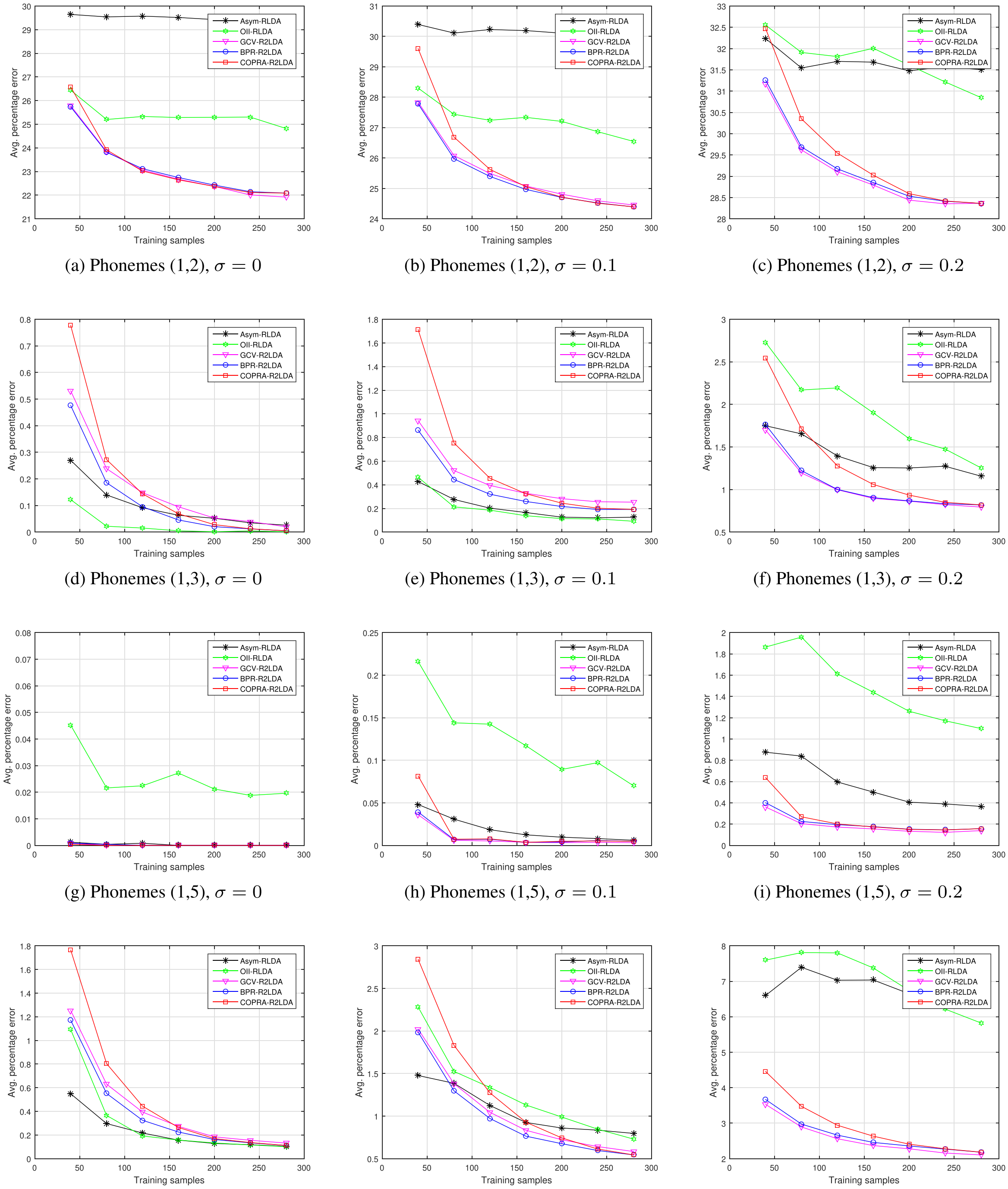
(a) MNIST (1,7), $\sigma = 0$ (b) MNIST (1,7), $\sigma = 1$ (c) MNIST (1,7), $\sigma = 2$

(d) MNIST (5,8), $\sigma = 0$ (e) MNIST (5,8), $\sigma = 1$ (f) MNIST (5,8), $\sigma = 2$

(g) MNIST (7,9), $\sigma = 0$ (h) MNIST (7,9), $\sigma = 1$ (i) MNIST (7,9), $\sigma = 2$

FIGURE 2. **MNIST data misclassification rates versus training data size for different test data noise levels.**

are, approximately, of similar order for scenarios with $n \approx p$. Each complexity expression includes a term of the form $\alpha k p^2$, with different α values for different methods. For a large α and/or a large number of test samples $k \gg p$, this term will dominate the complexity. For the RLDA methods, we have $\alpha = 1$. On the other hand, for the R2LDA methods, α takes the values l_{COPRA} and l_{BPR} and g_{GCV}, for the three methods respectively. These parameters are due to the computations involved in finding the regularization parameter γ_z each time a test data sample is classified. As an example, for $n \approx k \approx p$,



(a) Phonemes (1,2), $\sigma = 0$

(b) Phonemes (1,2), $\sigma = 0.1$

(c) Phonemes (1,2), $\sigma = 0.2$

(d) Phonemes (1,3), $\sigma = 0$

(e) Phonemes (1,3), $\sigma = 0.1$

(f) Phonemes (1,3), $\sigma = 0.2$

(g) Phonemes (1,5), $\sigma = 0$

(h) Phonemes (1,5), $\sigma = 0.1$

(i) Phonemes (1,5), $\sigma = 0.2$

(j) Phonemes (4,5), $\sigma = 0$

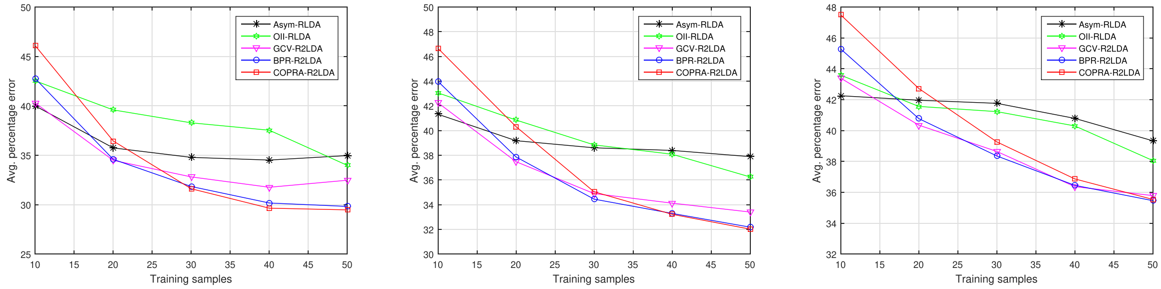
(k) Phonemes (4,5), $\sigma = 0.1$

(l) Phonemes (4,5), $\sigma = 0.2$

FIGURE 3. Phonemes data misclassification rates versus training data size for different test data noise levels.

an R2LDA algorithm with $\alpha \approx p$ would have a complexity $\mathcal{O}(p^4)$. Under the same conditions, an RLDA algorithm's complexity is $\mathcal{O}(p^3)$.

In addition to the time complexity, we also consider the runtimes of various algorithms observed during our experiments. We illustrate this using two examples. Fig. 6 compares the runtimes (in seconds) of various algorithms against the number of training samples for the Gaussian data used in Fig. 1. Fig. 6(a) and Fig. 6(b) plot the average runtime for a single test sample and 500 test samples, respectively. We observe that the COPRA-R2LDA is considerably slower than the other algorithms for both numbers of test data samples. Despite computing a new regularization parameter for each test data sample, BPR-R2LDA and GCV-R2LDA offer comparable runtimes to those of the benchmark RLDA methods.

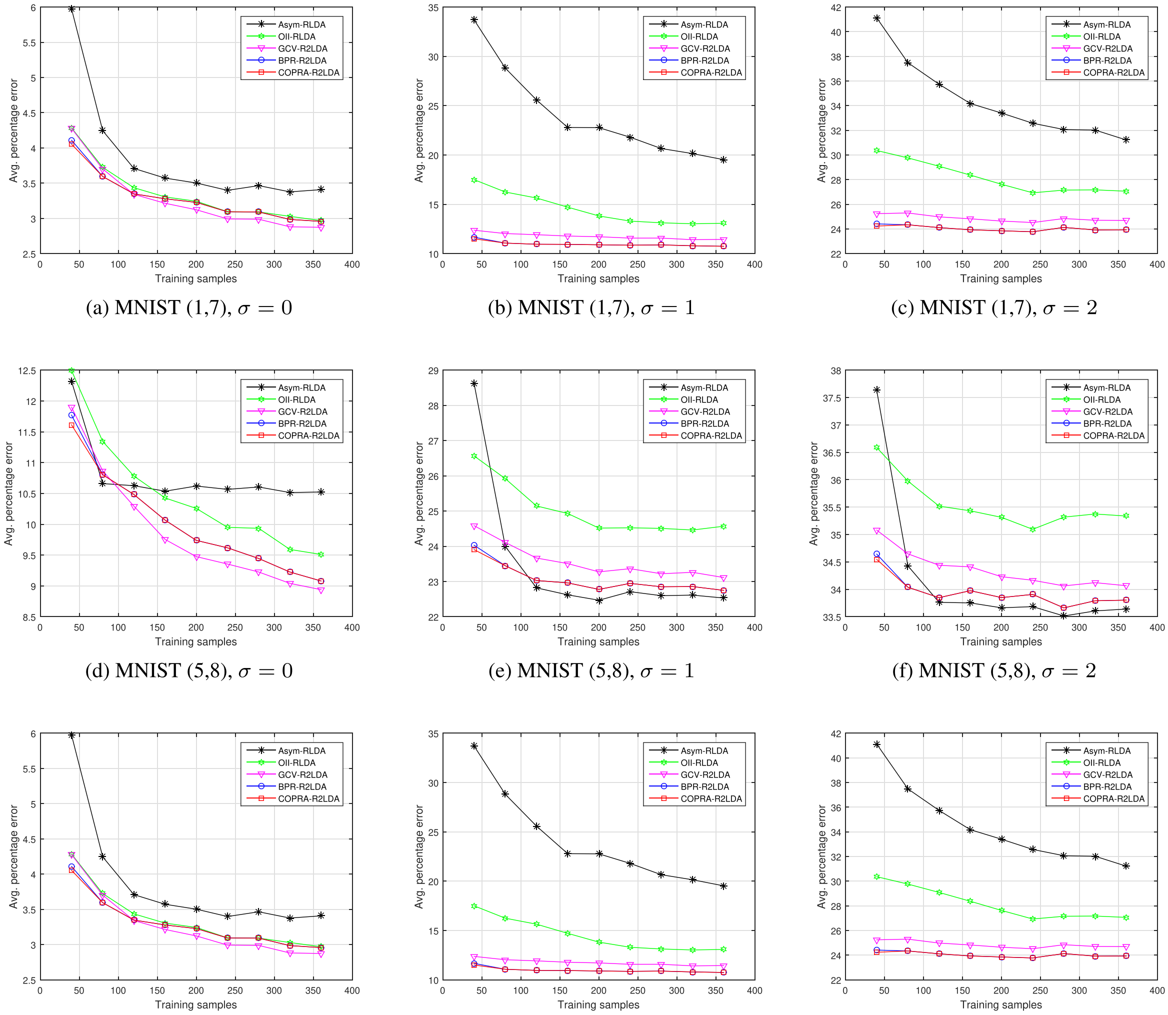


(a) Sonar, $\sigma = 0$

(b) Sonar, $\sigma = 0.1$

(c) Sonar, $\sigma = 0.2$

FIGURE 4. Sonar data misclassification rates versus training data size for different test data noise levels.



(a) MNIST (1,7), $\sigma = 0$

(b) MNIST (1,7), $\sigma = 1$

(c) MNIST (1,7), $\sigma = 2$

(d) MNIST (5,8), $\sigma = 0$

(e) MNIST (5,8), $\sigma = 1$

(f) MNIST (5,8), $\sigma = 2$

(g) MNIST (7,9), $\sigma = 0$

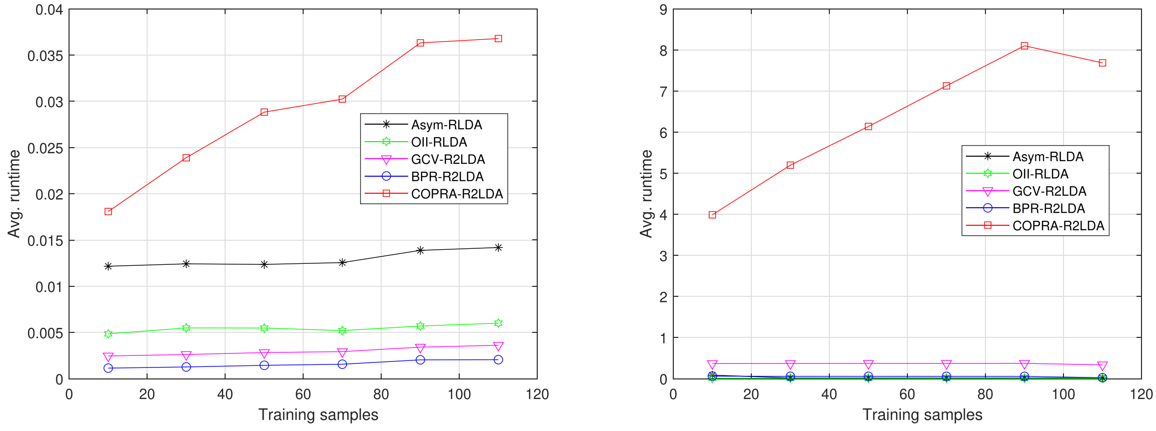
(h) MNIST (7,9), $\sigma = 1$

(i) MNIST(7,9), $\sigma = 2$

FIGURE 5. Reduced-dimension MNIST data misclassification rates versus training data size for different test data noise levels.

In Fig. 7, we show another example similar to Fig. 6 using the MNIST dataset. In this example, COPRA-R2LDA is faster than Asym-RLDA in the single-test case. Whereas, with 500 tests, COPRA-R2LDA becomes substantially slower than the rest of the algorithms. On the other hand, the runtimes of BPR-R2LDA and GCV-R2LDA stay relatively close to those of the RLDA methods when applied to 500 test samples, while offering the fastest runtimes in the single-test case. The slowness of the COPRA-R2LDA algorithm is attributed mainly to its large convergence time.

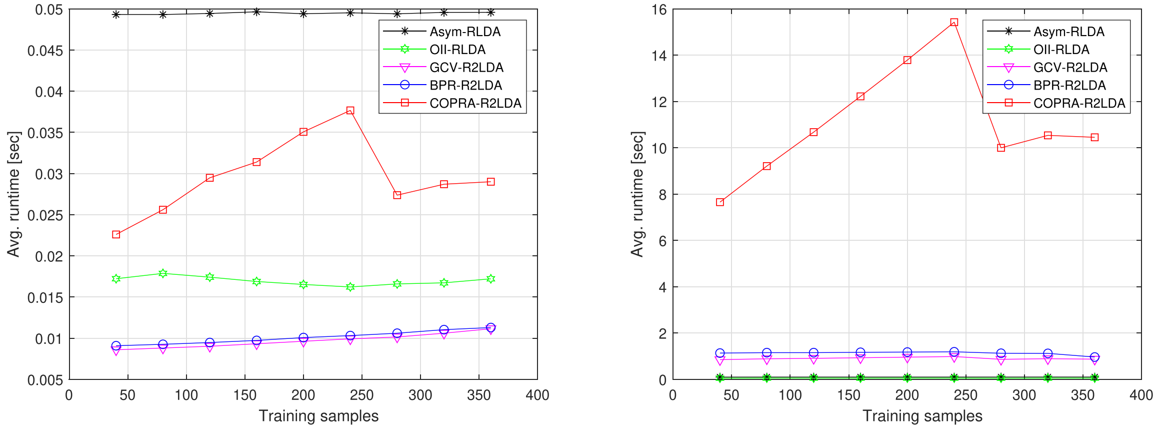
Based on the above discussions, we can conclude that, among the tested algorithms, BPR-R2LDA is the most



(a) One test

(b) 500 tests

FIGURE 6. Average runtime (in seconds) versus training data size for Gaussian data: (a) A single test sample, (b) 500 test samples.



(a) One test

(b) 500 tests

FIGURE 7. Average runtime (in seconds) versus training data size for the MNIST image pair (7, 9): (a) A single test sample, (b) 500 test samples.

TABLE 1. Time complexity summary.

No.	Algorithm	Complexity
1	Asym-RLDA	$\mathcal{O}(np^2 + p^3 + kp^2 + g_{\text{Asym}}p^2)$
2	OII-RLDA	$\mathcal{O}(np^2 + p^3 + kp^2)$
3	COPRA-R2LDA	$\mathcal{O}(np^2 + p^3 + l_{\text{COPRA}}kp^2)$
4	BPR-R2LDA	$\mathcal{O}(np^2 + p^3 + l_{\text{BPR}}kp^2)$
5	GCV-R2LDA	$\mathcal{O}(np^2 + p^3 + g_{\text{GCV}}kp^2)$

attractive classifier since it is much faster than COPRA-R2LDA and offers a more consistent classification performance than GCV-R2LDA.

V. CONCLUSION

We have presented novel regularized LDA classifiers based on a dual regularization scheme. The proposed R2LDA approach allows us to tune two regularization parameters independently. The first regularization parameter is computed offline from the training data. In contrast, the second regularization parameter is dynamically tuned to each test data sample. Based on synthetic and real datasets, results confirm our approach's effectiveness. The results also demonstrate the robustness of the proposed approach when noise is present in the test data. Although the proposed method is developed for binary classification, it can be easily extended to the multi-class case.

REFERENCES

[1] R. A. Fisher, ''The use of multiple measurements in taxonomic problems,'' *Ann. Eugenics*, vol. 7, no. 2, pp. 179–188, Sep. 1936.

[2] K. R. Varshney, ''Generalization error of linear discriminant analysis in spatially-correlated sensor networks,'' *IEEE Trans. Signal Process.*, vol. 60, no. 6, pp. 3295–3301, Jun. 2012.

[3] C. Avendano, S. Van Vuuren, and H. Hermansky, ''Data based filter design for RASTA-like channel normalization in ASR,'' in *Proc. 4th Int. Conf. Spoken Lang. Process. (ICSLP)*, 1996, pp. 2087–2090.

[4] 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, 2002.

[5] 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. Exp. Clin. Cancer Res.*, vol. 28, p. 149, Dec. 2009.

[6] 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.

[7] Z. Liu, K. Shi, K. Zhang, W. Ou, and L. Wang, "Discriminative sparse embedding based on adaptive graph for dimension reduction," *Eng. Appl. Artif. Intell.*, vol. 94, Sep. 2020, Art. no. 103758.

[8] P. J. Di Pillo, "The application of bias to discriminant analysis," *Commun. Statist.-Theory Methods*, vol. 5, no. 9, pp. 843–854, Jan. 1976.

[9] J. H. Friedman, "Regularized discriminant analysis," *J. Amer. Statist. Assoc.*, vol. 84, no. 405, pp. 165–175, 1989. [Online]. Available: http://www.jstor.org/stable/2289860

[10] Y. Guo, T. Hastie, and R. Tibshirani, "Regularized linear discriminant analysis and its application in microarrays," *Biostatistics*, vol. 8, no. 1, pp. 86–100, Jan. 2007, doi: 10.1093/biostatistics/kxj035.

[11] T. V. Bandos, L. Bruzzone, and G. Camps-Valls, "Classification of hyperspectral images with regularized linear discriminant analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 3, pp. 862–873, Mar. 2009.

[12] J. Ye and T. Xiong, "Computational and theoretical analysis of null space and orthogonal linear discriminant analysis," *J. Mach. Learn. Res.*, vol. 7, pp. 1183–1204, Dec. 2006. [Online]. Available: http://dl.acm.org/citation.cfm?id=1248547.1248590

[13] J. Ye, T. Xiong, Q. Li, R. Janardan, J. Bi, V. Cherkassky, and C. Kambhamettu, "Efficient model selection for regularized linear discriminant analysis," in *Proc. 15th ACM Int. Conf. Inf. Knowl. Manage. (CIKM)*, New York, NY, USA, 2006, pp. 532–539, doi: 10.1145/1183614.1183691.

[14] 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.

[15] B. Daniyar, J. Alex, and Z. Amin, "An efficient method to estimate the optimum regularization parameter in RLDA," *Bioinformatics*, vol. 32, 22, pp. 3461–3468, 2016.

[16] 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.

[17] K. Elkhalil, A. Kammoun, R. Couillet, T. Y. Al-Naffouri, and M.-S. Alouini, "A large dimensional study of regularized discriminant analysis," *IEEE Trans. Signal Process.*, vol. 68, pp. 2464–2479, 2020.

[18] 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.

[19] H. Sifaou, A. Kammoun, and M. S. Alouini, "High-dimensional linear discriminant analysis classifier for spiked covariance model," *J. Mach. Learn. Res.*, vol. 21, no. 112, pp. 1–24, 2020.

[20] T. W. Anderson, "Classification by multivariate analysis," *Psychometrika*, vol. 16, no. 1, pp. 31–50, Mar. 1951, doi: 10.1007/BF02313425.

[21] A. N. Tikhonov, "Solution of incorrectly formulated problems and the regularization method," *Sov. Math. Doklady*, vol. 4, no. 4, pp. 1035–1038, 1963.

[22] B. B. John, "Reviewed work: Solutions of ill-posed problems by A. N. Tikhonov, V. Y. Arsenin," *Math. Comput.*, vol. 32, no. 144, pp. 1320–1322, Oct. 1963.

[23] P. C. Hansen, *Discrete Inverse Problems: Insight and Algorithms*. Philadelphia, PA, USA: SIAM, 2010.

[24] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, Feb. 1970.

[25] 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.

[26] G. Wahba, *Spline Models for Observational Data*. Philadelphia, PA, USA: SIAM, 1990.

[27] A. Aries, Z. Nashed, and V. Morozov, *Methods for Solving Incorrectly Posed Problems*. New York, NY, USA: Springer, 2012. [Online]. Available: https://books.google.com.pk/books?id=z6beBwAAQBAJ

[28] 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. [Online]. Available: http://stacks.iop.org/0266-5611/24/i=5/a=055009

[29] 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.

[30] 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. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0165168418301658

[31] 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.

[32] 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.

[33] 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.

[34] C. Zarowski, *An Introduction to Numerical Analysis for Electrical and Computer Engineers*. Hoboken, NJ, USA: Wiley, 2004. [Online]. Available: https://books.google.com.pk/books?id=3AihEG52ImkC

[35] P. C. Hansen, "Regularization tools version 4.0 for MATLAB 7.3," *Numer. Algorithms*, vol. 46, no. 2, pp. 189–194, Nov. 2007.

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

[37] T. Hastie, A. Buja, and R. Tibshirani, "Penalized discriminant analysis," *Ann. Statist.*, vol. 23, no. 1, pp. 73–102, Feb. 1995, doi: 10.1214/aos/1176324456.

[38] R. P. Gorman and T. J. Sejnowski, "Analysis of hidden units in a layered network trained to classify sonar targets," *Neural Netw.*, vol. 1, no. 1, pp. 75–89, Jan. 1988. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0893608088900238

[39] Y. Saeys, I. Inza, and P. Larranaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, Oct. 2007, doi: 10.1093/bioinformatics/btm344.

ALAM ZAIB received the B.Sc. degree (Hons.) in electrical engineering from the University of Engineering and Technology (UET), Peshawar, Pakistan, in 2002, the M.Sc. degree in electrical engineering and information technology from the Universitat Politècnica de Catalunya (UPC), Barcelona, Spain, and the Karlsruhe Institute of Technology (KIT), Germany, in 2009, and the Ph.D. degree in electrical engineering from the King Fahd University of Petroleum and Minerals (KFUPM), Dhahran, Saudi Arabia, in 2016. From 2007 to 2009, he was an Erasmus Mundus Scholar in MERIT master program. Since 2010, he has been an Assistant Professor with the Department of Electrical and Computer Engineering, COMSATS University Islamabad (CUI), Abbottabad. His research interests include statistical signal processing, channel estimation for OFDM and massive MIMO systems, blind equalization, adaptive filtering, machine learning, and artificial neural networks.

TARIG BALLAL (Member, IEEE) received the B.Sc. degree (Hons.) in electrical engineering from the University of Khartoum, Khartoum, Sudan, in 2001, the M.Sc. degree in telecommunications from the Blekinge Institute of Technology, Karlskrona, Sweden, in 2005, and the Ph.D. degree from the School of Computer Science and Informatics, University College Dublin, Dublin, Ireland, in 2011. From April 2011 to July 2012, he worked as a Research Engineer with BiancaMed Ltd. and University College Dublin. Since September 2012, he has been as a Postdoctoral Fellow with the Electrical Engineering Department, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia, where he is currently as a Research Scientist. His current research interests include regularization and robust estimation methods, image and signal processing, machine learning, acoustic sensing, and tracking and localization.

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 the Technische Universität Dresden, Germany, in 2008. Since 2002, he has been associated as a Faculty Member with COMSATS University Islamabad (Abbottabad Campus). He is currently the Vice-Chancellor of the University of Engineering and Technology Mardan, Pakistan. His research interests include wireless communications and signal processing.

TAREQ Y. AL-NAFFOURI (Senior Member, IEEE) received the B.S. degree (Hons.) in mathematics and electrical engineering from the King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia, the M.S. degree in electrical engineering from the Georgia Institute of Technology, Atlanta, in 1998, and the Ph.D. degree in electrical engineering from Stanford University, Stanford, CA, USA, in 2004. He was a Visiting Scholar with the California Institute of Technology, Pasadena, CA, in 2005 and 2006. He was a Fulbright Scholar with the University of Southern California, in 2008. He is currently a Professor with the Electrical Engineering Department, King Abdullah University of Science and Technology (KAUST). He has over 300 publications in journal and conference proceedings and 20 issued/pending patents. His research interests include sparse, adaptive, and statistical signal processing and their applications to wireless communications and localization, machine learning, and network information theory. He was a recipient of the IEEE Education Society Chapter Achievement Award, in 2008, and the Al-Marai Award for Innovative Research in Communication, in 2009. From 2013 to 2018, he was an Associate Editor of the IEEE Transactions on Signal Processing.