Expert Systems With Applications 81 (2017) 88–107

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Noisy-free Length Discriminant Analysis with cosine hyperbolic framework for dimensionality reduction

K. Ramachandra Murthy*, Ashish Ghosh

Machine Intelligence Unit, Indian Statistical Institute, Kolkata 700108, India

ARTICLE INFO

Article history:
Received 12 August 2016
Revised 14 March 2017
Accepted 15 March 2017
Available online 21 March 2017

Keywords:
Dimensionality reduction
Subspace learning
Discriminant analysis
Relevant patterns
Cosine hyperbolic
Face recognition

ABSTRACT

Dimensionality Reduction (DR) is very useful and popular in many application areas of expert and intelligent systems, such as machine learning, finance, data and text mining, multimedia mining, image processing, anomaly detection, defense applications, bioinformatics and natural language processing. DR is widely applied for better data visualization and improving learning in all the above fields. In this manuscript, we propose a novel DR approach namely, Noisy-free Length Discriminant Analysis (NLDA) by developing Noisy-free Relevant Pattern Selection (NRPS). Traditional pattern selection methods discriminate boundary and non-boundary patterns with the help of class information and nearest neighbors. And these methods completely ignore noisy patterns which may degrade the performance of subsequent subspace learning. To overcome this, we develop Noisy-free Relevant Pattern Selection (NRPS), in which data instances are partitioned into boundary, non-boundary and noisy patterns. With the help of noisy-free boundary and non-boundary patterns, Noisy-free Length Discriminant Analysis (NLDA) has been proposed by developing new within and between-class scatters. These scatters model discriminations between lengths (L_2-norms) of different class instances by considering only boundary and non-boundary patterns, while ignoring noisy patterns. A cosine hyperbolic frame work has been developed to formulate the objective of NLDA. Moreover, NLDA can also model the discrimination of multimodal data as different class data may consist of different lengths. Experimental study conducted on the synthesized data, UCI, and leeds butterfly databases. Moreover, an experimental study over human and computer interaction, i.e., face recognition (one of the application areas of expert and intelligent systems), has been performed. And, these studies prove that the proposed method can produce better discriminated subspace compare to the state-of-the-art methods.

1. Introduction

Due to the rapid development of network, information, and multimedia technology during the past few decades, the high-dimensional data pre-processing has become a hot topic (Borges & Nievola, 2012; Cunningham & Ghahramani, 2015; Houari, Bounceur, Kechadi, Tari, & Euler, 2016; Song, Yang, Siadat, & Pechenizkiy, 2013). Dimensionality Reduction (DR) is a common pre-processing tool to make the input data more suitable for the learning models (supervised/unsupervised) in most of the machine learning and data mining techniques. DR can overcome the problem of "curse of dimensionality" (Fukunaga, 1990) which may occur in most of the high-dimensional data. High dimensionality also poses huge computational overhead for machine learning and data mining algorithms. DR is an effective methodology to tackle these problems and, it is a tool for visualizing and analyzing the complex relationships between the features/attributes. The goal of DR is to transform a higher-dimensional data space to a lower-dimensional one under the assumption that the intrinsic structure of the original data can be retained in the low-dimensional space (Bolón-Canedo et al., 2016; Choi, Kim, Plataniotis, & Ro, 2016; Oliva, Lee, Spolaor, Coy, & Wu, 2016; Perumal & Mouli, 2016).

Dimensionality Reduction (DR) is very useful and popular in many application areas of expert and intelligent systems, such as Machine Learning (ML), finance (Rada, 2008), text and data mining (He, 2013; Kim, Han, Lee, & Park, 2016; Romero & Ventura, 2007), multimedia mining (Jayaraman, Prakash, & Gupta, 2012), Image Processing (IP) (Oliva et al., 2017; Tomasoni, Saracoglu, & Paniagua, 2014), anomaly detection (F.Mazzarella et al., 2017), defense applications (Gilmore, 1985), bioinformatics (Ezziane, 2006) and Natural Language Processing (NLP). One of the major application of NLP (Gali, Istodor, & Fränti, 2017) is text/documents cate-

* Corresponding author.
E-mail addresses: k.ramachandra@isical.ac.in (K.R. Murthy), ash@isical.ac.in (A. Ghosh).

http://dx.doi.org/10.1016/j.eswa.2017.03.034
0957-4174/

gorization. News agencies and medical fields are those more sensitive to the need of categorizing documents as they have larger number of documents and corpora. A major difficulty with documents categorization is the high dimensionality of the input space. And, all the dimensions/features are not useful as most of them may be noisy or redundant features and they may degrade subsequent learning. Hence, DR can learn representative/discriminative features from the original input space and improve the quality of features. As an active research topic in IP and ML, human face recognition plays an increasingly important role in a wide range of application, such as criminal identification, credit card verification, surveillance systems, etc. A facial image is considered as a vector of pixels and is represented as a single point in the high-dimensional space. Here also, DR is applied to a great extent in order to extract useful facial features and to facilitate further processing such as face classification. Anomaly detection has remained one of the most difficult tasks in data mining due to the inherent difficulty in precisely defining and quantifying the notion of anomaly. Anomaly detection (Agovic, Banerjee, Ganguly, & Protopopescu, 2009) has to be typically customized to the application domain, since its definition is domain-dependent. A Typical domain is monitoring and management of high-volume feature-rich traffic in large networks that offers significant challenges in storage, transmission and computational costs (Huang, Sethu, & Kandasamy, 2016). The predominant approach to reducing these costs is based on leaning a low-dimensional subspace such that a certain large percentage of the useful information of the input data is preserved in the low-dimensional representation. Thus, DR is very much useful to learn discriminative features in different application of expert and intelligent systems.

Among the numerous DR approaches proposed to handle the problem of high-dimensional data processing, Principal Component Analysis (PCA) (Theodoridis & Koutroumbas, 2008; Zhang, Dubay, & Charest, 2015) and Multi-Dimensional Scaling (MDS) are the most classical methods. However, these algorithms are all inefficient for extracting the discriminative features of the data. Linear Discriminant Analysis (LDA) is a supervised DR method to extract discriminative features, which has been successfully applied to many practical applications (Theodoridis & Koutroumbas, 2008). LDA defines within and between-class scatter matrices, and selects the optimal subspace through maximizing the ratio of within (S_w) and between (S_b) class scatters. But, the within-class scatter matrix (S_w) is singular generally. This is known as the under sampled problem and commonly called Small Sampled Size (SSS) problem. It limits the application of LDA algorithm in many of the practical problems. The Maximum Margin Criterion (MMC) (Haifeng, Tao, & Keshu, 2006; Yang et al., 2009) can avoid the SSS problem by maximizing the distance between any two different classes and then realize between-class discrimination. This distance implies two aspects: (i) minimize the within-class distance with their variances, respectively and (ii) maximize the between-class distance with the mean value of each class. The matrix to be decomposed is $S_b - S_w$ in the final optimization model of MMC. An Exponential Discriminant Analysis (EDA) (Zhang, Fang, Tang, Shang, & Xu, 2010) has been proposed to overcome the SSS problem with the help of exponential framework. The advantages of EDA are that, compared with PCA + LDA, the EDA method can extract the most discriminant information that was contained in the null space of a within-class scatter matrix, and compared with another LDA extension, i.e., null-space LDA, the discriminant information that was contained in the non-null space of the within-class scatter matrix is not discarded. Furthermore, EDA is equivalent to transforming original data into a new space by distance diffusion mapping, and then, LDA is applied in such a new space (Zhang et al., 2010). Moreover, when there are outlier classes, LDA, MMC and EDA may project the data into an inappropriate low dimensional space because of its weak robustness to outliers. The structure of optimization function and the boundary make these methods trend to either within-class scatter minimization or between-class scatter maximization for DR. Thus, Angle Linear Discriminant Embedding (ALDE) (Liu, Feng, & Qiao, 2015) on the basis of angle measurement and utilizes the cosine of the angle (Co-Angle) to get the within and between-class scatter matrices, and thus avoids the problem of outlier classes.

One more limitation of LDA, MMC, EDA and ALDE is that, it assumes unimodal likelihoods which will not be able to preserve any complex structures of the data, that needed for classification. Few DR methods like, Linear Boundary Discriminant Analysis (LBDA) (Na, Park, & Choi, 2010), Local Fisher Discriminant Analysis (LFDA) (Huang, Li, & Liu, 2012; Sugiyama, 2007), Local Discriminative Gaussian (LDG) (Parrish & Gupta, 2012), Quadratic Mutual Information (QMI) (Bouzas, Arvanitopoulos, & Tefas, 2015), Stable Orthogonal Local Discriminant Embedding (SOLDE) (Gao, Ma, Zhang, Gao, & Liu, 2013), etc., discover the internal geometrical structure of the multimodal data. For improving the performance of LDA; LBDA (Na et al., 2010) considers boundary and non-boundary patterns for different samples based on their locations. Utilizing the principles of LDA; LBDA can find an optimal low-dimensional subspace which minimizes the scatter of non-boundary patterns and maximizes the scatter of boundary patterns. LFDA (Sugiyama, 2007), effectively combines the ideas of LDA and LPP, i.e., it maximizes between-class separability and preserves within-class local structure at the same time. Thus, LFDA is useful for dimensionality reduction of multimodal labeled data. LDG (Parrish & Gupta, 2012) is a supervised DR technique, whose objective function is an approximation to the leave-one-out training error of a local quadratic discriminant analysis classifier, and thus acts locally to each training point in order to find a mapping where similar data can be discriminated from dissimilar data. Moreover, it scales better for data sets with a large number of feature dimensions. And linear QMI (Bouzas et al., 2015) is based on the maximization of a non-parametric mutual information criterion between the feature vectors and their respective class labels. Furthermore, quadratic non-parametric implementations of mutual information are computationally efficient and do not require any prior assumptions about the class densities. Kernel QMI can directly be optimized inside the graph embedding framework and from the kernel method an equivalent linear version has been derived. A complete summary of pros and cons of the existing literation can be found in Table 1.

Most of the above traditional DR methods treat all the instances with equal importance and ignore spatial locations of the instances. Few patterns may be located far from the decision boundary and are well separated from the data of different classes; while others may mixed up near the boundary. Also there may be some patterns (noisy), which are closer to other class regions rather than their own class region and may effect the further processing of the data. Most of the pattern selection methods, wouldn't process the noisy-patterns which might effect the performance of subsequent subspace learning (Na et al., 2010; Na, Yun, Kim, & Choi, 2008; Shin & Cho, 2007). Thus, we propose Noisy-free Relevant Pattern Selection (NRPS) in order to assign different roles to patterns based on their locations and they are divided into three categories as, boundary patterns (the data near the decision boundary), non-boundary patterns (the data far from the decision boundary and closer to its class region) and noisy patterns (the data far from its class region and nearer to other class regions). With the help of NRPS, we propose a novel DR method Noisy-free Length Discriminant Analysis (NLDA) by investigating these three types of data in different ways. NLDA models the discrimination of the data with help of length (L_2-norm) differences between the patterns. The within-class scatter models the average compaction of each class pattern lengths and between-class scatter discriminates the

Table 1
Summary table of existing literature with pros and cons.

Methods	Pros	Cons
MDS	It is simple, robust and more flexible as it accepts coordinates as well as scalar products or Euclidean distances.	(i) When there are too many data points, structure becomes obscured (ii) Possibility of introducing artifacts (iii) Unsupervised methodology.
PCA	(i) Reduce the redundancy in the data. (ii) Reduction of noise since the maximum variation basis is chosen and so the small variations in the back-ground are ignored automatically.	(i)However, PCA relies on assumptions that are much too restrictive, especially when it comes to latent variable separation. (ii) PCA originally defined in order to process Gaussian distributions. (iii) It won't consider class discrimination information.
LDA	LDA achieves maximum discrimination using less features.	(i) LDA's objective suffers with small sampled size (SSS) problem. (ii) Assumes unimodal Gaussian likelihoods. (iii) Number of transformed features are less than number of classes. (iv) Subspace may be influenced by outlier classes.
MMC	(i) It has lower computational cost compare to LDA. (ii) And SSS problem has been avoided by MMC	(i) MMC disregards the discriminative information within the local structure of samples and performance is depended on choosing of a coefficient. (ii) Prone to outlier-classes.
EDA	(i) This one also overcomes the undersampled (SSS) problem. (ii) It has a diffusion effect on the distance between samples. (iii) The margin between different classes is enlarged.	(i) It also assumes Gaussian likelihoods. (ii) Computationally expensive compare to LDA. (iii) Subspace may be influenced by outlier classes.
ALDE	(i) ALDE can avoid SSS problem (ii) CO-Angle measure of ALDE eliminates the influence of outlier classes	It also assumes unimodal Gaussianity on the data set.
LFDA	(i) It takes local structures of the data into account so the multimodal data can be embedded appropriately. (ii) LFDA can reduce the dimensionality into an arbitrary dimensional space even with the rank deficiency of the between-class scatter matrix.	(i) The basis vectors of LFDA are statistically correlated and extracted features contain much redundancy. (ii) The overlapped information can distort the distribution of the features and even degrade the recognition performance.
LBDA	(i) LBDA concentrates on significant differences between different type of patterns. (ii) May efficiently deal multimodal data as deal with nearest neighbors.	(i) May suffer from SSS problem. (ii) Assumes Gaussianity (iii) Requires parameter tuning.
QMI	(i) Computationally efficient and (ii) Do not require any prior assumptions about the class densities.	QMI may not capture the discrimination between the classes, efficiently, if bandwidth of Parzen window estimators is not a scalar and may lead to overlapping classes in the subspace.
LDG	(i) Acts locally to each training point in order to find a mapping where similar data can be discriminated from dissimilar data. (ii) It applicable to multimodal data as it seeks a mapping where a quadratic boundary separates the classes.	Computationally very expensive for large data sets as it deals with leave-one-out cross validation error.

lengths of a class from those of the other classes. These scatter matrices are modelled using noisy-free boundary and non-boundary patterns, only. Also, cosine hyperbolic framework for formulating the NLDA's objective has been developed. The contributions of the manuscript are as follows:

- Noisy-free Relevant Pattern Selection (NRPS) method to remove noisy patterns with the help of a generalized threshold.
- Noisy-free Length Discriminant Analysis (NLDA) that models discrimination in terms of average length differences of the patterns using noise free boundary and non-boundary patterns.
- Cosine-hyperbolic framework to model NLDA's objective and to make the length differences between samples more significant.

Experimental study has been performed on synthetic data, UCI (Frank & Asuncion, 2010), leeds butterfly image recognition (Wang, Markert, & Everingham, 2009) and extended yale b face databases, and proposed method has been compared with state-of-the-art DR methods.

This paper is organized as follows: Section 2 discusses Noisy-free Relevant Pattern Selection and Section 3 introduces Noisy-free Length Discriminant Analysis and asymptotic time complexity analysis of NLDA has been discussed in Section 3.5. Experimental results are presented in Section 4 and the manuscript has been concluded in Section 5. Finally, Section 6 discusses further scope of NLDA.

2. Noisy-free Relevant Pattern Selection (NRPS)

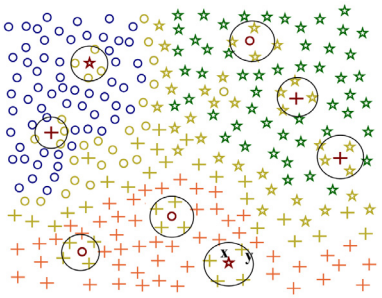
In rest of the manuscript, until and unless specified, vectors are denoted by lower-case letters in bold font and matrices by upper-case. And brackets are used to build matrices from lists of column vectors. *tr* and $\top$ denote trace and transpose of a vector or a matrix.

2.1. Motivation

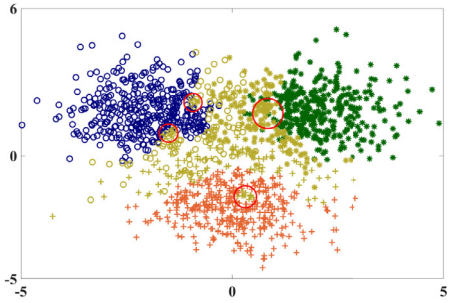
Neighborhood Property based Pattern Selection (NPPS) algorithm (Shin & Cho, 2007) has been proposed for pre-processing of subspace learning methods in order to select the boundary patterns according to a measure called Proximity (P). The proximity (P) of data patterns indicates how close they are placed to the boundary, i.e.,

$$P(\mathbf{x}, k) = -\sum_{i=1}^{\mathcal{L}} p_i(\mathbf{x}) log_2(p_i(\mathbf{x})), \tag{1}$$

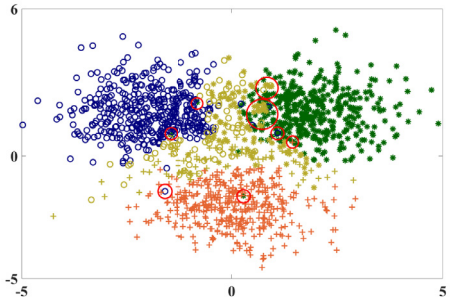
where $p_i(\mathbf{x}) = \frac{k_i}{k}$, k_i is the number of neighbors of $\mathbf{x}$ belonging to class c_i, k is the number of nearest neighbors and $\mathcal{L}$ is the number of classes. The selection of boundary patterns in NPPS is very strictly made, i.e., when the proximity (P) of a pattern is positive, it is considered as boundary pattern; otherwise, this pattern is considered as non-boundary one. According to this decision, only the patterns in the homogeneous region (the region in which all patterns are from the same class) can be considered as non-boundary patterns. However, small noise in the patterns distribution can convert the non-boundary region into boundary one. In order to deal with this problem, Relevant Pattern Selection (RPS) (Na et al., 2008) made few modifications like, the patterns and their neighborhoods are used in the determination of proximity, and accordingly non-boundary patterns are selected by using a non-zero threshold. This non-zero threshold depends on the number of neighborhood patterns and the number of classes (Na et al., 2008). In LBDA, this threshold is defined as $1 - 1/\mathcal{L}$; since as



(a) Boundary (yellow), noisy (red) and non-boundary (other colors) patterns.



(b) NPPS pattern selection ($Th = 0.0$) on random data.



(c) LBDA pattern selection ($Th = 1 - 1/\mathcal{L}$) on random data.

Fig. 1. Geometrical examples of NPPS and LBDA boundary and non-boundary patterns. Red circles indicate the location of the noisy patterns. Here, O - class 1 (blue), * - class 2 (green) and + - class 3 (orange). Also, boundary and noisy patterns patterns are colored with yellow and red, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

the number of classes $\mathcal{L}$ increases, the data become more probably mixed with other class data, and too much data may be classified as boundary patterns (Na et al., 2010).

In RPS (Na et al., 2008), it is suggested that the use of non-zero threshold may help to remove noisy patterns that force non-boundary into boundary region. But the selection of such a threshold is purely an empirical one. The threshold $1 - 1/\mathcal{L}$ $(= Th)$, proposed by Na et al. (2010), may not reduce the number of noisy patterns, efficiently. Let us consider a toy example of Fig. 1(a), in which $\mathcal{L} = 3$, $k = 5$ (k-nearest neighbors represented with the circles). Few patterns (i.e., noisy patterns) in the data are far away from their class region and nearer to the other class region and, they may dilute the boundary and non-boundary patterns. For most of these noisy patterns ($\mathbf{x}$ in Fig. 1(a)), k $(= 5)$ nearest neighbors ($\mathbf{y}$) are from other classes. Let $\mathbf{x} \in c_2$ be a noisy pattern as shown in Fig. 1(a) since it is located in the region of class c_3. For this pattern, $P(\mathbf{x}, k) = -0 * log_2(0) - (1/5) * log_2(1/5) - (4/5) * log_2(4/5) = 0.7219$ ($\because 0 * log_2(0) = 0$) and $Th = 1 - 1/3 = 0.6667$. Clearly, $P(\mathbf{x}, k) > Th$ and accordingly, the pattern $\mathbf{x}$ is marked as a boundary pattern. But this pattern is far from its class region and nearer to the other class region. Moreover, due to this pattern ($\mathbf{x}$, called as noisy pattern), its neighborhood patterns ($\mathbf{y}$) will be marked as boundary patters, since $P(\mathbf{y}, k) = -0 * log_2(0) - (1/5) * log_2(1/5) - (4/5) * log_2(4/5) = 0.7219$ which is same as the proximity of noisy pattern $\mathbf{x}$.

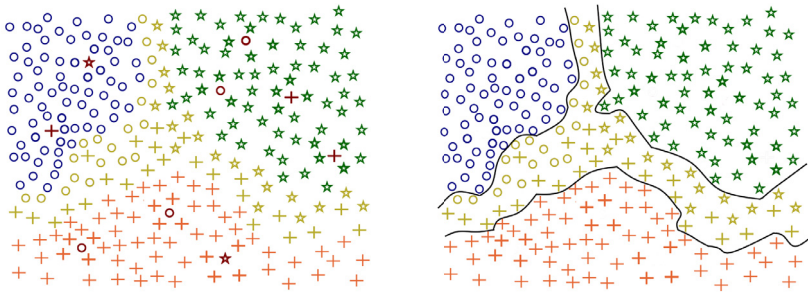
The reason is that the noisy pattern $\mathbf{x} \in c_2$ participate in the selection process of $\mathbf{y} \in c_3$, which is not desirable. If the effect of $\mathbf{x}$ is suppressed then the pattern $\mathbf{y}$ can be marked as non-boundary, which is more desirable. Consider another example (Fig. 1(b)) of randomly generated 3-class Gaussian data with parameters $\mu_1 = [1.75,\ 1.75]$, $\mu_2 = [-1.75,\ 1.75]$ and $\mu_3 = [0,\ -1.75]$, $\Sigma_i = [1\ \ 0; 0\ \ 1]$, $i = 1, 2, 3$; where μ_i and Σ_i are mean and covariance matrix of class c_i; $i = 1, 2$ and 3. Note that each class consists of 500 patterns. The parameters are set to $k = 11$ and $Th = 0.0$ (NPPS) *or* $1 - 1/\mathcal{L}$ (LBDA). It is clear, from the Fig. 1(b), that with threshold $Th = 0.0$ (NPPS), many noisy patterns (i.e. patterns closer to other class patterns) form clusters by forcing their neighborhoods as boundary patterns. Also, in case of $Th = 1 - 1/\mathcal{L} = 0.6667$ (LBDA), the noisy patterns and their neighborhoods are marked as boundary patterns as depicted in Fig. 1(c). This is not desirable because noisy patterns are closer to other class regions and their neighborhood patterns must treat as non-boundary patterns. Moreover, with LBDA's threshold (Th), some of the noisy patterns are marked as non-boundary (Fig. 1(c)) because the proximity of those patterns are $< 1 - 1/\mathcal{L} (= Th)$ and they participate in the subsequent subspace learning.

Therefore, with the thresholds $Th = 0.0$ *or* $1 - 1/\mathcal{L}$, noisy patterns are marked as boundary or non-boundary patterns and so these thresholds may not be a good choice to deal with such kind of the noisy data. Thus, NPPS (Na et al., 2008) and LBDA (Na et al., 2010) model the scatter matrices on the noisy boundary and non-boundary patterns, only Na et al. (2010). Thus, we proposed a novel method, Noisy-free Relevant Pattern Selection (NRPS), with the following advantages:

(i) A new threshold (Th'), which is better than 0.0 and $1 - 1/\mathcal{L}$ to filter the noisy patterns efficiently.
(ii) Dividing the input pattern set into a partition of three sets, namely, boundary, non-boundary and noisy pattern sets.

2.2. Proposed method (NRPS)

Let $\mathcal{X} \in \mathbb{R}^D$ is the set of input instances. And, $c(\mathbf{x})$ is the class label of an instance $\mathbf{x}$. The $\mathcal{N}_k(\mathbf{x})$ denotes set of $k-$nearest neighbors of a pattern $\mathbf{x}$ and $c_i \subseteq \mathcal{X}$ is the i^{th} class; $i = 1, 2, \ldots, \mathcal{L}$. As discussed above, few patterns are already well separated from the other data of different classes, while others are mixed up near the decision boundary. And, few patterns may be nearer to other class regions which dilutes relevant pattern selection. The formal definitions for three types of data are provided below:



(a) Noisy patterns (red patterns), boundary patterns (yellow) and non-boundary patterns (other than yellow).

(b) Boundary region without noisy patterns.

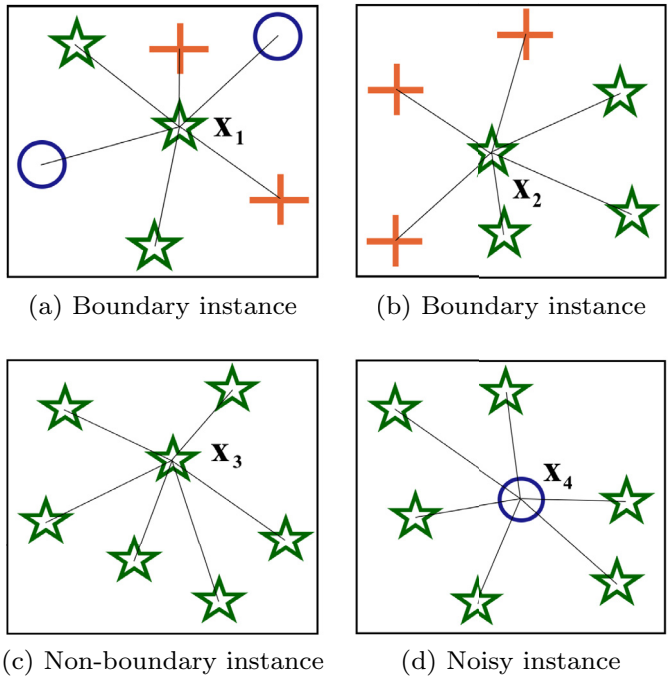
Fig. 2. Geometrical examples of noisy free RPS boundary and non-boundary regions. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Noisy Pattern. A pattern is said to be a 'Noisy Pattern', if its neighborhood tends to have more homogeneous neighbors from only one class which is different from its own class. That is, a pattern $\mathbf{x} \in c_i$ is said to be a 'Noisy Pattern' if $|\mathcal{N}_k(\mathbf{x}) \bigcap c_i| \ll |\mathcal{N}_k(\mathbf{x}) \bigcap c_j|$; for some j and $|\mathcal{N}_k(\mathbf{x}) \bigcap c_k| = 0;\ \forall\ k \neq i, j;$ $|.|$ is the cardinality of a set. We denote the set of noisy patterns as $\mathbb{N}_{oise}$.

Boundary Pattern. A pattern is said to be a 'Boundary Pattern', if it is nearer to boundary region and has more heterogenous neighbors (patterns from different classes). We denote the set of boundary patterns as $\mathbb{B}_{oundary}$.

Non-Boundary Pattern. A pattern is said to be a 'Non-Boundary Pattern', if it has all homogeneous neighbors (patterns from the same class) and is far away from boundary region. We denote the set of non-boundary patterns as $\mathbb{NB}_{oundary}$.

From the above definitions, a pattern belongs to $\mathbb{B}_{oundary}$, if it contains patterns from its own class and as well as from other classes. And it belongs to $\mathbb{NB}_{oundary}$ if it contains most of the patterns from its own class with in its nearest neighborhood. Finally, if a pattern's nearest neighborhood contains major patterns from only one class, other than its own class, then it will belong to the set $\mathbb{N}_{oise}$. The set of noisy patterns contains all patterns which form clusters in other class regions. Clearly, these noisy patterns are nearer to other class patterns and may degrade the selection of better boundary and non-boundary patterns. Our goal is to separate noisy patterns (Fig. 2(a) in which $k = 5$) from boundary and non-boundary (Fig. 2(b)) patterns to improve quality of further processing of the data.



(a) Boundary instance (b) Boundary instance

(c) Non-boundary instance (d) Noisy instance

Fig. 3. $k = 7$ and $\mathcal{L} = 3$. Here, O - class 1 (blue), * - class 2 (green) and + - class 3 (orange). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

2.2.1. Threshold selection

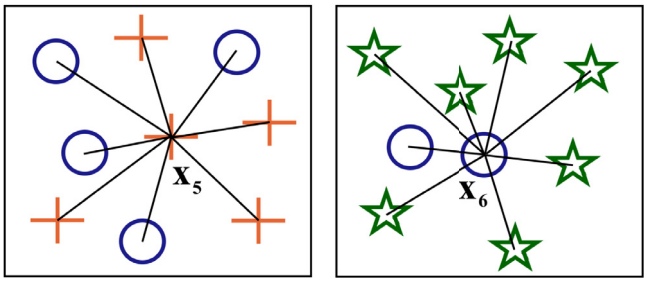
As mentioned earlier, threshold for proximity plays an important role in pattern selection. A good threshold filters better boundary and non-boundary patterns from noisy patterns. From Eq. (1), it is clear that proximity (P) is measured by a kind of entropy with in a nearest neighborhood of a pattern over different classes. The upper bound of $P(\mathbf{x}, k)$ can be derived as,

$$P(\mathbf{x}, k) = -\sum_{i=1}^{\mathcal{L}} p_i log_2(p_i)$$

$$\leq -\sum_{i=1}^{\mathcal{L}} \frac{1}{\mathcal{L}} log_2\left(\frac{1}{\mathcal{L}}\right) = log_2(\mathcal{L}).$$

Proximity of a pattern ($\mathbf{x}$) will attain this upper bound if the distribution of the classes is equi-probable with in $\mathcal{N}_k(\mathbf{x})$. If $P(\mathbf{x}, k) \approxeq log_2(\mathcal{L})$ then $\mathcal{N}_k(\mathbf{x})$ consists equal number of instances from all the classes, which leads heterogenous neighborhood. And, therefore, the pattern $\mathbf{x}$ may lie in the boundary region. If $P(\mathbf{x}, k) \approxeq 0$ then $\mathcal{N}_k(\mathbf{x})$ contains homogenous instances from the same class and $\mathbf{x}$ may lie in the non-boundary region. When $0 < P(\mathbf{x}, k) < log_2(\mathcal{L})$, the classification of the pattern $\mathbf{x}$ is very challenging. For example, let us assume $k = 7$ and $\mathcal{L} = 3$ as shown in the Fig. 3. From Fig. 3(a), it is clear that the instance $\mathbf{x_1} \in c_2$ can be tagged as boundary pattern as $\mathcal{N}_k(\mathbf{x_1})$ contains equal number of patterns from all the classes and $P(\mathbf{x_1}, k) = 1.5567$ (in this example the upper bound for proximity is 1.5850). Fig. 3(c) represents non-boundary pattern $\mathbf{x_3} \in c_2$ and it has homogenous (of same class) neighborhood. Note that, in this case $P(\mathbf{x_3}, k) = 0$. These two cases are easy to deal with. In Fig. 3(b), the pattern $\mathbf{x_2} \in c_2$ can also be boundary pattern for classes 2 and 3 as $\mathcal{N}_k(\mathbf{x_2})$ consists of equal (approximately) number of patterns from both the classes and $P(\mathbf{x_2}, k) = 0.9852$. Clearly, $0 < P(\mathbf{x_2}, k) < log_2(3)$. And, the pattern $\mathbf{x_4} \in c_1$ (noisy pattern) from Fig. 3(d) lies in the region of class c_2 as the neighbors are all from the class c_2. Note that, $P(\mathbf{x_4}, k) = 0.5917$. These two patterns ($\mathbf{x_2}$ and $\mathbf{x_4}$) with proximities 0.9852 and 0.5917 should be tagged as non-boundary and noisy patterns, respectively.

Consider another example with $k = 9$ and $\mathcal{L} = 3$ as in Fig. 4. The pattern $\mathbf{x_5} \in c_3$ ($P(\mathbf{x_5}, k) = 0.9911$) in Fig. 4(a) is a boundary pat-



(a) Boundary instance (b) Noisy instance

Fig. 4. $k = 9$ and $\mathcal{L} = 3$.

tern between classes c_1 and c_3, and $\mathbf{x_6} \in c_1$ ($P(\mathbf{x_6}, k) = 0.7642$) in Fig. 4(b) is a noisy pattern since only two patterns of c_1 are clustered in the region of c_2. These two types of patterns are also difficult to distinguish as boundary and noisy patterns. Moreover, the noisy patterns $\mathbf{x_4}$ and $\mathbf{x_6}$ will effect the proximity values of it's neighborhoods. Thus, a good threshold is required to distinguish the noisy patterns and their effect on the neighbors.

The upper bound $log_2(\mathcal{L})$ of proximity provides the intuition regarding the class distribution of patterns with in a neighborhood. If the class distribution is closer to uniform within the neighborhood of a pattern then the proximity of that pattern is approximately equal to the upper bound $log_2(\mathcal{L})$. That is, if neighborhood contains heterogeneous (different classes) patterns then the pattern will be treated as boundary one. If proximity is zero then the pattern will be treated as non-boundary. Therefore, NRPS defines a new threshold as,

$$Th' = Const * log_2(\mathcal{L});\ 0 < Const < 1.$$

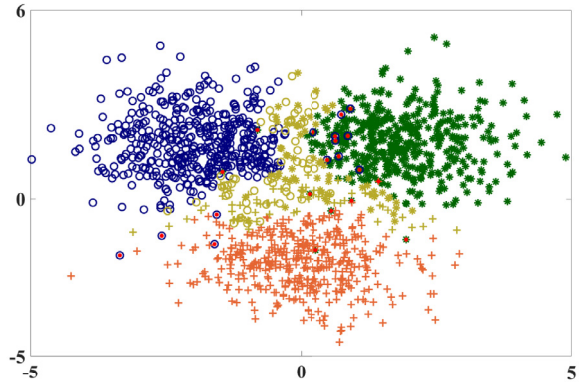
This threshold, Th' can give a better intuition regarding the class distribution as it is directly related to the definition of proximity. Thus, if $P(\mathbf{x}, k) > Th'$ then the pattern $\mathbf{x}$ can be treated as boundary one because the $\mathcal{N}_k(\mathbf{x})$ contains good amount of heterogenous patterns from different classes. It is clear from Figs. 3 and 4, that the patterns $\mathbf{x_1}$ (between c_1, c_2 and c_3), $\mathbf{x_2}$ (between c_2 and c_3) and $\mathbf{x_5}$ (between c_1 and c_3) will be marked as boundary one with the threshold $Th' = \frac{1}{2} * log_2(3) = 0.7925$ ($Const = \frac{1}{2}$). But the patterns $\mathbf{x_4}$ and $\mathbf{x_6}$ will not be marked as boundary patterns. If proximity of a pattern is zero (e.g., $\mathbf{x_3}$ in Fig 3(c)), then the pattern can be designated as non-boundary. And, if $0 < P(\mathbf{x}, k) < Th'$ then the pattern $\mathbf{x}$ is either a non-boundary or a noisy pattern (i.e., $\mathbf{x_4}$ and $\mathbf{x_6}$ in Figs. 3(d) and 4(b), respectively). The decision of these patterns is very important as there may be noisy patterns among them. Thus, in order to model this, we define

$$p_k(\mathbf{x}) = \left\{ p_i \middle| p_i \text{ is the probability of class } c_i \text{ in } \mathcal{N}_k(\mathbf{x}) \right\}.$$

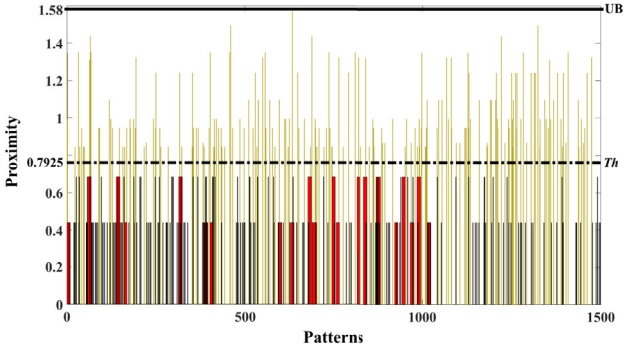
One can observe that, with in the neighborhoods of both $\mathbf{x_4}$ and $\mathbf{x_6}$, the probabilities of their own classes are non-zero minimums in comparison with probabilities of other classes. Note that, $\mathbf{x_4}$ and $\mathbf{x_6} \in c_1$ and $p_1 = \frac{1}{7}$ ($\frac{2}{9}$) in the neighborhood of $\mathbf{x_4}(\mathbf{x_6})$ and $0 < P(\mathbf{x_4}, 7), P(\mathbf{x_6}, 9) \leq 0.5 * log_2(\mathcal{L})$. Thus, these patterns can be marked as noisy patterns. With the help of these observations, NLDE defines the boundary, non-boundary and noisy sets mathematically as follows:

$$\begin{aligned}
\mathbb{B}_{oundary} &= \left\{ \mathbf{x} \in \mathcal{X} \middle| P(\mathbf{x}, k) > Th' \right\} \\
\mathbb{N}_{oise} &= \left\{ \mathbf{x} \in \mathcal{X} \middle| 0 < P(\mathbf{x}, k) \leq Th' \ \wedge \ p_{c(\mathbf{x})} == min\ p_k(\mathbf{x}) \right\} \\
\mathbb{NB}_{oundary} &= \left\{ \mathbf{x} \in \mathcal{X} \middle| P(\mathbf{x}, k) == 0 \ \vee \right. \\
&\qquad \left. \left(0 < P(\mathbf{x}, k) \leq Th' \ \wedge \ p_{c(\mathbf{x})} \neq min\ p_k(\mathbf{x})\right) \right\}. \qquad (2)
\end{aligned}$$

The second condition in the definition of $\mathbb{NB}_{oundary}$ removes the effect of noisy pattern in its neighborhood patterns. That is, for the



(a) Boundary and non-boundary patterns with noisy patterns on Gaussian random data from motivation section using NRPS (proposed method). Here $k = 11$ and $\mathcal{L} = 3$.



(b) Patterns Vs Proximity values. Red - noisy, Yellow - boundary, Black - non-boundary. UB - upper bound and Th - threshold value.

Fig. 5. NRPS on Gaussian random data. Here, O - class 1 (blue), * - class 2 (green) and + - class 3 (orange). Also, boundary and noisy patterns patterns are colored with yellow and red, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

neighborhood patterns (non-boundary patterns) of a noisy pattern, the probabilities $p_{c(\mathbf{x})}$ shouldn't be minimum (e.g., neighborhood patterns in the Figs. 3(d) and 4(b)). As mentioned before, noisy patterns and it's effect should be removed in order to improve the quality of relevant patterns. One can observe that from Fig. 5(a), the patterns that are far away from their class and nearer to other class regions marked as noisy patterns by Th'. These noisy patterns should be discarded in the further processing of relevant patterns. Also, it is clear from Fig. 5(a) that the boundary patterns are distributed along the boundary region and non-boundary patterns are the core patterns of each class. Fig. 5(b) depicts the proximity values of all the patterns for Gaussian data (from motivation section), $Th' = 0.5 * log_2(3) = 0.7925$ and the upper bound (UB). From this figure, one can note that, there are non-boundary patterns $\mathbf{x}$ (black color) with $0 < P(\mathbf{x}, k) \leq Th'$ and these patterns are the neighborhoods to the noisy patterns. Thus, these patterns are not effected by noisy ones and they are grouped into non-boundary patterns. The proposed algorithm (NRPS) divides the input data in to boundary, non-boundary and noisy patterns using Eq. (2) and the complete algorithm of NRPS can found in Algorithm 1.

2.2.2. Importance of $Th' = Const * log_2(\mathcal{L})$

The threshold in LBDA (Na et al., 2010) is defined as $Th = 1 - 1/\mathcal{L}$ since as the number of classes $\mathcal{L}$ increases data become mixed more with other class data and too much data may be tagged as boundary patterns. One can note that the threshold Th is always < 1 (Fig. 6) and may not remove noisy patterns, effectively. From Fig. 6, it is clear that $1 - 1/\mathcal{L} \to 1$ as $\mathcal{L} \to \infty$, i.e., $Th \in [0, 1]$. But, the proximity $P \in [0,\ log_2(\mathcal{L})]$; $log_2(\mathcal{L}) >> 1$ and thus as number

Algorithm 1: Noisy free Relevant Pattern Selection (NRPS).

Input: A set of input patterns $\mathcal{X} \in \mathbb{R}^{N\times D}$, k_{NRPS} - number of nearest neighbors.

Output: The sets $\mathbb{B}_{oundary}$, $\mathbb{NB}_{oundary}$ and $\mathbb{N}_{oise}$

$\mathbb{B}_{oundary} \leftarrow \emptyset$, $\mathbb{NB}_{oundary} \leftarrow \emptyset$ and $\mathbb{N}_{oise} \leftarrow \emptyset$
for $i \leftarrow 1$ ***to*** N **do**
  Compute $\mathcal{N}_{k_{NRPS}}(\mathbf{x_i})$ of the i^{th} pattern $\mathbf{x_i}$ from $\mathcal{X}$.
  $p_k(\mathbf{x_i}) \leftarrow \emptyset$
  $P(\mathbf{x_i}, k_{NRPS}) \leftarrow 0$
  for $l \leftarrow 1$ ***to*** $\mathcal{L}$ **do**
    $p_l \leftarrow \frac{|c_l \cap \mathcal{N}_{k_{NRPS}}(\mathbf{x_i})|}{k_{NRPS}}$
    $p_k(\mathbf{x_i}) \leftarrow p_k(\mathbf{x_i}) \cup p_l$
    $P(\mathbf{x_i}, k_{NRPS}) \leftarrow P(\mathbf{x_i}, k_{NRPS}) - p_l * log_2(p_l)$
  if $P(\mathbf{x_i}, k_{NRPS}) > Th$ **then**
    $\mathbb{B}_{oundary} \leftarrow \mathbb{B}_{oundary} \cup \{\mathbf{x_i}\}$
  else
    if $P(\mathbf{x_i}, k_{NRPS}) > 0$ & $p_{l(\mathbf{x_i})} == \min\ p_k(\mathbf{x_i})$ **then**
      $\mathbb{N}_{oise} \leftarrow \mathbb{N}_{oise} \cup \{\mathbf{x_i}\}$
    else
      if $P(\mathbf{x_i}, k_{NRPS}) == 0 \parallel \left(P(\mathbf{x_i}, k_{NRPS}) > 0 \text{ \& } p_{l(\mathbf{x_i})} != \min\ p_k(\mathbf{x_i})\right)$ **then**
        $\mathbb{NB}_{oundary} \leftarrow \mathbb{NB}_{oundary} \cup \{\mathbf{x_i}\}$
return $\mathbb{B}_{oundary}$, $\mathbb{NB}_{oundary}$ *and* $\mathbb{N}_{oise}$

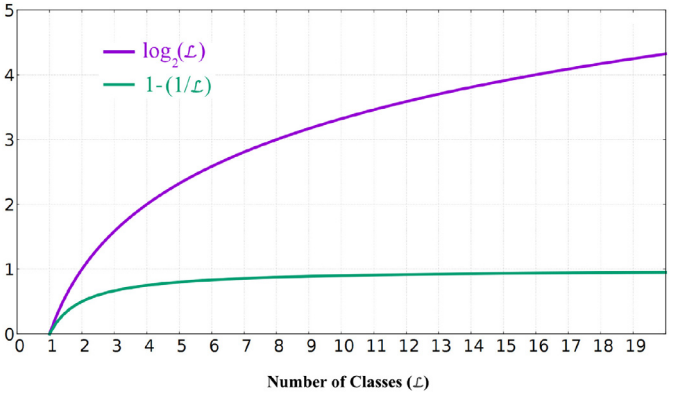


Fig. 6. Difference between thresholds.

of classes ($\mathcal{L}$) increases, more number of patterns will be classified as boundary patterns including noisy patterns. Thus, it may not be a good one to select relevant patterns. On the other hand, the upper bound $log_2(\mathcal{L}) \to \infty$ as $\mathcal{L} \to \infty$ (Fig. 6) and Th' can balance the classification of boundary and non-boundary patterns as it is related (upper bound) to the proximity. Thus, by varying *Const* in the interval $[0, 1]$, Th' can divide the input data into relevant groups, efficiently compare to Th.

3. Noisy-free Length Discriminant Analysis (NLDA)

3.1. Motivation

Most of the DR methods like, LDA, EDA, MMC, ALDE, etc., have been developed with the assumption that the data is from a unimodal Gaussian distribution, a property that often does not exist in real-world applications. Without this assumption, separability of different classes cannot be well characterized by inter and intra class scatters of these methods. They tend to give undesired results if samples in some classes form several separate clusters (i.e., multimodal), which is often observed in many practical applications.

The inter-class separability of the traditional methods maximize the distances ($\|\tilde{\boldsymbol{\mu}} - \tilde{\boldsymbol{\mu}}_i\|^2 = W^\top S_b W$) between global ($\tilde{\boldsymbol{\mu}} = W^\top \boldsymbol{\mu} \in \mathbb{R}^d$) and class means ($\tilde{\boldsymbol{\mu}}_i = W^\top \boldsymbol{\mu}_i \in \mathbb{R}^d;\ i = 1, 2, \ldots, \mathcal{L}$) of the projected data; where $S_b = \sum_i (\boldsymbol{\mu} - \boldsymbol{\mu}_i)(\boldsymbol{\mu} - \boldsymbol{\mu}_i)^\top$, $\boldsymbol{\mu}$ and $\boldsymbol{\mu}_i$ are global and i^{th} class means in input space, respectively. In case of multimodal data, the global and class means can overlap (or nearer) as shown in Fig. 7(a). Clearly, Fig. 7(a) contains two classes whose means $\boldsymbol{\mu}_i \in \mathbb{R}^D;\ i = 1, 2$ ($D \gg d$) are overlapped with global mean $\boldsymbol{\mu} \in \mathbb{R}^D$ and the above mentioned DR methods get failed to model the inter-class scatter ($\because S_b \cong 0$) on these kind of multimodal data. Thus, they completely depend on intra-class scatter (S_w) and obtain a subspace (or projection space) W as shown in Fig. 7(b). The subspace W produces overlapping classes and may lead to very poor performance. But, one can observe that, the length differences between the two class patterns are different for the multimodal data. That is, the length ($\|\bullet\|_2$) differences between $\mathbf{x_1} \in c_2$ and $\mathbf{x_3} \in c_1$; $\mathbf{x_1} \in c_2$ and $\mathbf{x_4} \in c_1$; $\mathbf{x_2} \in c_2$ and $\mathbf{x_3} \in c_1$; and $\mathbf{x_2} \in c_2$ and $\mathbf{x_4} \in c_1$, etc., are larger as shown in Fig. 7(a) and thus length discrimination between the classes can be used to model the separability within the multimodal data. From Fig. 7(b), it is clear that the optimal subspace W_{opt} will have maximum average length differences (irrespective of positive or negative) $\|\mathbf{y_1}\| - \|\mathbf{y_3}\|$, $\|\mathbf{y_2}\| - \|\mathbf{y_4}\|$, etc., between classes. Also, note that W_{opt} contains minimum average length differences of within class patterns, i.e., $\|\mathbf{y_1}\| - \|\mathbf{y_5}\|$, $\|\mathbf{y_2}\| - \|\mathbf{y_6}\|$, $\|\mathbf{y_3}\| - \|\mathbf{y_4}\|$, etc. are minimum. Even though the differences $\|\mathbf{y_5}\| - \|\mathbf{y_6}\|$, $\|\mathbf{y_1}\| - \|\mathbf{y_2}\|$ are high, W_{opt} seems to be the optimal one to maximize between and minimize within-class lengths (Fig. 7(b)). Thus, the subspace W_{opt} simplifies the decision boundary between the classes as the class c_2 patterns mapped far away from the class c_1 patterns. Thus, the average length discriminations between the classes may lead to better subspace for the multimodal data.

3.2. Problem formulation

In mathematical terms, the problem of DR can be stated as follows: given a random vector $\mathbf{x} \in \mathbb{R}^D$ find a lower dimensional representation $\mathbf{y}\left(= W^\top \mathbf{x}\right) \in \mathbb{R}^D$, that captures the discrimination between classes in the original data according to some objective; where $W \in \mathbb{R}^{D\times D}$ is an orthonormal projection matrix. And choose $d(\ <\ <D)$ number of projection vectors from the columns of W according to their contribution to the mapping error to reduce dimensionality. Next, formulation of the objective function for Generalized NLDA (GNLDA) has been modelled. For this purpose, we define the squared length differences in the projection space as,

$$\mathcal{F}\left(\|\mathbf{y_i}\|^2 - \|\mathbf{y_j}\|^2\right);\ i \neq j,$$

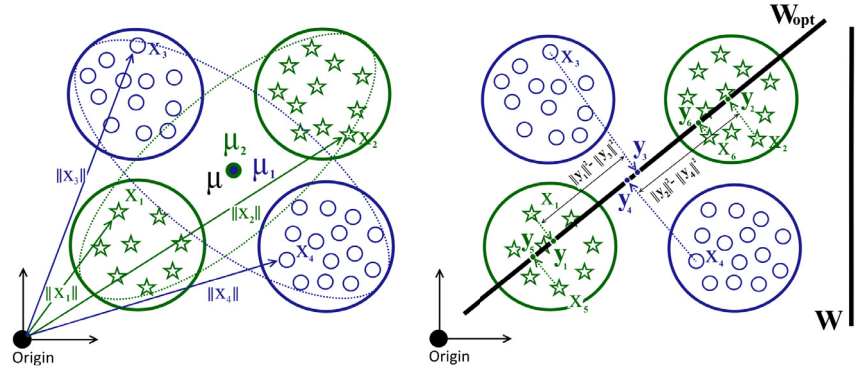
where $\mathcal{F}(\bullet)$ is a symmetric function about y-axis, i.e. $\mathcal{F}(x) = \mathcal{F}(-x);\ x \in \mathbb{R}$. The purpose of the symmetric function $\mathcal{F}$ is to take care of both positive and negative differences. With the help of this, the between ($\widehat{L}_b$) and within ($\widehat{L}_w$) class scatters are defined as,

$$\widehat{L}_b = \frac{1}{\mathcal{L}} \sum_{i=1}^{\mathcal{L}} \frac{1}{N_i} \sum_{\substack{j\in c_i;\\ k\notin c_i}} \mathcal{F}\left(\|\mathbf{y_j}\|^2 - \|\mathbf{y_k}\|^2\right),$$

$$\widehat{L}_w = \frac{1}{\mathcal{L}} \sum_{i=1}^{\mathcal{L}} \frac{1}{N_i} \sum_{\substack{j,k\in c_i;\\ j\neq k}} \mathcal{F}\left(\|\mathbf{y_j}\|^2 - \|\mathbf{y_k}\|^2\right),$$

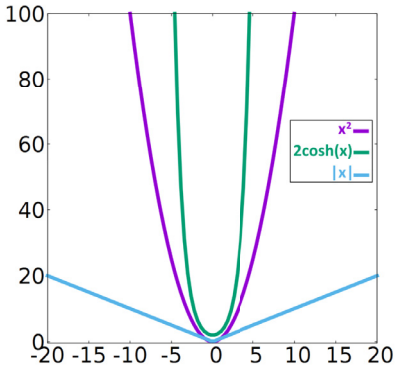
where N_i is the number of patterns in class i; $i = 1, 2, \ldots, \mathcal{L}$. Clearly, the $\widehat{L}_b$ models the between-class length differences and $\widehat{L}_w$ models the within-class length differences. And, the objective function of GNLDA is,

$$W_{opt} = \arg\max_{W^\top W = Id} \frac{\widehat{L}_b}{\widehat{L}_w}, \tag{3}$$



(a) An example, where traditional DR methods get failed and intuition of length discrimination.

(b) Intuition of proposed class separability for multimodal data.



(c) Symmetric functions.

Fig. 7. Motivation of NLDA and symmetric functions. Here, O - class 1 (blue) and * - class 2 (green).

where *Id* is an identity matrix. The above optimization function maximizes the between-class lengths and minimizes within-class lengths, so that W_{opt} can better model the discrimination of classes even for multimodal data (as explained in the above motivation section).

3.3. Cosine hyperbolic framework

The symmetric function $\mathcal{F}$ in the objective function (Eq. (3)) is very important one to deal with. One can use $\mathcal{F}(x) = x^2$, $\mathcal{F}(x) = |x|$, $\mathcal{F}(x) = cosh(x)$ (Fig. 7(c)), etc., in order to reformulate the objective of GNLDA. Among them the cosine hyperbolic function (*cosh*) may be the best one to characterize the length differences between the samples as the function is more steeper than the other symmetric functions (Fig. 7(c)). That means, the small length differences $\|\mathbf{y_i}\|^2 - \|\mathbf{y_j}\|^2 \approx 0$ will be highlighted properly and the larger differences $\|\mathbf{y_i}\|^2 - \|\mathbf{y_j}\|^2 \gg 0$ (or $\ll 0$) will be more significant with the help of cosine hyperbolic function compare to the other symmetric functions. The formulations of $\widehat{L}_b$ and $\widehat{L}_w$ with 2*cosh* are given as,

$$\widehat{L}_b = \frac{2}{\mathcal{L}} \sum_{i=1}^{\mathcal{L}} \frac{1}{N_i} \sum_{\substack{j \in c_i; \\ k \notin c_i}} cosh\left(\|\mathbf{y_j}\|^2 - \|\mathbf{y_k}\|^2\right), \text{ and}$$

$$\widehat{L}_w = \frac{2}{\mathcal{L}} \sum_{i=1}^{\mathcal{L}} \frac{1}{N_i} \sum_{\substack{j,k \in c_i; \\ j \neq k}} cosh\left(\|\mathbf{y_j}\|^2 - \|\mathbf{y_k}\|^2\right).$$

Note that, $\|\mathbf{y_j}\|^2 - \|\mathbf{y_k}\|^2 \geq 0$ (or ≤ 0) $\Rightarrow 2cosh\left(\|\mathbf{y_j}\|^2 - \|\mathbf{y_k}\|^2\right) \geq 2$. That means, all the cosine hyperbolic differences are always positive and moreover, they are significant. Since $2cosh(x) = \exp(x) + exp(-x)$, the between-class scatter matrix becomes,

$$\widehat{L}_b = \mathcal{S}\left(\|\mathbf{y_j}\|^2 - \|\mathbf{y_k}\|^2\right) + \mathcal{S}\left(\|\mathbf{y_k}\|^2 - \|\mathbf{y_j}\|^2\right); \tag{4}$$

where,

$$\mathcal{S}(t) = \frac{1}{\mathcal{L}} \sum_{i=1}^{\mathcal{L}} \frac{1}{N_i} \sum_{\substack{j \in c_i; \\ k \notin c_i}} exp(t).$$

Note that, each term of the above equation contains an exponential function. Both the terms in the above equations will incorporate the positive and negative effect of the difference $\|\mathbf{y_j}\|^2 - \|\mathbf{y_k}\|^2$. And, $exp(t) \geq 1$ whenever $t \geq 0$ or $exp(-t) \geq 1$ whenever $t \leq 0$. When the value of the decay function is greater than or equal to 1, the behavior of product $\prod$ is more significant than summation Σ (Zhang et al., 2010). So, to further enhance the significance of differences, one can change the summation (Σ) to product ($\prod$) and $\mathcal{S}(t)$ can be rewritten as,

$$\mathcal{S}'(t) = \frac{1}{\mathcal{L}} \prod_{i=1}^{\mathcal{L}} \frac{1}{N_i} \prod_{\substack{j \in c_i; \\ k \notin c_i}} exp(t),$$

$$= exp\left(\frac{1}{\mathcal{L}} \sum_{i=1}^{\mathcal{L}} \frac{1}{N_i} \sum_{\substack{j \in c_i; \\ k \notin c_i}} t \right).$$

Thus, the Eq. (4) can be reformulated as,

$$\widehat{L}_b = \mathcal{S}'\left(\|\mathbf{y_j}\|^2 - \|\mathbf{y_k}\|^2\right) + \mathcal{S}'\left(\|\mathbf{y_k}\|^2 - \|\mathbf{y_j}\|^2\right),$$

$$= 2cosh\left(\frac{1}{\mathcal{L}} \sum_{i=1}^{\mathcal{L}} \frac{1}{N_i} \sum_{\substack{j \in c_i; \\ k \notin c_i}} \|\mathbf{y_j}\|^2 - \|\mathbf{y_k}\|^2 \right).$$

Similarly, the within-class scatter can be reformulated with the help of cosine hyperbolic framework as,

$$\widehat{L}_w = 2cosh\left(\frac{1}{\mathcal{L}}\sum_{i=1}^{\mathcal{L}}\frac{1}{N_i}\sum_{\substack{j,k\in c_i;\\ j\neq k}}\|\mathbf{y_j}\|^2 - \|\mathbf{y_k}\|^2\right).$$

Note that,

$$\|\mathbf{y}_i\|^2 = \sum_{k=1}^{D} y_{ik}^2 = \sum_{k=1}^{D}\left(\mathbf{w_k}^\top\mathbf{x_i}\right)\left(\mathbf{x_i}^\top\mathbf{w_k}\right) = \sum_{k=1}^{D}\mathbf{w_k}^\top\left(\mathbf{x_i}\mathbf{x_i}^\top\right)\mathbf{w_k}$$
$$= tr\left(W^\top(\mathbf{x_i}\mathbf{x_i}^\top)W\right).$$

So,

$$\widehat{L}_b = 2cosh\left(tr\left(W^\top L_b W\right)\right);\ \widehat{L}_w = 2cosh\left(tr\left(W^\top L_w W\right)\right);$$

where

$$L_b = \frac{1}{\mathcal{L}}\sum_{i=1}^{\mathcal{L}}\frac{1}{N_i}\sum_{\substack{j\in c_i;\\ k\notin c_i}}\left(\mathbf{x_j}\mathbf{x_j}^\top - \mathbf{x_k}\mathbf{x_k}^\top\right)$$

$$L_w = \frac{1}{\mathcal{L}}\sum_{i=1}^{\mathcal{L}}\frac{1}{N_i}\sum_{\substack{j,k\in c_i;\\ j\neq k}}\left(\mathbf{x_j}\mathbf{x_j}^\top - \mathbf{x_k}\mathbf{x_k}^\top\right).$$

These reformulations of $\widehat{L}_b$ and $\widehat{L}_w$ will be used in the objective function GNLDA of Eq. (3). For simplicity purpose, first let us consider the following objective function of maximizing between-class scatter only,

$$W_{opt} = \arg\max_{W^\top W = Id} cosh\left(tr\left(W^\top L_b W\right)\right).$$

From the symmetricity of the cosine hyperbolic function, the above objective function acquires the maximum, if and only if $tr(W^\top L_b W)$ obtain the maximum. Note that, L_b is neither positive nor negative semi-definite. The maximum of $tr(W^\top L_b W)$ will be the eigenvectors $\{\mathbf{w_1^+}, \mathbf{w_2^+}, ..., \mathbf{w_{d_1}^+}\}$ or $\{\mathbf{w_1^-}, \mathbf{w_2^-}, ..., \mathbf{w_{d_2}^-}\}$ according to $\sum_{i=1}^{d_1}\lambda_{d_1}^+ \lesseqgtr \sum_{i=1}^{d_2}|\lambda_{d_2}^-|$, $(d_1 + d_2 \leq D)$; where λ^+ and λ^- are positive and negative eigenvalues, respectively. But, both the positive (λ^+) and negative (λ^-) eigenvalues are of equal importance as they contain similar amount of discrimination information as $cosh\left((\mathbf{w}^+)^\top L_b \mathbf{w}^+ \text{ or } (\mathbf{w}^-)^\top L_b \mathbf{w}^-\right) = \cosh(\lambda^+ \text{ or } \lambda^+) > 0$. The following theorem helps us to deal with the situation.

Theorem 1. *If* $\mathbf{w_1}, \mathbf{w_2}, ..., \mathbf{w_D}$ *are the eigenvectors correspond to the eigenvalues* $\lambda_1, \lambda_2, ..., \lambda_D$ *of* L_b *then* $\mathbf{w_1}, \mathbf{w_2}, ..., \mathbf{w_D}$ *are also the eigenvectors correspond to the eigenvalues* $cosh(\lambda_1), cosh(\lambda_2), ..., cosh(\lambda_D)$ *of* $cosh(L_b)$.

Proof.

$$\cosh(L_b) = \sum_{t=0}^{\infty}\frac{(-1)^t}{(2t+1)!}L_b^{2t+1}$$
$$= I - \frac{L_b^2}{2!} + \frac{L_b^4}{4!} - \frac{L_b^6}{6!} + \cdots + \frac{L_b^t}{t!} + \ldots.$$

And, $\mathbf{w_i};\ i = 1, 2, \ldots, D$ are the eigenvectors corresponds to the eigenvalues $\lambda_i;\ i = 1, 2, \ldots, D$, i.e.,

$$L_b\mathbf{w_i} = \lambda_i\mathbf{w_i}$$

Thus,

$$\cosh(L_b)\mathbf{w_i} = I\mathbf{w_i} - \frac{L_b^2\mathbf{w_i}}{2!} + \frac{L_b^4\mathbf{w_i}}{4!} - \frac{L_b^6\mathbf{w_i}}{6!} + \cdots + \frac{L_b^t\mathbf{w_i}}{t!} + \ldots$$
$$= \mathbf{w_i} - \frac{\lambda_i^2\mathbf{w_i}}{2!} + \frac{\lambda_i^4\mathbf{w_i}}{4!} - \frac{\lambda_i^6\mathbf{w_i}}{6!} + \cdots + \frac{\lambda_i^t\mathbf{w_i}}{t!} + \ldots$$
$$= \cosh(\lambda_i).\mathbf{w_i}.$$

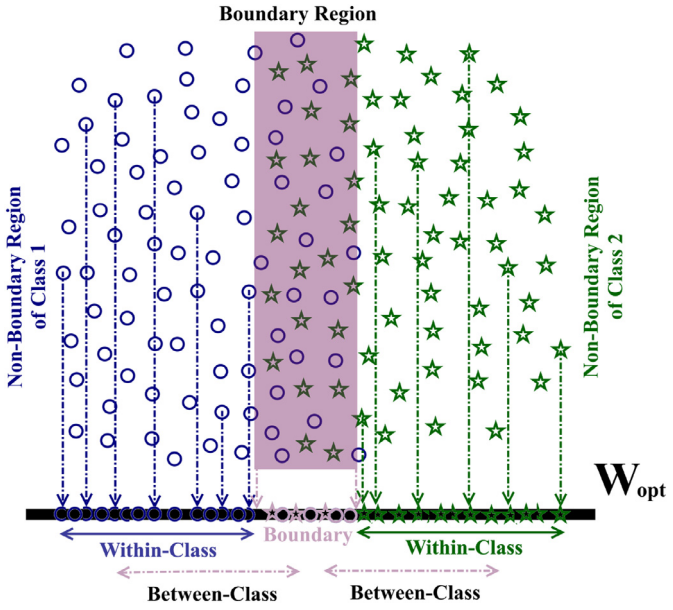


Fig. 8. Intuition behind NLDA.

That means, $\mathbf{w_i}$ is the eigenvector corresponds to the eigenvalue $\cosh(\lambda_i)$ of $\cosh(L_b)$, $\forall i = 1, 2, \ldots, D$. □

Note that, $cosh(\lambda_i) > 0;\ \lambda_i < 0$ *or* $\lambda_i \geq 0,\ \forall i = 1, 2, \ldots, D$. From the theorem, it is clear that X and $cosh(X)$ share the same eigenvectors and moreover, $cosh(X)$ have only non-zero eigenvalues. With the help of this theorem, the objective has been rewritten as,

$$W_{opt} = \arg\max_{W^\top W = Id} tr\left(W^\top \cosh(L_b) W\right).$$

Thus, both the positive and negative eigenvalues, and corresponding eigenvectors are also considered for the subspace formation. As $\cosh(x) \geq 1;\ x \in \mathbb{R}$, all the eigenvalues of $\cosh(S_b)$ are greater than or equal to 1, $\cosh(\lambda_i) \geq 1$. Thus, one can use $cosh(L_b)$ and $cosh(L_w)$ to reformulate the objective (Eq. (3)) of GNLDA as,

$$W_{opt} = \arg\max_{W^\top W = Id}\frac{tr\left(W^\top \cosh(L_b) W\right)}{tr(W^\top \cosh(L_w) W)}. \quad (5)$$

Note that the matrix $\cosh(L_w)$ is a full rank matrix and $cosh(L_w)^{-1}$ exists to avoid the small sample size problem. It can be solved by the following generalized eigenvalue decomposition method

$$\cosh(L_b)\mathbf{w_i} = \lambda_i \cosh(L_w)\mathbf{w_i}. \quad (6)$$

And $W_{opt} = [\mathbf{w_1}, \mathbf{w_2}, \ldots, \mathbf{w_d}]$ is the optimal solution whose columns are the eigenvectors $\mathbf{w_i}$ arranged in the non-decreasing order of first largest d eigenvalues $\mu_i;\ i = 1, 2, \ldots, d$.

3.4. Proposed method (NLDA)

A subspace (projection space) learning method is more efficient when the significant (insignificant) differences are emphasized (de-emphasized) with respect to the within and between-class scatterness (Na et al., 2010). In this work, firstly the effect of noisy patterns is removed and then the influence of boundary and non-boundary patterns on the proposed between and within-class scatter matrices is investigated. For better discrimination, non-boundary patterns should be projected far away from the boundary ones as boundary patterns represent the decision boundary region. Also, the projection of non-boundary patterns of each class as a group will lead to better compaction of the classes (Fig. 8).

From the Fig. 8, it is clear that by projecting non-boundary patterns far away from boundary patterns will leads to better discrimination and compaction of the classes. Thus, modelling between and within-class scatter matrices according to the locations of the data may generate better discriminant subspace. The modified between and within-class scatter matrices are formulated as,

$$\bar{L}_b = \frac{1}{N} \sum_{x_i \in \mathbb{B}_{oundary}} \sum_{x_j \in \mathbb{NB}_{oundary}} \left(\mathbf{x_i} \mathbf{x_i}^\top - \mathbf{x_j} \mathbf{x_j}^\top \right),$$

$$\bar{L}_w = \frac{1}{N} \sum_{i=1}^{\mathcal{L}} \sum_{\substack{j,k \in (c_i \wedge \mathbb{NB}_{oundary}); \\ j \neq k}} \left(\mathbf{x_j} \mathbf{x_j}^\top - \mathbf{x_k} \mathbf{x_k}^\top \right),$$

where N number of patterns in the input data. Clearly, the new between-class scatter matrix ($\bar{L}_b$) tries to model the separation between boundary and non-boundary patterns, and within-class scatter ($\bar{L}_w$) models the compaction of using non-boundary patterns, only. With the help of cosine hyperbolic framework and with similar kind of objective (Eq. (6)), the solution W_{opt} can be found by the following generalized eigenvalue decomposition method,

$$\cosh(\bar{L}_b) \mathbf{w_i} = \lambda_i \cosh(\bar{L}_w) \mathbf{w_i}. \tag{7}$$

The algorithm of NLDA can be found in Algorithm 2.

Algorithm 2: Noisy-free Length Discriminant Analysis (NLDA).

Input: A set of input patterns $\mathcal{X} \in \mathbb{R}^{N \times D}$, k_{NRPS} - number of nearest neighbors, d - desired number of features.
Output: The projection matrix W

$\{\mathbb{B}_{oundary}, \mathbb{NB}_{oundary}, \mathbb{N}_{oise}\} \leftarrow NRPS(\mathcal{X}, k_{NRPS})$
nb - number of non-boundary patterns, b - number of boundary patterns and $\mathcal{L}$ - number of classes
$\bar{L}_b \leftarrow 0$ and $\bar{L}_w \leftarrow 0$
for $i \leftarrow 1$ ***to*** b **do**
  $\mathbf{x_i} \in \mathbb{B}_{oundary}$
  for $j \leftarrow 1$ ***to*** nb **do**
    $\mathbf{x_j} \in \mathbb{NB}_{oundary}$
    $\bar{L}_b \leftarrow \bar{L}_b + \frac{1}{N}\left(\mathbf{x_i}.\mathbf{x_i}^\top - \mathbf{x_j}.\mathbf{x_j}^\top\right)$
for $i \leftarrow 1$ ***to*** $\mathcal{L}$ **do**
  $c_i \leftarrow i^{th}$ class
  $\widehat{c}_i \leftarrow c_i \cap \mathbb{NB}_{oundary}$
  for $j \leftarrow 1$ ***to*** $|\widehat{c}_i|$ **do**
    for $k \leftarrow j+1$ ***to*** $|\widehat{c}_i|$ **do**
      $\mathbf{x_j}, \mathbf{x_k} \in \widehat{c}_i$
      $\bar{L}_w \leftarrow \bar{L}_w + \frac{1}{N}\left(\mathbf{x_j}.\mathbf{x_j}^\top - \mathbf{x_k}.\mathbf{x_k}^\top\right)$
$\widehat{L}_b \leftarrow \cosh(\bar{L}_b)$; $\widehat{L}_w \leftarrow \cosh(\bar{L}_w)$
Solve the Eq. (7), and compute d largest eigenvectors $\mathbf{w_i}$; $i = 1, 2, \ldots, d$
return $W = [\mathbf{w_1}\ \mathbf{w_2} \ldots\ \mathbf{w_d}]$

3.5. Computational complexity

NLDA's computational complexity depends on mainly three aspects (i) computation of k-nearest neighbors for NRPS (ii) computation of cosine-hyperbolic of a matrix and (iii) finding eigenvalues and eigenvectors. The computation of k-nearest neighbors require $O(Nkd)$; N – number of training samples. And, $2cosh(A) = \exp(A) + exp(-A)$ and $exp(A)$ take $O(D^3)$ time. A wide variety of methods for computing exp(A) are analyzed in Moler and Loan (1978). The scaling and squaring method is one of the best methods for computing the matrix exponential (Higham, 2005). That means, the computation of $cosh(A)$ requires $2.O(D^3)$ time complexity. Most of the eigenvalue and eigenvector finders assume a tridiagonal matrix as an input. Householder reduction transforms the symmetric matrix to corresponding tridiagonal matrix in $4D^3/3$ operations (for eigenvalues only) and in $8D^3/3$ operations (for both eigenvalues and eigenvectors). The fastest method for eigen decomposition is the multiple relatively robust representations algorithm (Press, Teukolsky, Vetterling, & Flannery, 2007). In this method, inverse iteration determines the eigenvectors of a tridiagonal matrix in $O(D^2)$ operations. However, clustered eigenvalues lead to eigenvectors that are not properly orthogonal to one another. Using a procedure like Gram-Schmidt to orthogonalize the vectors requires $O(D^3)$ operations. The MRRR algorithm is a sophisticated version of inverse iteration that is $O(N^2)$ without requiring Gram-Schmidt. Therefore, the time complexity of NLDA is $O(Nkd + D^3)$.

4. Experimental study

We first prove that NLDA is efficient at handling multimodal data with the help of four (4) synthetic[1] Gaussian mixture random data. Next, twenty (20) UCI machine learning, leeds butterfly and extended yale b data sets were used to analyze the performance of the NLDA with the state-of-the-art methods. All the experiments have been performed using Matlab R2015b (version 8.6) on a Dell-Optiplex 990 machine with 16GB RAM and Intel (R) Core (TM) i7-2600 cpu @ 3.40 GHz, 3.40 GHz processor.

4.1. Study on multimodal synthesized data

In this experimental set-up, two dimensional multimodal random data have been randomly generated using the Gaussian mixture distribution. Four different data sets with different parameters of the above distribution are presented in Table 2. For each class of such data, 1000 patterns have been generated randomly. Projection vectors of different DR methods are plotted and shown in the Fig. 9. In all the figures, thick and thin vectors represent eigenvectors corresponding to largest and smallest eigenvalues, respectively. Five different separability measures (Sotoca, Mollineda, & Sanchez, 2006) namely, volume of overlap region (*overlap*), Thornton's Separability Index (*TSI*), Fraction of points on Boundary (*FB*), volume of local neighborhood (*volume*) and Non-parametric Separability (*NS*) measures have been used to evaluate the performance of the methods. The average (20 iterations) results of these measures for projected data of different DR methods on first principal direction (eigenvector corresponding to the largest eigenvalue) can be found in the Table 3.

4.1.1. Evaluation metrics (Sotoca et al., 2006)

This section consists of basic definitions for volume of overlap region (*overlap*), Thornton's Separability Index (*TSI*), Fraction of points on Boundary (*FB*), Non-parametric Separability (*NS*) and Volume of local neighborhood (*volume*) (Sotoca et al., 2006) which were used to evaluate the performances of different methods over synthetic data.

*4.1.1.1. Volume of overlap region (*overlap*).* This measure computes, for each feature f_k, the length of the overlap range normalized by the length of the total range in which all values of both classes are distributed. Then the volume of the overlap region for $\mathcal{L}$ classes is obtained as the sum over all class pairs of products of normalized

[1] NLDA mex file and synthetic data sets are available at http://www.isical.ac.in/~k.ramachandra/NLDA.html.

Table 2
Parameters of mixture Gaussian distribution that were used to generate random data for two classes. Here, for any class data, μ, Σ and P denote mean, covariance, and prior probabilities, respectively. And subscripts 1 and 2 denote class 1 and 2, respectively.

Name	Parameters
GM$_1$	$\mu_1 = [(0,\ 0),\ (6,\ 6)],\ \Sigma_1 = [[2\ 1; 1\ 2],\ [2\ -1; -1\ 2]],\ P_1 = (0.5,\ 0.5);$ $\mu_2 = [(0,\ 6),\ (6,\ 0)],\ \Sigma_2 = [[1\ 0; 0\ 1],\ [1\ 0; 0\ 1]], P_2 = (0.5,\ 0.5)$
GM$_2$	$\mu_1 = [(0,\ 0),\ (6,\ 6)],\ \Sigma_1 = [[2\ 0; 0\ 2],\ [2\ 0; 0\ 2]],\ P_1 = (0.5,\ 0.5);$ $\mu_2 = [(0,\ 6),\ (6,\ 0)],\ \Sigma_2 = [[1\ 0; 0\ 1],\ [1\ 0; 0\ 1]],\ P_2 = (0.5,\ 0.5)$
GM$_3$	$\mu_1 = [(0,\ 0),\ (6,\ 6)],\ \Sigma_1 = [[2\ -1; -1\ 2],\ [2\ 0; 0\ 2]],\ P_1 = (0.5,\ 0.5);$ $\mu_2 = [(0,\ 6),\ (6,\ 0)],\ \Sigma_2 = [[1\ 0; 0\ 1],\ [1\ 0; 0\ 1]],\ P_2 = (0.5,\ 0.5)$
GM$_4$	$\mu_1 = [(0,\ 0),\ (6,\ 6)],\ \Sigma_1 = [[2\ -1; -1\ 2],\ [2\ 1; 1\ 2]],\ P_1 = (0.5,\ 0.5);$ $\mu_2 = [(0,\ 6),\ (6,\ 0)],\ \Sigma_2 = [[1\ 0; 0\ 1],\ [1\ 0; 0\ 1]],\ P_2 = (0.5,\ 0.5)$

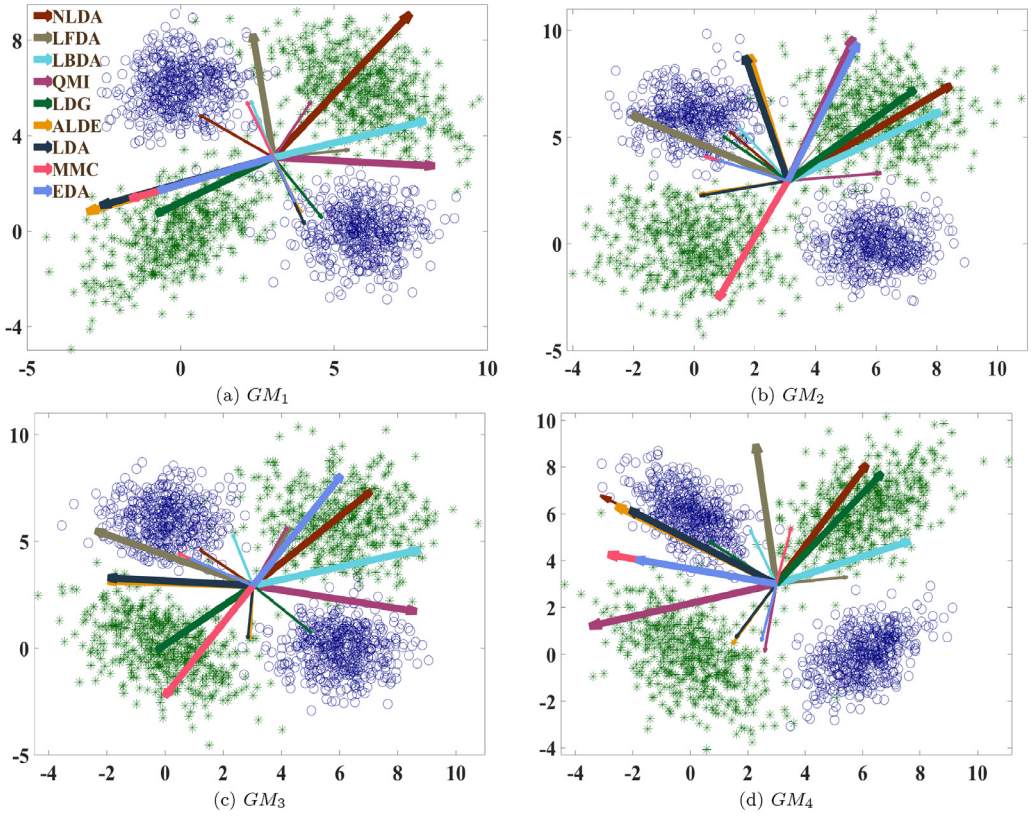


Fig. 9. Orthonormal eigenvectors of different DR methods. Thick and thin vectors represents the eigenvectors corresponding to the largest and smallest eigenvalues, respectively. o (blue) - class 1 and * (green) - class 2. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

lengths of overlapping ranges for all features.

$$overlap = \sum_{(c_i,c_j)} \prod_{k=1}^{d} \frac{min\ max_k - max\ min_k}{max\ max_k - min\ min_k},$$

where (c_i, c_j) goes through all pairs of classes, and

$$\begin{aligned} min\ max_k &= min\ \{max(f_k,\ c_i), max(f_k,\ c_j)\},\\ max\ min_k &= max\ \{min(f_k,\ c_i), min(f_k,\ c_j)\},\\ max\ max_k &= max\ \{max(f_k,\ c_i), max(f_k,\ c_j)\},\\ min\ min_k &= min\ \{min(f_k,\ c_i), min(f_k,\ c_j)\}. \end{aligned}$$

*4.1.1.2. Thornton's Separability Index (*TSI*).* *TSI* is defined as the fraction of a set of data points whose classification labels are the same as those of their nearest neighbours. Thus it is a measure of the degree to which inputs associated with the same output tend to cluster together. It may be written as

$$TSI = \frac{\sum_{i=1}^{N} \left(f(\mathbf{x_i}) + f(\mathbf{x_i'}) + 1\right) mod\ 2}{N}$$

where $\mathbf{x_i'}$ is the nearest neighbour of $\mathbf{x_i}$ and N is the number of patterns. It is intuitively obvious that *TSI* will be close to 1 for a set of patterns in which patterns with opposite labels exist in tight, well-separated clusters. As the clusters move closer together and patterns from opposing classes begin to overlap, the index begins to fall. If the centroids of the clusters coincide, or the patterns are uniformly distributed in the space without clustering, the nearest neighbour of a point will have no more than 50% probability of having the same class label as its neighbour, and the separability index will be close to 0.5. A regular intermeshed grid of alternately-labelled patterns (as would be generated by the exclusive-OR or parity problems) would have $TSI = 0$.

Table 3
Separability measures on randomly generated two class data for Gaussian mixture distribution. Here, Red and Blue color values represent best and second best results, respectively.

	FB	NS	TSI	Overlap	Volume	FB	NS	TSI	Overlap	Volume
	GM_1					GM_2				
NLDA	**0.0895**	**0.0139**	**0.9110**	**0.4246**	**0.0562**	**0.0889**	**0.0127**	**0.9096**	**0.4116**	**0.0579**
LFDA	0.3238	0.2016	0.6782	0.7358	0.0605	0.1935	0.0791	0.8069	0.6579	0.0598
LBDA	0.3664	0.2903	0.6267	0.7246	0.0715	0.2992	0.1499	0.6971	0.6888	0.0709
QMI	0.2052	0.0433	0.7971	0.5397	0.0604	0.2322	0.1055	0.7685	0.5548	0.0615
LDG	0.1003	0.0156	0.8984	0.4514	0.0572	0.1043	0.0153	0.8928	0.4373	0.0581
ALDE	0.2577	0.1359	0.7427	0.6599	0.0606	0.3155	0.1872	0.6878	0.7078	0.0605
LDA	0.2635	0.1410	0.7359	0.6687	0.0604	0.3159	0.1904	0.6831	0.7151	0.0603
MMC	0.2558	0.1486	0.7428	0.7071	0.0586	0.2333	0.1188	0.7671	0.6686	0.0581
EDA	0.2558	0.1486	0.7429	0.7071	0.0585	0.2333	0.1188	0.7669	0.6686	0.0579
	GM_3					GM_4				
NLDA	0.0950	0.0145	0.9044	0.4709	0.0524	**0.1419**	**0.0245**	**0.8585**	**0.4846**	**0.0556**
LFDA	0.2082	0.0645	0.7902	0.7944	0.0580	0.1952	0.0745	0.8079	0.6151	0.0579
LBDA	0.3648	0.2724	0.6358	0.7642	0.0748	0.3228	0.1420	0.6744	0.7168	0.0717
QMI	0.2588	0.0703	0.7397	0.6212	0.0604	0.3590	0.1551	0.6409	0.6681	0.0635
LDG	**0.0688**	**0.0094**	**0.9314**	**0.4276**	**0.0507**	0.1719	0.0260	0.8306	0.5145	0.0597
ALDE	0.2515	0.1998	0.7480	0.7163	0.0596	0.2974	0.1718	0.7050	0.7417	0.0597
LDA	0.2486	0.1791	0.7511	0.7235	0.0597	0.2967	0.1891	0.7019	0.7471	0.0593
MMC	0.1696	0.0694	0.8310	0.5197	0.0547	0.2217	0.1122	0.7822	0.7318	0.0573
EDA	0.1696	0.0694	0.8310	0.5197	0.0546	0.2217	0.1122	0.7822	0.7318	0.0572

4.1.1.3. Fraction of points on Boundary (FB). Friedman and Rafsky proposed a test on whether two samples are from the same distribution or not? It is thus useful for deciding if the patterns labeled as two classes form separable distributions. The method relies on computing a Minimum Spanning Tree (MST) that connect all the patterns to their nearest neighbors (regardless of class). Then the number of patterns connected to the opposite class by an edge in the MST are counted. These are considered to be the patterns lying next to the class boundary. The fraction of such patterns over all patterns in the data set is used as a measure (*FB*). Therefore, the smaller value represents less number of patterns nearer to the boundary.

4.1.1.4. Non-parametric Separability (NS). Let $\mathcal{N}_1^{=}(\mathbf{x_i})$ and $\mathcal{N}_1^{\neq}(\mathbf{x_i})$ be the intra-class nearest neighbor and the inter-class nearest neighbor of a given example $(\mathbf{x_i}, c_i)$, respectively. Then, *NS* can be computes as follows:

$$NS = \frac{\sum_{i=1}^{N} d(\mathcal{N}_1^{=}(\mathbf{x_i}),\ \mathbf{x_i})}{\sum_{i=1}^{N} d(\mathcal{N}_1^{\neq}(\mathbf{x_i}),\ \mathbf{x_i})},$$

$d(.)$ is the Euclidean distance. The proximity of patterns in opposite classes affects the error rate of a nearest neighbor classifier.

4.1.1.5. Volume of local neighborhood (volume). This measure represents the average volume occupied by the with-in class k - nearest neighbors of each instance. Let $\mathcal{N}_k^i(\mathbf{x})$ be the set of with-in class k - nearest neighbors of a given example $\mathbf{x}$ from class i, then the volume of this can be defined as follows:

$$volume(\mathbf{x}) = \prod_{h=1}^{d} \left(max\left(f_h, \mathcal{N}_k^i(\mathbf{x})\right) - min\left(f_h, \mathcal{N}_k^i(\mathbf{x})\right) \right),$$

where $max\left(f_h, \mathcal{N}_k^i(\mathbf{x})\right)$ and $min\left(f_h, \mathcal{N}_k^i(\mathbf{x})\right)$ represent the maximum and minimum values of feature f_h among the with-in class $k-$nearest neighbors of instance $\mathbf{x}$ from class i. From this, the average volume of with-in class local neighborhood can be expressed as,

$$volume = \frac{1}{\mathcal{L}} \sum_{i=1}^{\mathcal{L}} \left(\frac{1}{N_i} \sum_{\mathbf{x} \in c_i} volume(\mathbf{x}) \right).$$

It is clear that, the smaller value of *volume* represents better compaction of the classes.

Table 3 and Fig. 9 provide the experimental results of NLDA with the state-of-the-art methods, QMI, ALDE, LDG, EDA, LBDA, LFDA, MMC and LDA w.r.t the above mentioned separability measures. NLDA outperforms QMI, ALDE, EDA, LBDA, LFDA, MMC and LDA over all the separability measures, whereas LDG is the second best performer. The proposed NLDA showed 14.77%, 16.99%, 1.88% and 5.87% improvement for *FB, NS, TSI* and *overlap*, respectively, over LDG on GM_2. And for GM_1, NLDA's performance improved by 10.77%, 10.89%, 1.4%, 6.3% and 1.7% for *FB, NS, TSI, overlap* and *volume*, respectively, in comparison with LDG. Similarly, 17.45%, 5.77%, 3.36%, 5.81% and 6.87% improvement of NLDA over LDG for GM_4 dataset (Table 3). Whereas, on GM_3, LDG is the best and NLDA is the second best performers for the above measures. Also, geometrically (Fig. 9) one can note that, NLDA's principal direction (i.e., eigenvector corresponds to largest eigenvalue) projects classes far way from the other classes, i.e. it is simplifying the decision boundary between the classes. Clearly, NLDA outperforms others on GM_1, GM_2, and GM_4 data, and comparable on GM_3. Thus, it is proved that NLDA's subspace is better suitable for multimodal data compared to QMI, ALDE, LDG, EDA, LBDA, LFDA, MMC and LDA.

4.2. Study and analysis on UCI machine learning data

The list of twenty (20) UCI machine learning datasets, that are used for experimental purpose, can be found in Table 4. In this

Table 4
List of UCI machine learning data sets used for experimental purpose.

Datasets	Number of Patterns (N)	Dimensions (D)	Classes ($\mathcal{L}$)
Australian Credit Approval	690	14	2
Breast Cancer Wisconsin (Diagnostic)	569	32	2
Breast Cancer Wisconsin (Prognostic)	198	34	2
Climate Model Simulation Crashes	540	18	2
Alon Colon Cancer	62	2000	2
Connectionist Bench (Sonar, Mines, Rocks)	208	60	2
Hill-Valley	1212	101	2
Indian Liver Patient	583	10	2
Ionosphere	351	34	2
Madelon	4400	500	2
Musk (Version 2)	6598	168	2
Parkinsons	197	23	2
Planning Relax	182	13	2
Ringnorm	7400	21	2
Robot Execution Failures (LP1)	88	91	4
Robot Execution Failures (LP4)	117	91	3
Statlog (Landsat Satellite)	6435	36	6
Spambase	4601	57	2
SPECTF Heart	267	44	2
Vehicle Silhouettes	946	18	4

work, nearest neighborhood classifier and 10−fold cross validation approach have been adopted to compute the f-score, error rates and g-means of different DR methods. Throughout the experimental study, the parameters of NLDA have been varied as $k_{NRPS} \in \{7, 15, 25\}$ and $Const \in \{0.25, 0.5, 0.75\}$.

4.2.1. Evaluation metrics (Alejo, García, Sotoca, Mollineda, & Sánchez, 2007; Sokolova & Lapalme, 2009)

A classifier can be evaluated by computing the number of correctly classified examples (true positives), the number of correctly recognized examples that do not belong to the class (true negatives), and examples that either were incorrectly assigned to the class (false positives) or that were not recognized as class examples (false negatives). With the help these, one can define the metrics as follows:

$$precision = \frac{1}{\mathcal{L}} \sum_{i=1}^{\mathcal{L}} \frac{tp_i}{(tp_i + fp_i)},$$

$$recall = \frac{1}{\mathcal{L}} \sum_{i=1}^{\mathcal{L}} \frac{tp_i}{(tp_i + fn_i)},$$

$$F - Score\ (F_1) = \frac{2 * precision * recall}{precision + recall},$$

$$Error\ Rate\ (e) = \frac{1}{\mathcal{L}} \sum_{i=1}^{\mathcal{L}} \frac{fp_i + fn_i}{tp_i + fn_i + fp_i + tn_i},$$

$$g - mean\ (\mu_g) = \left(\prod_{i=1}^{\mathcal{L}} \frac{tp_i}{tp_i + fn_i + fp_i + tn_i} \right)^{\frac{1}{\mathcal{L}}},$$

$$Average\ Accuracy\ (\mu) = \frac{1}{\mathcal{L}} \sum_{i=1}^{\mathcal{L}} \frac{tp_i + tn_i}{tp_i + fn_i + fp_i + tn_i}$$

where tp_i, tn_i, fp_i and fn_i are true positive, true negative, false positive and false negative of class i; $i = 1, 1, \ldots, \mathcal{L}$, respectively. Precision is a measure of a classifier exactness where a low precision indicate a large number of false positives. Recall can be thought of as a measure of a classifier's completeness, where a low recall indicates many false negatives. And the F-score conveys the balance between the precision and recall. The error rate is the average per class classification error. Other measure g-mean is the geometric mean that is used to quantify the classifier performance in the class imbalance (Alejo et al., 2007).

4.2.2. Comparisons with state-of-the-art methods

We compared the proposed method, Noisy-free Length Discriminant Analysis (NLDA), with the state-of-the-art methods like ALDE (Liu et al., 2015), QMI (Bouzas et al., 2015), LDG (Parrish & Gupta, 2012), EDA (Zhang et al., 2010), LBDA (Na et al., 2010), LFDA (Sugiyama, 2007), MMC (Haifeng et al., 2006; Yang et al., 2009) and LDA (Theodoridis & Koutroumbas, 2008) over the 20 UCI machine learning datasets (Table 4). The average (of 30 iterations) results of $F - score\ (F_1)$, $ErrorRate\,(e)$ and $g - mean\ (\mu_g)$ of different methods over the 20 UCI machine learning datasets can be found in the Tables 5 and 6.

4.2.2.1. Analysis of $F - Scores$ (F_1). NLDA's F_1 results are improved by 1.8%, 2.66%, 2.2%, 20.04%, 2.85%, 2.35%, 4.55%, 2.05%, 3.09%, 2.64% and 1.79% over Breast Cancer Wisconsin (Diagnostic), Breast Cancer Wisconsin (Prognostic), Alon Colon Cance, Hill-valley, Madelon, Musk (Version 2), Ringnorm, Robot Execution Failures (LP4), Statlog (Landsat Satellite), Spambase and SPECTF Heart, respectively; compared to the second best methods LFDA, MMC, MMC, MMC, LFDA, MMC, LDG, LFDA, LFDA, LFDA, LDG and LFDA, respectively. And NLDA achieves best F_1 results over Australian Credit Approval, Connectionist Bench (Sonar, Mines, Rocks), Parksions and Robot Execution Failures (LP1); but comparable to state-of-the-art methods. On Climate Model Simulation Crashes, Indian Liver Patient, Ionosphere, Planning Relax and Vehicle Silhouettes; NLDA's performance not up to the mark. Overall, NLDA outperforms state-of-the-art methods on majority of the data in terms of F_1 measure.

4.2.2.2. Analysis of error rates (e). The error rates of NLDA over Breast Cancer Wisconsin (Diagnostic), Breast Cancer Wisconsin (Prognostic), Hill-Valley, Madelon, Musk (Version 2), Ringnorm, Robot Execution Failures (LP4), Spambase and SPECTF Heart showed 47.93%, 7.71%, 89% 7.31%, 10%, 27.42%, 20.13%, 48.53% and 1.9% improved compared to the second best performers, i.e., LFDA, MMC and LDG. For Australian Credit Approval, Alon Colon Cancer, Parkinsons, Planning Relax, Robot Execution Failures (LP1) and Statlog (Landsat Satellite); NLDA's performances are best as well as comparable to the other methods. NLDA's performances are not the first best, but are performing better on other datasets. Thus, average classification error in the NLDA's subspace (projection space) is very small compared to the other methods.

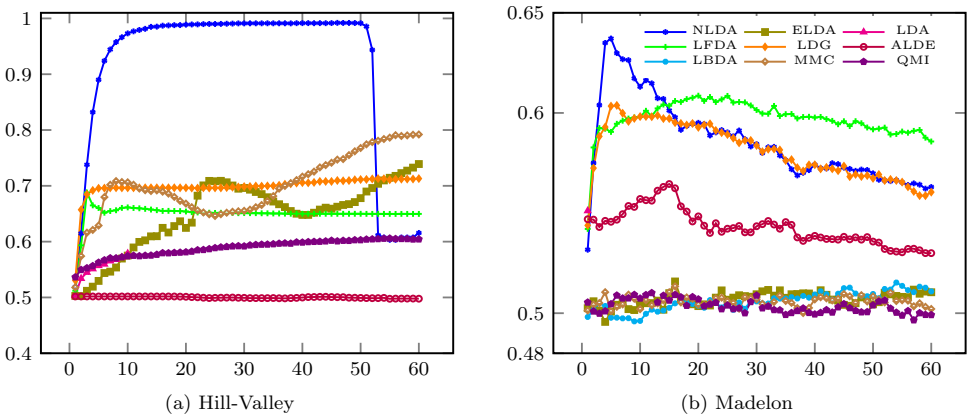
Table 5
Experimental results of different methods on UCI machine learning datasets. Red colored text represents the best values.

Methods	Evaluation Metrics	Australian Credit Approval	Breast Cancer Wisconsin (Diagnostic)	Breast Cancer Wisconsin (Prognostic)	Climate Model Simulation Crashes	Alon Colon Cancer	Connectionist Bench (Sonar, Mines, Rocks)	Hill-Valley	Indian Liver Patient	Ionosphere	Madelon	Best Rank Count
NLDA	F_1	**0.8186**	**0.9809**	**0.6657**	0.7389	**0.8556**	**0.8745**	**0.9921**	0.6237	0.8837	**0.6367**	**7/10**
	e	**0.1791**	**0.0214**	**0.2513**	0.0758	**0.2190**	0.1304	**0.0210**	0.3049	0.1011	**0.3628**	**6/10**
	μ_g	**0.4064**	**0.4930**	**0.2723**	0.1860	**0.4027**	**0.4343**	**0.4960**	0.2703	0.4121	**0.3180**	**7/10**
ELDA	F_1	**0.8190**	0.9529	0.6328	0.8039	0.7735	0.8639	0.7386	0.6034	**0.8994**	0.5153	**2/10**
	e	**0.1787**	0.0456	0.2731	**0.0589**	0.3012	0.1397	0.2610	0.3280	**0.0900**	0.4842	**3/10**
	μ_g	**0.4067**	0.4610	0.2560	**0.2150**	0.3665	0.4287	0.3689	0.2629	**0.4243**	0.2573	**3/10**
LDA	F_1	0.7863	0.9018	0.5702	0.7927	0.7693	0.6594	0.5780	0.5878	0.7419	0.5506	**-**
	e	0.2106	0.0914	0.3422	0.0610	0.2750	0.3341	0.4213	0.3356	0.2364	0.4489	**-**
	μ_g	0.3901	0.4349	0.2255	0.2082	0.3640	0.3257	0.2886	0.2515	0.3526	0.2751	**-**
LFDA	F_1	0.8012	0.9629	0.6350	0.7880	0.7543	**0.8787**	0.6872	0.6022	0.8822	0.6082	**1/10**
	e	0.1954	0.0411	0.2723	0.0594	0.3048	**0.1289**	0.3120	0.3287	0.1040	0.3914	**1/10**
	μ_g	0.3970	0.4646	0.2563	0.2074	0.3515	**0.4369**	0.3431	0.2628	0.4137	0.3038	**1/10**
LDG	F_1	0.8088	0.9467	0.6154	0.8077	0.7852	0.8647	0.7122	**0.6444**	0.8789	0.6034	**1/10**
	e	0.1886	0.0523	0.2922	0.0607	0.2694	0.1395	0.2870	**0.2940**	0.1066	0.3962	**1/10**
	μ_g	**0.4013**	0.4571	0.2497	**0.2159**	0.3683	0.4287	0.3556	**0.2846**	0.4133	0.3015	**3/10**
ALDE	F_1	**0.8122**	0.9487	0.6173	**0.8116**	0.7752	0.8656	0.5006	**0.6471**	0.8694	0.5640	**3/10**
	e	0.1854	0.0502	0.2834	**0.0570**	0.2786	0.1365	0.4983	**0.2946**	0.1166	0.4355	**2/10**
	μ_g	**0.4032**	0.4577	0.2451	**0.2164**	0.3625	0.4291	0.2497	**0.2874**	0.4086	0.2817	**3/10**
LBDA	F_1	**0.8139**	0.9521	0.6146	0.7172	0.7635	0.8596	-	0.6060	0.8532	0.5054	**1/10**
	e	0.1836	0.0486	0.2866	0.0819	0.2791	0.1416	-	0.3266	0.1246	0.4847	**-**
	μ_g	**0.4039**	0.4595	0.2453	0.1767	0.3482	0.4262	-	0.2645	0.3991	0.2524	**1/10**
MMC	F_1	**0.8176**	0.9540	0.6391	0.8028	0.8386	0.8617	0.7917	0.6015	**0.8967**	0.5128	**2/10**
	e	**0.1797**	0.0451	0.2731	0.0637	**0.2180**	0.1448	0.2080	0.3300	**0.0930**	0.4866	**3/10**
	μ_g	**0.4057**	0.4608	0.2607	**0.2171**	**0.4011**	0.4275	0.3955	0.2623	**0.4235**	0.2561	**4/10**
QMI	F_1	0.7931	0.9407	0.6152	0.6557	0.6563	0.8043	0.6049	0.6206	0.8885	0.5097	**-**
	e	0.2032	0.0580	0.2960	0.1149	0.3926	0.1957	0.3941	0.3167	**0.0989**	0.4899	**1/10**
	μ_g	0.3927	0.4536	0.2487	0.1599	0.3108	0.3965	0.3019	0.2736	0.4171	0.2546	**-**

Table 6
Experimental results of different methods on UCI machine learning datasets. Red colored text represents the best values.

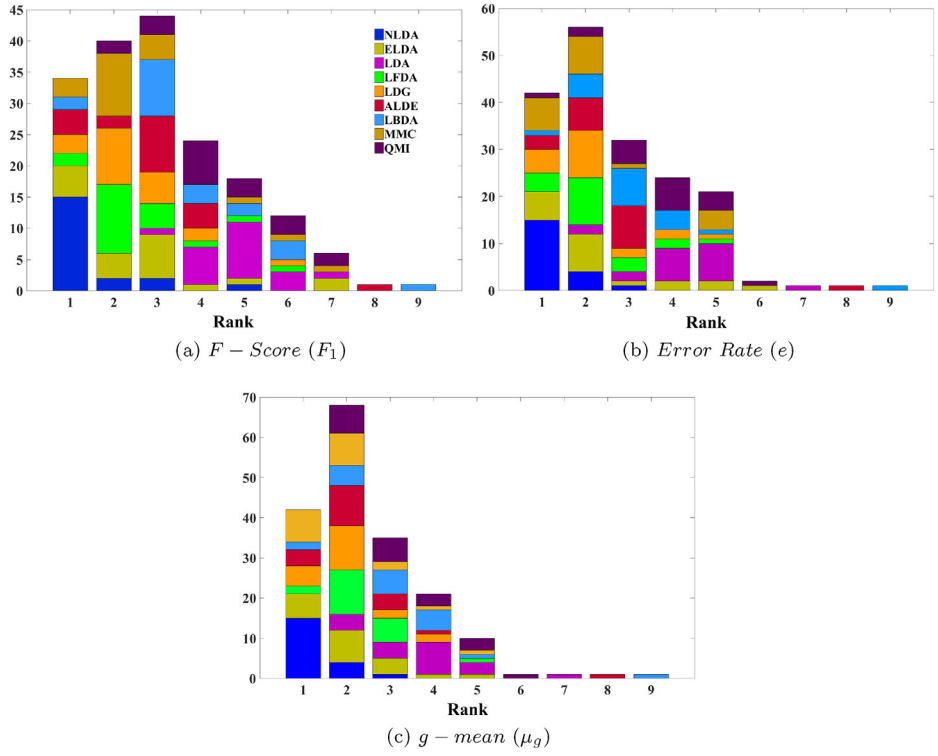
Methods	Evaluation Metrics	Musk (Version 2)	Parkinsons	Planning Relax	Ringnorm	Robot Execution Failures (LP1)	Robot Execution Failures (LP4)	Statlog (Landsat Satellite)	Spambase	SPECTF Heart	Vehicle Silhouettes	Best Rank Count
NLDA	F_1	**0.9798**	**0.9332**	0.5695	**0.9140**	**0.8901**	**0.9179**	**0.9313**	**0.9482**	**0.6863**	0.6988	**8/10**
	e	**0.0207**	**0.0766**	**0.3496**	**0.0948**	**0.0927**	**0.0762**	**0.0215**	**0.0386**	**0.2081**	0.1495	**9/10**
	μ_g	**0.3604**	**0.4041**	0.2342	**0.4901**	**0.2074**	**0.2521**	**0.1875**	**0.4798**	**0.2643**	0.1652	**8/10**
ELDA	F_1	0.9427	**0.9319**	**0.5817**	0.8667	0.8573	0.8195	0.8921	0.9198	0.5909	**0.7529**	**3/10**
	e	0.0303	**0.0785**	**0.3486**	0.1321	0.1055	0.1320	0.0312	0.0769	0.2908	**0.1236**	**3/10**
	μ_g	0.3426	**0.4036**	**0.2463**	0.4314	0.1990	0.2362	0.1375	0.4502	0.2314	**0.1819**	**3/10**
LDA	F_1	0.6707	0.6913	0.5379	0.6837	0.6671	0.8535	0.8665	0.8435	0.6251	0.5470	**-**
	e	0.1711	0.2326	0.3930	0.3162	0.1667	0.1183	0.0376	0.1496	0.2526	0.2256	**-**
	μ_g	0.2271	0.2867	0.2259	0.3418	0.1542	0.2389	0.1328	0.4121	0.2340	0.1284	**-**
LFDA	F_1	0.9425	**0.9385**	0.5721	0.8637	0.8534	0.8974	0.9004	0.9108	0.6684	0.7015	**1/10**
	e	0.0304	**0.0748**	0.3544	0.1347	0.1113	0.0954	**0.0286**	0.0851	0.2121	0.1479	**2/10**
	μ_g	0.3423	**0.4058**	0.2353	0.4298	0.1991	0.2465	0.1386	0.4447	0.2485	0.1660	**1/10**
LDG	F_1	0.9559	**0.9399**	0.5757	0.8685	**0.8907**	0.8368	0.8989	0.9218	0.6552	0.7009	**2/10**
	e	**0.023**	**0.0747**	0.3510	0.1306	**0.0948**	0.1210	**0.0290**	0.0750	0.2327	0.1484	**4/10**
	μ_g	0.3451	**0.4071**	0.2387	0.4329	**0.2066**	0.2319	0.1385	0.4513	0.2546	0.1657	**2/10**
ALDE	F_1	0.9454	**0.9316**	0.5680	0.8684	0.8476	0.8872	0.8917	0.9154	0.6563	0.6986	**1/10**
	e	0.0288	**0.0832**	0.3569	0.1307	0.1181	0.1015	0.0313	0.0811	0.2302	0.1496	**1/10**
	μ_g	0.3434	**0.4022**	0.2361	0.4327	0.1968	0.2491	0.1376	0.4481	0.2576	0.1665	**1/10**
LBDA	F_1	0.9202	**0.9309**	0.5660	0.8561	0.8499	0.8316	0.8919	0.9115	0.6421	0.6986	**1/10**
	e	0.0421	**0.0748**	0.3541	0.1425	0.1161	0.1196	0.0313	0.0843	0.2399	0.1493	**1/10**
	μ_g	0.3335	**0.4039**	0.2342	0.4263	0.1978	0.2290	0.1376	0.4446	0.2429	0.1652	**1/10**
MMC	F_1	0.9563	0.9293	0.5751	0.8666	0.8811	0.8354	0.8910	0.9194	0.5942	**0.7530**	**1/10**
	e	**0.0233**	0.0801	**0.3493**	0.1322	**0.0984**	0.1243	0.0315	0.0772	0.2901	**0.1236**	**4/10**
	μ_g	0.3490	**0.4029**	**0.2444**	0.4315	**0.2048**	0.2366	0.1374	0.4501	0.2356	**0.1819**	**4/10**
QMI	F_1	0.9096	0.9159	0.5596	0.8297	0.7992	0.8443	0.8800	0.9028	0.6010	0.6795	**-**
	e	0.0482	0.0857	0.3775	0.1677	0.1541	0.1191	0.0345	0.0928	0.2943	0.1589	**-**
	μ_g	0.3317	0.3956	0.2361	0.4117	0.1864	0.2334	0.1354	0.4411	0.2357	0.1610	**-**

4.2.2.3. Analysis of $g-means$ (μ_g). For Breast Cancer Wisconsin (Diagnostic), Breast Cancer Wisconsin (Prognostic), Hill-Valley, Madelon, Musk (Version 2), Ringnorm, Robot Execution Failures (LP4), Statlog (Landsat Satellite), Spambase and SPECTF Heart; the $g-means$ (μ_g) results of NLDA are improved by 6.11%, 4.45%, 25.4%, 4.67%, 3.27%, 13.21%, 1.2%, 35.71%, 6.32% and 2.6%, respectively, over other state-of-the-methods. And, NLDA achieving best as well as comparable with other methods over Australian Credit Approval, Alon Colon Cancer, Connectionist Bench (Sonar, Mines, Rocks), Parkinsons and Robot Execution Failures (LP1). For rest of the datasets, it is performing better but not the best one. Note that, LBDA fails to produce results on Hill-Valley since the threshold $Th = 1 - 1/\mathcal{L}$ leads to empty set of non-boundary patterns.



(a) Hill-Valley (b) Madelon

Fig. 10. Number of dimensions vs. average accuracy (μ) on UCI machine learning datasets for state-of-the-art methods.



(a) $F-Score$ (F_1) (b) $Error\ Rate$ (e)

(c) $g-mean$ (μ_g)

Fig. 11. Rankings of different methods on the three metrics for UCI machine learning datasets.

4.2.2.4. Rank Analysis. The Fig. 11 consists of the rankings for different methods over the 20 UCI machine learning methods. Rankings of the methods over $F-score$ (F_1) can be found in Fig. 11(a) and it is evident that NLDA is leading one in the first (1) rank. Note that, NLDA's least rank is five (5). ELDA is the second best method in the first rank, but it's least rank is seven (7). Also, ALDE is the third best method in the first rank and least rank is the eighth position. From, Fig. 11(b), it is clear that NLDA is achieving top three ranks over all the datasets in case of the metric $Error\,Rate\,(e)$. One can note that, ELDA and LFDA are the second and third performers over all the datasets. Even on $g-means$ (μ_g) (Fig. 11(c)), NLDA is the leading method compare to the other methods. Thus, NLDA is the leading method over all three metrics and it's performance is better than the other methods.

We also analyzed the performance of NLDA and comparative methods over Hill-Valley and Madelon using $Average\,Accuracy\,(\mu)$ up to 60 dimensions (features) as shown in the Fig. 10. For Hill-Valley (Fig. 10(a)) data, NLDA outperforms all other methods and achieves the highest accuracy value. Also, on Madelon (Fig. 10(b)), NLDA classification accuracy improved by 4% compare to all the other methods. And NLDA achieves best accuracy using only seven (7) transformed features. For all the above experimental study, one can conclude that NLDA's projection space is better suitable for supervised learning compare to the state-of-the-art methods.

4.2.3. Analysis of $Th' = Const * log_2(\mathcal{L})$

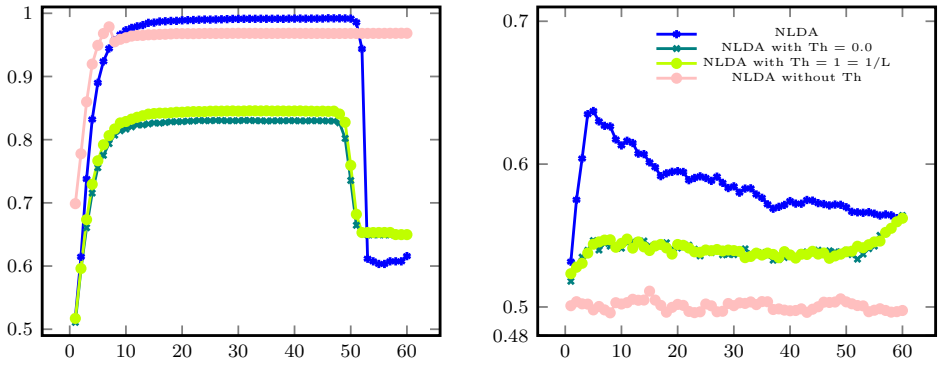
In this section, we analyse the importance of the proposed threshold $Th' = Const * log_2(\mathcal{L})$ compare to $Th = 0.0$ (NPPS) and $Th = 1 - 1/\mathcal{L}$ (LBDA). Also, we considered the method '*NLDA without Th*' which solves Eq. (6) without any relevant pattern selection. Table 7 depicts the results of three metrics $F-Score$ (F_1), $Error\,Rate\,(e)$ and $g-mean$ (μ_g) over 20 UCI machine learning data (Table 4). It is evident from the table that $Th' = Const * log_2(\mathcal{L})$ completely outperforms other thresholds over all the datasets. The results of '*NLDA* ($Th' = Const * log_2(\mathcal{L})$)' are superior to other thresholds for all the three metrics.

Also, the average classification accuracies across different dimensions (features) are displayed in the Fig. 12 for Hill-Valley and

Table 7
Experimental results on UCI machine learning data by varying thresholds.

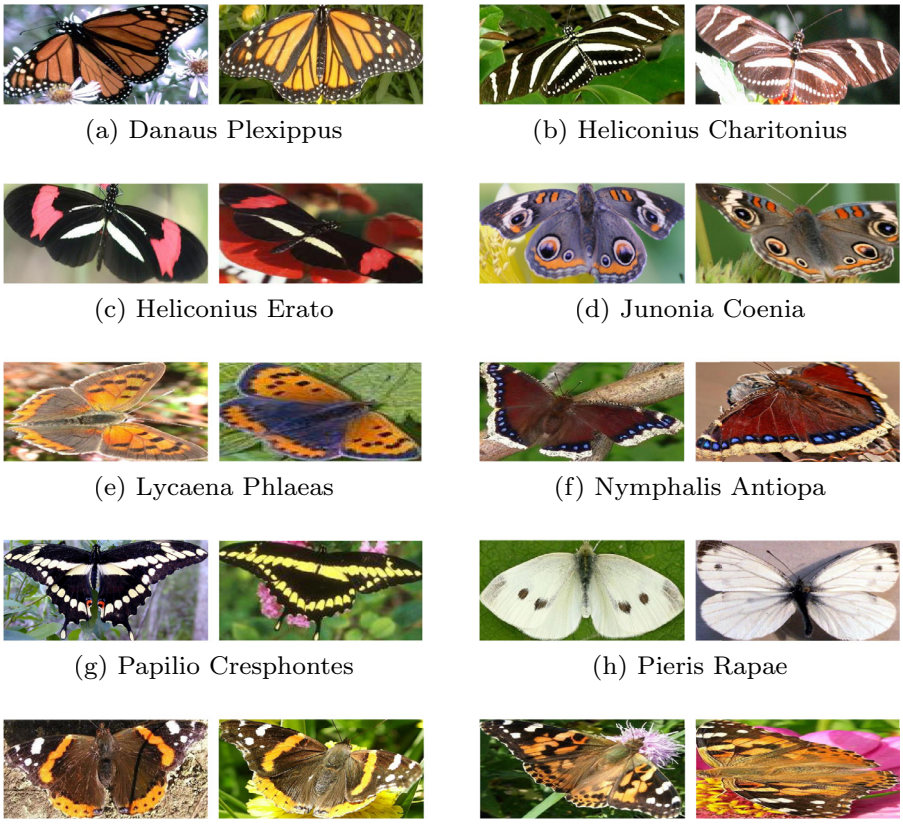
Methods	Evaluation Metrics	Australian Credit Approval	Breast Cancer Wisconsin (Diagnostic)	Breast Cancer Wisconsin (Prognostic)	Climate Model Simulation Crashes	Alon Colon Cancer	Connectionist Bench (Sonar, Mines, Rocks)	Hill-Valley	Indian Liver Patient	Ionosphere	Madelon
NLDA	$\mathbf{F_1}$	**0.8186**	**0.9809**	**0.6657**	**0.7389**	**0.8556**	**0.8745**	**0.9921**	**0.6237**	**0.8837**	**0.6367**
	e	**0.1791**	**0.0214**	**0.2513**	**0.0758**	**0.2290**	**0.1304**	**0.0210**	**0.3049**	**0.1011**	**0.3628**
	μ_g	**0.4064**	**0.4930**	**0.2723**	**0.1860**	**0.4027**	**0.4343**	**0.4960**	**0.2703**	**0.4121**	**0.3180**
NLDA with $Th = 0.0$	$\mathbf{F_1}$	0.8039	0.9479	0.6309	0.6830	0.7359	0.8606	0.8302	0.6112	0.8636	0.5632
	e	0.1933	0.0520	0.2843	0.0929	0.3165	0.1418	0.1807	0.3151	0.1165	0.4360
	μ_g	0.3990	0.4570	0.2576	0.1627	0.3458	0.4267	0.4148	0.2649	0.3996	0.2812
NLDA with $Th = 1 = 1/\mathcal{L}$	$\mathbf{F_1}$	0.8074	0.9473	0.6191	0.6823	0.7590	0.8606	0.8454	0.6071	0.8689	0.5615
	e	0.1899	0.0516	0.2881	0.0915	0.3097	0.1412	0.1610	0.3186	0.1112	0.4379
	μ_g	0.4007	0.4569	0.2506	0.1625	0.3545	0.4267	0.4225	0.2635	0.4025	0.2804
NLDA without *Th*	$\mathbf{F_1}$	0.7941	0.9486	0.6196	0.6593	0.7723	0.8673	0.9684	0.6061	0.8624	0.5106
	e	0.2022	0.0497	0.2893	0.1104	0.3204	0.1384	0.0097	0.3256	0.1247	0.4889
	μ_g	0.3933	0.4574	0.2499	0.1626	0.3646	0.4312	0.4992	0.2643	0.4100	0.2551

Methods	Evaluation Metrics	Musk (Version 2)	Parkinsons	Planning Relax	Ringnorm	Robot Execution Failures (LP1)	Robot Execution Failures (LP4)	Statlog (Landsat Satellite)	Spambase	SPECTF Heart	Vehicle Silhouettes
NLDA	$\mathbf{F_1}$	**0.9798**	**0.9332**	**0.5695**	**0.9140**	**0.8901**	**0.9179**	**0.9313**	**0.9482**	**0.6863**	**0.6988**
	e	**0.0207**	**0.0766**	**0.3496**	**0.0948**	**0.0927**	**0.0762**	**0.0215**	**0.0386**	**0.2081**	**0.1495**
	μ_g	**0.3604**	**0.4041**	**0.2342**	**0.4901**	**0.2074**	**0.2521**	**0.1875**	**0.4798**	**0.2643**	**0.1652**
NLDA with $Th = 0.0$	$\mathbf{F_1}$	0.9199	0.9255	0.5633	0.8241	0.8919	0.9053	0.8920	0.9111	0.6575	0.6975
	e	0.0419	0.0846	0.3519	0.1732	0.0866	0.0793	0.0312	0.0848	0.2310	0.1501
	μ_g	0.3332	0.4009	0.2325	0.4088	0.2060	0.2483	0.1376	0.4448	0.2535	0.1650
NLDA with $Th = 1 = 1/\mathcal{L}$	$\mathbf{F_1}$	0.9197	0.9333	0.5681	0.8224	0.8814	0.8970	0.8912	0.9105	0.6564	0.7186
	e	0.0420	0.0881	0.3566	0.1749	0.0922	0.0827	0.0314	0.0854	0.2299	0.1396
	μ_g	0.3333	0.4053	0.2408	0.4085	0.2048	0.2452	0.1374	0.4446	0.2498	0.1725
NLDA without *Th*	$\mathbf{F_1}$	0.8959	0.9317	0.5789	0.8265	0.7167	0.6960	0.8916	0.9109	0.5897	0.6965
	e	0.0566	0.0798	0.3557	0.1719	0.2397	0.2255	0.0313	0.0850	0.2954	0.1506
	μ_g	0.3290	0.4038	0.2431	0.4112	0.1577	0.1944	0.1375	0.4447	0.2291	0.1645



(a) Hill-Valley (b) Madelon

Fig. 12. Number of dimensions vs. average accuracy (μ) on UCI machine learning datasets for different thresholds.



(a) Danaus Plexippus (b) Heliconius Charitonius

(c) Heliconius Erato (d) Junonia Coenia

(e) Lycaena Phlaeas (f) Nymphalis Antiopa

(g) Papilio Cresphontes (h) Pieris Rapae

(i) Vanessa Atalanta (j) Vanessa Cardui

Fig. 13. Sample images of different classes from leeds butterfly dataset.

Madelon. In this case also, '*NLDA* ($Th' = Const * log_2(\mathcal{L})$)' outperforms other thresholds. On both the data, NLDA completely dominates rest of the three methods. Thus, one can conclude that the threshold $Th = Const * log_2(\mathcal{L})$ is the best one to select the relevant patterns and improving subspace learning. Note that, the threshold $Th = 1 - 1/\mathcal{L}$ leads to empty set of non-boundary patterns on Hill-Valley dataset, but due to cosine hyperbolic framework *NLDA* with $Th = 1 - 1/\mathcal{L}$ able to produce results for all the three metrics.

4.3. Image recognition on leeds butterfly (Wang et al., 2009)

This dataset contains images and textual descriptions for ten categories (species) of butterflies. The image dataset comprises 832 images in total, with the distribution ranging from 55 to 100 images per category. Images were collected from Google Images by querying with the scientific (Latin) name of the species as Danaus plexippus (Fig. 13(a)), Heliconius charitonius (Fig. 13(b)), Heliconius erato (Fig. 13(c)), Junonia coenia (Fig. 13(d)), Lycaena phlaeas (Fig. 13(e)), Nymphalis antiopa (Fig. 13(f)), Papilio cresphontes (Fig. 13(g)), Pieris rapae (Fig. 13(h)), Vanessa atalanta (Fig. 13(i)) and Vanessa cardui (Fig. 13(j)); and manually filtered for those depicting the butterfly of interest (Wang et al., 2009). The GIST descriptors (Oliva & Torralba, 2001) at eight orientations and four different scales are extracted to represent the visual contents of the "Leeds Butterfly" dataset, resulting a 512 dimensional vector for each image.

In this data set also, the above four evaluation metrics have been used to compare the proposed method NLDA with the different state-of-the-art methods. Table 8 presents the best average (about 30 iterations) results of F_1, e and μ_g for all the methods (up to 60 features) including proposed one (NLDA). It is evident from

Table 8
Experimental results of different methods on leeds butterfly dataset.

Methods	NLDA	ELDA	LDA	LFDA	LDG	ALDE	LBDA	MMC	QMI	NLDA with $Th = 0.0$	NLDA with $Th = 1 = 1/\mathcal{L}$	NLDA without Th
$\mathbf{F_1}$	**0.6505**	0.6289	0.4671	0.5744	0.6088	0.5098	0.4190	0.5828	0.3843	0.5087	0.5042	0.5539
e	**0.0711**	0.0755	0.1083	0.0874	0.0798	0.0996	0.1171	0.0846	0.1284	0.1002	0.0991	0.0895
μ_g	**0.0605**	0.0579	0.0407	0.0513	0.0553	0.0450	0.0373	0.0529	0.0327	0.0451	0.0448	0.0504

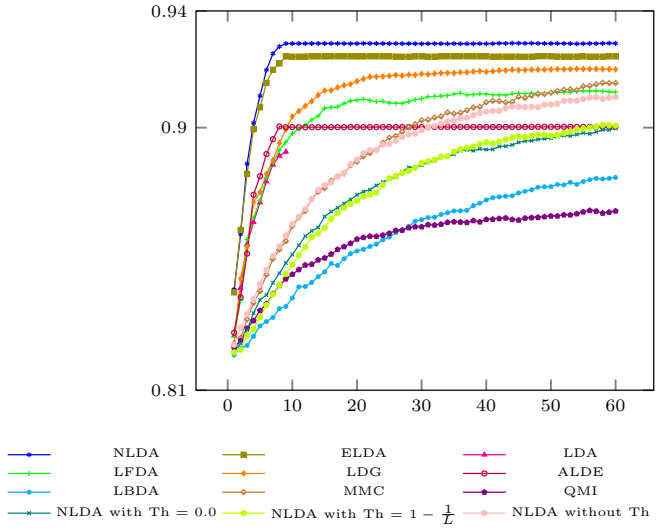


Fig. 14. Number of dimensions vs. average accuracy (μ) on leeds butterfly dataset for different thresholds.

Fig. 15. Sample images of extended yale b data sets.

the table that NLDA outperforms other state-of-the-art methods and NLDA with different thresholds. And ELDA and LDG are second and third best performers on the three metrics and QMI is the least performer. Also, Fig. 14 depicts the average accuracy values up to 60 dimensions (features) and NLDA achieves best accuracies throughout all the dimensions. Thus, NLDA shown better performance than the other methods and also the $Th' = Const * log_2(\mathcal{L})$ has proven to be better on leeds butterfly data in terms of image recognition.

4.4. Face recognition system

Face recognition (or facial recognition) is a biometric method of identifying an individual by comparing live capture or digital image data with the stored record for that person. Facial recognition systems are commonly used for security purposes but are increasingly being used in a variety of other applications. Face recognition is a preliminary step to a wide range of applications such as personal identity verification, video-surveillance, facial expression extraction, gender classification, advanced human and computer interaction, etc. These applications represent faces as points in high-dimensional image spaces and can employ Dimensionality Reduction (DR) to find a more meaningful representation, therefore, addressing the issue of the “curse of dimensionality”.

In order to verify the DR and discrimination capabilities of NLDA’s subspace in case of face recognition, Extended Yale B face recognition data set (Fig. 15) has been employed. This dataset now has 38 individuals and around 64 near frontal images under different illuminations per individual. These images were captured un-

Table 9
Experimental results of different methods on extended yale b face recognition.

Methods	F_1	e	μ_g
NLDA	**0.7381**	**0.0143**	**0.0198**
ELDA	0.6173	0.0210	0.0154
LDA	0.5293	0.0254	0.0133
LFDA	0.3398	0.0354	0.0082
LDG	0.5062	0.0265	0.0127
ALDE	0.5578	0.0238	0.0141
LBDA	0.3794	0.0330	0.0091
MMC	0.7191	0.0152	0.0185
QMI	0.3706	0.0441	0.0096
NLDA without Th	0.4719	0.0284	0.0117

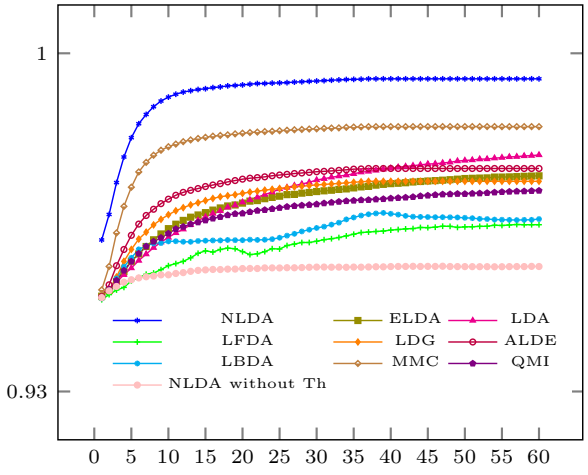


Fig. 16. Number of dimensions vs. average accuracy (μ) on extended yale b face recognition.

der various controlled indoor lighting conditions with variations in illumination. The goal is to test the ability of NLDA's subspace in recognizing faces with different illumination conditions. They were cropped, normalized and scaled to the size of 29×26.

In this experimental study, hold-out cross validation method has been employed in such a way that 10% data used for training and the rest for testing. And the average results of F_1, e, μ_g and accuracy for 30 iterations have been showed in the Table 9 and Fig. 16. It is evident from this results that NLDA's subspace is more discriminative compare to the other methods. Therefore, the extracted features of NLDA consists of exactness and completeness (F_1) for the subsequent classifier, of less classification error (e) and of more balance (μ_g). Moreover, from the Fig. 16, it is clear that NLDA has been achieved best classification accuracy by out performing others. MMC and ELDA are occupied second and third best performances, respectively. And rest of the comparative methods are not up to the mark due to their limitations (mentioned in Table 1). Note that, NLDA with $Th = 0.0$ and NLDA with $Th = 1 - 1/\mathcal{L}$ are unable to generating sufficient non-boundary patterns due to the small thresholds and leads to ill posed problems that can't solved.

Thus, in summary, four different categories of data (synthetic, UCI, leeds butterfly and extended yale b) have been used to evaluate the performances of proposed method. The geometrical study (Fig. 9) on synthetic data proved that NLDA (with $Th' = Const * log_2(\mathcal{L})$) is efficient in dealing the multimodal data since the different classes have different lengths. Moreover, the results (Table 3) of evaluation metrics, *TSI, FB, volume, NS* and *overlap*, highlighted that NLDA's subspace/projection space is simplified in case of multimodality. The results of F_1, e, μ_g and μ over UCI machine learning data (Tables 5, 6 and Figs. 10, 11) asses that NLDA's features are more balanced (μ_g) and discriminative compare to the others. This phenomena also evident from the experiments on leeds butterfly (Table 8 and Fig. 14) and extended yale b data (Table 9 and Fig. 16). Hence, NLDA is consistently performing better over all the data since it's formulation using lengths (motivation of Noisy-free Length Discriminant Analysis section) with the help of noisy free relevant patterns. Whereas, the performances of the methods ELDA, LDA, LFDA, LDG, ALDE, LBDA, MMC and QMI are pulled back by their limitations such as SSS problem, Gaussianity assumption, outliers, redundancy, overlapped features, etc. (mentioned in Table 1). Moreover, the analysis of NLDA without and with different thresholds, $Th = 0$, $Th = 1 - 1/\mathcal{L}$ and Th' (Tables 7–9 and Figs. 12, 14, 16) has been evaluated over UCI, leeds and extended yale b data. And, it is shown that Th' is more robust and efficient, in removing noise and, generating better boundary and non-boundary patters that are suitable for subspace learning. Therefore, NLDA with NRPS is efficient in handling multimodal data and it's subspace preserves better discrimination compare to other state-of-the-art methods.

5. Conclusion and discussion

In this manuscript, a novel method Noisy-free Relevant Pattern Selection (NRPS) has been proposed with the help of new threshold. NRPS partitions the set of patterns into boundary, non-boundary and noise pattern sets. With the help of these boundary and non-boundary pattern sets, a novel DR method namely, Noisy-free Length Discriminant Analysis (NLDA) has been developed. NLDA models the squared length differences between the patterns of different classes to form the between and within-class scatter matrices. And cosine hyperbolic framework has been developed to incorporate the positive and negative norm differences. The experimental study on synthetic data proved the efficiency of the NLDA to multimodal datasets. Also, NLDA compared with eight state-of-the-methods on UCI machine learning and leeds butterfly datasets, and it is evident that NLDA outperforms these methods. Thus, NLDA's subspace is more suitable for supervised learning when compared to other methods.

Clearly, NLDA models, theoretically, within and between class scatters with the help of average squared distances with cosine hyperbolic framework. Moreover, the proposed method considers the noisy free boundary and non-boundary patterns to model the inter and intra class separability, respectively. Whereas, most of the existing methods build the between class scatters, theoretically, using Fisher's criterion that leads to poor performance over multimodal data sets. And also, these methods didn't filter out the noisy patterns and may lead to complex subspace. Thus, NLDA's performance is consistency and good compare to the other methods.

The performance of NLDA mainly depends on the *Const* value in $Th = Const * log_2(\mathcal{L})$. In this manuscript, we considered *Const* to be from the set {0.25, 0.5, 0.75}. But, tuning this parameter over the interval $[0, 1]$ for different data sets is highly difficult and computationally expensive. Also, selecting a k_{NRPS} value for dividing, accurately, input data into boundary, non-boundary and noisy patterns is highly challenging and time consuming process. Moreover, balancing these two parameter values, as they dependent on each other, require a lot of human intervention.

6. Further scope of research

There is a plenty of scope to extended this work. Some of the possible directions are discusses below.

- The $Const \in [0, 1]$ in Th' can be made a data driven parameter, so that it can adopt to the underlying data.
- Our future work will be focused on maximizing

$$W_{opt} = \max_{W^\top W = Id} \frac{\widehat{L}_b}{\widehat{L}_w};$$

where

$$\widehat{L}_b = \frac{2}{\mathcal{L}} \sum_{i=1}^{\mathcal{L}} \frac{1}{N_i} \sum_{\substack{j \in c_i; \\ k \notin c_i}} cosh\left(\|\mathbf{y_j}\|^2 - \|\mathbf{y_k}\|^2\right), \text{ and}$$

$$\widehat{L}_w = \frac{2}{\mathcal{L}} \sum_{i=1}^{\mathcal{L}} \frac{1}{N_i} \sum_{\substack{j,k \in c_i; \\ j \neq k}} cosh\left(\|\mathbf{y_j}\|^2 - \|\mathbf{y_k}\|^2\right),$$

with the help of iterative optimization methodologies (Bertsekas, 1999) with convergence analysis.

- A non-linear version of NLDA (Lee & Verleysen, 2007) and Noisy-free Relevant Pattern Selection can be developed using kernel trick

  $k(\mathbf{x_i}, \mathbf{x_j}) = \phi(\mathbf{x_i})^\top \phi(\mathbf{x_j})$;

  where $\phi : X \to \mathcal{H}$ is a mapping from the input space X to a Hilbert space $\mathcal{H}$ and k is a Mercers kernel.
- Also, one can explore the behavior of different symmetric functions $\mathcal{F}(.)$ to better model the NLDA.
- The concept of length differences ($\mathcal{F}(\|y_i\|^2 - \|y_j\|^2)$) can be used in manifold learning (Lee & Verleysen, 2007) in order to preserve the curvature of the manifold along with the neighborhoods. Also, it can be used in multi-manifold learning for improving linear separability as different manifolds consist of different lengths.

Acknowledgment

K. Ramachandra Murthy is grateful to Council of Scientific & Industrial Research (CSIR), India for providing him a Research Associateship [No. 9/93(159) /2014 EMR-I].

References

Agovic, A., Banerjee, A., Ganguly, A., & Protopopescu, V. (2009). Anomaly detection using manifold embedding and its applications in transportation corridors. *Intelligent Data Analysis - Knowledge Discovery from Data Streams, 13*(3), 435–455.

Alejo, R., García, V., Sotoca, J. M., Mollineda, R. A., & Sánchez, J. S. (2007). Improving the performance of the RBF neural networks trained with imbalanced samples. In *International work-conference on artificial neural networks* (pp. 162–169). Springer.

Bertsekas, D. P. (1999). *Nonlinear programming* (2nd). Athena scientific Belmont.

Bolón-Canedo, V., Fernández-Francos, D., Peteiro-Barral, D., Alonso-Betanzos, A., Guijarro-Berdiñas, B., & Sánchez-Maroño, N. (2016). A unified pipeline for online feature selection and classification. *Expert Systems with Applications, 55*, 532–545.

Borges, H. B., & Nievola, J. C. (2012). Comparing the dimensionality reduction methods in gene expression databases. *Expert Systems with Applications, 39*(12), 10780–10795.

Bouzas, D., Arvanitopoulos, N., & Tefas, A. (2015). Graph embedded nonparametric mutual information for supervised dimensionality reduction. *IEEE Transactions on Neural Networks and Learning Systems, 26*(5), 951–963.

Choi, J. Y., Kim, D. H., Plataniotis, K. N., & Ro, Y. M. (2016). Classifier ensemble generation and selection with multiple feature representations for classification applications in computer-aided detection and diagnosis on mammography. *Expert Systems with Applications, 46*, 106–121.

Cunningham, J. P., & Ghahramani, Z. (2015). Linear dimensionality reduction: Survey, insights, and generalizations. *Journal of Machine Learning Research, 16*, 2859–2900.

Ezziane, Z. (2006). Applications of artificial intelligence in bioinformatics: A review. *Expert Systems with Applications, 30*(1), 2–10.

Frank, A., & Asuncion, A. (2010). UCI machine learning repository. http://archive.ics.uci.edu/ml.

Fukunaga, K. (1990). *Introduction to statistical pattern recognition* (2nd). New York, San Francisco, London: Academic Press.

Gali, N., Mariescu-Istodor, R., & Fränti, P. (2017). Using linguistic features to automatically extract web page title. *Expert Systems with Applications, 79*, 296–312.

Gao, Q., Ma, J., Zhang, H., Gao, X., & Liu, Y. (2013). Stable orthogonal local discriminant embedding for linear dimensionality reduction. *IEEE Transactions on Image Processing, 22*(7), 2521–2531.

Gilmore, J. F. (1985). Military applications of expert systems. *Future Generation Computer Systems, 1*(6), 403–410.

Haifeng, L., Tao, J., & Keshu, Z. (2006). Efficient and robust feature extraction by maximum margin criterion. *IEEE Transactions on Neural Networks, 17*(1), 157–165.

He, W. (2013). Improving user experience with case-based reasoning systems using text mining and web 2.0. *Expert Systems with Applications, 40*(2), 500–507.

Higham, N. J. (2005). The scaling and squaring method for the matrix exponential revisited. *SIAM Journal on Matrix Analysis and Applications, 26*(4), 1179–1193.

Houari, R., Bounceur, A., Kechadi, M. T., Tari, A. K., & Euler, R. (2016). Dimensionality reduction in data mining: A copula approach. *Expert Systems with Applications, 64*, 247–260.

Huang, H., Li, J., & Liu, J. (2012). Gene expression data classification based on improved semi-supervised local Fisher discriminant analysis. *Expert Systems with Applications, 39*(3), 2314–2320.

Huang, T., Sethu, H., & Kandasamy, N. (2016). A new approach to dimensionality reduction for anomaly detection in data traffic. *IEEE Transactions on Network and Service Management, 13*(3), 651–665.

Jayaraman, U., Prakash, S., & Gupta, P. (2012). An efficient color and texture based iris image retrieval technique. *Expert Systems with Applications, 39*(5), 4915–4926.

Kim, J., Han, M., Lee, Y., & Park, Y. (2016). Futuristic data-driven scenario building: Incorporating text mining and fuzzy association rule mining into fuzzy cognitive map. *Expert Systems with Applications, 57*, 311–323.

Lee, J. A., & Verleysen, M. (2007). *Nonlinear dimensionality reduction*. New York, London: Springer.

Liu, S., Feng, L., & Qiao, H. (2015). Scatter balance: An angle-based supervised dimensionality reduction. *IEEE Transactions on Neural Networks and Learning Systems, 26*(2), 277–289.

Mazzarella, F., Vespe, M., Alessandrini, A., Tarchi, D., Aulicino, G., & Vollero, A. (2017). A novel anomaly detection approach to identify intentional {AIS} on-off switching. *Expert Systems with Applications, 78*, 110–123.

Moler, C., & Loan, C. V. (1978). Nineteen dubious ways to compute the exponential of a matrix. *SIAM Review, 20*, 801–836.

Na, J. H., Park, M. S., & Choi, J. Y. (2010). Linear boundary discriminant analysis. *Pattern Recognition, 43*(3), 929–936.

Na, J. H., Yun, S. M., Kim, M., & Choi, J. Y. (2008). Relevant pattern selection for subspace learning. In *19th international conference on pattern recognition* (pp. 1–4).

Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision, 42*(3), 145–175.

Oliva, D., Hinojosa, S., Cuevas, E., Pajares, G., Avalos, O., & Gálvez, J. (2017). Cross entropy based thresholding for magnetic resonance brain images using crow search algorithm. *Expert Systems with Applications*. (in press)

Oliva, J. T., Lee, H. D., Spolaor, N., Coy, C. S. R., & Wu, F. C. (2016). Prototype system for feature extraction, classification and study of medical images. *Expert Systems with Applications, 63*, 267–283.

Parrish, N., & Gupta, M. R. (2012). Dimensionality reduction by local discriminative gaussians. In *Proceedings of the 29th international conference on machine learning*. Edinburgh, Scotland, UK: Omnipress.

Perumal, R. S., & Mouli, P. V. S. S. R. C. (2016). Dimensionality reduced local directional pattern (DR-LDP) for face recognition. *Expert Systems with Applications, 63*, 66–73.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (2007). *Numerical recipes : The art of scientific computing* (3rd). New York, NY, USA: Cambridge University Press.

Rada, R. (2008). Expert systems and evolutionary computing for financial investing: A review. *Expert Systems with Applications, 34*(4), 2232–2240.

Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications, 33*(1), 135–146.

Shin, H., & Cho, S. (2007). Neighborhood property based pattern selection for support vector machines. *Neural Computation, 19*(3), 816–855.

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management, 45*(4), 427–437.

Song, M., Yang, H., Siadat, S. H., & Pechenizkiy, M. (2013). A comparative study of dimensionality reduction techniques to enhance trace clustering performances. *Expert Systems with Applications, 40*(9), 3722–3737.

Sotoca, J. M., Mollineda, R. A., & Sanchez, J. S. (2006). A meta-learning framework for pattern classification by means of data complexity measures. *Inteligencia Artificial, 10*(29), 31–38.

Sugiyama, M. (2007). Dimensionality reduction of multimodal labeled data by local Fisher discriminant analysis. *Journal of Machine Learning Research, 8*, 1027–1061.

Theodoridis, S., & Koutroumbas, K. (2008). *Pattern recognition* (4th). Academic Press.

Tomasoni, F., Saracoglu, B. H., & Paniagua, G. (2014). A decision-making algorithm for automatic flow pattern identification in high-speed imaging. *Expert Systems with Applications, 41*(8), 3935–3943.

Wang, J., Markert, K., & Everingham, M. (2009). Learning models for object recognition from natural language descriptions. In *Proceedings of the British machine vision conference: vol. 1* (p. 2).

Yang, W., Wang, J., Ren, M., Yang, J., Zhang, L., & Liu, G. (2009). Feature extraction based on Laplacian bidirectional maximum margin criterion. *Pattern Recognition, 42*(11), 2327–2334.

Zhang, S., Dubay, R., & Charest, M. (2015). A principal component analysis model-based predictive controller for controlling part warpage in plastic injection molding. *Expert Systems with Applications, 42*(6), 2919–2927.

Zhang, T., Fang, B., Tang, Y. Y., Shang, Z., & Xu, B. (2010). Generalized discriminant analysis: A matrix exponential approach. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 40*(1), 186–197.