Linear Discriminant Analysis for the Small Sample Size Problem: An Overview

Alok Sharma[1,2,3], Kuldip K. Paliwal[2],
[1]Laboratory of DNA Information Analysis, University of Tokyo, Japan.
[2]School of Engineering, Griffith University, Australia.
[3]School of Engineering and Physics, University of the South Pacific, Fiji.

Abstract

Dimensionality reduction is an important aspect in the pattern classification literature, and linear discriminant analysis (LDA) is one of the most widely studied dimensionality reduction technique. The application of variants of LDA technique for solving small sample size (SSS) problem can be found in many research areas e.g. face recognition, bioinformatics, text recognition, etc. The improvement of the performance of variants of LDA technique has great potential in various fields of research. In this paper, we present an overview of these methods. We covered the type, characteristics and taxonomy of these methods which can overcome SSS problem. We have also highlighted some important datasets and software/packages.

Introduction

In a pattern classification (or recognition) system, an object (or pattern) which is characterized in terms of a feature vector is assigned a class label from a finite number of predefined classes. For this, the pattern classifier is trained using a set of training vectors (called the training dataset) and its performance is evaluated by classifying the feature vectors from the test dataset (which is normally different from the training dataset). In many pattern classification problems, the dimensionality of the feature vector is very large. It is therefore imperative to reduce the dimensionality of the feature space for improving the robustness (or generalization capability) and computational complexity of the pattern classifier. Different methods used for dimensionality reduction can be grouped into two categories: feature selection methods and feature extraction methods. Feature selection methods retain only few useful features and discards less important (or low ranked) features. Feature extraction methods reduce the dimensionality by constructing a few features from the large number of original features through their linear (or non-linear) combination. There are two popular feature extraction techniques reported in the literature for reducing the dimensionality of the feature space. These are principal component analysis (PCA) and linear discriminant analysis (LDA). PCA is an unsupervised technique, while LDA is a supervised technique. In general, LDA outperforms PCA in terms of classification performance.

The LDA technique finds an orientation $\mathbf{W}$ that transforms high dimensional feature vectors belonging to different classes to a lower dimensional feature space such that the projected feature vectors of a class on this lower dimensional space are well separated from the feature vectors of other classes. If the dimensionality reduction is from d-dimensional ($\mathbb{R}^d$) space to h-dimensional ($\mathbb{R}^h$) space (where $h < d$), then the size of the orientation matrix $\mathbf{W}$ is $\mathbb{R}^{d\times h}$; i.e., it has h column vectors. The orientation matrix $\mathbf{W}$ is obtained by maximizing the Fisher's criterion function; in other words by the eigenvalue decomposition (EVD) of $\mathbf{S}_W^{-1}\mathbf{S}_B$, where $S_W \in \mathbb{R}^{d\times d}$ is within-class scatter matrix and $\mathbf{S}_B \in \mathbb{R}^{d\times d}$ is between-class scatter matrix. For a c-class problem, the value of h will be $\min(c-1, d)$. If the dimensionality d is very large compared to the number of training vectors n, then $\mathbf{S}_W$ becomes singular and the evaluation of eigenvalues and eigenvectors of $\mathbf{S}_W^{-1}\mathbf{S}_B$ becomes impossible. This drawback is considered to be the main problem of LDA and is commonly known as the small sample size (SSS) problem (Fukunaga, 1990).

Over last several years, the discriminant analysis research is centered on developing algorithms that can solve SSS problem. In this overview, we focus on the LDA based techniques that can solve SSS problems. For brevity, we refer these techniques as LDA-SSS techniques. We provide taxonomy, characteristics and usage of these LDA-SSS techniques. The objective is to make the readers aware of the benefits and importance of these methods in the pattern classification applications. In addition, we have also highlighted some existing software/packages or programs useful for the LDA-SSS problem and mentioned about some of the commonly used datasets. Since these packages are not available from one place, we have developed Matlab functions for various LDA-SSS methods and it can be downloaded from our website (*<http: link will be provided upon the acceptance of the paper*>).

Linear discriminant analysis

As mentioned earlier, the LDA technique finds an orientation $\mathbf{W}$ that reduces high dimensional feature vectors belonging to different classes to a lower dimensional feature space such that the projected feature vectors of a class on this lower dimensional space are well separated from the feature vectors of other classes. This technique is illustrated in Figure 1, where two-dimensional feature vectors are reduced to one-dimensional feature vector. The feature vectors belong to three different classes namely C1, C2 and C3. An orientation is to be found where the projected feature vectors (on a line) of a class are to be maximally separated from the feature vectors of other classes. It can be observed that orientation $\hat{\mathbf{W}}$ does not separate projected feature vectors quite well. However, rotating the line further to orientation $\mathbf{W}$ and projecting two-dimensional feature vectors on this orientation separate the projected feature vectors of a class with other classes. Thus, the orientation $\mathbf{W}$ is a better selection than the orientation $\hat{\mathbf{W}}$. The value of $\mathbf{W}$ can be obtained by maximizing the Fisher's criterion function $J(\mathbf{W})$. This criterion function depends on three factors: orientation $\mathbf{W}$, within-class scatter matrix ($\mathbf{S}_W$) and between-class scatter matrix ($\mathbf{S}_B$). If the dimensionality reduction is from d-dimensional space to h-dimensional space, then the size of orientation matrix $\mathbf{W}$ is $d \times h$, and $\mathbf{W}$ has $h \leq \min(c-1, d)$ (where c is the number of classes) column vectors known as the basis vectors.

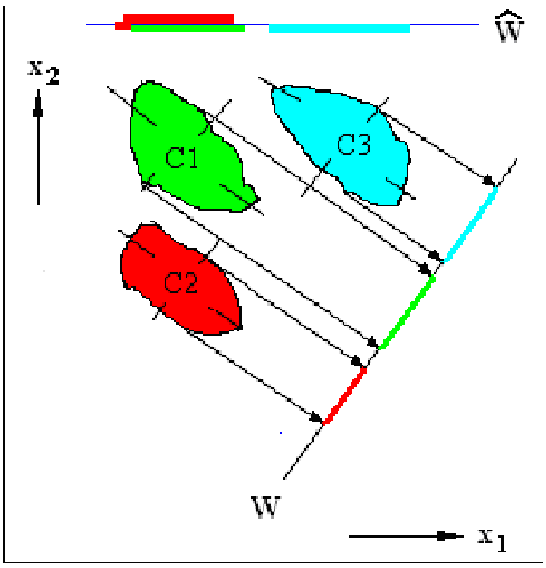

Figure 1: An illustration of LDA technique

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

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

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

If $\boldsymbol{\mu}_j$ is the centroid of $\mathbf{X}_j$ and $\boldsymbol{\mu}$ is the centroid of $\mathbf{X}$, then the total scatter matrix $\mathbf{S}_T \in \mathbb{R}^{d\times d}$, within-class scatter matrix $\mathbf{S}_W \in \mathbb{R}^{d\times d}$ and between-class scatter matrix $\mathbf{S}_B \in \mathbb{R}^{d\times d}$ are defined as (Duda and Hart, 1973, Sharma and Paliwal, 2006; 2008b)

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

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

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

where $\mathbf{S}_T = \mathbf{S}_B + \mathbf{S}_W$. The Fisher's criterion as a function of $\mathbf{W}$ can be given as

$$J(\mathbf{W}) = |\mathbf{W}^{\mathrm{T}}\mathbf{S}_B\mathbf{W}| / |\mathbf{W}^{\mathrm{T}}\mathbf{S}_W\mathbf{W}|$$

where $|\cdot|$ is the determinant. The orientation matrix $\mathbf{W}$ is the solution of eigenvalue problem

$$\mathbf{S}_W^{-1}\mathbf{S}_B\mathbf{w}_i = \lambda_i \mathbf{w}_i$$

where $\mathbf{w}_i$ (for $i = 1 \ldots h$) are the column vectors of $\mathbf{W}$ that correspond to the largest eigenvalues (λ_i). There are several other criterion function also used which provide equivalent results (Fukunaga, 1990).

In the conventional LDA technique, $\mathbf{S}_W$ needs to be non-singular. However, in the SSS case, this scatter matrix becomes singular. To overcome this problem, various LDA-SSS methods have been proposed in the literature. The next section discusses variants of LDA technique.

Variants of LDA technique (LDA-SSS) for solving SSS problem

In LDA-SSS, there are four informative spaces namely, null space of $\mathbf{S}_w$ ($\mathbf{S}_w^{null}$), range space of $\mathbf{S}_W$ ($\mathbf{S}_W^{range}$), range space of $\mathbf{S}_B$ ($\mathbf{S}_B^{range}$) and null space of $\mathbf{S}_B$ ($\mathbf{S}_B^{null}$). The computations of these spaces are very expensive and different methods use different strategies to tackle the computational problem. A popular way of reducing the computational complexity is by doing a preprocessing step. The preprocessing step is described as follows. It is known that the null space of $\mathbf{S}_T$ does not contain any discriminant information (Huang et al., 2002). 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, $\mathbf{U}_1 \in \mathbb{R}^{d \times r_t}$, is used as a transformation matrix. In the reduced dimensional space the scatter matrices is 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.

These four informative spaces are illustrated in Figure 2 after carrying out the preprocessing step[1]; i.e., the data is first transformed to the range space of $\mathbf{S}_T$. Let the transformed spaces are depicted by $\mathbf{S}_w^{null}$, $\mathbf{S}_w^{range}$, $\mathbf{S}_b^{null}$ and $\mathbf{S}_b^{range}$. In Figure 2, the symbols r_w, r_b and r_t are the rank of matrices $\mathbf{S}_W$, $\mathbf{S}_B$ and $\mathbf{S}_T$, respectively. If the samples in training set are linearly independent then $r_t = r_w + r_b$ and their values will be $r_t = n - 1$, $r_w = n - c$ and $r_b = c - 1$. Further, the dimensionality of spaces $\mathbf{S}_w^{null}$ and $\mathbf{S}_b^{range}$ will be identical. Similarly, the dimensionality of spaces $\mathbf{S}_w^{range}$ and $\mathbf{S}_b^{null}$ will be identical.

These four individual spaces contain significant discriminant information useful for classification. This is illustrated in Figure 3[2] where the classification performance obtained by the individual spaces is shown. Among these spaces, $\mathbf{S}_b^{null}$ is the least effective space, but it still contains some discriminant information. Different combinations of these spaces are used in the literature for finding the orientation

---

[1] These four spaces can also be represented in Figure 2 without performing a preprocessing step. In that case, r_t in the figure will be replaced by the dimensionality d and the size of the spaces will change accordingly.

[2] For this experiment, first we project the original feature vectors onto the range space of $\mathbf{S}_T$ matrix as a pre-processing step. Then all the spaces are utilized individually to do dimensionality reduction and to classify a test feature vector, the nearest neighbor classifier is used. To obtain performance in terms of average classification accuracy, k-fold cross-validation process has been applied, where $k = 5$. The details of the datasets have been given later in Section 'Datasets'.

matrix $\mathbf{W}$. The following four combinations have been used most in the literature: 1) $\mathbf{S}_W^{range}$ and $\mathbf{S}_B^{range}$, 2) $\mathbf{S}_W^{null}$ and $\mathbf{S}_B^{range}$, 3) $\mathbf{S}_W^{null}$, $\mathbf{S}_W^{range}$ and $\mathbf{S}_B^{range}$, and 4) $\mathbf{S}_W^{null}$, S_W^{range}, $\mathbf{S}_B^{range}$ and S_B^{null}. Based on these distinct combinations, we categorize the following LDA-SSS techniques into one of the four categories: null LDA (NLDA) (Chen et al., 2000), PCA + NLDA (Huang et al., 2002), orthogonal LDA (OLDA) (Ye 2005), uncorrelated LDA (ULDA) (Ye et al., 2004), QR-NLDA (Chu and Thye, 2010), fast NLDA (FNLDA) (Sharma and Paliwal, 2012a), discriminant common vector LDA (CLDA) (Cevikalp et al., 2005), direct LDA (DLDA) (Yu and Yang, 2001), kernel DLDA (KDLDA) (Lu et al., 2003a), parameterized DLDA (PDLDA) (Song et al., 2007), improved DLDA (IDLDA) (Paliwal and Sharma, 2010), pseudoinverse LDA (PILDA) (Tian et al., 1986), fast PILDA (FPILDA) (Liu et al., 2007), improved PILDA (IPILDA) (Paliwal and Sharma, 2012), LDA/QR (Ye and Li, 2005), approximate LDA (ALDA) (Paliwal and Sharma, 2011), PCA+LDA (Swets and Weng, 1996; Belhumer et al., 1997), regularized LDA (RLDA) (Friedman, 1989; Lu et al., 2003b, 2005; Zhao et al., 1998, 1999, 2003), eigenfeature regularization (EFR) (Jiang et al., 2008), extrapolation of scatter matrices (ELDA) (Sharma and Paliwal, 2010), maximum uncertainty LDA (MLDA) (Thomaz et al., 2005), penalized LDA (PLDA) (Hastie et al., 1995), two-stage LDA (TSLDA) (Sharma and Paliwal, 2012b), maximum margin criterion LDA (MMC-LDA) (Li et al.,

2003) and improved RLDA (IRLDA) (Sharma et al., 2013).

The classification accuracies of several of these methods have been computed on three datasets (for description of datasets please refer to Section Datasets) and 2-fold cross-validation results are shown in Table 1 and their average classification performance over 3 datasets is shown in Figure 4.. The nearest neighbor classifier has been used for classification purpose.

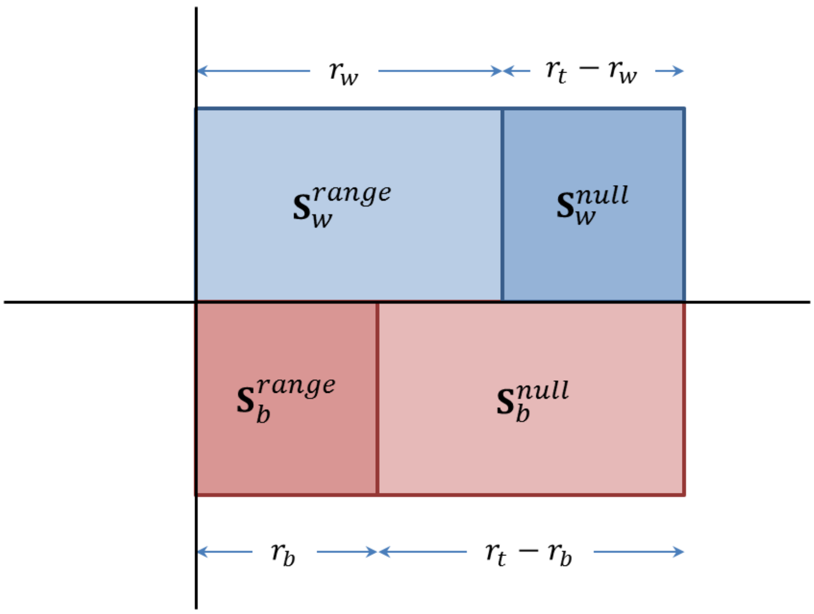

Figure 2: An illustration of all the four spaces of LDA when SSS problem exist.

Table 2 shows the categorization (or taxonomy) of these LDA-SSS methods. It should be noted that different LDA-SSS techniques use different combinations of spaces and the performance of a given LDA-SSS technique depends on the particular combination it uses. In addition, it depends in what manner these spaces are combined. Four categories are depicted (types 1-4). Most of the techniques fall under the first three categories. The fourth category (type-4) has not been fully explored in the literature. Figure 5 depicts average classification performance of all types over 3 face recognition datasets. Further characterization of these categories is discussed in the following subsections.

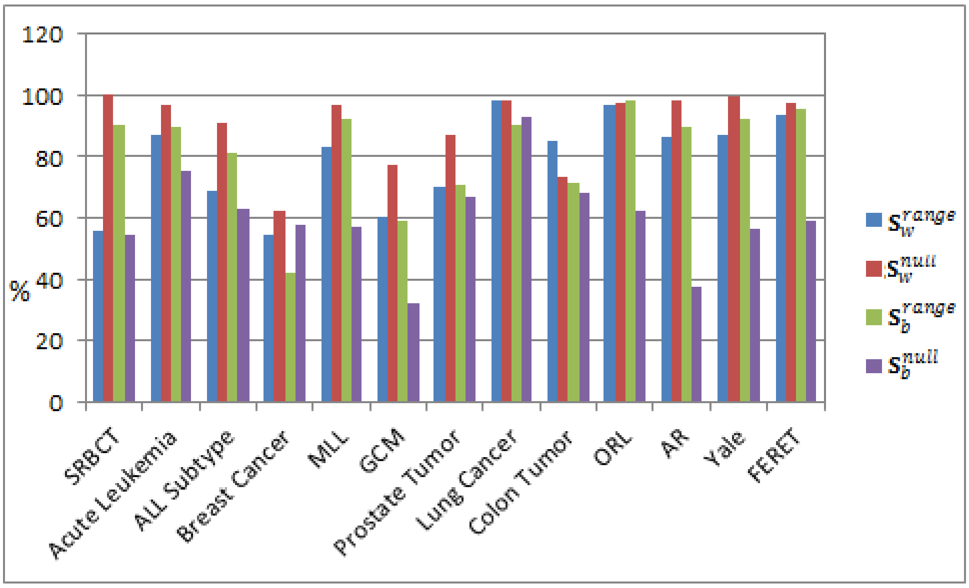

Figure 3: Average classification accuracy over k-fold cross-validation $(k = 5)$ using spaces $\mathbf{S}_w^{null}$, S_w^{range}, $\mathbf{S}_b^{range}$ and S_b^{null}.

Table 1: Classification accuracies (in percentage) of several LDA based techniques.

Techniques	ORL	AR	FERET	*Average*
DLDA	89.5	80.8	92.9	87.7
OLDA	91.5	80.8	97.1	89.8
PCA+LDA	86.0	83.4	95.7	88.4
RLDA	91.5	75.4	97.3	88.1
MLDA	92.0	76.2	97.8	88.7
EFR	92.3	81.8	97.7	90.6
TSLDA	92.3	87.7	97.7	92.6
PILDA	91.0	82.1	96.1	89.7
FPILDA	91.0	82.1	96.1	89.7
NLDA	91.5	80.8	97.1	89.8

ULDA	88.3	89.6	97.1	91.7
QR-NLDA	91.5	80.8	97.1	89.8
FNLDA	91.5	80.8	97.1	89.8
CLDA	91.5	80.8	97.1	89.8
IPILDA	87.5	87.9	97.1	90.8
ELDA	90.8	87.0	97.1	91.6
ALDA	91.3	72.1	96.7	86.7
IDLDA	91.5	72.7	96.9	87
IRLDA	92.0	81.9	97.7	90.5

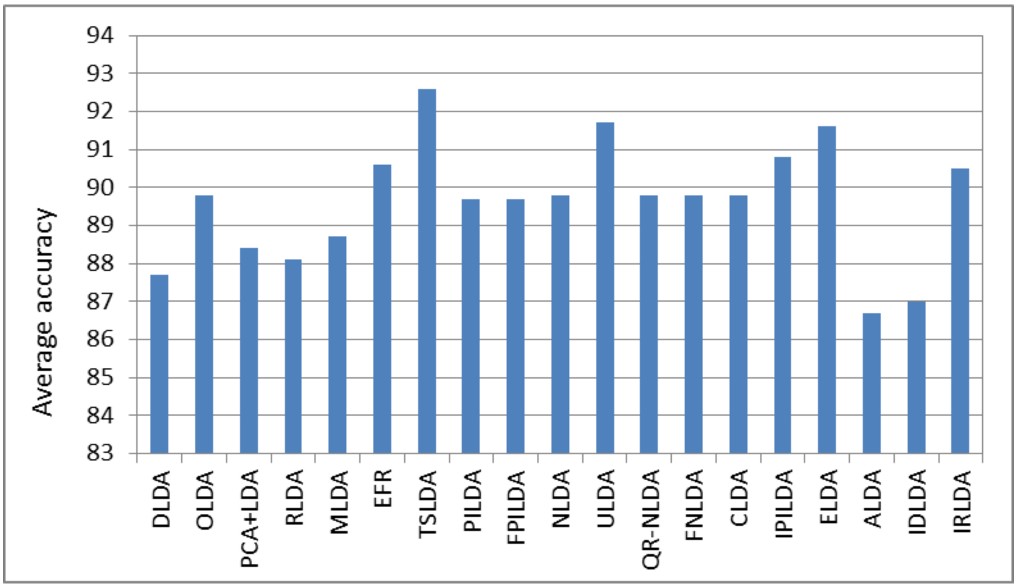

Figure 4: Average classification accuracies (in percentage) of several LDA based techniques over 3 datasets.

Table 2: Taxonomy for LDA based algorithms used for solving SSS problem

TYPE-1 $\mathbf{S}_W^{range} + \mathbf{S}_B^{range}$	TYPE-2 $\mathbf{S}_W^{null} + \mathbf{S}_B^{range}$	TYPE-3 $\mathbf{S}_W^{null} + \mathbf{S}_W^{range} + \mathbf{S}_B^{range}$	TYPE-4 All spaces
DLDA	NLDA	RLDA	TSLDA
KDLDA	PCA+NLDA	ALDA	
PDLDA	OLDA	EFR	
PILDA	ULDA	ELDA	
FPILDA	QR-NLDA	MLDA	
LDA/QR	FNLDA	IDLDA	
PCA+LDA	CLDA	PLDA	
MMC-LDA	IPILDA	IRLDA	

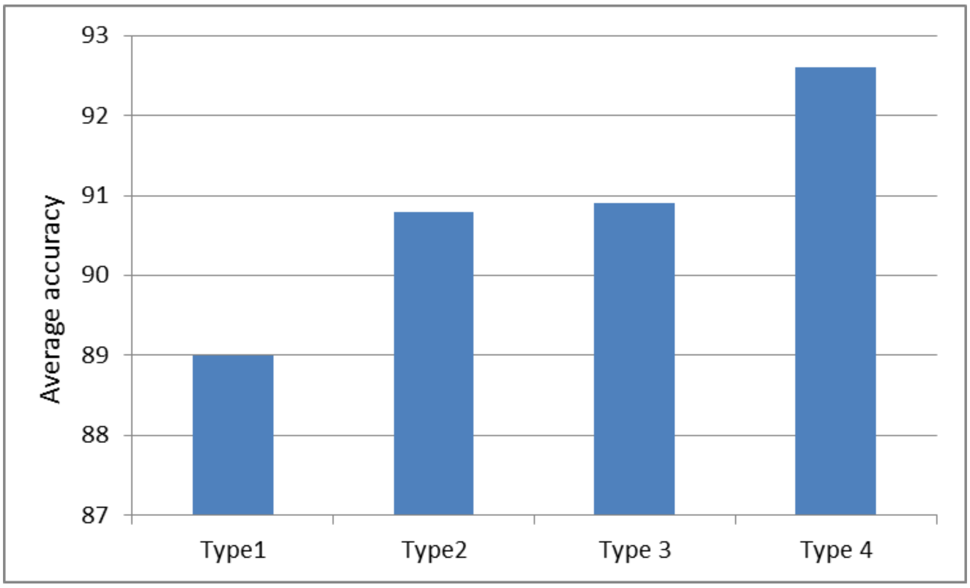

Figure 5: Average classification accuracy of best 3 methods of a particular type over three face recognition datasets (ORL, AR and FERET). (For Type 4 only 1 method has been selected). It can be observed that as the type increases the average performance improves. However, the improvement is based on how effectively the spaces are utilized in the computation of the orientation matrix.

Type-1 based techniques

LDA-SSS techniques of type-1 category employ $\mathbf{S}_W^{range}$ and $\mathbf{S}_B^{range}$ spaces to compute the orientation matrix $\mathbf{W}$ and therefore discard $\mathbf{S}_W^{null}$ and $\mathbf{S}_B^{null}$. This could, however, affect the classification performance adversely as the discarded spaces have significant discrimination information. Some of these methods compute $\mathbf{W}$ in two stages (e.g. DLDA) and some in one stage (e.g. PILDA). In general, type-1 methods are economical in computing the orientation matrix. However, their performances are not as good as that of other types of methods.

Type-2 based techniques

Techniques in this category utilize $\mathbf{S}_W^{null}$ and $\mathbf{S}_B^{range}$ spaces and discard the other two spaces. It has seen empirically (in Figure 3) that for most of the datasets, $\mathbf{S}_W^{null}$ contains more discriminant information than other spaces for classification performance. Therefore, employing $\mathbf{S}_W^{null}$ in a discriminant technique would enable to compute better orientation matrix $\mathbf{W}$ compared to Type-1 based techniques. However, since these techniques discard the other two spaces, its classification performance is suboptimal. The theory of many of these techniques are different, but they produce almost similar performance in terms of classification accuracy. The computational complexity of some of the type-2 methods is high. Nonetheless, they show encouraging classification performances.

Type-3 based techniques

To compute the orientation matrix $\mathbf{W}$, the techniques in this category utilize the three spaces; i.e., $\mathbf{S}_W^{null}$, $\mathbf{S}_W^{range}$ and $\mathbf{S}_B^{range}$. All the three spaces contain significant discrimination information and since Type-3 techniques employ more spaces than the previous two categories (Type-1 and Type-2), intuitively it would give a better classification performance. However, different strategies of combining these three spaces would result in different level of generalization capability. These methods require higher computational complexity. But produce encouraging performance if all the three spaces are effectively utilized.

Type-4 based techniques

It has been seen (in Figure 3) that though $\mathbf{S}_B^{null}$ is the least effective space, it still contains some discrimination information useful for classification. If $\mathbf{S}_B^{null}$ can also be used appropriately with the other spaces for the computation of orientation matrix $\mathbf{W}$, then classification performance can be further improved. So far very few techniques have been explored in this category. The computational complexity in this category is very high but they can produce good classification performance provided that all the spaces are utilized effectively.

This section illustrated the four informative spaces for solving SSS problem. Based on the utilization of different spaces, various techniques can be categorized into 4 types. However, it is possible that performance of techniques in a given type can vary. This is because various techniques (of a particular type) apply the spaces for computing the orientation matrix in different ways. Therefore, how effectively spaces are utilized can vary the performance of techniques (this can be observed from Table 1 where techniques of a particular type vary in performances). Nonetheless, in general utilizing spaces effectively would improve the performance (as shown in Figure 5 for best 3 methods).

Review of LDA based techniques for solving SSS problem

In this section, we review some of the common LDA based techniques for solving SSS problem. 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, approximation of inverse of $\mathbf{S}_W$ matrix has been used to compute the orientation matrix $\mathbf{W}$. There are various techniques to compute this inverse in the literature in different ways. Here we review some of the techniques:

<u>Fisherface (PCA+LDA) technique</u>

In Fisherface method, d-dimensional features are firstly reduced to h-dimensional feature space by the application of PCA and then LDA is applied to further reduce features to k dimensions. There are several criteria for determining the value of h (Belhumeur et al., 1997; Swets and Weng, 1996). One way is to select $h = n - c$ as the rank of $\mathbf{S}_W$ is $n - c$ (Belhumeur et al., 1997). The advantage of this method is that it overcome SSS problem. However, the drawback is that some discriminant information has been lost in the PCA application to $n - c$ dimensional space.

<u>Direct LDA</u>

Direct LDA (DLDA) is an important dimensionality reduction technique for solving small sample size problem (Yu and Yang, 2001). In the DLDA method, 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 $\mathbf{S}_B$; i.e.,

$$\mathbf{U}_r^{\mathrm{T}}\mathbf{S}_B\mathbf{U}_r = \mathbf{\Lambda}_{\mathrm{B}}^2,$$

or $\mathbf{\Lambda}_B^{-1}\mathbf{U}_r^{\mathrm{T}}\mathbf{S}_B\mathbf{U}_r\mathbf{\Lambda}_B^{-1} = \mathbf{I}_{b\times b}$,

where $\mathbf{U}_r$ corresponds to the range space of $\mathbf{S}_B$ (i.e., $\mathbf{\Lambda}_B$) and $b = rank(\mathbf{S}_B)$.

In the second stage, the dimensionality of this transformed samples is further transformed by some regulating matrices; i.e., the transformation matrix $\mathbf{U}_r\mathbf{\Lambda}_B^{-1}$ is used to transform $\mathbf{S}_W$ matrix as

$$\hat{\mathbf{S}}_W = \mathbf{\Lambda}_B^{-1}\mathbf{U}_r^{\mathrm{T}}\mathbf{S}_W\mathbf{U}_r\mathbf{\Lambda}_B^{-1} = \mathbf{F}\mathbf{\Sigma}_w^2\mathbf{F}^{\mathrm{T}},$$

or $\mathbf{\Sigma}_w^{-1}\mathbf{F}^{\mathrm{T}}\mathbf{\Lambda}_B^{-1}\mathbf{U}_r^{\mathrm{T}}\mathbf{S}_W\mathbf{U}_r\mathbf{\Lambda}_B^{-1}\mathbf{F}\mathbf{\Sigma}_w^{-1} = \mathbf{I}_{b\times b}$

Therefore, the orientation matrix of DLDA technique can be given as $\mathbf{W} = \mathbf{U}_r\mathbf{\Lambda}_B^{-1}\mathbf{F}\mathbf{\Sigma}_w^{-1}$. The benefit of DLDA technique is that it does not require PCA transformations to reduce the dimensionality as required by other techniques like Fisherface (or PCA+LDA) technique (Swets and Weng, 1996; Belhumeur et al., 1997).

Regularized LDA

When the dimensionality of feature space is very large compared to the number of training samples available, then the $\mathbf{S}_W$ matrix becomes singular. To overcome this singularity problem in the regularized LDA (RLDA) method, a small perturbation to the $\mathbf{S}_W$ matrix has been added (Friedman, 1989; Zhao et al., 1999; Dai and Yuen, 2007). This makes the $\mathbf{S}_W$ matrix non-singular. The regularization can be applied as follows:

$$(\mathbf{S}_W + \delta\mathbf{I})^{-1}\mathbf{S}_B\mathbf{w}_i = \lambda_i\mathbf{w}_i$$

where $\delta > 0$ is a perturbation term or regularization parameter. The addition of δ in the regularized method helps to incorporate both the null space and range space of $\mathbf{S}_W$. However, the drawback is that there is no direct way of evaluating the parameter as it requires heuristic approaches to evaluate it and a poor choice of δ can degrade the generalization performance of the method. The parameter δ has been added just to perform the inverse operation feasible and it has no physical meaning.

<u>Null LDA technique</u>

In the null LDA (NLDA) technique (Chen et al., 2000), the h column vectors of the orientation $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_h]$ are taken to be in the null space of the within-class scatter matrix $\mathbf{S}_W$; i.e., $\mathbf{w}_i^T\mathbf{S}_W\mathbf{w}_\mathrm{i} = 0$ for $i = 1 \ldots h$. In addition, these column vectors have to satisfy the condition $\mathbf{w}_i^T\mathbf{S}_B\mathbf{w}_\mathrm{i} \neq 0$ for $i = 1 \ldots h$.

Since the dimensionality of the null space of $\mathbf{S}_W$ is $d - (n - c)$, we will have $d - (n - c)$

linearly independent vectors satisfying the two above mentioned conditions. Since $d-(n-c)$ is greater than h, Chen et al. (2000) have used eigen analysis of $\mathbf{S}_B$ matrix to select h leading eigenvectors from these $d-(n-c)$ vectors to form the orientation matrix $\mathbf{W}$. Thus, in the null space method $\mathbf{W}$ is found by maximizing $|\mathbf{W}^{\mathrm{T}}\mathbf{S}_B\mathbf{W}|$ subject to the constraint $|\mathbf{W}^{\mathrm{T}}\mathbf{S}_W\mathbf{W}| = 0$, i.e.,

$$\mathbf{W} = \arg\max_{|\mathbf{W}^{\mathrm{T}}\mathbf{S}_W\mathbf{W}|=0} |\mathbf{W}^{\mathrm{T}}\mathbf{S}_B\mathbf{W}|$$

The null LDA technique finds the orientation $\mathbf{W}$ in two stages. In the first stage, it computes $\mathbf{W}$ such that $\mathbf{S}_W\mathbf{W} = 0$: i.e., data is projected on the null space of $\mathbf{S}_W$ and throws the range space of $\mathbf{S}_W$. Then in the second stage it finds $\mathbf{W}$ that satisfies $\mathbf{S}_B\mathbf{W} \neq 0$ and maximizes $|\mathbf{W}^{\mathrm{T}}\mathbf{S}_B\mathbf{W}|$. The second stage is commonly implemented through the PCA method applied on $\mathbf{S}_B$. When the dimensionality d of the original feature space is very large in comparison to sample size n, the evaluation of null space becomes nearly impossible as the eigenvalue decomposition of such a large $d \times d$ matrix will lead to serious computational problems. This is a major problem. There are two main techniques in this respect suggested in the literature for computing the orientation $\mathbf{W}$. In the first technique, a pre-processing step is introduced where the PCA technique is applied to reduce the dimensionality from d to $n-1$ by removing the null space of $\mathbf{S}_T$. In the reduced $n-1$ dimensional space it is possible to compute the null space of $\mathbf{S}_W$. This pre-processing step is then followed by the two steps of the null space LDA method (Huang et al., 2002). In the second technique, no pre-processing is necessary but the required null space of $\mathbf{S}_W$ is computed in the first stage by first finding the range space of $\mathbf{S}_W$, then projecting the data onto this range space followed by subtracting it from the original data. After this step, the PCA method is applied to carry out the second stage. It can be seen that in both the techniques range space of $\mathbf{S}_W$ was thrown which could have some discriminant information for classification.

<u>Orthogonal LDA</u>

Orthogonal LDA (OLDA) method (Ye, 2005) has shown to be equivalent to the null LDA method under a mild condition; i.e., when the training vectors are linearly independent (Ye and Xiong, 2006). In his method, the orientation matrix $\mathbf{W}$ is obtained by simultaneously diagonalizing scatter matrices. Therefore, a matrix A_1 can be found which diagonalizes all scatter matrices; i.e.,

$\mathbf{A}_1^{\mathrm{T}}\mathbf{S}_B\mathbf{A}_1 = \mathbf{\Sigma}_B$, $\mathbf{A}_1^{\mathrm{T}}\mathbf{S}_W\mathbf{A}_1 = \mathbf{\Sigma}_W$ and $\mathbf{A}_1^{\mathrm{T}}\mathbf{S}_T\mathbf{A}_1 = \mathbf{I}_T$,

where $\mathbf{A}_1 = \mathbf{U}_1\mathbf{\Sigma}_T^{-1}\mathbf{P}$, $\mathbf{U}_1$ is range space of $\mathbf{S}_T$, $\mathbf{\Sigma}_T$ is eigenvalues of $\mathbf{S}_T$ and $\mathbf{\Sigma}_T^{-1}\mathbf{U}_1^{\mathrm{T}}\mathbf{H}_B = \boldsymbol{P\Sigma Q}^{\mathrm{T}}$ ($\mathbf{S}_B = \mathbf{H}_B\mathbf{H}_B^{\mathrm{T}}$). The orientation matrix $\mathbf{W}$ can be found by orthogonalizing matrix $\mathbf{A}_1$; i.e., $\mathbf{A}_1 = \mathbf{QR}$, where $\mathbf{W} = \mathbf{Q}$.

In this method, the dimensionality is reduced from $\mathbb{R}^d$ to $\mathbb{R}^{c-1}$. The computational complexity of OLDA method is better than null LDA method and is estimated to be $14dn^2 + 4dnc + 2dc^2$ flops (where c is the number of classes).

QR-NLDA

Chu and Thye (2010) proposed a new implementation of null LDA method by doing QR decomposition. This is faster method than OLDA. Their approach requires approximately $4dn^2 + 2dnc$ computations.

Fast NLDA

Fast NLDA (FNLDA) method (Sharma and Paliwal, 2012a) is an alternative method of NLDA. It assumes that the training vectors are linearly independent. In this method, 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 $c-1$. This method is so far the fastest method of performing null LDA operation. The fast computation is achieved by using random matrix multiplication with scatter matrices. The computational complexity of FNLDA is $dn^2 + 2dnc$.

Pseudoinverse method

In the pseudoinverse LDA (PILDA) method (Tian et al., 1986), the inverse of within-class scatter matrix $\mathbf{S}_W$ is estimated by its pseudoinverse and then the conventional eigenvalue problem is solved to compute the orientation matrix $\mathbf{W}$. In this method, a pre-processing step is used where feature vectors are projected on the range space of $\mathbf{S}_T$ to reduce the computational complexity (Huang et al., 2002). After the pre-processing step, the reduced dimensional within-class scatter matrix $\hat{\mathbf{S}}_W$ is decomposed as

$\hat{\mathbf{S}}_W = \mathbf{U}_w \mathbf{D}_w^2 \mathbf{U}_w^{\mathrm{T}}$, where $\mathbf{U}_w \in \mathbb{R}^{t\times t}, \mathbf{D}_w \in \mathbb{R}^{t\times t}, t = rank(\mathbf{S}_T)$

$\mathbf{D}_w = \begin{bmatrix} \mathbf{\Lambda}_w & 0 \\ 0 & 0 \end{bmatrix}$ and $\mathbf{\Lambda}_w \in \mathbb{R}^{w\times w}$ (w is the rank of $\mathbf{S}_W$ such that $w < t$).

Let the eigenvectors corresponding to the range space of $\hat{\mathbf{S}}_W$ is $\mathbf{U}_{wr}$ and the eigenvectors corresponding to the null space of $\hat{\mathbf{S}}_W$ is $\mathbf{U}_{wn}$, i.e., $\mathbf{U}_w = [\mathbf{U}_{wr}, \mathbf{U}_{wn}]$, then the pseudoinverse of $\mathbf{S}_W$ can be expressed as

$$\hat{\mathbf{S}}_W^{+} = \mathbf{U}_{wr} \mathbf{\Lambda}_w^{-2} \mathbf{U}_{wr}^{\mathrm{T}}$$

The orientation matrix $\mathbf{W}$ can now be computed by solving the following conventional eigenvalue problem

$$\hat{\mathbf{S}}_W^+ \hat{\mathbf{S}}_B \mathbf{w}_i = \lambda_i \mathbf{w}_i$$

where $\hat{\mathbf{S}}_B$ is the between-class scatter matrix in the reduced space. It can be observed that the null space of within-class scatter matrix is discard which would sacrifice some discriminant information.

Eigenfeature regularization

In eignefeature regularization (EFR) method (Jiang et al., 2008), $\mathbf{S}_W$ 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. Since the extrapolation has been done by an estimation function, it cannot be guaranteed to be optimal in dimensionality reduction.

Extrapolation LDA

In extrapolation LDA (ELDA) method (Sharma and Paliwal, 2010), the null space of $\mathbf{S}_W$ matrix is regularized by extrapolating eigenvalues of $\mathbf{S}_W$ using exponential fitting function. This method utilizes range space information and null space information of $\mathbf{S}_W$ matrix.

Maximum uncertainty LDA

The maximum uncertainty LDA (MLDA) method is based on maximum entropy covariance selection approach that overcomes singularity and instability of $\mathbf{S}_W$ matrix (Thomaz and Gillies, 2005). The MLDA is constructed by replace $\mathbf{S}_W$ with its estimate in the Fisher criterion function. This is computed by updating less reliable eigenvalues of $\mathbf{S}_W$.

Two stage LDA

The two stage LDA (TSLDA) method (Sharma and Paliwal, 2012b) exploits all the 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. It has been shown that all four spaces contain some discriminant information which is useful for classification.

Applications of the LDA-SSS techniques

In many applications the number of features or dimensionality is much larger than the number of training samples. In these applications, LDA-SSS techniques have been successfully applied. Some of the applications of LDA-SSS techniques are described as follows:

Face recognition

Face recognition system comprises of two main steps: feature extraction (including face detection) and face recognition (Zhao et al., 2003; Sanderson and Paliwal, 2003). In feature extraction step, an image of a face (of size $m \times n$) is normally represented by the illumination levels of $m \times n$ pixels (giving a feature vector of dimensionality $d = mn$) and in the recognition step an unknown face image is identified/verified. Several LDA-SSS techniques have been applied for this application (e.g. Swets and Weng, 1996; Zhao et al., 1998, 1999).

Cancer classification

The DNA microarray data for cancer classification consists of large number of genes (dimensions) compared to the number of tissue samples or feature vectors. The high dimensionality of the feature space degrades the generalization performance of the classifier and increases its computational complexity. This situation, however, can be overcome by first reducing the dimensionality of feature space, followed by classification in the lower-dimensional feature space. Different methods used for dimensionality reduction can be grouped into two categories: feature selection methods and feature extraction methods. Feature selection methods (e.g. Golub et al., 1999; Furey et al., 2000; Mak and Kung, 2006; Cui et al., 2010; Sharma et al., 2011, 2012d, 2012e, 2012f) retain only a few useful features and discard others. Feature extraction methods construct a few features from the large number of original features through their linear (or nonlinear) combination. A number of papers have been reported for the cancer classification task using the microarray data (Dudoit et al., 2002; Li et al., 2003, 2005; Moghaddam et al., 2006; Sharma and Paliwal, 2008a).

Text document classification

In the text document classification, a free text document is categorized to a pre-defined category based on its contents (Aas and Eikvil, 1999). The text document is a collection of words. To represent a given text document as a feature vector, a finite dictionary of words is chosen and frequencies of these words (e.g. monogram, bigram etc.) are used as features. Dimensionality reduction and classification techniques are applied for the categorization of a document. The LDA-SSS techniques have also been applied to text document classification (e.g. Ye, 2005; Ye and Xiong, 2006).

Datasets

In this section we cover some of the commonly used datasets for LDA related methods. Three types of datasets have been depicted. These are face recognition data, DNA microarray gene expression data and text data. The description of datasets is given in Table 4[3].

Table 4: Description of datasets

Dataset	Description
Face recognition	
AR (Martinez, 2002)	Contains over 4000 color images of 126 people's faces (70 men and 56 women). Images are with frontal illumination, occlusions and facial expressions. http://www2.ece.ohio-state.edu/~aleix/ARdatabase.html
ORL (Samaria and Harter, 1994	Contains 400 images of 40 people having 10 images per subject. The images were taken at different times, varying the lighting, facial expressions (open / closed eyes, smiling / not smiling) and facial details (glasses / no glasses). http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html
FERET (Phillips et al., 2000)	Contains 14126 images from 1199 individuals. Images of human heads with views ranging from frontal to left and right profiles. http://www.itl.nist.gov/iad/humanid/feret/feret_master.html
Yale (Belhumeur et al. 1997)	Contains 165 images of 15 subjects. There are 11 images per subject, one for each of the following facial expressions or configurations: center-light, with glasses, happy, left-light, with no glasses, normal, right-light, sad, sleepy, surprised and wink. http://cvc.yale.edu/projects/yalefaces/yalefaces.html
Cancer classification	
Acute leukemia (Golub et al., 1999)	Consists of DNA microarray gene expression data of human acute leukemias for cancer classification. Two types of acute leukemias data are provided for classification namely acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML). The dataset is subdivided into 38 training samples and 34 test samples. The training set consists of 38 bone marrow samples (27 ALL and 11 AML) over 7129 probes. The testing set consists of 34 samples with 20 ALL and 14 AML, prepared under different experimental conditions. All the samples have 7129 dimensions and all are numeric.
ALL subtyoe (Yeoh et al. 2002)	Consists of 12558 genes of subtypes of acute lymphoblastic leukemia. The dataset is subdivided into 215 training samples and 112 testing samples. These train and test sets belong to seven classes namely T-ALL, E2A-PBX1, BCR-ABL, TEL-AML1, MLL, hyperdiploid >50 chromosomes and other (contains diagnostic samples that did not fit into any of the former six classes). The training samples per class are 28, 18, 9, 52, 14, 42 and 52 respectively. The test samples per class are 15, 9, 6, 27, 6, 22 and 27 respectively.
Breast cancer (van't Veer, 2002)	This is a 2 class problem with 78 training samples (34 relapse and 44 non-relapse) and 19 testing samples (12 relapse and 7 non-relapse) of relapse and non-relapse. The dimension of breast cancer dataset is 24481.
GCM (Ramaswamy et al., 2001)	This Global Cancer Map (GCM) dataset has 14 classes with 144 training samples and 46 testing samples. There are 16063 number of gene expression levels in this dataset.
MLL (Armstrong et al., 2002)	This dataset has 3 classes namely ALL, MLL and AML leukemia. The training data contains 57 leukemia samples (20 ALL, 17 MLL and 20 AML) whereas the testing data contains 15 samples (4 ALL, 3 MLL and 8 AML). The dimension of MLL dataset is 12582.
Lung adenocarcinoma (Beer et al., 2002)	Consists of 96 samples each having 7129 genes. This is a three class classification problem. Out of 96 samples, 86 are primary lung adenocarcinomas, including 67 stage I tumor and 19 stage III tumor. An addition of 10 non-neoplastic lung samples are provided.
Lung (Gordon et al., 2002)	Contains gene expression levels of malignant mesothelioma (MPM) and adenocarcinoma (ADCA) of the lung. There are 181 tissue samples (31 MPM and 150 ADCA). The training set contains 32

[3] For more datasets on face see Ralph Gross (2005), Zhao et al. (2003) and http://www.face-rec.org/databases/. For bio-medical data see Kent Ridge Bio-medical Repository (http://datam.i2r.a-star.edu.sg/datasets/krbd/).

	of them, 16 MPM and 16 ADCA. The rest of 149 samples are used for testing. Each sample is described by 12533 genes.
Prostate (Singh et al., 2002)	This is a 2-class problem with tumor class versus normal class. It contains 52 prostate tumor samples and 50 non-tumor samples (or normal). Each sample is described by 12600 genes. A separate test contains 25 tumor and 9 normal samples.
SRBCT (Khan et al., 2002)	Consists of 83 samples with each having 2308 genes. This is a four class classification problem. The tumors are Burkitt lymphoma (BL), the Ewing family of tumors (EWS), neuroblastoma (NB) and rhabdomyosarcoma (RMS). There are 63 samples for training and 20 samples for testing. The training set consists of 8, 23, 12 and 20 samples of BL, EWS, NB and RMS respectively. The testing set consists of 3, 6, 6 and 5 samples of BL, EWS, NB and RMS respectively.
Colon tumor (Alon et al., 1999)	Contains 2 classes of colon tumor samples. A total of 62 samples are given out of which 40 are tumor biopsies (labelled as 'negative') and 22 are normal (labelled as 'positive'). Each sample has 2000 genes. The dataset does not have separate training and testing sets.
Ovarian cancer (Petricoin III et al., 2002)	Contains 253 samples of ovarian cancer (162 samples) and non-ovarian cancer (91 samples). The dimension of feature vector is 15154. These 15154 identities are normalized prior to processing.
Central nervous system (Pomeroy et al., 2002)	This is a two class problem with 60 patient samples of central nervous system embryonal tumor. There are 21 survivors and 39 failures which contribute to 60 samples. There are 7129 genes of the samples of the dataset.
Lung cancer 2 (Bhattacharjee et al., 2001)	This is a 5-class problem with a total of 203 normal lung and snap-frozen lung tumors. The dataset includes 139 samples of lung adenocarcinoma, 20 samples of pulmonary carcinoids, 6 samples of small-cell lung carcinomas, 21 samples of squamous cell lung carcinomas and 17 normal lung samples. Each sample has 12600 genes.
Text document classification	
Reuters-21578 (Lewis, 1999)	Contains 22 files. The first 21 files contain 1000 documents and the last file contains 578 documents.
TREC (2000)	Large collection of text data.
Dexter (Blake and Merz, 1998)	Collection of text classification in a bag-of-word representation. This dataset has sparse continuous input variables.

Packages

In this section we list some of the packages available. This is shown in Table 5. We have also developed in our laboratory a LDA-SSS package (written in Matlab), which provides the Matlab functions for computing $\mathbf{S}_W^{null}$, $\mathbf{S}_W^{range}$, $\mathbf{S}_B^{range}$ and $\mathbf{S}_B^{null}$, and implementation of several LDA-SSS techniques such as DLDA, PILDA, FPILDA, PCA+LDA, NLDA, OLDA, ULDA, QR-NLDA, FNLDA, CLDA, IPILDA, ALDA, EFR, ELDA, MLDA, IDLDA and TSLDA.

Table 5: Packages

Code/package	Description
WEKA (Witten and Frank, 2000) (A Java based data mining tool with open source machine learning software)	Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from person's Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes. http:/www.cs.waikato.ac.nz/ml/weka/
LDA-SSS	This is Matlab based package, it contains several algorithms related to LDA-SSS. The following techniques/functions are in this package: DLDA, PILDA, FPILDA, PCA+LDA, NLDA, OLDA, ULDA, QR-NLDA, FNLDA, CLDA, IPILDA, ALDA, EFR, ELDA, MLDA, IDLDA and TSLDA *<LINK WILL BE PROVIDED UPON ACCEPTANCE OF THE PAPER>*
Dimensionality reduction techniques	This package is mainly written in Matlab. It includes a number of dimensionality reduction

	techniques listed as follows: multi-label dimensionality reduction, generalized low rank approximations of matrices, ULDA, OLDA, kernel discriminant analysis via QR (KDA/QR) and approximate KDA/QR. http://www.public.asu.edu/~jye02/Software/index.html
MASS package	The MASS package is based on R and contains functions for performing linear and quadratic discriminant function analysis. http://www.statmethods.net/advstats/discriminant.html
DTREG	DTREG is a tool for modeling business and medical data with categorical variables. This includes several predictive modeling methods (e.g., multilayer perceptron, probabilistic neural networks, LDA, PCA, factor analysis, linear regression, decision trees, SVM etc.) http://www.dtreg.com/index.htm
dChip	dChip software is for analysis and visualization of gene expression and SNP microarrays. This has interface with R software. It is capable of doing probe-level analysis, high-level analysis (including gene filtering, hierarchical clustering, variance and correlation analysis, classifying samples by LDA, PCA etc.) and SNP array analysis. https://sites.google.com/site/dchipsoft/home
XLSTAT	XLSTAT is data analysis and statistical solution for Microsoft Excel. The XLSTAT statistical analysis add-in offers a wide variety of functions (including discriminant analysis) to enhance the analytical capabilities of Excel for data analysis and statistics requirements. http://www.xlstat.com/en/

Conclusion

In this paper, we reviewed LDA-SSS algorithms for dimensionality reduction. Some of these algorithms provide the-state-of-the-art performance in many applications. We discuss and categorize LDA-SSS algorithms into 4 distinct categories based on the combination of spaces of scatter matrices. We have also highlighted some datasets and software/packages useful to investigate the SSS problem. The LDA-SSS package written in our laboratory has been made available (<*LINK will be provided*>).

Reference

Aas, K., Eikvil, L., Text categorization: a survey, Norwegian Computing Center Report NR 941, 1999.

Alon U., Barkai N., Notterman D.A., Gish K., Ybarra S., Mack D., Levine A.J., Broad patterns of gene expression revealed by clustering of tumor and normal colon tissues probed by oligonucleotide arrays, PNAS, vol. 96, pp 6745-6750, 1999.

Armstrong, S.A., Staunton, J.E., Silverman, L.B., Pieters, R., den Boer, M.L., Minden, M.D., Sallan, S.E., Lander, E.S., Golub, T.R., and Korsemeyer, S.J., MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. Nature

Genetics, vol. 30, pp 41-47, 2002.

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

Belhumeur, P.N. Hespanha J.P. and Kriegman, D.J. 'Eigenfaces vs. fisherfaces: recognition using class specific linear projection, IEEE Trans. Pattern Anal. Machine Intell., vol. 19, no. 7, pp. 711–720, 1997.

Bhattacharjee, A., Richards, W.G., Staunton, J., Li C., Monti, S., Vasa, P., Ladd, C., Beheshti, J., Bueno, R., Gillette, M., Loda, M., Weber, G., Mark, E.J., Lander, E.S., Wong, W., Johnson, B.E., Golub, T.R., Sugarbaker, D.J. and Meyerson, M., Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma sub-classes, PNAS, vol. 98, no. 24, pp 13790-13795, 2001.

Blake, C.L., Merz , C.J., UCI repository of machine learning databases, http://www.ics.uci.edu/_mlearn, Irvine, CA, University of Calif., Dept. of Information and Comp. Sci., 1998.

Cevikalp, H. Neamtu, M. Wilkes, M. A. Barkana, Discriminative common vectors for face recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 1, pp. 4–13, 2005.

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

Chu, D. Thye, G.S. A new and fast implementation for null space based linear discriminant analysis, Pattern Recognition, vol. 43, pp. 1373–1379, 2010.

Cui, X. Zhao, H., Wilson, J., Optimized Ranking and Selection Methods for Feature Selection with Application in Microarray Experiments, J. Biopharmaceutical Statistics, vol. 20, no. 2, pp. 223-239, 2010.

Dabney, A.R., Classification of microarrays to nearest centroids, Bioinformatics, vol. 21, no. 22, pp. 4148-4154, 2005.

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

Dudoit, S. Fridlyand, J. Speed, T.P., Comparison of discrimination methods for the classification of tumors using gene expression data, J. Am. Stat Assoc., vol. 97, no. 457, pp. 77-87, 2002.

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

Fukunaga, K. Introduction to Statistical Pattern Recognition. Academic Press Inc., Hartcourt Brace Jovanovich, Publishers. 1990.

Furey TS, Cristianini N, Du®y N, Bednarski DW, Schummer M, Haussler D, Support vector machine classi¯cation and validation of cancer tissue samples using microarray expression data, Bioinformatics, vol. 16, pp. 906-914, 2000.

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

Gordon, G.J., Jensen, R.V., Hsiao, L.-L., Gullans, S.R., Blumenstock, J.E., Ramaswamy, S., Richards, W.G., Sugarbaker, D.J., Bueno, R.: Translation of Microarray Data Into Clinically Relevant Cancer Diagnostic Tests Using Gene Expression Ratios in Lung Cancer and Mesothelioma, Cancer Research, vol. 62, pp 4963-4967, 2002.

Hastie, T. Buja, A., Tibshirani, R., Penalized discriminant analysis. *Annals of Statistics*, 23, pp. 73–102, 1995

Huang, R. Liu, Q., Lu, H., Ma, S., Solving the small sample size problem of LDA, Proceedings of ICPR 3 (2002) 29–32. 2002.

Jiang, X., Mandal, B., Kot, A., Eigenfeature regularization and extraction in face recognition. IEEE Trans. Patt. Anal. Machine Intell., vol. 30, no. 3, pp. 383–394, 2008.

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

Lewis, D.D., Reuters-21578 text categorization test collection distribution 1.0. http://www.daviddlewis.com/resources/testcollections/reuters21578/, 1999.

Li, H., Zhang, K., Jiang, T., Robust and accurate cancer classification with gene expression profiling, Proc. IEEE Comput. Syst. Bioinform. Conf., pp. 310-321, 2005.

Li, H. Jiang, T. Zhang, K., Efficient and robust feature extraction by maximum margin criterion, in: Advances in Neural Information Processing Systems, 2003.

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.

Lu, J., Plataniotis, K., Venetsanopoulos, A., Face recognition using kernel direct discriminant analysis algorithms. IEEE Trans. Neural Networks, vol. 14, no. 1, pp. 117–126, 2003a.

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

Lu, J. Plataniotis, K.N. Venetsanopoulos, A.N., "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.

Mak MW, Kung SY, A solution to the curse of dimensionality problem in pairwise scoring techniques, Int. Conf. on Neural Info. Process. (ICONIP'06), pp. 314-323, 2006.

Martinez, A.M., 2002. Recognizing imprecisely localized, partially occluded, and expression variant faces from a single sample per class. IEEE Trans. Pattern Anal. Machine Intell. 24 (6), 748–763.

Moghaddam, B., Weiss, Y., Avidan, S., Generalized spectral bounds for sparse LDA, Int. Conf. Mach. Learn., ICML'06, pp. 641-648, 2006.

Paliwal, K.K. and Sharma, A., Improved pseudoinverse linear discriminant analysis method for dimensionality reduction, International Journal of Pattern Recognition and Artificial Intelligence, vol. 26, no. 1, pp. 1250002-1 – 1250002-9, 2012.

Paliwal, K.K. and Sharma, A., 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.

Paliwal, K.K. and Sharma, A., Improved direct LDA and its application to DNA microarray gene expression data, Pattern Recognition Letters, vol. 31, pp. 2489-2492, 2010.

Petricoin III, E.F., Ardekani, A.M., Hitt, B.A., Levine, P.J., Fusaro, V.A., Steinberg, M.S., Mills, G.B., Simone, C., Fishman, D.A., Kohn, E.C. and Liotta, L.A., Use of proteomic patterns in serum to identify ovarian cancer, The LANCET, vol. 359, pp 572-577, 2002.

Phillips, P.J., Moon, H., Rauss, P.J., Rizvi, S., The FERET evaluation methodology for face recognition algorithms. IEEE Trans. Patt. Anal. Mach. Intell., vol. 22, no. 10, pp. 1090-1104, 2000.

Pomeroy, S.L., Tamayo, P., Gaasenbeek, M., Sturla, L.M., Angelo, M., McLaughlin, M.E., Kim, J.Y.H., Goumnerova, L.C., Black, P.McL., Lau, C., Allen, J.C., Zagzag, D., Olson, J.M., Curran, T., Wetmore, C., Biegel, J.A., Poggio, T., Mukherjee, S., Rifkin, R., Califano, A., Stolovitzky, G., Louis, D.N., Mesirov, J.P., Lander, E.S. and Golub, T.R., Gene expression-based classification and outcome prediction of central nervous system embryonal tumors, Nature, vol. 415, pp 436-442, 2002.

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

Samaria, F., Harter, A., Parameterization of a stochastic model for human face identification, In Proc. of the Second IEEE Workshop Appl. of Comp. Vis., pp. 138-142, 1994.

Sanderson, C., Paliwal, K.K., Fast features for face authentication under illumination direction changes, Pattern Recognition Letters, vol. 24, pp. 2409-2419, 2003

Sharma, A., Paliwal, K.K., 'Class-dependent PCA, LDA and MDC: a combined classifier for pattern classification', Pattern Recognition, vol. 39, issue 7, pp. 1215-1229, 2006.

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

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

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

Sharma, A., Imoto, S., Miyano, S., Sharma, V., "Null space based feature selection method for gene expression data", International Journal of Machine Learning and Cybernetics, 2011, DOI: 10.1007/s13042-011-0061-9.

Sharma, A., Paliwal, K.K., A new perspective to null linear discriminant analysis method and its fast implementation using random matrix multiplication with scatter matrices, Pattern Recognition, vol. 45, pp. 2205-2213, 2012a.

Sharma, A., Paliwal, K.K., A two-stage linear discriminant analysis for face-recognition, Pattern Recognition Letters, vol. 33, pp. 1157-1162, 2012b.

Sharma, A., Imoto, S., Miyano, S., "A top-r feature selection algorithm for microarray gene expression data", IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 9, no. 3, pp. 754-764, 2012d.

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, 2012e.

Sharma, A., Imoto, S., Miyano, S., "A filter based feature selection algorithm using null space of covariance matrix for DNA microarray gene expression data", Current Bioinformatics, vol. 7, no. 3, (6 pages), 2012f.

Sharma, A., Paliwal, K.K., Imoto, S., Miyano, S., "A feature selection method using improved regularized linear discriminant analysis", Machine Vision and Applications, DOI: 10.1007/s00138-013-0577-y, 2013.

Singh D., Febbo, P.G., Ross, K., Jackson, D.G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A.A., D'Amico, A.V., Richie, J.P., Lander, E.S., Loda, M., Kantoff, P.W., Golub, T.R., Sellers W.R.: Gene Expression Correlates of Clinical Prostate Cancer Behavior, Cancer Cell, vol. 1, pp 203-209, 2002.

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

Swets, D.L., Weng, J., Using discriminative eigenfeatures for image retrieval, IEEE Trans. on Patt. Anal. and Mac. Intelligence, vol. 18, no. 8, pp. 831-836, 1996.

Thomaz, C.E., Kitani, E.C., and Gillies, D.F., A maximum uncertainty LDA-based approach for limited sample size problems with application to face recognition. Proc. 18th Brazilian Symp. On Computer Graphics and Image Processing, (IEEE CS Press), pp. 89–96, 2005

Tian, Q. Barbero, M. Gu Z. H. and Lee, S. H., Image classification by the Foley-Sammon transform, Opt. Eng., vol. 25, no. 7, pp. 834-840, 1986.

van 't Veer, L.J., Dai, H., van de Vijver, M.J., He, Y.D, Hart, A.M.H, Mao, M., Peterse, H.L., van der Kooy, K., Marton, M.J., Witteveen A.T., Schreiber, G.J., Kerkhoven, R.M., Roberts, C., Linsley, P.S., Bernards, R. and Friend, S.H.: Gene expression profiling predicts clinical outcome of breast cancer, Letters to Nature, Nature, vol. 415, pp. 530-536, 2002.

Witten, I.H., Frank, E., Data mining: practical machine learning tools with java implementations, Morgan Kaufmann, San Francisco, CA, 2000. http:/www.cs.waikato.ac.nz/ml/weka/

Ye, J. Characterization of a family of algorithms for generalized discriminant analysis on undersampled problems, Journal of Machine Learning Research, vol. 6, pp. 483–502, 2005.

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

Ye, J., Li, Q., A Two-Stage Linear Discriminant Analysis via QR-Decomposition, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 6, pp. 929-941, 2005.

Yeoh EJ, Ross ME, Shurtleff SA, Williams WK, Patel D, Mahfouz R, Behm FG, Raimondi SC, Relling MV, Patel A, Cheng C, Campana D, Wilkins D, Zhou X, Li J, Liu H, Pui CH, Evans WE, Naeve C, Wong L, Downing JR.: 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.

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

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

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

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