A Doubly Regularized Linear Discriminant Analysis Classifier with Automatic Parameter Selection

Alam Zaib*, Tarig Ballal†, Shahid Khattak* Tareq Y. Al-Naffouri†
* COMSATS University Islamabad (CUI), Abbottabad Campus, Abbottabad, Pakistan
† King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia

***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, regularized LDA (RLDA) methods have been proposed. However, the classification performance of these methods vary depending on the size of training and test data. In this paper, we propose a doubly regularized LDA classifier that we denote as R2LDA. In the proposed R2LDA approach, two regularization operations are carried out; one involving only the training data set, while the other also includes the given test data sample. The proposed R2LDA algorithm, unlike the classical RLDA techniques, caters for errors due to training data as well as the possible noise in the test data. Choosing the two regularization parameters in R2LDA can be automated through existing methods based on least squares (LS). Particularly, we show that a constrained perturbation regularization approach (COPRA) is well suited for the regularization parameter selection task needed for the proposed R2LDA classifier. Results obtained from both synthetic and real data demonstrate the consistency and effectiveness of the proposed R2LDA-COPRA classifier, especially in scenarios involving noisy test data.**

***Index Terms*—Linear discriminant analysis, LDA, regularization, covariance matrix estimation, data classification**

I. Introduction

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

The performance of LDA based classifiers depends heavily on accurate estimation of the class statistics in the form of sample covariance matrices and mean vectors. These estimates are fairly accurate 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 the limitation in the 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 [7], giving the name regularized LDA (RLDA) [8], [9], [10]. The performance of RLDA classifiers is ultimately dictated by the choice of the regularization parameter. It is essential to judiciously set the value of the regularization parameter to reap the full benefits of RLDA. Towards this end, various regularization techniques have been proposed, e.g., cross-validation [11] has been one of the classical techniques for estimating the ridge parameter as evidenced in [12], [13], [14], [5], [15]. However, the search mechanisms of these methods lead to high computational complexity. In addition, they are not based on performance optimizing criteria.

Recently, an optimal regularization method that minimizes the asymptotic classification error was derived in [16], [17]. The method is based on recent results from random matrix theory. In [18], the latter method was extended to a more general class of discriminant analysis based classifiers, with LDA obtained as a special case. Despite being elegant approaches, both [17] and [18] require a grid search mechanism to find the best value of the regularization parameter. In [19], an improved RLDA classifier is proposed which avoids the grid search but is limited to spiked-model covariance structures. It is worth mentioning here that these theoretical results strongly rely on the Gaussian assumption, and so they might not apply equally well to real data. Moreover, the performance of the above mentioned approaches deteriorates significantly when the test data is contaminated with noise that is not observed during the training stage.

Focusing on binary classification, this paper presents a doubly regularized LDA (R2LDA) classifier by expressing the LDA score function as an inner product of two vectors which are linearly related to the mean vectors and the data covariance matrices. These vectors are estimated by using a perturbation regularization approach [20] where the regularization parameters can be selected to be optimal in the mean-squared-error (MSE) sense. The proposed method takes care of the ill-conditioning of the sample covariance matrices and the uncertainties in the training or the test data. In addition, the proposed method has the following distinctive features:

- Two regularization parameters are calculated based on both the training and test data. These parameters can be tuned independently to cope with the different perturbations including those 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 [17], [18], [19], which rely on the Gaussian assumption.

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 a prior probability that $\mathbf{x}$ belongs to a class $\mathcal{C}_i$ and assume that the class conditional densities $\mathbb{P}\left(\mathbf{x}|\mathbf{x}\in\mathcal{C}_i\right), i=0,1$ are Gaussian with mean vectors $\mathbf{m}_i \in \mathbb{R}^{p\times 1}$ and non-negative covariance matrices $\mathbf{\Sigma}_i \in \mathbb{R}^{p\times p}$.

LDA employs the Bayesian discriminant rule, which assigns $\mathbf{x}$ to the class with 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 classes $\mathcal{C}_0$ and $\mathcal{C}_1$, respectively, where n_i is the number of samples in class $\mathcal{C}_i$ and $n = n_0 + n_1$ is the total number of training samples. The LDA score function reads [21]

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

where $(.)^{\mathrm{T}}$ is the matrix transpose operation. The unbiased mean vector estimates $\hat{\mathbf{m}}_i$ and the pooled sample covariance matrix $\hat{\mathbf{\Sigma}}$ are given by

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

where the sample covariance matrices $\hat{\mathbf{\Sigma}}_i$ are defined as

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

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 covariance matrix $\hat{\mathbf{\Sigma}}$. In many practical setups where n is comparable to p, $\hat{\mathbf{\Sigma}}$ becomes ill conditioned, or even singular. To get around this issue, $\hat{\mathbf{\Sigma}}^{-1}$ in (1) is replaced with a regularized matrix $\mathbf{H} = (\mathbf{I} + \gamma\hat{\mathbf{\Sigma}})^{-1}$, where $\gamma \in \mathbb{R}^{+}$ and $\mathbf{I}$ is the identity matrix of dimension p. This replacement results in the RLDA score function [17], [16]

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

In this work, we employ a different form of regularization to that in (1). In the proposed regularized LDA classifier, we apply two separate regularization operations, which help in accounting for errors in the training data and providing robustness against error contributions that are present in the test data.

III. The proposed R2LDA classifier

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, especially when the error statistics of the test data deviate from those of the training data. To address this issue, we reformulate the LDA score function (1) as

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

where $\mathbf{x}' := \mathbf{x} - \frac{1}{2}\hat{\mathbf{m}}^{+}$, $\hat{\mathbf{m}}^{+} := \hat{\mathbf{m}}_0 + \hat{\mathbf{m}}_1$, $\hat{\mathbf{m}}^{-} := \hat{\mathbf{m}}_0 - \hat{\mathbf{m}}_1$, $\mathbf{z} := \hat{\mathbf{\Sigma}}^{-\frac{1}{2}}\mathbf{x}'$, and $\mathbf{b} := \hat{\mathbf{\Sigma}}^{-\frac{1}{2}}\hat{\mathbf{m}}^{-}$. 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 in (6). To this end, we utilize the linear models

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

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

where $\mathbf{v}_x$ and $\mathbf{v}_m$ are additive noise vectors. Each of (7) and (8) can be represented by the model

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

where $\mathbf{v}$ represents model noise. To simplify our derivations, we make the following assumptions:

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

In Section V, we will see that these simplifying assumptions still work for different classification examples.

Focusing on (9), regularization methods, commonly named ridge regression or Tikhonov regularization [22], [23], [24], can be applied to obtain a stabilized estimate of $\mathbf{c}$. This estimate can be expressed in a closed form as [10]

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

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

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

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

Now, it only remains to set the values of the regularization parameters γ_z and γ_b. In the following section, we present a robust method to compute the regularization parameter for the regularized least-squares (RLS) solution in (10).

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 $\mathbf{x}'$. In addition, $\mathbf{x}'$ also has errors coming from the test data. By carrying out two independent estimations to obtain regularized estimates of $\mathbf{z}$ and $\mathbf{b}$ in (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 the key advantage of proposed R2LDA method over the classical RLDA based on (5) which uses a single regularization operation that involves only the training data. It will also become clearer that the proposed R2LDA still uses the statistics from the training data only, which is fundamental requirement of any machine learning algorithm.

IV. Regularization Parameter Selection

Several methods have been proposed in the literature for selecting the regularization parameter γ required in (10), e.g., the L-curve [25], the generalized cross-validation (GCV) [26], and the quasi-optimal method [27], [28], to mention a few. These methods use different criteria which results in different values of the regularization parameter (see [29]).

In this work, we adopt the constrained perturbation regularization approach (COPRA) [20], which allows for regularization parameter selection in a way that optimizes the MSE. We adapt this algorithm to the setting of the problem in hand. COPRA works by introducing an artificial perturbation in the linear model to improve the singular-value structure of the resulting model matrix $\hat{\boldsymbol{\Sigma}}$, and hence, is well suited to the naturally perturbed model in hand. To proceed, we start by replacing $\hat{\boldsymbol{\Sigma}}^{\frac{1}{2}}$ in (9) by a perturbed version to obtain the model

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

where $\boldsymbol{\Delta} \in \mathbb{C}^{m\times n}$ is an unknown perturbation matrix which is norm bounded by a positive number λ, i.e., $\|\boldsymbol{\Delta}\|_2 \leq \lambda$. One can consider Δ to be a way to perturb $\hat{\boldsymbol{\Sigma}}^{\frac{1}{2}}$ to make the solution of (12) stable [20]. The perturbation Δ can also be thought of as a genuine error in the model due to the noisy nature of $\hat{\boldsymbol{\Sigma}}^{\frac{1}{2}}$, which is the case for (9). To obtain an estimate of $\mathbf{c}$, we consider the minimization of the worst-case residual error. Namely, we pursue the following optimization:

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

Interestingly, as shown in [30], [20], [31], the min-max problem (13) can be converted to a minimization problem whose solution is given by (10) with the constraint

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

We observe that the solution of (13) depends on the bound λ (in addition to the other linear system parameters) and is agnostic to the structure of the perturbation matrix $\boldsymbol{\Delta}$. Note that both λ and $\hat{\mathbf{c}}$ are unknown. However, we can substitute (10) and the EVD of $\hat{\boldsymbol{\Sigma}}$ in (14) and manipulate to obtain

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

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

$$\mathbb{E}\left(\mathbf{y}\mathbf{y}^{\mathrm{H}}\right) = \mathbf{U}\mathbf{D}\mathbf{U}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{cc}}\mathbf{U}\mathbf{D}\mathbf{U}^{\mathrm{T}} + \sigma_v^2\mathbf{I}. \tag{16}$$

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

$$\hat{\boldsymbol{\Sigma}} = [\mathbf{U}_1\ \ \mathbf{U}_2]\begin{bmatrix}\mathbf{D}_1^2 & \mathbf{0}\\ \mathbf{0} & \mathbf{D}_2^2\end{bmatrix}\begin{bmatrix}\mathbf{U}_1^{\mathrm{T}}\\ \mathbf{U}_2^{\mathrm{T}}\end{bmatrix} \simeq \mathbf{U}_1\mathbf{D}_1^2\mathbf{U}_1^{\mathrm{T}}, \tag{17}$$

where $\mathbf{D}_1$ and $\mathbf{D}_2$ are diagonal matrices containing n_s most significant and n–n_s least significant eigenvalues, respectively. This partitioning is introduced as a general case form. For the special case where all eigenvalues are significant, we set $n_s = n$ and no partitioning is required. A threshold based approach to find the point of this partitioning is recommended in [20]. However, a simple and intuitive rule is used here to determine the value of n_s as the smaller value of p and n, i.e., $n_s = \min(n, p)$. The rationale behind (17) will be explained subsequently (see Remark 2).

Now, we substitute (16) and (17) in (15) and manipulate to obtain (18, shown on the top of the next page). Next, we proceed by eliminating the dependency of λ on the unknowns σ_v^2 and $\boldsymbol{\Sigma}_{\mathbf{cc}}$ in (18) by using the MSE criterion. The MSE of the RLS estimator (10) can be written as [10]

$$\begin{aligned}\text{MSE} = \text{tr}\left(\mathbb{E}\left\{(\mathbf{c}-\hat{\mathbf{c}})(\mathbf{c}-\hat{\mathbf{c}})^{\mathrm{H}}\right\}\right) &= \sigma_v^2\text{tr}\left(\mathbf{D}^2\left(\mathbf{D}^2+\gamma\mathbf{I}\right)^{-2}\right)\\ &+ \gamma^2\text{tr}\left(\left(\mathbf{D}^2+\gamma\mathbf{I}\right)^{-2}\mathbf{U}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{cc}}\mathbf{U}\right).\end{aligned} \tag{19}$$

By differentiating (19), the regularization parameter γ that minimizes the MSE, is given by

$$\frac{\partial\,(\text{MSE})}{\partial\ \gamma} = 0 \implies \gamma \simeq \frac{n\sigma_v^2}{\text{tr}\left(\boldsymbol{\Sigma}_{\mathbf{cc}}\right)}. \tag{20}$$

By substituting (20) in (18), we obtain (21, shown on the top of the next page), which shows a bound λ that does not depend on the statistics of $\mathbf{c}$ or that of the noise. Note that the derivation of (16)–(18) is largely based on the Assumptions 1–3. Ultimately, by using (21), we can eliminate λ from (15) to obtain (22), where $\mathbf{d} = \mathbf{U}^{\mathrm{T}}\mathbf{y}$ and $\beta = n/n_s$. Equation (22), which is non-linear in γ, can be solved by using Newton's Method [32] 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 [20].

Remark 2: Equation (22) is based on the contribution of only the significant eigenvalues of $\hat{\boldsymbol{\Sigma}}$ which occupy the diagonal of the matrix $\mathbf{D}_1^2$. In this case, if Newton's method iterations start from a small initial value of γ, the (diagonal) matrix inversion operation required to compute the right-hand side of (22) will be numerically stable since the diagonal elements of $\mathbf{D}_1^2$ are not overly small. This highlights the benefit of partitioning and truncation of the insignificant eigenvalues in (17).

A. Summary of the Proposed R2LDA-COPRA Algorithm

The main steps involved in the proposed R2LDA algorithm based on COPRA are summarized as follows:

1. *Estimate the class statistics;* $\hat{\mathbf{m}}_i$, $\hat{\boldsymbol{\Sigma}}_i$ *and* $\hat{\boldsymbol{\Sigma}}$ *based on the training data by using (2) and (3).*
2. *Compute* $\hat{\mathbf{m}}^+$, $\hat{\mathbf{m}}^-$ *and the EVD of* $\hat{\boldsymbol{\Sigma}}$ *to determine* $\mathbf{D}_1$ *and* $\mathbf{U}_1$ *corresponding to the* n_s *most significant eigenvalues.*

$$\lambda^2 \left(\mathrm{tr}\left(\left(\mathbf{D}_1^2+\gamma\mathbf{I}\right)^{-2} \left(\mathbf{D}_1^2+\frac{n_s\sigma_v^2}{\mathrm{tr}\left(\boldsymbol{\Sigma}_{\mathbf{cc}}\right)}\mathbf{I}\right)\right) + \frac{(n-n_s)n_s\sigma_v^2}{\gamma^2 \mathrm{tr}\left(\boldsymbol{\Sigma}_{\mathbf{cc}}\right)} \right) \simeq \mathrm{tr}\left(\mathbf{D}_1^2 \left(\mathbf{D}_1^2+\gamma\mathbf{I}\right)^{-2} \left(\mathbf{D}_1^2+\frac{n_s\sigma_v^2}{\mathrm{tr}\left(\boldsymbol{\Sigma}_{\mathbf{cc}}\right)}\mathbf{I}\right)\right) \tag{18}$$

$$\lambda^2 \left(\mathrm{tr}\left(\left(\mathbf{D}_1^2+\gamma\mathbf{I}\right)^{-2} \left(\frac{n}{n_s}\mathbf{D}_1^2+\gamma\mathbf{I}\right)\right) + \frac{(n-n_s)}{\gamma} \right) \simeq \mathrm{tr}\left(\mathbf{D}_1^2 \left(\mathbf{D}_1^2+\gamma\mathbf{I}\right)^{-2} \left(\frac{n}{n_s}\mathbf{D}_1^2+\gamma\mathbf{I}\right)\right) \tag{21}$$

$$\begin{aligned} &\mathrm{tr}\left(\mathbf{D}_1^2 \left(\mathbf{D}_1^2+\gamma\mathbf{I}\right)^{-2} \mathbf{d}\mathbf{d}^{\mathrm{H}}\right) \mathrm{tr}\left(\left(\mathbf{D}_1^2+\gamma\mathbf{I}\right)^{-2} \left(\beta\mathbf{D}_1^2+\gamma\mathbf{I}\right)\right) + \frac{(n-n_s)}{\gamma} \mathrm{tr}\left(\mathbf{D}_1^2 \left(\mathbf{D}_1^2+\gamma\mathbf{I}\right)^{-2} \mathbf{d}\mathbf{d}^{\mathrm{H}}\right) \\ &\qquad - \mathrm{tr}\left(\left(\mathbf{D}_1^2+\gamma\mathbf{I}\right)^{-2} \mathbf{d}\mathbf{d}^{\mathrm{H}}\right) \mathrm{tr}\left(\mathbf{D}_1^2 \left(\mathbf{D}_1^2+\gamma\mathbf{I}\right)^{-2} \left(\beta\mathbf{D}_1^2+\gamma\mathbf{I}\right)\right) = 0 \end{aligned} \tag{22}$$

3. *Set* $\mathbf{y} = \hat{\mathbf{m}}^-$ *in (22) and solve using Newton's method to obtain* γ_b.
4. *For a given test sample, compute* $\mathbf{x}'$. *Then repeat step* 3 *by setting* $\mathbf{y} = \mathbf{x}'$ *to obtain* γ_z.
5. *Compute the R2LDA score function given in (11) and classify the given test sample according to (4).*

By using (19), COPRA guarantees that we obtain the best (in terms of MSE) regularized estimates of $\mathbf{z}$ and $\mathbf{b}$ required to form our R2LDA score function in (11). This does not guarantee optimal classification performance based on (11). However, our results show that the proposed R2LDA algorithm still outperforms classical RLDA classifiers of the form given by (5). It is worth mentioning here that COPRA can be replaced with other regularization methods to compute the regularization parameters γ_b and γ_z. Further, the proposed R2LDA algorithm uses only the statistics from the training data (step 1). The computations in step 4 and step 5 use the given test sample and not the test data or the noise statistics.

V. Results

We demonstrate the performance of the proposed R2LDA classification against the RLDA techniques of the asymptotic error estimator (Asym)[17] and the improved error estimator (Impr)[19]. We also consider GCV [26] and bounded perturbation regularization (BPR) [33] as alternatives to COPRA in selecting the two regularization parameters of the R2LDA classifier. We consider both synthetic and real data for performance comparison.

The synthetic data was generated using a Gaussian data model with class covariance matrices and mean vectors defined as: $\boldsymbol{\Sigma}_0$, which is of dimensionality $p \times p = 100 \times 100$ and has 1 on the main diagonal and 0.1 as off-diagonal elements; $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_0 + \mathbf{I}$ and $\mathbf{m}_1 = -\mathbf{m}_0$, where $\mathbf{m}_0 = [a, a, ..., a]^{\mathrm{T}}$. The parameter a was chosen according to Mahalanobis distance (δ) between classes defined as, $\delta^2 = (\mathbf{m}_0 - \mathbf{m}_1)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{m}_0 - \mathbf{m}_1)$ [17]. We set $\delta^2 = 9$. A training set $\mathcal{S}_i$ of size n_i for the class i was generated in each trial. We set $n_0 = n_1$. For the test data, we generated an independent set of samples for each class. A total of 500 training trials were carried out, each followed by 500 test trials.

For real data, we use the MNIST dataset of 20×20 grayscale images of handwritten digits [34], and the phonemes dataset considered in [35]. The later 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) [35], which is widely used in speech recognition. The MNIST images are vectorized to result in data of dimensionally equal to $p = 400$. For binary classification, selected pairs of images were used. On the other hand, we used only two phonemes transcribed as: "sh"as in "she"and "dcl"as in "dark", for binary classification. Real data results were obtained from 100 training attempts. In each attempt, the training samples were chosen randomly from the dataset. Each trained model was tested using 500 examples, which were also randomly selected from the dataset.

For both the synthetic and real datasets, zero-mean Gaussian noise with standard deviation σ_n was added during the *test* phase. The properties of the noise were not known by the proposed R2LDA classifier, nor were they known by any of the benchmarks.

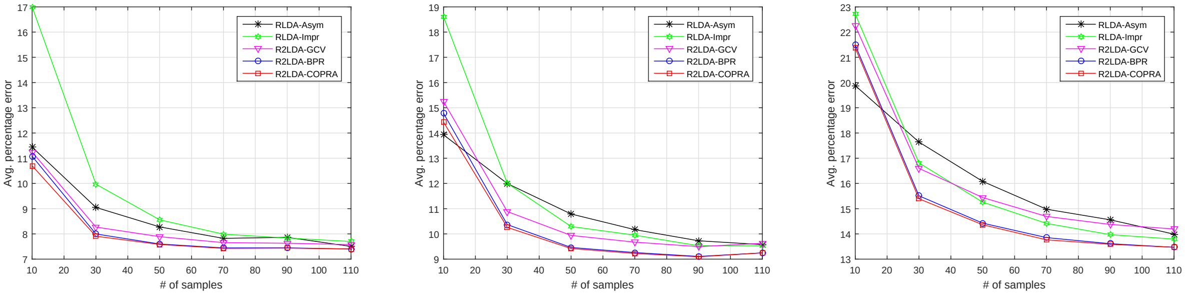
A. Discussion

Figs.1–3 shows classification error versus the size of the training data n for different scenarios. Fig.1 presents the results for the (synthetic) Gaussian data, while Fig.2 and Fig.3 present results for the MNIST and phonemes datasets, respectively. The MNIST results are based on images/digits pair examples of (1,7), (5,8) and (7,9). From these results in Figs.1–3, we observe the following:

- On average, the R2LDA methods outperform the RLDA methods.
- The R2LDA remains more consistent and stable than the RLDA methods as the level of noise in the test data increases. This is more visible with the MNIST and phonemes datasets.
- Amongst the R2LDA classifiers, R2LDA-COPRA is the most consistent. R2LDA-GCV and R2LDA-BPR falters occasionally as in Fig.2(a) and Fig.2(g).

VI. Conclusions

We presented a novel regularized LDA classifier based on a dual regularization approach to provide robustness against both training and test data perturbations. In the proposed classifier, the regularization parameters are obtained by solving a nonlinear equation using Newton's method. Results based on both synthetic and real data demonstrate the effectiveness of our method, especially when noise is present in the test data. Although the proposed method is presented for binary classification, it can be easily extended to multi-class problems.

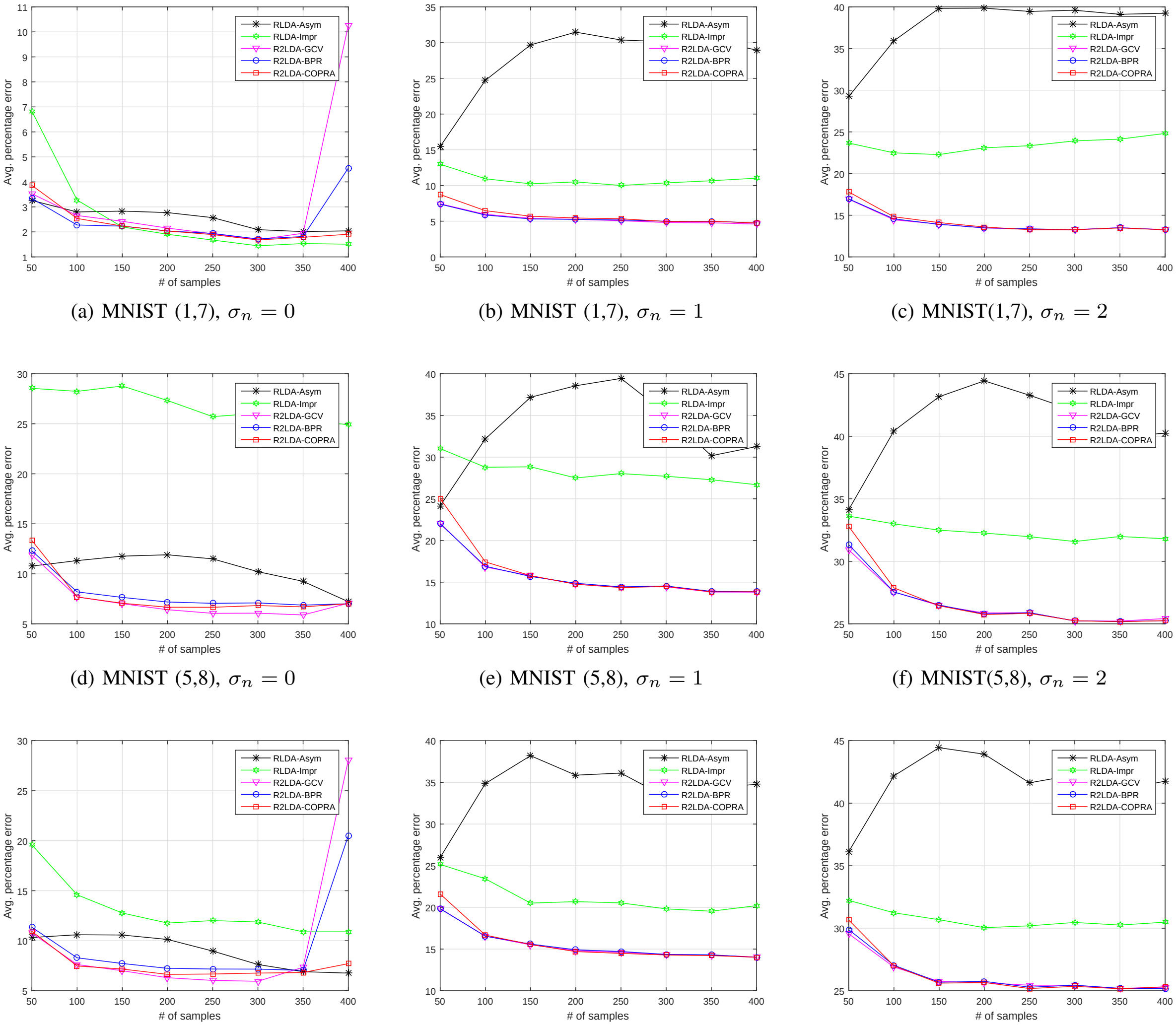


(a) Gaussian, $\sigma_n = 0$

(b) Gaussian, $\sigma_n = 1$

(c) Gaussian, $\sigma_n = 2$

Fig. 1: Gaussian data Misclassification rate versus training data size for different test data noise levels.



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

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

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

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

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

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

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

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

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

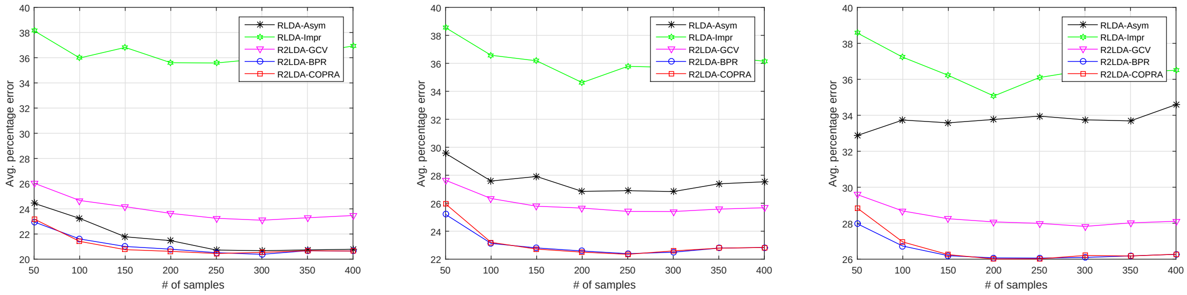
Fig. 2: MNIST data Misclassification rate versus training data size for different test data noise levels.

REFERENCES

[1] R. A. Fisher, “The use of multiple measurements in taxonomic problems,” *Annals of Eugenics*, vol. 7, no. 7, pp. 179–188, 1936.

[2] K. R. Varshney, “Generalization error of linear discriminant analysis in spatially-correlated sensor networks,” *IEEE Transactions on Signal Processing*, vol. 60, no. 6, pp. 3295–3301, June 2012.

[3] C. Avendano, S. Van Vuuren, and H. Hermansky, “Data based filter design for rasta-like channel normalization in asr,” in *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*,



(a) Phonemes, $\sigma_n = 0$

(b) Phonemes, $\sigma_n = 1$

(c) Phonemes, $\sigma_n = 2$

Fig. 3: Phoneme data Misclassification rate versus training data size for different test data noise levels.

vol. 4, Oct 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," 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," vol. 28, pp. 1–8, 2009.

[6] D. L. Swets and J. J. Weng, "Using discriminant eigenfeatures for image retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 8, pp. 831–836, Aug 1996.

[7] P. J. D. Pillo, "The application of bias to discriminant analysis," *Communications in Statistics - Theory and Methods*, vol. 5, no. 9, pp. 843–854, 1976.

[8] A. E. Hoerl, "Application of ridge analysis to regression problems," *Chemical Engineering Progress*, vol. 58, no. 3, pp. 54–59, 1962.

[9] A. E. Hoerl and R. W. Kennard, "Ridge regression: Applications to nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 69–82, Feb 1970.

[10] ——, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, pp. 55–67, 1970.

[11] J. H. Friedman, "Regularized discriminant analysis," *Journal of the American Statistical Association*, vol. 84, no. 405, pp. 165–175, 1989. [Online]. Available: http://www.jstor.org/stable/2289860

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

[13] T. V. Bandos, L. Bruzzone, and G. Camps-Valls, "Classification of hyperspectral images with regularized linear discriminant analysis," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 47, no. 3, pp. 862–873, March 2009.

[14] 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

[15] J. Ye, T. Xiong, Q. Li, R. Janardan, J. Bi, V. Cherkassky, and C. Kambhamettu, "Efficient model selection for regularized linear discriminant analysis," in *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, ser. CIKM '06. New York, NY, USA: ACM, 2006, pp. 532–539. [Online]. Available: http://doi.acm.org/10.1145/1183614.1183691

[16] A. Zollanvari and E. R. Dougherty, "Generalized consistent error estimator of linear discriminant analysis," *IEEE Transactions on Signal Processing*, vol. 63, no. 11, pp. 2804–2814, June 2015.

[17] 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.

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

[19] H. Sifaou, A. Kammoun, and M.-S. Alouini, "Improved LDA classifier based on spiked models," in *2018 IEEE 19th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, June 2018.

[20] M. A. Suliman, T. Ballal, and T. Y. Al-Naffouri, "Perturbation-based regularization for signal estimation in linear discrete ill-posed problems," *Signal Processing*, vol. 152, pp. 35–46, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0165168418301658

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

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

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

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

[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. [Online]. Available: http://dx.doi.org/10.1137/0914086

[26] G. Wahba, *Spline Models for Observational Data*. Philadelphia: Society for Industrial and Applied Mathematics, 1990.

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

[28] F. Bauer and M. Rei, "Regularization independent of the noise level: an analysis of quasi-optimality," *Inverse Problems*, vol. 24, no. 5, p. 055009, 2008. [Online]. Available: http://stacks.iop.org/0266-5611/24/i=5/a=055009

[29] F. Bauer and M. A. Lukas, "Original article: Comparing parameter choice methods for regularization of ill-posed problems," *Math. Comput. Simul.*, vol. 81, no. 9, pp. 1795–1841, May 2011. [Online]. Available: http://dx.doi.org/10.1016/j.matcom.2011.01.016

[30] S. Chandrasekaran, G. H. Golub, M. Gu, and A. H. Sayed, "Parameter estimation in the presence of bounded data uncertainties," *SIAM J. Matrix Analysis and Applications*, vol. 19, pp. 235–252, Jan 1998. [Online]. Available: https://doi.org/10.1137/S0895479896301674

[31] T. Ballal and T. Y. Al-Naffouri, "Improved linear least squares estimation using bounded data uncertainty," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICAAS)*, April 2015, pp. 3427–3431.

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

[33] T. Ballal, M. A. Suliman, and T. Y. Al-Naffouri, "Bounded perturbation regularization for linear least squares estimation," *IEEE Access*, vol. 5, pp. 27 551–27 562, 2017.

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

[35] T. Hastie, A. Buja, and R. Tibshirani, "Penalized discriminant analysis," *Ann. Statist.*, vol. 23, no. 1, pp. 73–102, 02 1995. [Online]. Available: https://doi.org/10.1214/aos/1176324456