A Deterministic Approach to Regularized Linear Discriminant Analysis

Alok Sharma[1,2], Kuldip K. Paliwal[1]
[1]School of Engineering, Griffith University, Australia
[2]School of Engineering and Physics, University of the South Pacific, Fiji

Abstract

The regularized linear discriminant analysis (RLDA) technique is one of the popular methods for dimensionality reduction used for small sample size problems. In this technique, regularization parameter is conventionally computed using a cross-validation procedure. In this paper, we propose a deterministic way of computing the regularization parameter in RLDA for small sample size problem. The computational cost of the proposed deterministic RLDA is significantly less than the cross-validation based RLDA technique. The deterministic RLDA technique is also compared with other popular techniques on a number of datasets and favorable results are obtained.

1. Introduction

Linear discriminant analysis (LDA) is a popular technique for dimensionality reduction and feature extraction. Dimensionality reduction is a pre-requisite for many statistical pattern recognition techniques. It is primarily applied for improving generalization capability and reducing computational complexity of a classifier. In LDA the

dimensionality is reduced from d-dimensional space to h-dimensional space (where $h < d$) by using a transformation $\mathbf{W} \in \mathbb{R}^{d \times h}$. The transformation (or orientation) matrix $\mathbf{W}$ is found by maximizing the Fisher's criterion: $J(\mathbf{W}) = \left|\mathbf{W}^{\mathrm{T}}\mathbf{S}_B\mathbf{W}\right|/|\mathbf{W}^{\mathrm{T}}\mathbf{S}_W\mathbf{W}|$, where $\mathbf{S}_W \in \mathbb{R}^{d \times d}$ is within-class scatter matrix and $\mathbf{S}_B \in \mathbb{R}^{d \times d}$ is between-class scatter matrix. Under this criterion, the transformation of feature vectors from higher dimensional space to lower dimensional space is done in such a manner that the between-class scatter in the lower dimensional space is maximized and within-class scatter is minimized. The orientation matrix $\mathbf{W}$ is computed by the eigenvalue decomposition (EVD) of $\mathbf{S}_W^{-1}\mathbf{S}_B$ [1].

In many pattern classification applications, the matrix $\mathbf{S}_W$ becomes singular and its inverse computation becomes impossible. This is due to the large dimensionality of feature vectors compared to small number of vectors available for training. This is known as small sample size (SSS) problem [2]. There exist several techniques that can overcome this problem [3]-[11],[19]-[34]. Among these techniques, regularized LDA (RLDA) technique [3] is one of the pioneering methods for solving SSS problem. The RLDA technique has been widely studied in the literature [12]-[14]. It has been applied in areas like face recognition [13],[14] and bioinformatics [15].

.

In the RLDA technique, the $\mathbf{S}_W$ matrix is regularized to overcome the singularity problem of $\mathbf{S}_W$. This regularization can be done in various ways. For example, Zhao et al. [12][16][17] have done this by adding a small positive constant α (known as regularization parameter) to the diagonal elements of matrix $\mathbf{S}_W$; i.e., the matrix $\mathbf{S}_W$ is approximated by $\mathbf{S}_W + \alpha\mathbf{I}$ and the orientation matrix is computed by EVD of $(\mathbf{S}_W + \alpha\mathbf{I})^{-1}\mathbf{S}_B$. The performance of RLDA technique depends on the choice of the regularization parameter α. In the past studies [18], this parameter is chosen rather heuristically, for example, by applying cross-validation procedure on the training data. In the cross-validation based RLDA technique (denoted here as CV-RLDA), the training data is divided into two subsets: training subset and validation subset. The cross-validation procedure searches over a finite range of α values and finds an α value in this range that maximizes the classification accuracy over the validation subset. In the cross-validation procedure, the estimate of α depends on the range over which it is explored. For a given dataset, its classification accuracy can vary depending upon the range of α being explored. Since many values of α have to be searched in this range, the computational cost of this procedure is quite high. In addition, the cross-validation procedure used in the CV-RLDA technique is biased towards the classifier used.

In order to address these drawbacks of CV-RLDA technique, we explore a deterministic way for finding the regularization parameter α. This would provide a unique value of the regularization parameter on a given training data. We call this approach as the deterministic RLDA (DRLDA) technique. This technique avoids the use of the heuristic (cross-validation) procedure for parameter estimation and improves the computational efficiency. We show that this deterministic approach computes the regularization parameter by maximizing the Fisher's criterion and its classification performance is quite promising compared to other LDA techniques.

2. Related work

In a SSS problem, the within-class scatter matrix $\mathbf{S}_W$ becomes singular and its inverse computation becomes impossible. In order to overcome this problem, generally inverse computation of $\mathbf{S}_W$ is avoided or approximated for the computation of orientation matrix $\mathbf{W}$. There are several techniques that can overcome this SSS problem. One way to solve this problem is by estimating the inverse of $\mathbf{S}_W$ by its pseudoinverse and then the conventional eigenvalue problem can be solved to compute the orientation matrix $\mathbf{W}$. This was the basis of Pseudoinverse LDA (PILDA) technique [20]. Some improvements of PILDA have also been presented in [28, 31]. In Fisherface (PCA+LDA) technique,

d-dimensional features are firstly reduced to h-dimensional feature space by the application of PCA [2][52][53] and then LDA is applied to further reduce features to k dimensions. There are several ways for determining the value of h in PCA+LDA technique [4][5]. In the Direct LDA (DLDA) technique [7], the dimensionality is reduced in two stages. In the first stage, a transformation matrix is computed to transform the training samples to the range space of $\mathbf{S}_B$, and in the second stage, the dimensionality of this transformed samples is further transformed by some regulating matrices. The Improved DLDA technique [11], addresses drawbacks of DLDA technique. In the improved DLDA technique, first $\mathbf{S}_W$ is decomposed into its eigenvalues and eigenvectors instead of $\mathbf{S}_B$ matrix as of DLDA technique. Here, both its null space and range space information are utilized by approximating $\mathbf{S}_W$ by a well deterministic substitution. Then $\mathbf{S}_B$ is diagonalized using regulating matrices. For the Null LDA (NLDA) technique [6], the orientation $\mathbf{W}$ is computed in two stages. In the first stage, the data is projected on the null space of $\mathbf{S}_W$ and in the second stage it finds $\mathbf{W}$ that maximizes $|\mathbf{W}^{\mathrm{T}}\mathbf{S}_B\mathbf{W}|$. In orthogonal LDA (OLDA) technique [8], the orientation matrix $\mathbf{W}$ is obtained by simultaneously diagonalizing scatter matrices. It has shown that OLDA is equivalent to NLDA under a mild condition [8]. The Uncorrelated LDA (ULDA) technique [21], is a slight variation of OLDA technique. In ULDA, the orientation

matrix $\mathbf{W}$ is made uncorrelated. The fast NLDA (FNLDA) technique [25], is an alternative method of NLDA. In this technique, the orientation matrix is obtained by using the relation $\mathbf{W} = \mathbf{S}_T^{+}\mathbf{S}_B\mathbf{Y}$, where $\mathbf{Y}$ is a random matrix of rank $c-1$, and c is the number of classes. This technique is so far the fastest technique of performing null LDA operation. In extrapolation LDA (ELDA) technique [32], the null space of $\mathbf{S}_W$ matrix is regularized by extrapolating eigenvalues of $\mathbf{S}_W$ using exponential fitting function. This technique utilizes range space information and null space information of $\mathbf{S}_W$ matrix. The two stage LDA (TSLDA) technique [34], exploits all four informative spaces of scatter matrices. These spaces are included in two separate discriminant analyses in parallel. In the first analysis, null space of $\mathbf{S}_W$ and range space of $\mathbf{S}_B$ are retained. In the second analysis, range space of $\mathbf{S}_W$ and null space of $\mathbf{S}_B$ are retained. In eigenfeature regularization (EFR) technique [10], $\mathbf{S}_W$ is regularized by extrapolating its eigenvalues in its null space. The lagging eigenvalues of $\mathbf{S}_W$ is considered as noisy or unreliable which are replaced by an estimation function. The General Tensor Discriminant Analysis (GTDA) technique [48] has been developed for image recognition problems. This work focuses on the representation and pre-processing of appearance-based models for human gait sequences. Two models were presented: gabor gait and tensor gait. In [49], authors proposed a constrained empirical risk

minimization framework for distance metric learning (DML) to solve SSS problem. In Double Shrinking Sparse Dimension Reduction technique [50], the SSS problem is solved by penalizing the parameter space. A detailed explanation regarding LDA is given in [51] and an overview regarding SSS based LDA techniques is given in [47]. There are other techniques which can solve SSS problem and applied in various fields of research [54]-[62]. In this paper, we focus on regularize LDA (RLDA) technique. This technique overcomes SSS problem by utilizing a small perturbation to the $\mathbf{S}_W$ matrix. The details of RLDA have been discussed in the next section.

3. Regularized linear discriminant techniques for SSS problem

In the RLDA technique, the within-class scatter matrix $\mathbf{S}_W$ is approximated by adding a regularization parameter to make it a non-singular matrix [3]. There are, however, different ways to perform regularization (see for details, [3][12]-[14][16][17][30][33]). In this paper we adopted Zhao's model [12][16][17] to approximate $\mathbf{S}_W$ by adding a positive constant in the following way $\hat{\mathbf{S}}_W = \mathbf{S}_W + \alpha\mathbf{I}$[1]. This will make within-class scatter matrix a non-singular matrix and then its inverse computation would be possible. The RLDA technique computes the orientation matrix $\mathbf{W}$ by EVD of $\hat{\mathbf{S}}_W^{-1}\mathbf{S}_B$.

[1] In the Friedman's model [3], $\mathbf{S}_W$ is estimated as $\hat{\mathbf{S}}_W = (1-\alpha)\mathbf{S}_W + \alpha\mathbf{I}$. We have compared Zhao's model and Friedman's model of CV-RLDA and found that Zhao's model exhibits comparatively better generalization capability (see Appendix-I for details). Furthermore, we have considered Zhao's model because it is relatively simpler for establishing deterministic approach of computing α (in DRLDA).

Thus, this technique uses null space of $\mathbf{S}_W$, range space of $\mathbf{S}_W$ and range space of $\mathbf{S}_B$ in one step (i.e., simultaneously).

In the RLDA technique, a fixed value of regularization parameter can be used, but it may not give the best classification performance as shown in Appendix-II. Therefore, the regularization parameter α is normally computed by the cross-validation procedure. The cross-validation procedure (e.g. leave-one-out or k-fold) employs a particular classifier to estimate α and is conducted on the training set (which is different from the test set). We briefly describe below the leave-one out cross-validation procedure used in the CV-RLDA technique. Let $[a, b]$ be the range of α to be explored and α_0 be any value in this range. Consider a case when n training samples are available. The training set is first subdivided into training subset (consisting of $n-1$ samples) and validation subset (consisting of 1 sample). For this particular subdivision of training set, the following operations are required: 1) computation of scatter matrices $\mathbf{S}_B$, $\mathbf{S}_W$ and $\hat{\mathbf{S}}_W = \mathbf{S}_W + \alpha_0\mathbf{I}$ for $n-1$ samples in the training subset; 2) EVD of $\hat{\mathbf{S}}_W^{-1}\mathbf{S}_B$ to compute orientation matrix $\mathbf{W}$; and 3) classification of the left out sample (from the validation subset) by the classifier to obtain the classification accuracy. These computational operations are carried out for $n-1$ subdivisions of the training set and the average

classification accuracy over the $n-1$ runs is computed. This average classification accuracy is obtained for a particular value of α (namely α_0). All the above operations will be repeated for other values of α in the range $[a, b]$ to get the highest average classification accuracy. From this description, it is obvious that the cross-validation procedure used in the CV-RLDA technique has the following drawbacks:

- Since the cross-validation procedure repeats the above-mentioned computational operations many times for different values of α, its computation complexity is extremely large.
- Since the cross-validation procedure used in the CV-RLDA technique searches the α parameter over a finite range $[a, b]$, it may not estimate its optimum value. In order to estimate its optimum value, one has to investigate all possible values of α in the range of $(0, \infty)$. However, it is an impossible task (as it will take infinite amount of computation time). Thus, the α value computed by this procedure depends on two factors: 1) the range over which it is searched, and 2) the fineness of the search procedure.
- The cross-validation procedure estimates the regularization parameter in CV-RLDA for a particular classifier. Thus, the estimated value is specific to the classifier and cannot be generalized to other classifiers.

In our proposed DRLDA technique, we use a deterministic approach to estimate α parameter by maximizing the modified Fisher's criterion. The proposed technique is described in the next section.

4. DRLDA technique

4.1 Notations

Let $\mathfrak{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ denotes n training samples (or feature vectors) in a d-dimensional space having class labels $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$, where $\omega \in \{1,2, \ldots, c\}$ and c is the number of classes. The set $\mathfrak{X}$ can be subdivided into c subsets $\mathfrak{X}_1$, $\mathfrak{X}_2$,…, $\mathfrak{X}_c$ where $\mathfrak{X}_j$ belongs to class j and consists of n_j number of samples such that:

$$n = \sum_{j=1}^{c} n_j$$

and $\mathfrak{X}_j \subset \mathfrak{X}$ and $\mathfrak{X}_1 \cup \mathfrak{X}_2 \cup \ldots \cup \mathfrak{X}_c = \mathfrak{X}$.

If $\boldsymbol{\mu}_j$ is the centroid of $\mathfrak{X}_j$ and $\boldsymbol{\mu}$ is the centroid of $\mathfrak{X}$, then the total scatter matrix $\mathbf{S}_T$, within-class scatter matrix $\mathbf{S}_W$ and between-class scatter matrix $\mathbf{S}_B$ are defined as [1][35][36]

$$\mathbf{S}_T = \sum_{\mathbf{x} \in \mathfrak{X}} (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}},$$

$$\mathbf{S}_W = \sum_{j=1}^{c} \sum_{\mathbf{x} \in \mathfrak{X}_j} (\mathbf{x} - \boldsymbol{\mu}_j)(\mathbf{x} - \boldsymbol{\mu}_j)^{\mathrm{T}},$$

and $\mathbf{S}_B = \sum_{j=1}^{c} n_j \left(\boldsymbol{\mu}_j - \boldsymbol{\mu}\right)\left(\boldsymbol{\mu}_j - \boldsymbol{\mu}\right)^{\mathrm{T}}$.

Since for SSS problem $d > n$, the scatter matrices will be singular. It is known that the null space of $\mathbf{S}_T$ does not contain any discriminant information [19]. Therefore, the dimensionality can be reduced from d-dimensional space to r_t-dimensional space (where r_t is the rank of $\mathbf{S}_T$) by applying PCA as a pre-processing step. The range space of $\mathbf{S}_T$ matrix, $\mathbf{U}_1 \in \mathbb{R}^{d \times r_t}$, will be used as a transformation matrix. In the reduced dimensional space the scatter matrices will be given by: $\mathbf{S}_w = \mathbf{U}_1^{\mathrm{T}} \mathbf{S}_W \mathbf{U}_1$ and $\mathbf{S}_b = \mathbf{U}_1^{\mathrm{T}} \mathbf{S}_B \mathbf{U}_1$. After this procedure $\mathbf{S}_w \in \mathbb{R}^{r_t \times r_t}$ and $\mathbf{S}_b \in \mathbb{R}^{r_t \times r_t}$ are reduced dimensional within-class scatter matrix and reduced dimensional between-class scatter matrix, respectively.

4.2 Deterministic approach to regularized LDA

In the SSS problem, $\mathbf{S}_w$ matrix becomes singular and its inverse computation becomes impossible. In order to overcome this drawback, the RLDA technique adds a small positive constant α to all the diagonal elements of matrix $\mathbf{S}_w$ to make it non-singular; i.e., $\mathbf{S}_w$ is replaced by $\hat{\mathbf{S}}_w = \mathbf{S}_w + \alpha \mathbf{I}$. In this section, we describe a procedure to compute the regularization parameter α deterministically. In RLDA, the modified Fisher's criterion is given as follows:

$$\hat{J}(\mathbf{w}, \alpha) = \frac{\mathbf{w}^{\mathrm{T}} \mathbf{S}_b \mathbf{w}}{\mathbf{w}^{\mathrm{T}} (\mathbf{S}_w + \alpha \mathbf{I}) \mathbf{w}} \qquad (1)$$

where $\mathbf{w} \in \mathbb{R}^{r_t \times 1}$ is the orientation vector. Let us denote a function

$$f = \mathbf{w}^{\mathrm{T}} \mathbf{S}_b \mathbf{w} \tag{2}$$

and a constraint function

$$g = \mathbf{w}^{\mathrm{T}} (\mathbf{S}_w + \alpha \mathbf{I}) \mathbf{w} - b = 0 \tag{3}$$

where $b > 0$ is any constant. To find the maximum of f under the constraint, let us define a function $F = f - \lambda g$, where λ is Lagrange's multiplier ($\lambda \neq 0$). By setting its derivative to zero, we get

$$\frac{\partial F}{\partial \mathbf{w}} = 2\mathbf{S}_b \mathbf{w} - \lambda (2\mathbf{S}_w \mathbf{w} + 2\alpha \mathbf{w}) = 0$$

or $$(\frac{1}{\lambda} \mathbf{S}_b - \mathbf{S}_w) \mathbf{w} = \alpha \mathbf{w}. \tag{4}$$

Substituting value of $\alpha \mathbf{w}$ from equation (4) into equation (3), we get

$$g = \mathbf{w}^{\mathrm{T}} \mathbf{S}_w \mathbf{w} + \mathbf{w}^{\mathrm{T}} (\frac{1}{\lambda} \mathbf{S}_b - \mathbf{S}_w) \mathbf{w} - b = 0$$

or $$\mathbf{w}^{\mathrm{T}} \mathbf{S}_b \mathbf{w} = \lambda b. \tag{5}$$

Also from equation (3), we get

$$\mathbf{w}^{\mathrm{T}} (\mathbf{S}_w + \alpha \mathbf{I}) \mathbf{w} = b. \tag{6}$$

Dividing equation (5) by equation (6), we get

$$\lambda = \frac{\mathbf{w}^{\mathrm{T}} \mathbf{S}_b \mathbf{w}}{\mathbf{w}^{\mathrm{T}} (\mathbf{S}_w + \alpha \mathbf{I}) \mathbf{w}} \tag{7}$$

The right-hand side of equation (7) is same as the criterion $\hat{J}(\mathbf{w}, \alpha)$ defined in equation (1). The left-hand side of equation (7) is the Lagrange's multiplier (in equation (4)).

Since our aim is to maximize the modified Fisher's criterion $\hat{J}(\mathbf{w},\alpha)$, we must set λ equal to maximum of $\hat{J}(\mathbf{w},\alpha)$. However, it is not possible to find the maximum of $\hat{J}(\mathbf{w},\alpha)$ as α is not known to us. So, as an approximation we set λ equal to the maximum of the original Fisher's criterion $(\mathbf{w}^{\mathrm{T}}\mathbf{S}_b\mathbf{w}/\mathbf{w}^{\mathrm{T}}\mathbf{S}_w\mathbf{w})$. In order to maximize the original Fisher's criterion, we must have eigenvector $\mathbf{w}$ to correspond to the maximum eigenvalue of $\mathbf{S}_w^{-1}\mathbf{S}_b$. Since in SSS problem $\mathbf{S}_w$ is singular and non-invertible, we approximate the inverse of $\mathbf{S}_W$ by its pseudoinverse and carry out the EVD of $\mathbf{S}_w^{+}\mathbf{S}_b$ to find the highest (or leading) eigenvalue, where $\mathbf{S}_w^{+}$ is the pseudoinverse of $\mathbf{S}_w$. Thus, if λ_{max} denotes the highest eigenvalue of $\hat{J}(\mathbf{w},\alpha)$, then

$$\begin{aligned}\lambda_{max} &= \max\left(\frac{\mathbf{w}^{\mathrm{T}}\mathbf{S}_b\mathbf{w}}{\mathbf{w}^{\mathrm{T}}(\mathbf{S}_w+\alpha\mathbf{I})\mathbf{w}}\right)\\ &\approx \max\left(\frac{\mathbf{w}^{\mathrm{T}}\mathbf{S}_b\mathbf{w}}{\mathbf{w}^{\mathrm{T}}\mathbf{S}_w\mathbf{w}}\right)\\ &\approx \text{largest eigenvalue of } \mathbf{S}_w^{+}\mathbf{S}_b \qquad (8)\end{aligned}$$

Thereby, the evaluation of α can be carried out from equation (4) by doing EVD of $(\frac{1}{\lambda}\mathbf{S}_b-\mathbf{S}_w)$, where $\lambda=\lambda_{max}$. The EVD of $(\frac{1}{\lambda}\mathbf{S}_b-\mathbf{S}_w)$ will give $r_b=rank(\mathbf{S}_b)$ eigenvalues. Since the highest eigenvalue will correspond to the most discriminant eigenvector, α is the highest eigenvalue. Therefore, if EVD of $(\frac{1}{\lambda}\mathbf{S}_b-\mathbf{S}_w)$ is given by

$$(\tfrac{1}{\lambda}\mathbf{S}_b-\mathbf{S}_w)=\mathbf{E}\mathbf{D}_{bw}\mathbf{E}^{\mathrm{T}} \qquad (9)$$

where $\mathbf{E}\in\mathbb{R}^{r_t\times r_t}$ is a matrix of eigenvectors and $\mathbf{D}_{bw}\in\mathbb{R}^{r_t\times r_t}$ is a diagonal matrix of

corresponding eigenvalues. Now the α parameter can be computed as

$$\alpha = \max \mathbf{D}_{bw} \tag{10}$$

After evaluating α, orientation vector $\mathbf{w}$ can be obtained by performing the EVD of $(\mathbf{S}_w + \alpha\mathbf{I})^{-1}\mathbf{S}_b$; i.e., from

$$(\mathbf{S}_w + \alpha\mathbf{I})^{-1}\mathbf{S}_b\mathbf{w} = \gamma\mathbf{w}. \tag{11}$$

From the r_b eigenvectors obtained by this EVD, h $(\leq r_b)$ eigenvectors corresponding to h highest eigenvalues are used to form the orientation matrix $\mathbf{W}$.

It can be shown from Lemma 1 that for DRLDA technique, its maximum eigenvalue is approximately equal to the highest (finite) eigenvalue of Fisher's criterion.

Lemma 1: *The highest eigenvalue of DRLDA is approximately equivalent to the highest (finite) eigenvalue of Fisher's criterion.*

Proof 1: From equation 11,

$$\mathbf{S}_b\mathbf{w}_j = \gamma_j(\mathbf{S}_w + \alpha\mathbf{I})\mathbf{w}_j, \tag{12}$$

where α is the maximum eigenvalue of $(1/\lambda_{max}\mathbf{S}_b - \mathbf{S}_w)$ (from equation 4); $\lambda_{max} \geq 0$ is approximately the highest eigenvalue of Fisher's criterion $\mathbf{w}^{\mathrm{T}}\mathbf{S}_b\mathbf{w}/\mathbf{w}^{\mathrm{T}}\mathbf{S}_w\mathbf{w}$ (since λ_{max} is the largest eigenvalue of $\mathbf{S}_w^{+}\mathbf{S}_b$) [46]; $j = 1 \ldots r_b$ and $r_b = rank(\mathbf{S}_b)$. Substituting

$\alpha\mathbf{w} = (1/\lambda_{max}\mathbf{S}_b - \mathbf{S}_w)\mathbf{w}$ (from equation 4, where $\lambda = \lambda_{max}$) into equation 12, we get,

$$\mathbf{S}_b\mathbf{w}_m = \gamma_m\mathbf{S}_w\mathbf{w}_m + \gamma_m(1/\lambda_{max}\mathbf{S}_b - \mathbf{S}_w)\mathbf{w}_m,$$

or $$(\lambda_{max} - \gamma_m)\mathbf{S}_b\mathbf{w}_m = 0.$$

where $\gamma_m = \max(\gamma_j)$ and $\mathbf{w}_m$ is the corresponding eigenvector. Since $\mathbf{S}_b\mathbf{w}_m \neq 0$ (from equation 5), $\gamma_m = \lambda_{max}$ and $\gamma_j < \lambda_{max}$, where $j \neq m$. This concludes the proof.

Corollary 1: The value of regularization parameter is non-negative; i.e., $\alpha \geq 0$ for $r_w \leq r_t$, where $r_t = rank(\mathbf{S}_T)$ and $r_w = rank(\mathbf{S}_w)$.

Proof. Please see Appendix-III.

The summary of the DRLDA technique is given in Table 1[2].

Table 1: DRLDA technique

Step 1.	Pre-processing stage: apply PCA to find the range space $\mathbf{U}_1 \in \mathbb{R}^{d\times r_t}$ of total scatter matrix $\mathbf{S}_T$ and transform original d-dimensional data space to r_t-dimensional data space, where $r_t = rank(\mathbf{S}_T)$. Find reduced-dimensional between-class scatter matrix $\mathbf{S}_b = \mathbf{U}_1^{\mathrm{T}}\mathbf{S}_B\mathbf{U}_1$ and reduced-dimensional within-class scatter matrix $\mathbf{S}_w = \mathbf{U}_1^{\mathrm{T}}\mathbf{S}_W\mathbf{U}_1$, where $\mathbf{S}_b \in \mathbb{R}^{r_t\times r_t}$ and $\mathbf{S}_w \in \mathbb{R}^{r_t\times r_t}$.
Step 2.	Find the highest eigenvalue λ_{max} by performing EVD of $\mathbf{S}_w^{+}\mathbf{S}_b$.
Step 3.	Compute EVD of $(1/\lambda_{max}\mathbf{S}_b - \mathbf{S}_w)$ to find its highest eigenvalue α.
Step 4.	Compute EVD of $(\mathbf{S}_w + \alpha\mathbf{I})^{-1}\mathbf{S}_b$ to find h eigenvectors $\mathbf{w}_j \in \mathbb{R}^{r_t\times 1}$ corresponding to the leading eigenvalues, where $1 \leq h \leq r_b$ and $r_b = rank(\mathbf{S}_b)$. The eigenvectors $\mathbf{w}_j$ are column vectors of the orientation matrix $\mathbf{W}' \in \mathbb{R}^{r_t\times h}$.
Step 5.	Find orientation matrix $\mathbf{W} \in \mathbb{R}^{d\times h}$ in a d-dimensional space; i.e., $\mathbf{W} = \mathbf{U}_1\mathbf{W}'$.

[2] Matlab code will be provided upon acceptance of the paper on our website.

The computational requirement for Step 1 of the technique (Table 1) would be $O(dn^2)$; for Step 2 would be $O(n^3)$; for Step 3 would be $O(n^3)$; for Step 4 would be $O(n^3)$; and, for Step 5 would be $O(dn^2)$. Therefore, the total estimated for SSS case ($d \gg n$) would be $O(dn^2)$.

5. Experimental setup and results

Experiments are conducted to illustrate the relative performance of the DRLDA technique with respect to other techniques for the following two applications: 1) face recognition and 2) cancer classification. For face recognition, two commonly known datasets namely ORL dataset [37] and AR dataset [38] are utilized. The ORL dataset contains 400 images of 40 people having 10 images per person. We use the dimensionality of the original feature space to be 5152. The AR dataset contains 100 classes. We use a subset of AR dataset with 14 face images per class. We use the dimensionality of feature space to be 4980. For cancer classification, 6 commonly available datasets are used. All the datasets used in the experimentation are described in Table 2. For some datasets, number of training samples and test samples are predefined by their donors (Table 2). For these datasets, we use test samples to evaluate the classification performance. For some datasets, the training and test samples are not

predefined. For these datasets we carried out k-fold cross-validation procedure[3] to compute the classification performance, where $k = 3$.

The DRLDA technique is compared with the following techniques: Null LDA (NLDA) [6], cross-validation based RLDA (CV-RLDA), Pseudo-inverse LDA (PILDA) [20], Direct LDA (DLDA) [7], Fisherface or PCA+LDA [4][5], Uncorrelated LDA (ULDA) [21] and eigenfeature regularization (EFR) [10]. All the techniques are used to find the orientation matrix $\mathbf{W} \in \mathbb{R}^{d \times c-1}$, thereby, transforming the original space to $c-1$ dimensional space, where c is the number of classes. Then nearest neighbour classifier (NNC) using Euclidean distance measure is used for classifying a test feature vector.

The setting up of CV-RLDA technique in our experiments is described as follows: the regularization parameter α of CV-RLDA is computed by using leave-one-out cross-validation procedure on the training set. This is done in two steps. In the first step, we perform a coarse search for α by dividing the pre-selected range $[10^{-4}, 1] * \lambda_W$ (where λ_W is the maximum eigenvalue of $\mathbf{S}_W$) into 10 equal intervals and finding the

[3] In the k-fold cross-validation procedure [39], we first partition all the available samples randomly into k roughly equal segments. Then hold out one segment as validation data and the remaining $k-1$ segments as training data. Using the training data, we applied a discriminant technique to obtain orientation matrix and the validation data to compute classification accuracy. This partitioning of samples and computation of classification accuracy is carried out k times to evaluate average classification accuracy.

interval whose center value gives the best classification performance on the training set. In the second step, this interval is further divided into 10 subintervals for fine search and the center value of the subinterval that gives the best classification performance is used as the final value of the regularization parameter. Thus, in this procedure, a total of 20 α values are investigated. The regularization parameters computed by CV-RLDA technique on various datasets are shown in Appendix-IV.

The classification accuracy on all the datasets using the above mentioned techniques are shown in Table 3 (the highest classification accuracies obtained are depicted in bold fonts). It can be seen from Table 3 that out of 8 datasets used, the number of times the highest classification accuracy obtained by NLDA is 2, CV-RLDA is 5, PILDA is 1, DLDA is 1, PCA+LDA is 3, ULDA is 2, EFR is 4 and DRLDA is 6. In particular, DRLDA performs better than CV-RLDA for most of the datasets shown in Table 2 (it outperforms CV-RLDA for 3 out of 8 datasets, shows equal classification accuracy for 3 datasets and is inferior to CV-RLDA in the remaining 2 datasets). Note that the CV-RLDA technique when implemented in an ideal form (i.e., when α is searched in the range $(0,\infty)$ with infinitely small step size) should give in principle better results than the DRLDA technique. Since it is not possible for practical reasons (i.e.,

computational cost is infinitely large), a finite range is used in CV-RLDA technique. As a result, DRLDA technique is performing here better in terms of classification accuracy for majority of datasets. In addition, the computational cost of CV-RLDA technique (with α being searched in the finite range) is considerably higher than the DRLDA technique as shown in Table 4. Here, we measure the CPU time taken by its 'Matlab' program on a Sony computer (*core i7 processor at 2.8GHz*).

Table 2: Datasets used in the experimentation

Datasets	**Class**	**Dimension**	**Number of available samples**	**Number of training samples**	**Number of test samples**
Acute Leukemia [40]	2	7129	72	38	34
ALL subtype [41]	7	12558	327	215	112
GCM [42]	14	16063	198	144	54
Lung Adenocarcinoma [43]	3	7129	96 ([67,19,10])*	-	-
MLL [44]	3	12582	72	57	15
SRBCT [45]	4	2308	83	63	20
Face ORL [37]	40	5152	400 (10/class)	-	-
Face AR [38]	100	4980	1400 (14/class)	-	-

* The values in the square parenthesis indicate number of samples per class.

Table 3: Classification accuracy (in percentage) on datasets using various techniques.

Database	**NLDA**	**CV-RLDA**	**PILDA**	**DLDA**	**PCA+LDA**	**ULDA**	**EFR**	**DRLDA**
Acute Leukemia	97.1	97.1	73.5	97.1	**100.0**	97.1	**100.0**	**100.0**
ALL subtype	86.6	**95.5**	62.5	93.8	80.7	82.1	86.6	93.8
GCM	70.4	**74.1**	46.3	59.3	59.3	66.7	68.5	70.4
Lung Adeno.	81.7	81.7	74.2	72.0	81.7	80.7	83.9	**86.0**
MLL	**100.0**	**100.0**	80.0	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**
SRBCT	**100.0**	**100.0**	85.0	80.0	**100.0**	**100.0**	**100.0**	**100.0**
Face ORL	96.9	**97.2**	96.4	96.7	92.8	92.5	96.7	**97.2**
Face AR	95.7	96.3	**97.3**	96.3	94.9	95.8	**97.3**	**97.3**

Table 4: The comparison of cputime (in seconds) of CV-RLDA and DRLDA techniques.

Database	**CV-RLDA CPU Time**	**DRLDA CPU Time**
Acute Leukemia	4.68	0.07
ALL subtype	1021.9	1.90
GCM	265.0	1.26
Lung Adeno.	57.9	0.48
MLL	13.6	0.24
SRBCT	17.0	0.08

Face ORL	7396.1	7.41
Face AR	739,380	89.9

Furthermore, various techniques using artificial data are experimented. For this, we have created a 2-class problem with initial dimensions d =10, 25, 30, 50, and 100. In order to have ill-posed problem, we generated only 3 samples per class. The dimensionality is reduced from d to 1 for all the techniques and then nearest neighbour classifier is used to evaluate the performance in terms of classification accuracy. For each dimension d, the data is created 100 times to compute average classification accuracy. Table 5 depicts the average classification accuracy over 100 runs. It can be observed from Table 5 that EFR technique is not able to perform because of scarce samples. The DRLDA technique and CV-RLDA technique are performing similar. Pseudoinverse technique (PILDA) is performing the lowest as there is not enough information in the range space of scatter matrices.

Table 5: Classification accuracy (in percentage) on artificial dataset using various techniques.

Dimension	NLDA	CV-RLDA	PILDA	DLDA	PCA+LDA	ULDA	EFR	DRLDA
10	84.3	87.2	66.2	87.8	85.7	84.3	-	87.2
25	95.0	96.7	58.2	96.3	93.7	95.0	-	97.2
30	96.0	97.8	52.8	95.8	96.2	96.0		98.0
50	98.8	99.2	49.5	99.2	98.7	98.8	-	99.2
100	100	100	50	99.5	99.8	100	-	100

We have also carried out sensitivity analysis with respect to the classification accuracy. For this purpose, we use Acute Leukemia dataset as a prototype and contaminated the dataset by adding Gaussian noise. We then applied techniques again to evaluate classification performance by using nearest neighbor classifier. The generated noise levels are 1%, 2%, 5% and 10% of the standard deviation of the original feature values. The noisy data has been generated 10 times to compute average classification accuracy. The results are shown in Figure 1. It can be observed from Figure 1 that for low level noise the degradation in classification performance is not enough. But when the noise

level increases the classification accuracy deteriorates. The performance of PILDA and DLDA techniques are lower than other techniques. However, most of the techniques try to maintain the discriminant information in the noisy environment.

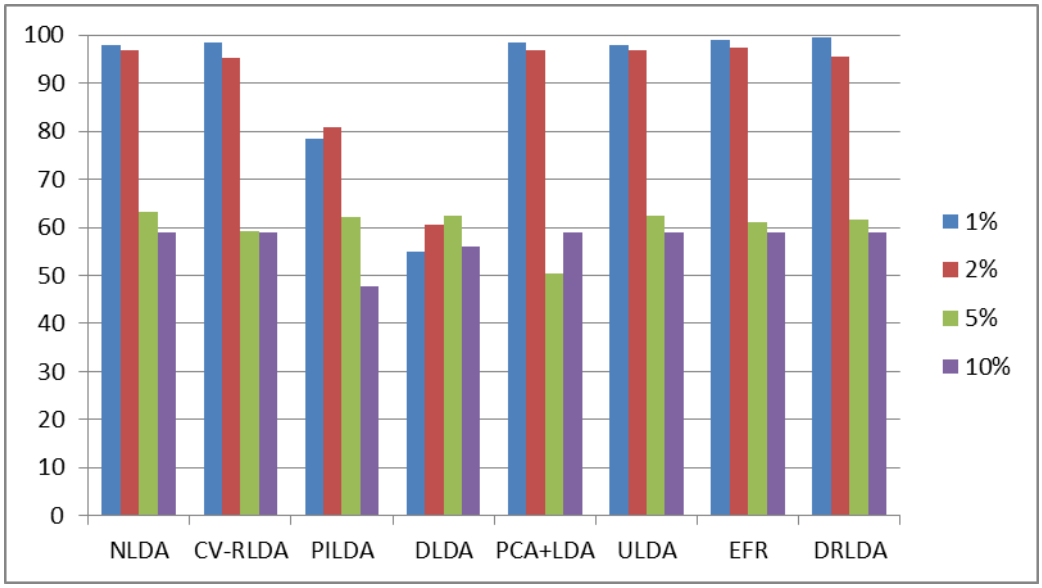

Figure 1: Sensitivity analysis of various techniques on Acute Leukemia dataset at different noise levels. The y-axis depicts average classification accuracy and x-axis depicts the techniques used. The noise levels are 1%, 2%, 5% and 10%.

6. Discussion

In order to compare the performance in terms of classification accuracy we compared 7 well known techniques with DRLDA. These techniques compute the orientation matrix $\mathbf{W}$ by utilizing different combinations of informative spaces (i.e., null space of $\mathbf{S}_W$, range space of $\mathbf{S}_W$ and range space of $\mathbf{S}_B$). Each informative space contains a certain level of discriminant information. Theoretically, it is effective to utilize all the informative

spaces for the computation of orientation matrix for better generalization capability. How well a technique is combining these spaces would determine its generalization capability. It has been shown that usually the null space of $\mathbf{S}_W$ contains more discriminant information than the range space of $\mathbf{S}_W$ [6][8][22][34]. Therefore, it is likely that a technique that utilizes null space of $\mathbf{S}_W$ effectively, may perform better (in generalization capability) than a technique which does not use the null space of $\mathbf{S}_W$.

From the techniques that we have used the NLDA technique employs null space of $\mathbf{S}_W$ and range space of $\mathbf{S}_B$. Whereas PILDA, DLDA and PCA+LDA techniques employ range space of $\mathbf{S}_W$ and range space of $\mathbf{S}_B$. Provided the techniques extract the maximum possible information from the spaces they employed then NLDA should beat PILDA, DLDA and PCA+LDA techniques. From Table 3, we can see that NLDA is outperforming PILDA in 7 out of 8 cases. Comparing the classification accuracies of NLDA and DLDA, we can see that NLDA is outperforming DLDA in 4 out of 8 cases and in 2 cases the performance are identical. In a similar way NLDA is surpassing PCA+LDA in 4 out of 8 cases and in 3 cases the performance are identical. On the other hand, the ULDA technique also employs the same spaces as of NLDA technique, however, the classification performance of ULDA is inferior to NLDA (only in 1 out of 8

cases ULDA is beating NLDA). This means that orthogonal $\mathbf{W}$ is more effective than uncorrelated $\mathbf{W}$.

The other three techniques (CV-RLDA, EFR and DRLDA) employ three spaces; namely, null space of $\mathbf{S}_W$, range space of $\mathbf{S}_W$ and range space of $\mathbf{S}_B$. Intuitively, these three techniques contain more discriminant information than above mentioned 5 techniques. However, different strategies of using the three spaces would result in different level of generalization capabilities. In CV-RLDA, the estimation of regularization parameter α depends upon the range of α values being explored (which is restricted due to limited computation time), the cross-validation procedure (e.g. leave-one-out, k-fold) being employed and the classifier used. On the other hand, EFR and DRLDA techniques do not have this problem. The EFR technique utilizes an intuitive model for extrapolating the eigenvalues of range space of $\mathbf{S}_W$ to the null space of $\mathbf{S}_W$. This way it captures all the spaces. However, the model used for extrapolation is rather arbitrary and it is not necessary that it is an optimum model. The DRLDA technique captures the information from all the spaces by deterministically finding the optimal α parameter from the training samples. From Table 3, it can be observed that EFR is surpassing CV-RLDA in 3 out of 8 cases and exhibiting identical classification accuracies in 2 cases. Similarly,

DRLDA is outperforming CV-RLDA in 3 out of 8 cases and giving equal results in 3 cases. From Table 3 and Table 4, we can also observe that though the classification accuracy of CV-RLDA is high (which depends on the search of the regularization parameter), its computational time is extremely large.

Thus we have shown that DRLDA technique is performing better than other LDA techniques for the SSS problem. We can intuitively explain its better performance as follows. In the DRLDA technique, we are maximizing the modified Fisher's criterion; i.e., the ratio of between-class scatter and within-class scatter (see equation 1). To get the α parameter, we are maximizing the difference between the between-class scatter and within-class scatter (see equation 4). Thus, we are combining two different philosophies of LDA mechanism in our DRLDA technique and this is helping us in getting better performance.

7. Conclusion

The paper presented a deterministic approach of computing regularized LDA. It avoids the use of the heuristic (cross-validation) procedure for computing the regularization parameter. The technique has been experimented on a number of datasets and

compared with several popular techniques. The DRLDA technique exhibits highest classification accuracy for 6 out of 8 datasets and its computational cost is significantly less than CV-RLDA technique.

Appendix-I

In this appendix, the generalization capabilities of Zhao's model and Friedman's model of CV-RLDA are demonstrated on several datasets. In order to do this, first we project the original feature vectors onto the range space of total-scatter matrix as a pre-processing step. Then we employ reduced dimensional within-class scatter matrix $\hat{\mathbf{S}}_w$ for the two models of CV-RLDA (see Section 4.1 for details about reduced dimensional matrices). In the first model of CV-RLDA, $\mathbf{S}_w$ is approximated as $\hat{\mathbf{S}}_w = \mathbf{S}_w + \alpha\mathbf{I}$ and in the second model $\mathbf{S}_w$ is approximated as $\hat{\mathbf{S}}_w = (1-\alpha)\mathbf{S}_w + \alpha\mathbf{I}$. For brevity, we refer the former model of CV-RLDA as CV-RLDA-1 and the latter model as CV-RLDA-2. Table A1 depicts the classification performance of these two models. The details of the datasets and the selection of the regularization parameter α can be found in Section 4.

It can be seen from Table A1 that CV-RLDA-1 exhibits relatively better classification performance than CV-RLDA-2.

Table A1: Classification accuracy (in percentage) on test set using CV-RLDA-1 and CV-RLDA-2 techniques.

Database	CV-RLDA-1	CV-RLDA-2
Acute Leukemia	97.1	97.1
ALL subtype	95.5	86.6
GCM	74.1	70.4
MLL	100.0	100.0
SRBCT	100.0	100.0

Appendix-II

In this appendix, for RLDA technique we show the sensitivity of classification accuracy when selecting the regularization parameter, α. For this purpose we use four values of α. These are $\delta = [0.001, 0.01, 0.1, 1]$, where $\alpha = \delta * \lambda_W$ and λ_W is the maximum eigenvalue of within-class scatter matrix. We applied 3-fold cross-validation procedure on a number of datasets and shown the results in Table A2.

Table A2: Classification accuracy (in percentage) using 3-fold cross-validation procedure (the highest classification accuracies obtained are depicted in bold fonts).

Database	$\delta = 0.001$	$\delta = 0.01$	$\delta = 0.1$	$\delta = 1$
Acute Leukemia	98.6	98.6	98.6	**100**
ALL subtype	**90.3**	**90.3**	86.0	69.2
GCM	72.7	74.3	**76.5**	59.0
Lung Adeno.	81.7	80.7	**85.0**	80.7
MLL	**95.7**	**95.7**	**95.7**	**95.7**

SRBCT	**100.0**	**100.0**	**100.0**	96.2
Face ORL	**96.9**	**96.9**	**96.9**	**96.9**
Face AR	95.8	**97.9**	96.3	81.8

It can be observed from the table that the different values of the regularization parameter give different classification accuracies and therefore, the choice of the regularization parameter affects the classification performance. Thus, it is important to select the regularization parameter correctly to get the good classification performance.

To do this, a cross-validation approach is usually opted. The α parameter is searched in the pre-defined range and the value of α which gives the best classification performance on the training set is selected. It is assumed that the optimum value of α will give the best generalization capability; i.e., the best classification performance on the test set.

Appendix III

Corollary 1: The value of regularization parameter is non-negative; i.e., $\alpha \geq 0$ for $r_w \leq r_t$, where $r_t = rank(\mathbf{S}_T)$ and $r_w = rank(\mathbf{S}_w)$.

Proof 1: From equation 1, we can write

$$J = \frac{\mathbf{w}^{\mathrm{T}}\mathbf{S}_b\mathbf{w}}{\mathbf{w}^{\mathrm{T}}(\mathbf{S}_w+\alpha\mathbf{I})\mathbf{w}}, \quad \text{A1}$$

where $\mathbf{S}_b \in \mathbb{R}^{r_t \times r_t}$ and $\mathbf{S}_w \in \mathbb{R}^{r_t \times r_t}$. We can rearrange the above expression as

$$\mathbf{w}^{\mathrm{T}}\mathbf{S}_b\mathbf{w} = J\mathbf{w}^{\mathrm{T}}(\mathbf{S}_w + \alpha\mathbf{I})\mathbf{w} \quad \text{A2}$$

The eigenvalue decomposition (EVD) of $\mathbf{S}_W$ matrix (assuming $r_w < r_t$) can be given as $\mathbf{S}_w = \mathbf{U}\boldsymbol{\Lambda}^2\mathbf{U}^{\mathrm{T}}$, where $\mathbf{U} \in \mathbb{R}^{r_t \times r_t}$ is an orthogonal matrix, $\boldsymbol{\Lambda}^2 = \begin{bmatrix} \boldsymbol{\Lambda}_w^2 & 0 \\ 0 & 0 \end{bmatrix} \in \mathbb{R}^{r_t \times r_t}$ and $\boldsymbol{\Lambda}_w = diag(q_1^2, q_2^2, \dots, q_{r_w}^2) \in \mathbb{R}^{r_w \times r_w}$ are diagonal matrices (as $r_w < r_t$). The eigenvalues $q_k^2 > 0$ for $k = 1,2, \dots, r_w$. Therefore,

$$\mathbf{S}_w' = (\mathbf{S}_w + \alpha\mathbf{I}) = \mathbf{U}\mathbf{D}\mathbf{U}^{\mathrm{T}}, \text{ where } \mathbf{D} = \boldsymbol{\Lambda}^2 + \alpha\mathbf{I}$$

$$\text{or } \mathbf{D}^{-1/2}\mathbf{U}^{\mathrm{T}}\mathbf{S}_w'\mathbf{U}\mathbf{D}^{-1/2} = \mathbf{I} \qquad \text{A3}$$

The between class scatter matrix $\mathbf{S}_b$ can be transformed by multiplying $\mathbf{U}\mathbf{D}^{-1/2}$ on the right side and $\mathbf{D}^{-1/2}\mathbf{U}^{\mathrm{T}}$ on the left side of $\mathbf{S}_b$ as $\mathbf{D}^{-1/2}\mathbf{U}^{\mathrm{T}}\mathbf{S}_b\mathbf{U}\mathbf{D}^{-1/2}$. The EVD of this matrix will give

$$\mathbf{D}^{-1/2}\mathbf{U}^{\mathrm{T}}\mathbf{S}_b\mathbf{U}\mathbf{D}^{-1/2} = \mathbf{E}\mathbf{D}_b\mathbf{E}^{\mathbf{T}}, \qquad \text{A4}$$

where $\mathbf{E} \in \mathbb{R}^{r_t \times r_t}$ is an orthogonal matrix and $\mathbf{D}_b \in \mathbb{R}^{r_t \times r_t}$ is a diagonal matrix. Equation A4 can be rearranged as

$$\mathbf{E}^{\mathbf{T}}\mathbf{D}^{-1/2}\mathbf{U}^{\mathrm{T}}\mathbf{S}_b\mathbf{U}\mathbf{D}^{-1/2}\mathbf{E} = \mathbf{D}_b, \qquad \text{A5}$$

Let the leading eigenvalue of $\mathbf{D}_b$ is γ and its corresponding eigenvector is $\mathbf{e} \in \mathbf{E}$. Then equation A5 can be rewritten as

$$\mathbf{e}^{\mathbf{T}}\mathbf{D}^{-1/2}\mathbf{U}^{\mathrm{T}}\mathbf{S}_b\mathbf{U}\mathbf{D}^{-1/2}\mathbf{e} = \gamma, \qquad \text{A6}$$

The eigenvector $\mathbf{e}$ can be multiplied right side and $\mathbf{e}^{\mathrm{T}}$ on left side of equation A3, we get

$$\mathbf{e}^{\mathrm{T}}\mathbf{D}^{-1/2}\mathbf{U}^{\mathrm{T}}\mathbf{S}_w'\mathbf{U}\mathbf{D}^{-1/2}\mathbf{e} = 1 \qquad \text{A7}$$

It can be seen from equations A3 and A5 that matrix $\mathbf{W} = \mathbf{U}\mathbf{D}^{-1/2}\mathbf{E}$ diagonalizes both $\mathbf{S}_b$ and $\mathbf{S}_w'$, simultaneously. Also vector $\mathbf{w} = \mathbf{U}\mathbf{D}^{-1/2}\mathbf{e}$ simultaneously gives γ and unity eigenvalues in equations A6 and A7. Therefore, $\mathbf{w}$ is a solution of equation A2. Substituting $\mathbf{w} = \mathbf{U}\mathbf{D}^{-1/2}\mathbf{e}$ in equation A2, we get

$J = \gamma$; i.e., $\mathbf{w}$ is a solution of equation A2.

From Lemma 1, the maximum eigenvalue of expression $(\mathbf{S}_W + \alpha\mathbf{I})^{-1}\mathbf{S}_b\mathbf{w} = \gamma\mathbf{w}$ is $\gamma_m = \lambda_{max} > 0$ (i.e., real, positive and finite). Therefore, the eigenvectors corresponding to this positive γ_m should also be in real hyperplane (i.e., the components of the vector $\mathbf{w}$ have to have real values). Since $\mathbf{w} = \mathbf{U}\mathbf{D}^{-1/2}\mathbf{e}$ with $\mathbf{w}$ to be in real hyperplane, we must have $\mathbf{D}^{-1/2}$ to be real.

Since $\mathbf{D} = \mathbf{\Lambda}^2 + \alpha\mathbf{I} = diag(q_1^2 + \alpha, q_2^2 + \alpha, \dots, q_{r_w}^2 + \alpha, \alpha, \dots, \alpha)$, we have
$\mathbf{D}^{-1/2} = diag(1/\sqrt{q_1^2 + \alpha}, 1/\sqrt{q_2^2 + \alpha}, \dots, 1/\sqrt{q_{r_w}^2 + \alpha}, 1/\sqrt{\alpha}, \dots, 1/\sqrt{\alpha})$.
Therefore, the elements of $\mathbf{D}^{-1/2}$, must satisfy $1/\sqrt{q_k^2 + \alpha} > 0$ and $1/\sqrt{\alpha} > 0$ for $k = 1,2, \dots, r_w$ (note $r_w < r_t$); i.e., α cannot be negative or $\alpha > 0$. Furthermore, if $r_w = r_t$ then matrix $\mathbf{S}_w$ will be a non-singular matrix and its inverse will exist. In this case, regularization is not required and therefore $\alpha = 0$. Thus, $\alpha \geq 0$ for $r_w \leq r_t$. This concludes the proof.

Appendix IV

In this appendix, we show computed value of CV-RLDA technique. The value of α is computed by first doing a coarse search on a predefined range to find a coarse value. After this, a fine search is conducted using this coarse value to get the regularization parameter. In this experiment, we use $\alpha = \delta * \lambda_w$ where $\delta = [10^{-4}, 1]$ and λ_w is the highest eigenvalue of within-class scatter matrix. The values are depicted in Table A3. In addition, we have also shown regularization parameters computed by DRLDA technique as a reference.

Table A3: Computed values of regularization parameter for CV-RLDA and DRLDA on various datasets

Database	**CV-RLDA**	**CV-RLDA**	**DRLDA**
	$\boldsymbol{\delta}$	$\boldsymbol{\alpha}$	$\boldsymbol{\alpha}$
Acute Leukemia	0.0057	935.3	6.54×10^{9}
ALL subtype	0.5056	5.17×10^{5}	1.11×10^{11}
GCM	0.0501	2.42×10^{4}	1.34×10^{9}
MLL	0.0057	2621.5	2.98×10^{10}
SRBCT	0.1056	33.01	5715.2

Reference

[1] R.O. Duda and P.E. Hart, Pattern classification and scene analysis, Wiley, New York, 1973.

[2] K. Fukunaga, Introduction to statistical pattern recognition. Academic Press Inc., Hartcourt Brace Jovanovich, Publishers. 1990.

[3] J.H. Friedman, "Regularized discriminant analysis", *Journal of the American Statistical Association*, vol. 84, no. 405, pp. 165-175, 1989.

[4] D.L. Swets, and J. Weng, "Using discriminative eigenfeatures for image retrieval", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1996, 18, (8), pp. 831-836.

[5] P.N. Belhumeur, J.P. Hespanha and D.J. Kriegman, "Eigenfaces vs. fisherfaces: recognition using class specific linear projection", *IEEE Trans. Pattern Analysis and Machine Inteligence.*, vol. 19, no. 7, pp. 711–720, 1997.

[6] L.-F. Chen, H.-Y.M. Liao, M.-T. Ko, J.-C. Lin, and G.-J. Yu, "A new LDA-based face recognition system which can solve the small sample size problem", *Pattern Recognition*, vol. 33, pp. 1713-1726, 2000.

[7] H. Yu and J. Yang, "A direct LDA algorithm for high-dimensional data-with application to face recognition", *Pattern Recognition*, vol. 34, pp. 2067-2070, 2001.

[8] J. Ye, "Characterization of a family of algorithms for generalized discriminant analysis on undersampled problems", *Journal of Machine Learning Research*, vol. 6, pp. 483-502, 2005.

[9] A. Sharma and K.K. Paliwal, "A gradient linear discriminant analysis for small sample sized problem", *Neural Processing Letters*, vol. 27, pp 17-24, 2008.

[10] X. Jiang, B. Mandal and A. Kot, "Eigenfeature regularization and extraction in face recognition", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 3, pp. 383-394, 2008.

[11] K.K. Paliwal and A. Sharma, "Improved direct LDA and its application to DNA gene

microarray data", *Pattern Recognition Letters*, vol. 31, issue 16, pp. 2489-2492, 2010.

[12] W. Zhao, R. Chellappa, P.J. Phillips, Subspace linear discriminant analysis for face recognition, Technical Report CAR-TR-914, CS-TR-4009, University of Maryland at College Park, USA, 1999.

[13] D.Q. Dai and P.C., Yuen, "Regularized discriminant analysis and its application to face recognition", *Pattern Recognition*, vol. 36, no. 3, pp. 845-847, 2003.

[14] D.Q. Dai and P.C., Yuen, "Face recognition by regularized discriminant analysis", *IEEE Transactions of SMC Part B*, vol. 37, issue 4, pp. 1080-1085, 2007.

[15] Y. Guo, T. Hastie, and R. Tibshirani, "Regularized discriminant analysis and its application in microarrays", *Biostatistics*, vol. 8, no. 1, pp. 86-100, 2007.

[16] W. Zhao, R. Chellappa, A. Krishnaswamy, "Discriminant analysis of principal components for face recognition", Proc. Thir Int. Conf. on Automatic Face and Gesture Recognition, pp. 336-341, Nara, Japan, 1998.

[17] W. Zhao, R. Chellappa, and P.J. Phillips, "Face recognition: a literature survey", ACM Computing Surveys, vol. 35, no. 4, pp. 399-458, 2003.

[18] T. Hastie, R. Tibshirani and J. Friedman, The elements of statistical learning, Springer, NY, USA, 2001.

[19] R. Huang, Q. Liu, H. Lu, and S. Ma**,** "Solving the Small Sample Size Problem of LDA", *Proceedings of ICPR*, vol. 3, pp. 29-32, 2002.

[20] Q. Tian, M. Barbero, Z.H. Gu and S.H. Lee, 'Image classification by the Foley-Sammon transform', Optical Engineering, vol. 25, no. 7, pp. 834-840, 1986.

[21] J. Ye, R. Janardan, Q. Li, and H. Park, "Feature extraction via generalized uncorrelated linear discriminant analysis" *The Twenty-First International Conference on Machine Learning*, pp. 895–902, 2004.

[22] K.K. Paliwal and A. Sharma, "Approximate LDA technique for dimensionality reduction in the small sample size case", *Journal of Pattern Recognition Research*, vol. 6, no. 2, pp. 298-306, 2011.

[23] J. Yang, D. Zhang and J.-Y. Yang, "A generased K-L expansion method which can deal with small samples size and high-dimensional problems", Pattern Analysis Application, vol. 6, pp. 47-54, 2003.

[24] D. Chu and G.S. Thye, "A new and fast implementation for null space based linear discriminant analysis', *Pattern Recognition*, vol. 43, pp. 1373-1379, 2010.

[25] A. Sharma and K.K. Paliwal, "A new perspective to null linear discriminant analysis method and its fast implementation using random matrix multiplication with scatter matrices", *Pattern Recognition*, vol. 45, issue 6, pp. 2205-2212, 2012.

[26] J. Ye and T. Xiong, "Computational and theoretical analysis of null space and orthogonal

linear discriminant analysis", *Journal of Machine Learning Research*, vol. 7, pp. 1183-1204, 2006.

[27] J. Liu, S.C. Chen, X.Y. Tan, "Efficient Pseudo-inverse Linear Discriminant Analysis and its Nonlinear Form for Face Recognition", *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 21, no. 8, pp. 1265-1278, 2007.

[28] K.K. Paliwal and A. Sharma, "Improved Pseudoinverse Linear Discriminant Analysis Method for Dimensionality Reduction", *International Journal of Pattern Recognition and Artificial Intelligence*, 2011, DOI No: 10.1142/S0218001412500024.

[29] J. Lu, K. Plataniotis and A. Venetsanopoulos, "Face recognition using kernel direct discriminant analysis algorithms", *IEEE Transactions on Neural Networks*, vol. 14, no. 1, pp. 117-126, 2003.

[30] J. Lu, K. Plataniotis and A. Venetsanopoulos, "Regularization studies of linear discriminant analysis in small sample size scenarios with application to face recognition", *Pattern Recognition Letters*, vol. 26, no. 2, pp. 181-191, 2005.

[31] F. Song, D. Zhang, J. Wang, H. Liu and Q. Tao, "A parameterized direct LDA and its application to face recognition", *Neurocomputing*, vol. 71, pp. 191-196, 2007.

[32] A. Sharma, A. and K.K. Paliwal, "Regularisation of eigenfeatures by extrapolation of scatter-matrix in face-recognition problem", *Electronics Letters*, IEE, vol. 46, no. 10, pp 450-475, 2010.

[33] J. Lu, K.N. Plataniotis, A.N. Venetsanopoulos, "Regularized discriminant analysis for the small sample", *Pattern Recognition Letters*, vol. 24, pp. 3079-3087, 2003.

[34] A. Sharma and K.K. Paliwal, "A two-stage linear discriminant analysis for face-recognition", *Pattern Recognition Letters*, vol. 33, issue 9, pp. 1157-1162, 2012.

[35] A. Sharma, and K.K. Paliwal, "Rotational linear discriminant analysis technique for dimensionality reduction", *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, no. 10, pp 1336-1347, 2008.

[36] A. Sharma, K.K. Paliwal, Onwubolu, G.C., "Class-dependent PCA, LDA and MDC: a combined classifier for pattern classification", Pattern Recognition, vol. 39, issue 7, 2006, pp. 1215-1229.

[37] F. Samaria and A. Harter, "Parameterization of a stochastic model for human face identification", *Proc. Second IEEE Workshop Applications of Comp. Vision*, pp. 138-142, 1994.

[38] A.M. Martinez, "Recognizing imprecisely localized, partially occluded, and expression variant faces from a single sample per class", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 6, pp. 748-763, 2002.

[39] E. Alpaydin, Introduction to machine learning, MIT Press, 2004.

[40] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield and E.S. Lander, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring", *Science*, vol. 286, 531-537, 1999.

[41] E.J. Yeoh , M.E. Ross, S.A. Shurtleff , W.K. Williams, D. Patel, R. Mahfouz, F.G. Behm, S.C. Raimondi, M.V. Relling, A. Patel, C. Cheng, D. Campana, D. Wilkins, X. Zhou, J. Li, H. Liu, C.H. Pui, W.E. Evans, C. Naeve, L. Wong, J.R. Downing, "Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling", *Cancer*, vol. 1, no. 2, pp 133-143, 2002.

[42] S. Ramaswamy, P. Tamayo, R. Rifkin, S. Mukherjee, C.-H. Yeang, M. Angelo, C. Ladd, M. Reich, E. Latulippe, J.P. Mesirov, T. Poggio, W. Gerald, M.. Loda, E.S. Lander and T.R. Golub, "Multiclass cancer diagnosis using tumor gene expression signatures", *Proc. Natl. Acad. Sci. USA*, vol. 98, no. 26, pp 15149-15154, 2001.

[43] D.G. Beer, S.L.R. Kardia, C.-C. Huang, T.J. Giordano, A.M. Levin, D.E. Misek, L. Lin, G. Chen, T.G. Gharib, D.G. Thomas, M.L. Lizyness, R. Kuick, S. Hayasaka, J.M.G. Taylor, M.D. Iannettoni, M.B. Orringer and S. Hanash, "Gene-expression profiles predict survival of patients with lung adenocarcinoma", *Nature Medicine*, vol. 8, pp. 816-824, 2002

[44] S.A. Armstrong, J.E. Staunton, L.B. Silverman, R. Pieters, M.L. den Boer, M.D. Minden, S.E. Sallan, E.S. Lander, T.R. Golub, and S.J. Korsemeyer, "MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia", *Nature Genetics*, vol. 30, pp 41-47, 2002.

[45] J. Khan, J.S. Wei, M. Ringner, L.H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C.R. Antonescu, C. Peterson and P.S. Meltzer, "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural network", *Nature Medicine*, vol. 7, pp. 673-679, 2001.

[46] Liu, J., Chen, S.C., Tan, X.Y., Efficient pseudo-inverse linear discriminant analysis and its nonlinear form for face recognition, Int. J. Patt. Recogn. Artif. Intell. vol. 21, no. 8, pp. 1265-1278, 2007.

[47] Sharma, A., and Paliwal, K.K., Linear discriminant analysis for small sample size problem: an overview, Int. J. Mach. Learn. & Cyber., 2014, DOI 10.1007/s13042-013-0226-9.

[48] Tao, D., Li, X., Wu, X. and Maybank, S., General tensor discriminant analysis and gabor features for gait recognition, IEEE Trans. Patt. Anal. Mach. Intell., vol. 29, issue 10, pp. 1700-1715, 2007.

[49] Bian, W. and Tao, D., Constrained empirical risk minimization framework for distance metric learning, IEEE Trans. Neural Net. And Learning Sys., vol. 23, no. 8, pp. 1194-1205, 2012.

[50] Zhou, T. and Tao, D., Double shrinking sparse dimension reduction, IEEE Trans. on Image. Proc., vol. 22, issue 1, pp. 244-257, 2013.

[51] Tao, D., Li, X., Wu, X. and Maybank, S.J., Geometric mean for subspace selection, IEEE Trans. Patt. Anal. Mach. Learn., vol. 31, issue 2, pp. 260-274, 2009.

[52] Sharma, A., Paliwal, K.K., Imoto, S. and Miyano, S., Principal component analysis using QR decomposition, Int. Jnr. of Mach. Learn. And Cybernetics, vol. 4, no. 6, pp. 679-683, 2013.

[53] Sharma, A. and Paliwal, K.K., Fast principal component analysis using fixed-point algorithm, vol. 28, no. 10, pp. 1151-1155, 2007.

[54] Wang, S.-J., Chen, H.-L., Peng, X.-J. and Zhou, C.-G., Exponential locality preserving projections for small sample size problem, Neurocomputing, vol. 74, issue 17, pg. 3654-3662, 2011.

[55] Zhang, L., Zhou, W. and Chang, P.-C., Generalized nonlinear discriminant analysis and its small sample size problems, Neurocomputing, vol. 74, issue 4, pp. 568-574, 2011.

[56] Huang, H., Liu, J., Feng, H. and He, T., Ear recognition based on uncorrelated local Fisher discriminant analysis, Neurocomputing, vol. 74, issue 17, pp. 3103-3113, 2011.

[57] Huerta, E.B., Duval, B. and Hao, J.-K., A hybrid LDA and genetic algorithm for gene selection and classification of microarray data, Neurocomputing, vol. 73, issue 13-15, pp. 2375-2383, 2010.

[58] Sharma, A., Imoto, S., Miyano, S., A top-r feature selection algorithm for microarray gene expression data, IEEE/ACM Trans. on Comp. Biol. and Bioinformatics, vol. 9, issue 3, pp. 754-764, 2012.

[59] Sharma, A., Imoto, S.,Miyano, S., A between-class overlapping filter-based method for transcriptome data analysis, Journal of Bioinformatics and Computational Biology, vol. 10, no. 5, pp. 1250010-1 1250010-20, 2012.

[60] Sharma, A., Imoto, S., Miyano, S., Sharma, V., Null space based feature selection method for gene expression data, Int. Jnr. of Mach. Learn. and Cybernetics, vol. 3, issue 4, pp. 269-276, 2012.

[61] Sharma, A. and Paliwal, K.K., Cancer classification by gradient LDA technique using microarray gene expression data, Data & Knowledge Engineering, vol. 66, issue 2, pp 338-347, 2008.

[62] Yang, W and Wu, H., Regularized complete linear discriminant analysis, Neurocomputing, vol. 137, pp. 185-191, 2014.