Two-dimensional Bhattacharyya bound linear discriminant analysis with its applications

Yan-Ru Guo[1] · Yan-Qin Bai[1] · Chun-Na Li[2] · Lan Bai[3] · Yuan-Hai Shao[2]

Abstract

The recently proposed L2-norm linear discriminant analysis criterion based on Bhattacharyya error bound estimation (L2BLDA) was an effective improvement over linear discriminant analysis (LDA) and was used to handle vector input samples. When faced with two-dimensional (2D) inputs, such as images, converting two-dimensional data to vectors, regardless of the inherent structure of the image, may result in some loss of useful information. In this paper, we propose a novel two-dimensional Bhattacharyya bound linear discriminant analysis (2DBLDA). 2DBLDA maximizes the matrix-based between-class distance, which is measured by the weighted pairwise distances of class means and minimizes the matrix-based within-class distance. The criterion of 2DBLDA is equivalent to optimizing the upper bound of the Bhattacharyya error. The weighting constant between the between-class and within-class terms is determined by the involved data that make the proposed 2DBLDA adaptive. The construction of 2DBLDA avoids the small sample size (SSS) problem, is robust, and can be solved through a simple standard eigenvalue decomposition problem. The experimental results on image recognition and face image reconstruction demonstrate the effectiveness of 2DBLDA.

Keywords Feature extraction · Dimensionality reduction · Two-dimensional linear discriminant analysis · Robust linear discriminant analysis · Bhattacharyya error bound

1 Introduction

Feature extraction plays an important role in pattern recognition. As a powerful supervised feature extraction method, linear discriminant analysis (LDA) [1] has been successfully applied in many problems, such as face recognition [2, 3], text mining [4, 5], image retrieval [6, 7], gait recognition [8], and microarrays [9, 10].

However, classical LDA is a vector (or one-dimensional) 1D based method. When input data are naturally of matrix (or two-dimensional) 2D form, such as images, two issues may arise. First, converting 2D data to 1D data may produce high-dimensional vectors and hence may lead to the small sample size (SSS) problem [11]. For example, a 32×32 face image corresponds to a 1024-dimensional vector. Second, during the transformation from 2D data to 1D data, the underlying spatial (structural) information

is destroyed. Therefore, useful discriminant information may be lost [12, 13]. To handle these problems, many image-as-matrix methods have been developed [14, 15]. In contrast to the image-as-vector methods, the image-as-matrix methods treat an image as a two-order tensor, and their objective functions are expressed as functions of the image matrix instead of the high-dimensional image vector. The representative image-as-matrix method is two-dimensional LDA (2DLDA) [16]. 2DLDA constructed the within-class scatter matrix and between-class scatter matrix by using the original image samples represented in matrix form rather than converting matrices to vectors beforehand. Compared to LDA, 2DLDA can alleviate the SSS problem when a mild condition is satisfied [17] and can preserve the original structure of the input matrix.

Thereafter, some modifications and improvements of 2DLDA were studied by many researchers. Due to the squared L2-norm nature of 2DLDA, it was sensitive to noise and outliers. To improve the robustness of 2DLDA, robust replacements of the L2-norm were studied, including the L1-norm [18–21], nuclear norm [22, 23], Lp-norm [24, 25], and Schatten Lp-norm, $0 < p < 1$ [26]. Some of the studies focused on extracting the

✉ Chun-Na Li
  na1013na@163.com

Extended author information available on the last page of the article.

discriminative transformations on both sides of the matrix samples. The authors in [27, 28] implemented 2DLDA on matrices in sequence or independently and then combined left-and right side transformations to achieve bilateral dimensionality reduction. Li et al. [25] used iterative schemes to extract transformations on both sides. Extensions to other machine learning problems and real applications were also investigated. For example, Wang et al. [29] proposed a convolutional 2DLDA for nonlinear dimensionality reduction, and Xiao et al. [30] studied a two-dimensional quaternion sparse discriminant analysis that met the requirements of representing RGB and RGB-D images.

Although 2DLDA can ease the SSS problem, it may still face the singularity issue theoretically as LDA since it needs to solve a generalized eigenvalue problem. Recently, a novel vector-based L2-norm linear discriminant analysis criterion based on Bhattacharyya error bound estimation (L2BLDA) [31] was proposed. Compared to LDA, L2BLDA solved a simple standard eigenvalue decomposition problem rather than a generalized eigenvalue decomposition problem, which avoided the singularity issue and had robustness. In fact, minimizing the Bhattacharyya error [32] bound is a reasonable way to establish classification [33]. In this paper, inspired by L2BLDA, to cope with the SSS problem and improve the robustness of 2DLDA, we first derive a Bhattacharyya error upper bound for matrix input classification and then propose a novel two-dimensional linear discriminant analysis by minimizing this Bhattacharyya error upper bound, called 2DBLDA. The proposed 2DBLDA has the following characteristics:

- 2DBLDA is proposed for the novel two-dimensional matrix input problem. The 2DBLDA criterion is proven to be an upper bound of the theoretical framework of the Bhattacharyya error bound optimality. We have proved that optimizing this upper bound of the Bhattacharyya error can lead to an optimal discriminant direction. Therefore, the rationality of the 2DBLDA optimization problem is guaranteed theoretically.
- The weighting constant of the between-class distance and the within-class distance of 2DBLDA is adaptive to the involved data that is calculated according to input data. This constant not only helps the objective of 2DBLDA achieve the minimum error bound but also makes the proposed 2DBLDA adaptive without tuning any parameters. By considering the above weighted between-class distance information, 2DBLDA could achieve robustness.
- Unlike 2DLDA, 2DBLDA is solved effectively through a standard eigenvalue decomposition problem, which does not involve the inverse of a matrix and hence avoids the SSS problem.

- To observe the discriminant ability of our method, we consider the accuracy of different databases, plot the variation of the accuracy with dimension reduction, and measure the reconstruction performance of the face image. The experimental results on image recognition and face reconstruction demonstrate the effectiveness of 2DBLDA.

The paper is organized as follows. Section 2 briefly introduces LDA, L2BLDA and 2DLDA. Section 3 proposes our 2DBLDA and gives the corresponding theoretical analysis. Section 4 compares 2DBLDA with its related approaches. Section 5 discusses the relationship between our 2DBLDA and related methods and analyses the experimental results. Finally, the concluding remarks are given in Section 6. The proof of the Bhattacharyya error upper bound of 2DBLDA is given in the Appendix.

The notations of this paper are given as follows. We consider a supervised learning problem in the $d_1 \times d_2$-dimensional matrix space $\mathbb{R}^{d_1 \times d_2}$. The training dataset is given by $T = \{(\mathbf{X}_1, y_1), ..., (\mathbf{X}_N, y_N)\}$, where $\mathbf{X}_l \in \mathbb{R}^{d_1 \times d_2}$ is the l-th input matrix sample and $y_l \in \{1, ..., c\}$ is the corresponding label, $l = 1, ..., N$. Assume that the i-th class contains N_i samples, $i = 1, \ldots, c$. Then, we have $\sum_{i=1}^{c} N_i = N$. We further write the samples in the i-th class as $\{\mathbf{X}_{is}\}$, where $\mathbf{X}_{is}$ is the s-th sample in the i-th class, $i = 1, \ldots, c$, $s = 1, \ldots, N_i$. Let $\overline{\mathbf{X}} = \frac{1}{N} \sum_{l=1}^{N} \mathbf{X}_l$ be the mean of all matrix samples and $\overline{\mathbf{X}}_i = \frac{1}{N_i} \sum_{s=1}^{N_i} \mathbf{X}_{is}$ be the mean of matrix samples in the i-th class. For a matrix $\mathbf{Q} = (\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_n) \in \mathbb{R}^{m \times n}$, its Frobenius norm (F-norm) $||\mathbf{Q}||_F$ is defined as $||\mathbf{Q}||_F = \sqrt{\sum_{i=1}^{n} ||\mathbf{q}_i||_2^2}$. The F-norm is a natural generalization of the vector L2-norm on matrices.

2 Related work

2.1 Linear discriminant analysis

Linear discriminant analysis (LDA) finds a projection transformation matrix $\mathbf{W}$ such that the ratio of between-class distance to within-class distance is maximized in the projected space. For data in $\mathbb{R}^n$, LDA finds an optimal $\mathbf{W} \in \mathbb{R}^{n \times r}$, $r \leq n$, such that the most discriminant information of the data is retained in $\mathbb{R}^r$ by solving the following problem:

$$\max_{\mathbf{W}} \frac{\mathrm{tr}(\mathbf{W}^T \mathbf{S}_b \mathbf{W})}{\mathrm{tr}(\mathbf{W}^T \mathbf{S}_w \mathbf{W})}, \tag{1}$$

where $\mathrm{tr}(\cdot)$ is the trace operation of a matrix, and the between-class scatter matrix $\mathbf{S}_b$ and the within-class scatter

matrix $\mathbf{S}_w$ are defined by

$$\mathbf{S}_b = \frac{1}{N} \sum_{i=1}^{c} N_i (\overline{\mathbf{x}}_i - \overline{\mathbf{x}})(\overline{\mathbf{x}}_i - \overline{\mathbf{x}})^T \qquad (2)$$

and

$$\mathbf{S}_w = \frac{1}{N} \sum_{i=1}^{c} \sum_{s=1}^{N_i} (\mathbf{x}_{is} - \overline{\mathbf{x}}_i)(\mathbf{x}_{is} - \overline{\mathbf{x}}_i)^T, \qquad (3)$$

where $\overline{\mathbf{x}}_i \in \mathbb{R}^n$ is the mean of the samples in the i-th class, $\overline{\mathbf{x}} \in \mathbb{R}^n$ is the mean of the whole data, and $\mathbf{x}_{is} \in \mathbb{R}^n$ is the s-th sample of the i-th class. The optimization problem (1) is equivalent to the generalized problem $\mathbf{S}_b\mathbf{w} = \lambda\mathbf{S}_w\mathbf{w}$, where $\lambda \neq 0$, with its solution $\mathbf{W} = (\mathbf{w}_1, \ldots, \mathbf{w}_r)$ given by the first r largest eigenvalues of $(\mathbf{S}_w)^{-1}\mathbf{S}_b$ in case $\mathbf{S}_w$ being nonsingular.

2.2 L2-norm linear discriminant analysis criterion via the Bhattacharyya error bound estimation

As an improvement over LDA, the L2-norm linear discriminant analysis criterion based on Bhattacharyya error bound estimation (L2BLDA) [31] is a recently proposed vector-based weighted linear discriminant analysis. In the vector space $\mathbb{R}^n$, by minimizing an upper bound of the Bhattacharyya error, the optimization problem of L2BLDA is formulated as

$$\begin{aligned} \min_{\mathbf{W}} \quad & -\frac{1}{N} \sum_{i<j} \sqrt{N_i N_j} ||\mathbf{W}^T(\overline{\mathbf{x}}_i - \overline{\mathbf{x}}_j)||_2^2 + \Delta \sum_{i=1}^{c} \sum_{s=1}^{N_i} ||\mathbf{W}^T(\mathbf{x}_{is} - \overline{\mathbf{x}}_i)||_2^2 \\ \text{s.t.} \quad & \mathbf{W}^T\mathbf{W} = \mathbf{I}, \end{aligned}$$
$$(4)$$

where $\mathbf{W} \in \mathbb{R}^{n \times r}$, $r \leq n$, $P_i = \frac{N_i}{N}$, $P_j = \frac{N_j}{N}$, $\overline{\mathbf{x}}_i \in \mathbb{R}^n$ is the mean of the samples in the i-th class, $\mathbf{x}_{is} \in \mathbb{R}^n$ is the s-th sample of the i-th class, $\Delta = \frac{1}{4} \sum_{i<j}^{c} \sqrt{P_i P_j} ||\overline{\mathbf{x}}_i - \overline{\mathbf{x}}_j||_2^2$, and $\mathbf{I} \in \mathbb{R}^{r \times r}$ is the identity matrix.

L2BLDA is solved through the following standard eigenvalue decomposition problem:

$$\begin{aligned} \min_{\mathbf{W}} \quad & \text{tr}(\mathbf{W}^T\mathbf{S}\mathbf{W}) \\ \text{s.t.} \quad & \mathbf{W}^T\mathbf{W} = \mathbf{I}, \end{aligned} \qquad (5)$$

where

$$\mathbf{S} = -\frac{1}{N} \sum_{i<j} \sqrt{N_i N_j}(\overline{\mathbf{x}}_i - \overline{\mathbf{x}}_j)(\overline{\mathbf{x}}_i - \overline{\mathbf{x}}_j)^T + \Delta \sum_{i=1}^{c} \sum_{s=1}^{N_i} (\mathbf{x}_{is} - \overline{\mathbf{x}}_i)(\mathbf{x}_{is} - \overline{\mathbf{x}}_i)^T. \qquad (6)$$

Then, $\mathbf{W} = (\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_r)$ is obtained by the r orthogonormal eigenvectors that correspond to the first r nonzero smallest eigenvectors of $\mathbf{S}$. After obtaining the

optimal $\mathbf{W}$, a new sample $\mathbf{x} \in \mathbb{R}^n$ is projected into $\mathbb{R}^r$ by $\mathbf{W}^T\mathbf{x}$.

2.3 Two-dimensional linear discriminant analysis

Different from LDA or L2BLDA, which works on vector samples, two-dimensional linear discriminant analysis (2DLDA) [16, 17] operates on matrix samples. 2DLDA defines the between-class scatter matrix and the within-class scatter matrix directly on the 2D data set T as

$$\mathbf{S}_b = \frac{1}{N} \sum_{i=1}^{c} N_i (\overline{\mathbf{X}}_i - \overline{\mathbf{X}})(\overline{\mathbf{X}}_i - \overline{\mathbf{X}})^T \qquad (7)$$

and

$$\mathbf{S}_w = \frac{1}{N} \sum_{i=1}^{c} \sum_{s=1}^{N_i} (\mathbf{X}_{is} - \overline{\mathbf{X}}_i)(\mathbf{X}_{is} - \overline{\mathbf{X}}_i)^T. \qquad (8)$$

Then 2DLDA solves the following optimization problem:

$$\max_{\mathbf{W}} \quad \frac{\text{tr}(\mathbf{W}^T\mathbf{S}_b\mathbf{W})}{\text{tr}(\mathbf{W}^T\mathbf{S}_w\mathbf{W})} = \frac{\sum\limits_{i=1}^{c} N_i \|\mathbf{W}^T(\overline{\mathbf{X}}_i - \overline{\mathbf{X}})\|_F^2}{\sum\limits_{i=1}^{c} \sum\limits_{s=1}^{N_i} \|\mathbf{W}^T(\mathbf{X}_{is} - \overline{\mathbf{X}}_i)\|_F^2}, \qquad (9)$$

where $\mathbf{W} = (\mathbf{w}_1, \ldots, \mathbf{w}_r) \in \mathbb{R}^{d_1 \times r}$, $r \leq d_1$. $i = 1, \ldots, c$, $j = 1, \ldots, N_i$. (9) can be solved through the generalized eigenvalue problem $\mathbf{S}_b\mathbf{w} = \lambda\mathbf{S}_w\mathbf{w}$ in case $\mathbf{S}_w$ is nonsingular, and its solution is the r eigenvectors corresponding to the first largest r nonzero eigenvalues. After obtaining optimal $\mathbf{W}$, a new sample $\mathbf{X} \in \mathbb{R}^{d_1 \times d_2}$ is projected into $\mathbb{R}^{r \times d_2}$ by $\mathbf{W}^T\mathbf{X}$. Note that 2DLDA will still encounter the singularity problem when $\mathbf{S}_w$ is not of full rank.

3 Two-dimensional Bhattacharyya bound linear discriminant analysis

3.1 The derivation of a Bhattacharyya error bound estimation

In this section, we derive a new two-dimensional linear discriminant analysis criterion by minimizing a Bhattacharyya error bound.

From the viewpoint of minimizing the probability of classification error, the Bayes classifier is the best classifier [1], and its error rate, known as the Bayes error, is defined as

$$\epsilon = 1 - \int \max_{i \in \{1,2,\ldots,c\}} \{P_i p_i(\mathbf{X})\} d\mathbf{X}, \qquad (10)$$

where $\mathbf{X}$ is a sample, P_i is the prior probability, and $p_i(\mathbf{X})$ is the probability density function of the i-th class of the data. Computing the Bayes error is usually hard, and therefore minimizing its upper bound is often considered

an alternative effective method [35–37]. Among various bounds, the Bhattacharyya error [32] is a close upper bound to the Bayes error, which is given by

$$\epsilon_B = \sum_{i<j}^{c} \sqrt{P_i P_j} \int \sqrt{p_i(\mathbf{X}) p_j(\mathbf{X})} d\mathbf{X}. \quad (11)$$

Under the background of two-dimensional supervised dimensionality reduction, if we can derive a relatively close upper bound of ϵ_B, we may obtain a reasonable dimensionality reduction model. In fact, under some basic assumptions, we can obtain an upper bound of ϵ_B, as shown in the following proposition.

Proposition 1 *Assume P_i and $p_i(X)$ are the prior probability and the probability density function of the i-th class for the training data set T, respectively, and the data samples in each class are independent and identically normally distributed. Let $p_1(X), p_2(X), \ldots, p_c(X)$ be the Gaussian functions given by $p_i(X) = \mathcal{N}(X|\overline{X}_i, \Sigma_i)$, where $\overline{X}_i$ and Σ_i are the class mean and the class covariance matrix, respectively. We further suppose $\Sigma_i = \Sigma$, $i = 1, 2, \ldots, c$, where Σ is the covariance matrix of the data set T, and $\overline{X}_i$ and Σ can be estimated accurately from T. Then for arbitrary projection vector $w \in \mathbb{R}^{d_1}$, the Bhattacharyya error bound ϵ_B defined by (11) on the data set $\widetilde{T} = \{\widetilde{X}_i | \widetilde{X}_i = w^T X_i \in \mathbb{R}^{1 \times d_2}\}$ satisfies the following:*

$$\epsilon_B \leq -\frac{a}{8} \sum_{i<j}^{c} \sqrt{P_i P_j} ||w^T(\overline{X}_i - \overline{X}_j)||_2^2 + \frac{a}{8} \Delta \sum_{i=1}^{c} \sum_{s=1}^{N_i} ||w^T(X_{is} - \overline{X}_i)||_2^2$$
$$+ \sum_{i<j}^{c} \sqrt{P_i P_j}, \quad (12)$$

where $\Delta = \frac{1}{4} \sum_{i<j}^{c} \sqrt{P_i P_j} ||\overline{X}_i - \overline{X}_j||_F^2$, and $a > 0$ is some constant.

Proof See the Appendix. □

3.2 The proposed two-dimensional Bhattacharyya bound linear discriminant analysis

Proposition 1 gives a reasonable upper bound of ϵ_B. After obtaining an upper error bound, it is natural to minimize it. Therefore, we minimize the upper bound of ϵ_B in (12), that is, the right side of (12). In fact, by minimizing it, we can easily obtain a novel two-dimensional Bhattacharyya bound linear discriminant analysis (2DBLDA) as follows:

$$\min_{w^T w=1} -\frac{1}{N} \sum_{i<j} \sqrt{N_i N_j} ||w^T(\overline{\mathbf{X}}_i - \overline{\mathbf{X}}_j)||_2^2 + \Delta \sum_{i=1}^{c} \sum_{s=1}^{N_i} ||w^T(\mathbf{X}_{is} - \overline{\mathbf{X}}_i)||_2^2 \quad (13)$$

where $\Delta = \frac{1}{4} \sum_{i<j}^{c} \sqrt{P_i P_j} ||\overline{\mathbf{X}}_i - \overline{\mathbf{X}}_j||_F^2$, $\mathbf{w} \in \mathbb{R}^{d_1}$, $P_i = \frac{N_i}{N}$.

By applying (13), we can project a $d_1 \times d_2$ sample $\mathbf{X}$ into a $1 \times d_2$ sample $\widetilde{\mathbf{X}}$ by $\widetilde{\mathbf{X}} = \mathbf{w}^T \mathbf{X}$. However, it does not usually contain enough discriminant information in the $1 \times d_2$ space, and we may need $r \geq 1$ projection vectors $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_r$ that constitute a projection matrix $\mathbf{W} = (\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_r) \in \mathbb{R}^{d_1 \times r}$ and project $\mathbf{X}$ into a $r \times d_2$ space by $\widetilde{\mathbf{X}} = \mathbf{W}^T \mathbf{X}$.

In general, we consider the following 2DBLDA

$$\min_{\mathbf{W}} -\frac{1}{N} \sum_{i<j} \sqrt{N_i N_j} ||\mathbf{W}^T(\overline{\mathbf{X}}_i - \overline{\mathbf{X}}_j)||_F^2 + \Delta \sum_{i=1}^{c} \sum_{s=1}^{N_i} ||\mathbf{W}^T(\mathbf{X}_{is} - \overline{\mathbf{X}}_i)||_F^2$$
$$\text{s.t.} \quad \mathbf{W}^T \mathbf{W} = \mathbf{I}, \quad (14)$$

where $\mathbf{W} \in \mathbb{R}^{r \times d_1}$, $r \leq d_1$. We now give the geometric meaning of 2DBLDA. Minimizing the first term in (14) will make the means of two different classes far from each other in the projected space, which guarantees the between-class separativeness. Here, the coefficients $\frac{1}{N} \sqrt{N_i N_j}$ in the first term weight distance pairs between different class means. Minimizing the second term in (14) forces each sample around its own class mean in the projected space. The weighting constant Δ in front of the second term balances the between-class importance and within-class importance while also ensuring a minimum error bound according to the proof of Proposition 1. We can observe that 2DBLDA is adaptive to different data since Δ is determined by the given data set. To ensure minimum redundancy in the projected space, we also consider an orthogonormal constraint $\mathbf{W}^T \mathbf{W} = \mathbf{I}$ on discriminant directions.

It is easily seen that we can solve 2DBLDA through the following standard eigenvalue decomposition problem:

$$\min_{\mathbf{W}} \quad tr(\mathbf{W}^T \mathbf{S} \mathbf{W})$$
$$\text{s.t.} \quad \mathbf{W}^T \mathbf{W} = \mathbf{I}, \quad (15)$$

where

$$\mathbf{S} = -\frac{1}{N} \sum_{i<j} \sqrt{N_i N_j} (\overline{\mathbf{X}}_i - \overline{\mathbf{X}}_j)(\overline{\mathbf{X}}_i - \overline{\mathbf{X}}_j)^T + \Delta \sum_{i=1}^{c} \sum_{s=1}^{N_i} (\mathbf{X}_{is} - \overline{\mathbf{X}}_i)(\mathbf{X}_{is} - \overline{\mathbf{X}}_i)^T. \quad (16)$$

Then, we obtain the optimal solution as $\mathbf{W} = (\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_r)$, where $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_r$ are the r orthogonormal eigenvectors corresponding to the first r smallest nonzero eigenvectors of $\mathbf{S}$.

4 Experiments

In this section, we compare the proposed 2DBLDA with 2DPCA [34], 2DPCA-L1 [12], 2DLDA [16] and L1-2DLDA [18,

19]. The learning parameter δ of L1-2DLDA is selected optimally from the set {0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1} by grid search.

We experimented on three image databases for image recognition and one face image database for face reconstruction. In the experiments, after applying a dimensionality reduction method on training data and then obtaining a projection matrix, the test data are projected to lower dimensional space using this projection matrix. For image recognition, the nearest neighbours classifier is employed to obtain classification accuracy. In addition, when the data classes are unbalanced, area under the ROC curve (AUC) and G-mean are used as the performance measurement index. For face reconstruction, the mean reconstruction error is used for performance evaluation. All the methods will be carried out on a PC with P4 2.3 GHz CPU by Matlab 2017b.

4.1 Image recognition

4.1.1 Performance on three image databases

The Yale database[1] is a human face database that contains 165 images of 15 individuals, and each individual includes 11 images. The database is considered to evaluate the performance of methods when facial expression and lighting conditions are changed.

Columbia Object Image Library (Coil100)[2] is a database of colour images of 100 objects. The objects were placed on a motorized turntable against a black background. The turntable was rotated 360 degrees to vary object pose with respect to a fixed colour camera. Images of the objects were taken at pose intervals of 5 degrees. The database contains 900 images of 100 objects, with each object containing 9 images.

The COVID database[3] has 349 CT images containing clinical findings of COVID-19 from 216 patients and 397 non-COVID CT scans. The images are collected from COVID-19 related papers from medRxiv[4], bioRxiv[5], Lancet, etc. In our experiment, 195 COVID-19 images and 195 non-COVID-19 images were randomly extracted.

We resize each image to 32×32 for all the above three databases. Since the number of samples in some classes of the image data used in the experiment is relatively small, to avoid the chance that images in these classes may not be selected due to random cross-validation, for each class, we

randomly select 60% of the data samples as the training set, and deem the rest as the test set. Therefore, this strategy makes sure both train and test data set contain samples from every class. First, we obtain all projection matrices from the training data and then compute the test classification accuracy on the projected test data. Since 2DPCA, 2DLDA and 2DBLDA have no parameters, the result of one run is the final result. For L1-2DLDA, there is one parameter, and we adopt ten-fold cross validation on the training set to find its optimal parameter. Then, this optimal parameter is used to run L1-2DLDA ten times on the test set to eliminate the influence of random initialization, and the average accuracy of these ten accuracies is adopted. Similarly, for 2DPCA-L1, since its performance is affected by the initialization projections, we repeat the method ten times and adopt mean accuracy along with standard variance. The results on these databases are listed in Table 1, and the best accuracies are shown in bold figures. From the table, we see that our 2DBLDA has comparable performance compared to other methods. The 2DPCA-L1 and L1-2DLDA obviously have the highest computational burden. In contrast, 2DBLDA costs the least CPU time than 2DPCA and 2DLDA.

To further see the superiority of our 2DBLDA, we artificially pollute the training data by adding each training sample with a rectangle block occlusion at a random location. We set the occlusion area ratio to 10%, 20%, 30%, 40%. For convenience, we denote these four data sets as $\text{Yale}_{b0.1}$, $\text{Yale}_{b0.2}$, $\text{Yale}_{b0.3}$ and $\text{Yale}_{b0.4}$, where the subscript "b" represents block occlusion and the number next to it means occlusion ratio. For the Coil100 data and COVID data, we add random rectangular Gaussian noise of mean 0 and variance 0.2 that covers 10%, 20%, 30%, 40% areas of each training image at a random position. Denote these eight data sets as $\text{Coil}_{g0.1}$, $\text{Coil}_{g0.2}$, $\text{Coil}_{g0.3}$, $\text{Coil}_{g0.4}$, $\text{COVID}_{g0.1}$, $\text{COVID}_{g0.2}$, $\text{COVID}_{g0.3}$ and $\text{COVID}_{g0.4}$, where the subscript "g" represents Gaussian noise, and the number next to it means noise ratio. Some noise samples are shown in Fig. 1.

The classification results on the noise datasets are listed in Table 2. From the table, we have the following observations : (i) All methods are affected by noise, and their corresponding accuracies are lower than those of the original data. In general, the larger the noise area is, the lower the accuracy is. (ii) The proposed 2DBLDA has the highest average accuracy on all noise data. (iii) L1-2DLDA and 2DPCA perform better than 2DPCA-L1 and 2DLDA. (iv) L1-2DLDA can achieve the optimal accuracy when δ is relatively small. (v) For CPU time, we see that 2DPCA-L1 and L1-2DLDA have the same computing time level but are all slower than 2DPCA and 2DLDA, and that 2DLDA and 2DBLDA run the fastest since they obtain all the discriminant vectors once for all.

Table 1 Comparison of mean accuracy (%), CPU time (second) for different methods on the original three databases

Data set	2DPCA AC/Time	2DPCA-L1 AC/Time	2DLDA AC/Time	L1-2DLDA AC/Time(δ)	2DBLDA AC/Time
Yale	**85.00**	82.67±0.86	83.33	**85.00**±0.00	**85.00**
	1.6835	9.2095	1.4929	10.4038(0.001)	1.4174
Coil100	74.00	67.93±1.26	72.00	73.37±0.46	**74.33**
	30.4082	67.3805	29.1930	71.5989(0.01)	28.7575
COVID	**95.00**	90.63±0.93	94.38	93.75±0.51	94.38
	6.0727	21.2971	5.2936	20.6759(0.05)	5.9291
Average accuracy	**84.67**	80.41	83.24	84.04	84.57

The optimal parameter δ of L1-2DLDA is shown in bracket
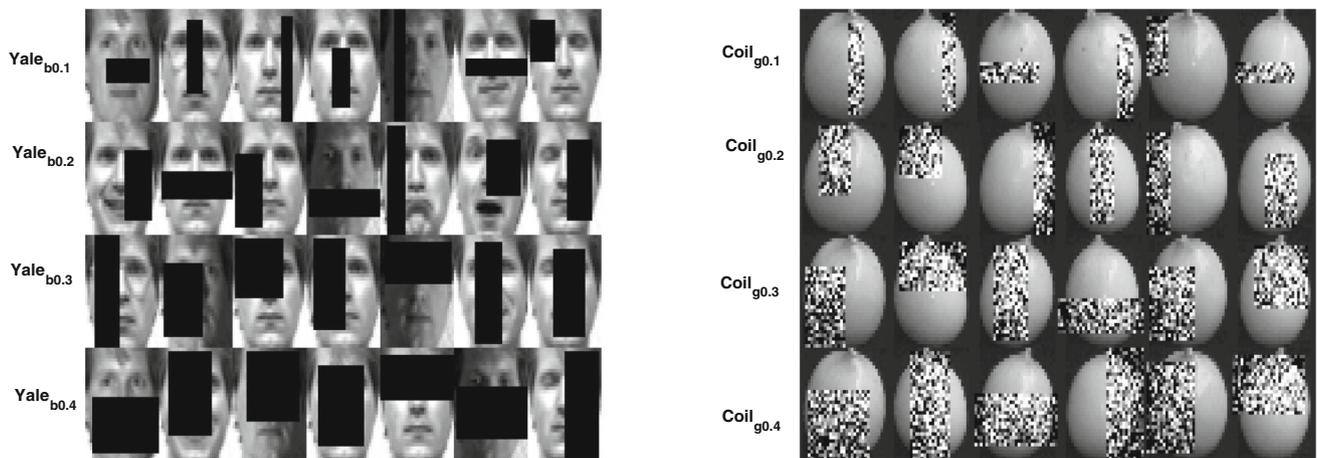
(a) Yale

(b) Coil100
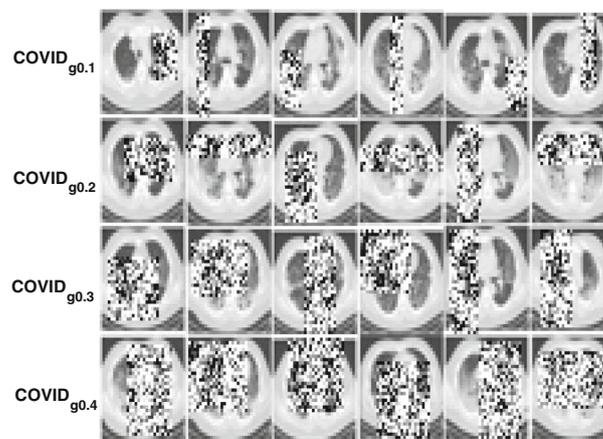
(c) COVID

Fig. 1 Noise samples from the three databases

Table 2 Comparison of mean accuracy (%), CPU time (second) for different methods on noise databases

Data set	2DPCA AC/Time	2DPCA-L1 AC/Time	2DLDA AC/Time	L1-2DLDA AC/Time(δ)	2DBLDA AC/Time
Yale$_{b0.1}$	76.67	59.33±2.63	76.66	77.33±1.17	**78.33**
	2.0922	10.0383	1.8348	11.1539(0.1)	1.6212
Yale$_{b0.2}$	**76.67**	55.83±3.07	70.00	73.67±2.33	**76.67**
	2.7252	11.0812	2.3906	12.1323(0.005)	2.0293
Yale$_{b0.3}$	63.33	50.67±7.58	63.33	64.67±2.33	**65.00**
	3.4940	12.3434	3.1027	13.4336(0.05)	2.6751
Yale$_{b0.4}$	56.67	49.67±3.31	56.67	51.83±3.46	**60.00**
	4.6380	13.9008	4.1114	14.9425(0.005)	3.6111
Coil$_{g0.1}$	71.67	63.77±3.39	70.67	71.37±1.11	**72.33**
	36.3437	84.5664	36.3030	87.0550(0.001)	35.3140
Coil$_{g0.2}$	67.00	51.43±1.75	68.67	67.10±0.88	**70.00**
	43.2088	86.9822	43.1494	91.3349(0.05)	41.0304
Coil$_{g0.3}$	63.33	48.93±3.43	63.33	61.80±2.54	**68.33**
	48.2347	90.4744	45.5738	94.9467(0.1)	43.4944
Coil$_{g0.4}$	58.00	43.70±0.64	60.00	57.73±1.84	**61.00**
	52.1063	93.9737	47.4356	98.0803(0.005)	45.9146
COVID$_{g0.1}$	**94.38**	91.44±0.42	93.75	94.11±0.03	**94.38**
	6.7681	23.4892	6.8060	22.6747(0.1)	6.2296
COVID$_{g0.2}$	93.13	91.25±0.00	93.13	92.94±0.93	**93.75**
	6.8413	23.5693	6.8593	23.9786(0.01)	6.3514
COVID$_{g0.3}$	**91.25**	90.88±0.67	90.00	90.88±0.72	**91.25**
	7.0305	23.9982	7.0121	24.6301(0.005)	6.5660
COVID$_{g0.4}$	90.63	88.88±0.92	87.50	92.32±0.78	**92.50**
	7.2641	24.5959	6.9099	25.4767(0.1)	6.6628
Average accuracy	75.23	65.48	74.48	74.65	**76.96**

The optimal parameter δ of L1-2DLDA is shown in bracket

4.1.2 The influence of the reduced dimension

To observe the discriminant ability of the dimensionality method, we measure feature ranking by observing the effect of sample classification in projection space and plot the accuracy variation along with the reduced dimensions in Figs. 2 and 3. Figure 2 depicts the variation of accuracies along dimensions on the original three databases, and Fig. 3 depicts the corresponding results on noise databases.

The results show the following: (i) With the increase of the number of reduced dimensions, the accuracies of 2DPCA and our 2DBLDA first achieve their highest and then have a relatively steady trend, while other methods vary greatly. (ii) Regardless of on the original data or the noise data, the proposed 2DBLDA has the highest accuracy under the optimal reduced dimension. (iii) All the methods are greatly influenced by the reduced dimension, and it is necessary to choose an optimal reduced dimension. (iv) In addition, the optimal reduced dimension of 2DBLDA is not too large compared to other methods in general.

4.1.3 The influence of the unbalanced classes

In this subsection, we verify the influence of our algorithm on unbalanced classes. To construct unbalanced data, different numbers of images are randomly selected from each class to form the training set, and the remaining data are deemed as the test set. In specific, for the COVID database, we randomly select 60% of the sample number for each class from COVID-19 images and non-COVID-19 images in a ratio of 1:1.5 as the training set. Notably, the training set and test set we construct are unbalanced. To test the robustness, as before, we pollute the training images with a black rectangular block, which covers 10%, 20%, 30% and 40% of each image at a random position. In this situation, we use AUC and G-mean to measure the performance of all methods, which are both designed for unbalanced data. The results on original databases and noise databases are demonstrated in Figs. 4 and 5. From Figs. 4 and 5, we can see that the proposed 2DBLDA has the highest AUC and G-mean of all databases. Though
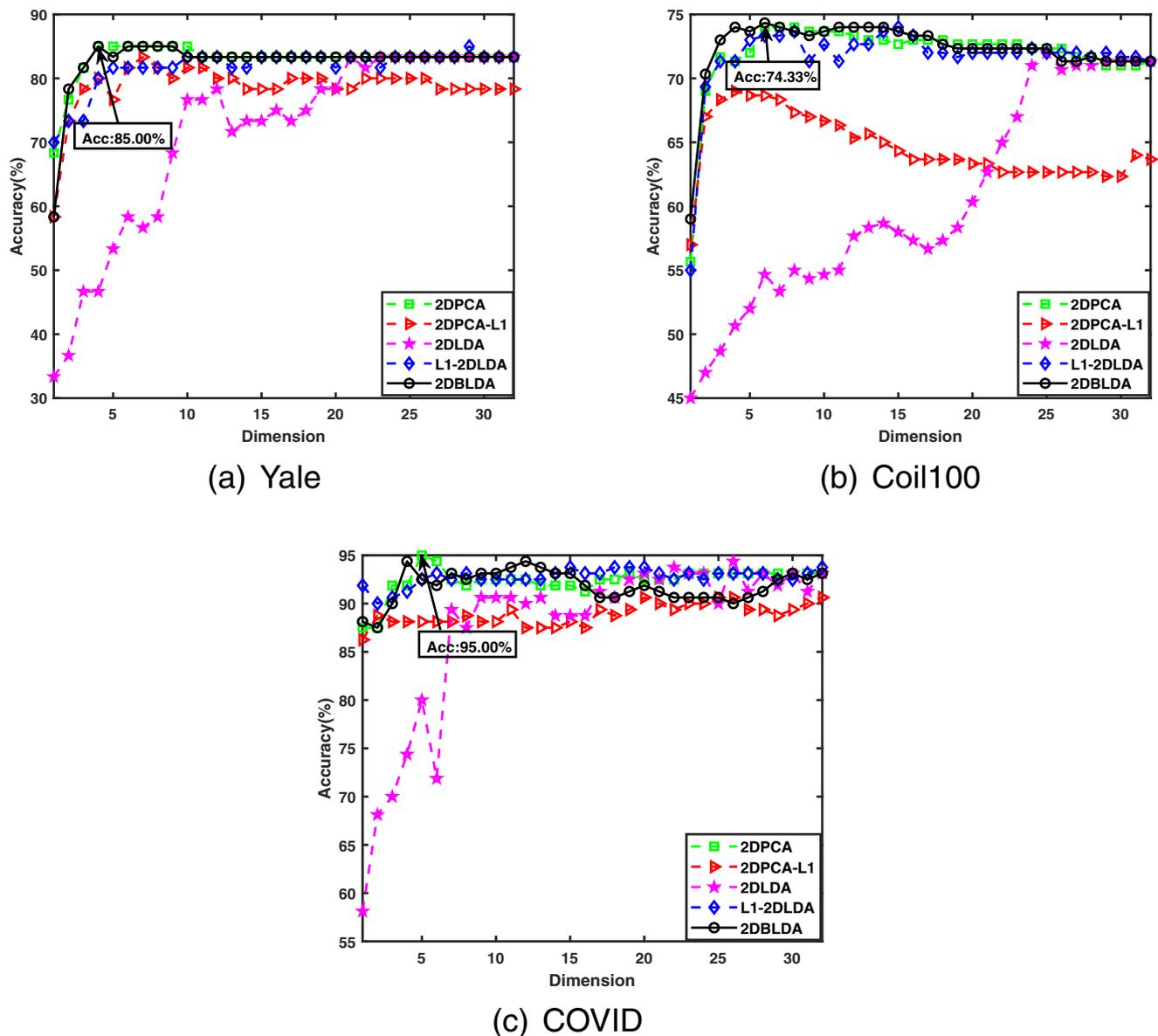
Fig. 2 Accuracies of all methods on the original three databases

the larger the noise area is the lower the performance is for all algorithms, when the block percentage increases, 2DBLDA and L1-2DLDA are less affected by noise, while the performance of other methods decreases dramatically and the proposed 2DBLDA is the best. The result is in fact consistent with the formulation of 2DBLDA, where its weighted between-class distance information and weighting constant of the between-class distance and the within-class distance make contribution to its good performance on unbalanced problems. The result also shows that compared to other methods, our 2DBLDA is more adaptive and robust to different data.

4.2 Face Reconstruction

In this part, the proposed 2DBLDA and other methods are applied to face reconstruction on the Indian female database. The Indian females database contains 242 human face images of 22 female individuals, and each individual has 11 different images. The original images are resized to 32×32 pixels.

We introduce face image reconstruction. For a given image $\mathbf{X} \in \mathbb{R}^{d_1 \times d_2}$, suppose we have obtained a projection matrix $\mathbf{W} = (\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_r) \in \mathbb{R}^{d_1 \times r}$, $r \leq d_1$. Then $\mathbf{X}$ is projected into the $r \times d_2$-dimensional space by
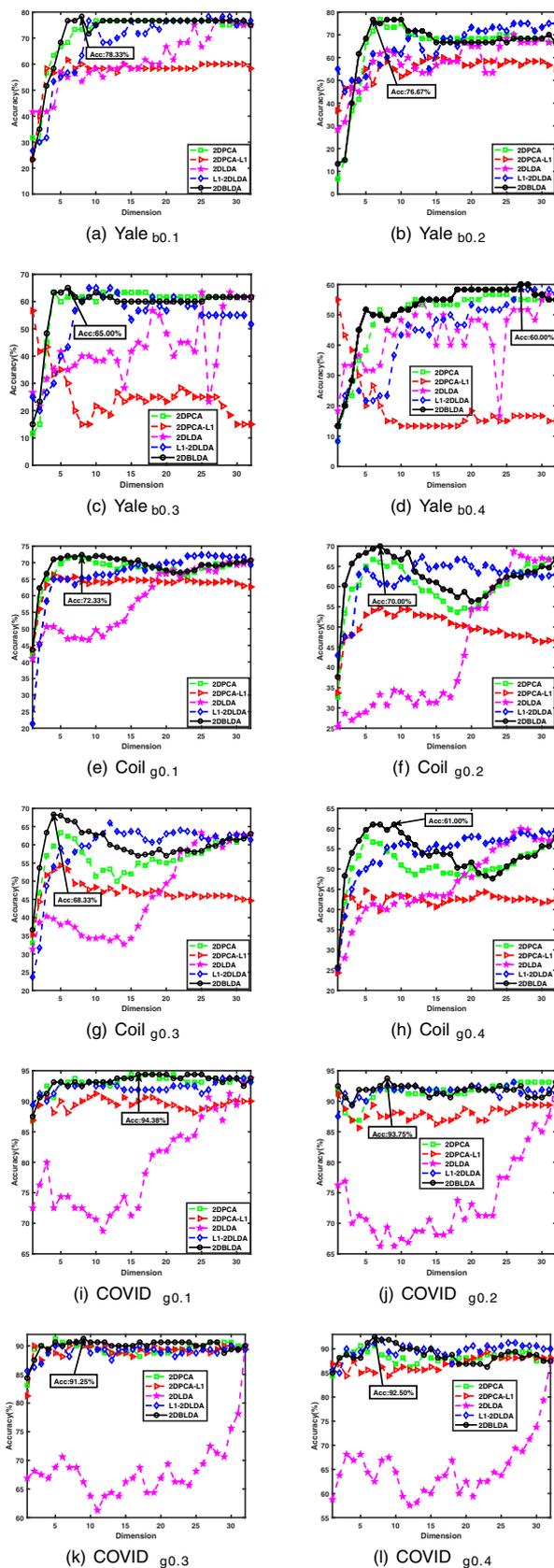
Fig. 3 Accuracies of all methods on three databases with different levels of noise

$\widetilde{\mathbf{X}} = \mathbf{W}^T \mathbf{X}$. Since $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_r$ are orthonormal, then the reconstructed image of $\mathbf{X}$ can be obtained by $\widehat{\mathbf{X}} = \mathbf{W}\widetilde{\mathbf{X}} = \mathbf{W}\mathbf{W}^T \mathbf{X}$. To measure the reconstruction performance, we use the average reconstruction error (ARE) as a performance indicator, which is defined as

$$\bar{e}_r = \frac{1}{N} \sum_{i=1}^{N} ||\mathbf{X}_i - \mathbf{W}\mathbf{W}^T \mathbf{X}_i||_F, \qquad (17)$$

where $r = 1, 2, \ldots, d_1$.

We first experiment on the original data and compute the ARE for each method. The variation in ARE along different dimensions is shown in Fig. 6 (a). From the figure, we see that when the dimension is less than 15, our 2DBLDA performs the best, especially when the dimension is greater than 5. When the dimension is greater than 15, 2DPCA is comparable or slightly better than our 2DBLDA, but both of these methods almost achieve steady performance. The result shows that 2DBLDA can achieve good performance for low dimensions. The other three methods obviously perform worse than our 2DBLDA and 2DPCA on all the dimensions. When $r = 15$, we demonstrate the reconstructed face images for 7 random individuals in Fig. 6b. We can visually see that 2DBLDA and 2DPCA have the best reconstruction performance.

To further evaluate the effectiveness of the proposed 2DBLDA, we add two different types of noise to the data. The first type of noise is Gaussian noise with mean 0 and variance 0.05 that covers 30% of the area of each image. The ARE of each method under different dimensions is plotted in Fig. 6c. On Gaussian noise data, we see that our 2DBLDA outperforms other methods on almost all the reduced dimensions, and 2DPCA is comparable to our 2DBLDA only when the dimension is greater than 27, indicating that the proposed 2DBLDA can achieve fairly good performance by employing only a small number of reduced dimensions. We then add the second type of noise, dummy noise, to the data. Here, the dummy noise is the image that is generated from the discrete uniform distribution on [0,1] and is of the same size as the original image. An additional 100 dummy images are added to the whole database. After the projection matrix is obtained on these polluted data, it is used to reconstruct human face images. The result in Fig. 6e demonstrates that our 2DBLDA has the lowest ARE on these databases for all the dimensions, and when the dimension is greater than 20, it has a rather low ARE. The reconstructed face images when $r = 15$ shown in Fig. 6f also support the above argument.

5 Discussion

To further clarify the contribution of our method, we discuss the differences between the proposed 2DBLDA and its two
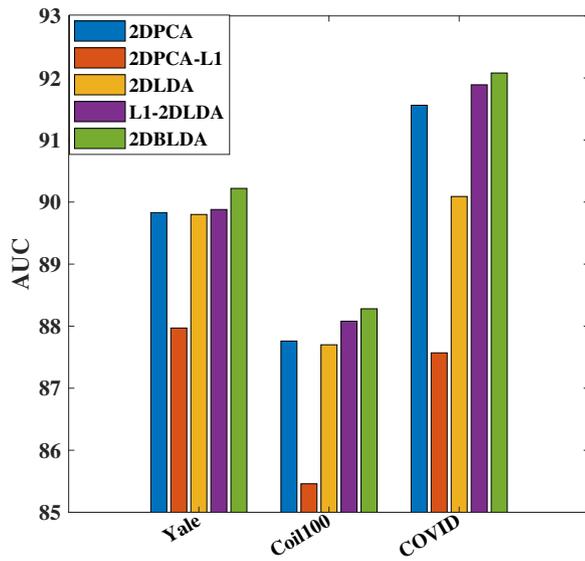
(a) Original databases
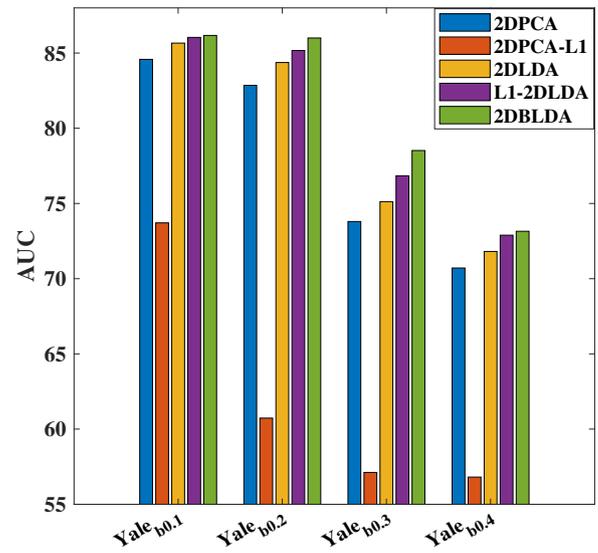
(b) Yale $_{noise}$
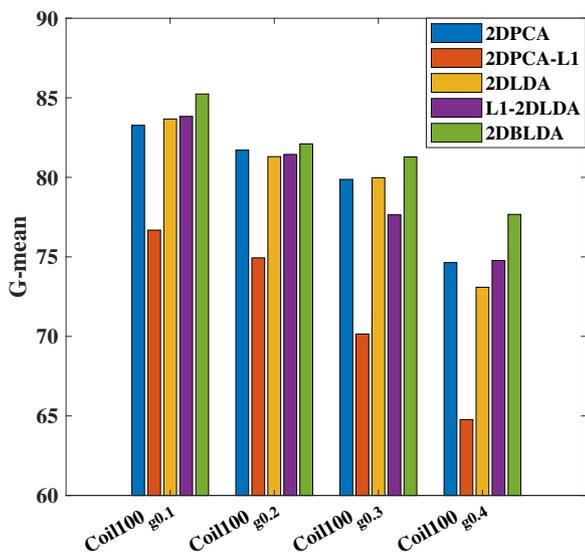
(c) Coil100 $_{noise}$

(d) COVID $_{noise}$

Fig. 4 AUC of all methods on different databases

closely related methods, RLp2DLDA and L2BLDA, and give a detailed analysis of the experimental results.

5.1 Relationship between RLp2DLDA, L2BLDA and 2DBLDA

(i) *Difference From RLp2DLDA:* The formulation of 2DBLDA is different from any existing 2D linear discriminant analysis method, and the 2DBLDA

criterion is derived by minimizing an upper bound of the theoretical framework of the Bhattacharyya error bound optimality. Although robust bilateral Lp-norm two-dimensional linear discriminant analysis (RLp2DLDA) is also derived from some upper bound of the Bhattacharyya error, they have different formulations since they have different error bounds. In fact, the bound for 2DBLDA may be closer than the bound of RLp2DLDA, which can be observed from two

Fig. 5 G-mean of all methods on different databases

aspects. First, when deriving its bound, RLp2DLDA ignores the term $\sqrt{P_i P_j}$ and replaces it by 1, which obviously magnifies the upper bound. In contrast, our 2DBLDA keeps this term and fully explores this weighting information, which leads to one of good properties of 2DBLDA, that is, robustness. Second, RLp2DLDA also magnifies its upper bound when using the Lp-norm ($0 < p < 1$) rather than the L2-norm. Therefore, this results in two advantages of our 2DBLDA over RLp2DLDA: one is that 2DBLDA obtains a meaningful weighting parameter that does

not need tuning, and the other is that 2DBLDA can simply solve its optimization problem through a standard eigenvalue problem, while RLp2DLDA solves its optimization problem through an iteration technique without proving its convergence.

(ii) *Difference From L2BLDA:* Compared to the vector-based robust Bhattacharyya bound linear discriminant analysis through an adaptive algorithm (L2BLDA), the proposed 2DBLDA is a matrix-based dimensionality reduction method. Although 2DBLDA is a
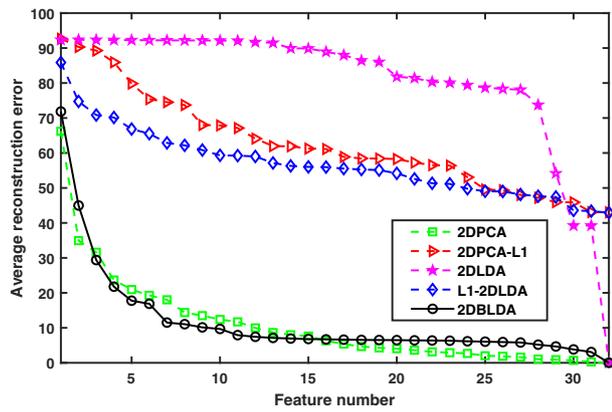
(a) Original database
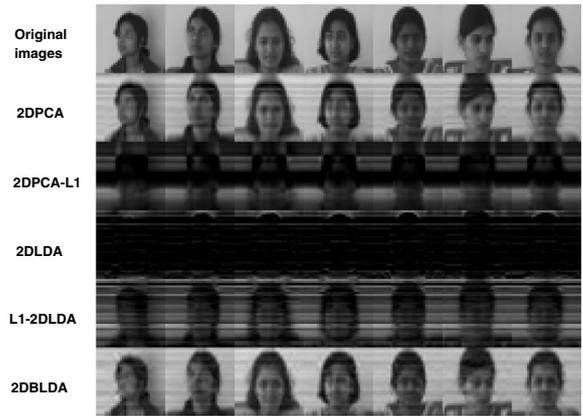
(b) Reconstructed faces
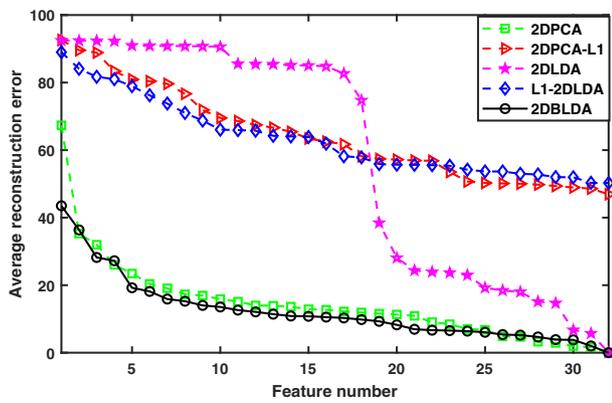
(c) Gaussian noise database
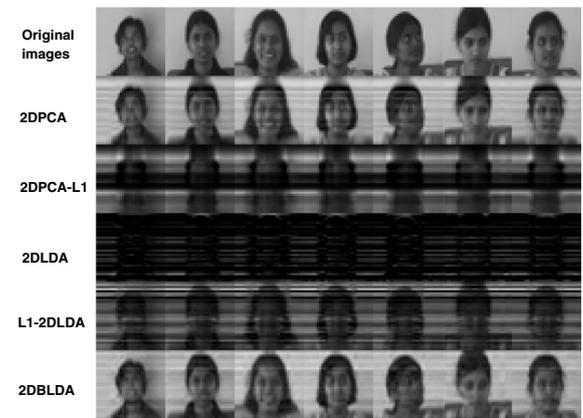
(d) Reconstructed faces
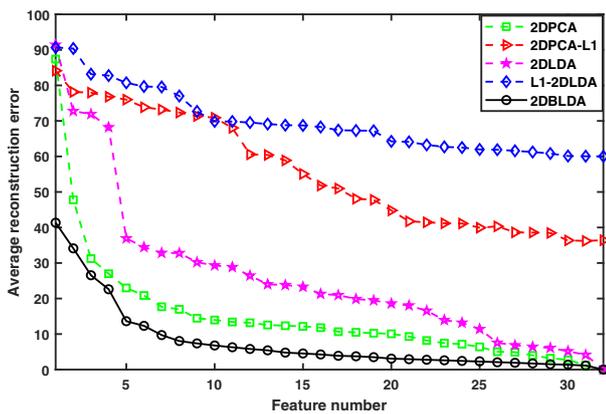
(e) Dummy noise database
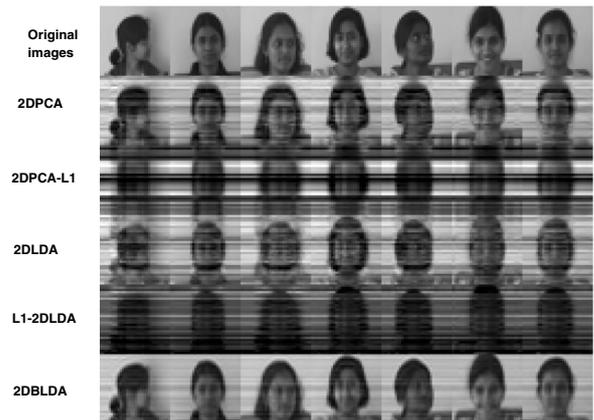
(f) Reconstructed faces

Fig. 6 Face reconstruction results on different databases

generalization of L2BLDA, it is not so direct from view of the derivation of its upper bound. In fact, the derivation procedure of the Bhattacharyya error bound of 2DBLDA is not exactly the same as that of L2BLDA. In addition, 2DBLDA can more effectively deal with the matrix input without vectoring it first, which improves the computing efficiency, especially when computing the scatter matrices.

5.2 Experimental results summary

(i) To study the performance of 2DBLDA, we give the variation of accuracies under different databases and different noise levels. The time of 2DBLDA is also investigated in Tables 1-2. Experimental results show that 2DBLDA runs fast and improves the robustness of 2DLDA.

(ii) To compare the behavior of 2DBLDA and other related methods under different reduced dimensions, we plot the accuracy variation along with the reduced dimensions in Figs. 2-3. The results demonstrate that compared to other methods, the proposed 2DBLDA obtains better classification results under its optimal reduced dimension.

(iii) To see the application ability of 2DBLDA in unbalanced classes, we experiment on three original and different noise image databases. From the results in Fig. 4 and Fig. 5, we see that the proposed 2DBLDA has the best performance compared to other methods.

(iv) To observe the behavior of the proposed method visually, we reconstruct face images by the obtained projection matrix. Original and polluted Indian female databases are used for face reconstruction. By choosing an appropriate reduced dimension but not necessarily too large, the proposed 2DBLDA can obtain good face reconstruction performance.

6 Conclusion

This paper proposed a novel two-dimensional linear discriminant analysis via Bhattacharyya upper bound optimality (2DBLDA). Different from the existing 2DLDA, optimizing the criterion of 2DBLDA was equivalent to optimizing the upper bound of the Bhattacharyya error, leading to maximizing a weighted between-class distance and minimizing the within-class distance, where these two distances were weighted by a meaningful adaptive constant that can be computed directly from the involved data. The 2DBLDA had no parameters to be tuned and could be effectively solved by a standard eigenvalue decomposition problem. Experimental results on image recognition and face image reconstruction demonstrated the superiority of the proposed method. Our MATLAB code can be downloaded from http://www.optimal-group.org/Resources/Code/2DBLDA.html.

However, a drawback of 2DBLDA is that its classification performance degrades when the class distribution of the samples is inconsistent. A TAISL technique could be used to handle this issue [38]. Since sparse learning could make the data have better interpretation after dimensionality reduction [20], one of the future studies also includes considering a sparse model. In the end, applying our algorithm to track fault detection is worth studying [39, 40].

Appendix

Proof of Proposition 1: We first note that $p_i(\widetilde{\mathbf{X}}) = \mathcal{N}(\widetilde{\mathbf{X}}|\overline{\widetilde{\mathbf{X}}}_i, \widetilde{\boldsymbol{\Sigma}})$, where $\overline{\widetilde{\mathbf{X}}}_i = \mathbf{w}^T\overline{\mathbf{X}}_i \in \mathbb{R}^{1 \times d_2}$ is the i-class mean, and $\widetilde{\boldsymbol{\Sigma}}$ is the covariance matrix in the $1 \times d_2$ projected space. Denote

$$\mathbf{D} = \begin{pmatrix} \mathbf{w}^T\mathbf{X}_1 \\ \vdots \\ \mathbf{w}^T\mathbf{X}_N \end{pmatrix}^T \in \mathbb{R}^{d_2 \times N} \text{ and } \overline{\widetilde{\mathbf{X}}}_\mathbf{I} = \begin{pmatrix} \mathbf{w}^T\overline{\mathbf{X}}_{t_1} \\ \vdots \\ \mathbf{w}^T\overline{\mathbf{X}}_{t_N} \end{pmatrix}^T \in \mathbb{R}^{d_2 \times N}. \quad (18)$$

Then $\widetilde{\boldsymbol{\Sigma}} = (\mathbf{D} - \overline{\widetilde{\mathbf{X}}}_\mathbf{I})(\mathbf{D} - \overline{\widetilde{\mathbf{X}}}_\mathbf{I})^T$.

According to [1], we have

$$\int \sqrt{p_i(\widetilde{\mathbf{X}})p_j(\widetilde{\mathbf{X}})} = e^{-\frac{1}{8}(\overline{\widetilde{\mathbf{X}}}_i - \overline{\widetilde{\mathbf{X}}}_j)\widetilde{\boldsymbol{\Sigma}}^{-1}(\overline{\widetilde{\mathbf{X}}}_i - \overline{\widetilde{\mathbf{X}}}_j)^T}. \quad (19)$$

The upper bound of the error ϵ_B can be estimated as

$$
\begin{aligned}
\epsilon_B &= \sum_{i<j}^{c} \sqrt{P_i P_j} e^{-\frac{1}{8}(\overline{\widetilde{\mathbf{X}}}_i - \overline{\widetilde{\mathbf{X}}}_j)\widetilde{\boldsymbol{\Sigma}}^{-1}(\overline{\widetilde{\mathbf{X}}}_i - \overline{\widetilde{\mathbf{X}}}_j)^T} \\
&= \sum_{i<j}^{c} \sqrt{P_i P_j} e^{-\frac{1}{8}||(\overline{\widetilde{\mathbf{X}}}_i - \overline{\widetilde{\mathbf{X}}}_j)\widetilde{\boldsymbol{\Sigma}}^{-\frac{1}{2}}||_2^2} \\
&\leq \sum_{i<j}^{c} \sqrt{P_i P_j} \left(1 - \frac{a}{8}||(\overline{\widetilde{\mathbf{X}}}_i - \overline{\widetilde{\mathbf{X}}}_j)\widetilde{\boldsymbol{\Sigma}}^{-\frac{1}{2}}||_2^2\right) \\
&= \sum_{i<j}^{c} \sqrt{P_i P_j} - \frac{a}{8}\sum_{i<j}^{c} \sqrt{P_i P_j} \cdot ||(\mathbf{w}^T\overline{\mathbf{X}}_i - \mathbf{w}^T\overline{\mathbf{X}}_j)\widetilde{\boldsymbol{\Sigma}}^{-\frac{1}{2}}||_2^2 \quad (20) \\
&\leq \sum_{i<j}^{c} \sqrt{P_i P_j} - \frac{a}{8}\sum_{i<j}^{c} \sqrt{P_i P_j} \cdot \frac{||(\mathbf{w}^T\overline{\mathbf{X}}_i - \mathbf{w}^T\overline{\mathbf{X}}_j)||_2^2}{||\widetilde{\boldsymbol{\Sigma}}^{\frac{1}{2}}||_F^2} \\
&\leq \sum_{i<j}^{c} \sqrt{P_i P_j} - \frac{a}{8}\sum_{i<j}^{c} \sqrt{P_i P_j} \cdot ||\mathbf{w}^T(\overline{\mathbf{X}}_i - \overline{\mathbf{X}}_j)||_2^2 \\
&\quad + \frac{a}{8}\sum_{i<j}^{c} \sqrt{P_i P_j} \cdot \Delta'_{ij}||\widetilde{\boldsymbol{\Sigma}}^{\frac{1}{2}}||_F^2,
\end{aligned}
$$

where $\Delta'_{ij} = \frac{1}{4}||\overline{\mathbf{X}}_i - \overline{\mathbf{X}}_j||_F^2$, $a > 0$ is some constant. For the first inequality of (20), note that the real value function $f(z) = e^{-z}$ is concave when $z \in [0, b]$, $b > 0$; therefore, $e^{-z} \leq 1 - \frac{1-e^{-b}}{b}z$. By taking $a = \frac{1-e^{-b}}{b}$ and noting $\overline{\widetilde{\mathbf{X}}}_i = \mathbf{w}^T\overline{\mathbf{X}}_i$, the first inequality is obtained. For the second inequality, we first note that for any $\mathbf{z} \in \mathbb{R}^{1 \times d_2}$ and an invertible $\mathbf{A} \in \mathbb{R}^{d_2 \times d_2}$, $||\mathbf{z}||_2 = ||(\mathbf{z}\mathbf{A})\mathbf{A}^{-1}||_2 \leq ||\mathbf{z}\mathbf{A}||_2 \cdot ||\mathbf{A}^{-1}||_F$, which implies $||\mathbf{z}\mathbf{A}||_2 \geq \frac{||\mathbf{z}||_2}{||\mathbf{A}^{-1}||_F}$. By taking $\mathbf{z} = \mathbf{w}^T\overline{\mathbf{X}}_i - \mathbf{w}^T\overline{\mathbf{X}}_j$ and $\mathbf{A} = \widetilde{\boldsymbol{\Sigma}}^{-\frac{1}{2}}$, we get the second inequality. For the last inequality, since $||\mathbf{w}||_2 = 1$,

$$||\mathbf{w}^T(\overline{\mathbf{X}}_i - \overline{\mathbf{X}}_j)||_2^2 \leq ||\mathbf{w}||_2^2 \cdot ||\overline{\mathbf{X}}_i - \overline{\mathbf{X}}_j||_F^2 = ||\overline{\mathbf{X}}_i - \overline{\mathbf{X}}_j||_F^2$$

and $\frac{1}{||\widetilde{\boldsymbol{\Sigma}}^{\frac{1}{2}}||_F^2}\left(1 - \frac{1}{||\widetilde{\boldsymbol{\Sigma}}^{\frac{1}{2}}||_F^2}\right) \leq \frac{1}{4}$, we have

$$\left(||\mathbf{w}^T(\overline{\mathbf{X}}_i - \overline{\mathbf{X}}_j)||_2^2 - \frac{||\mathbf{w}^T(\overline{\mathbf{X}}_i - \overline{\mathbf{X}}_j)||_2^2}{||\widetilde{\boldsymbol{\Sigma}}^{\frac{1}{2}}||_F^2}\right) \cdot \frac{1}{||\widetilde{\boldsymbol{\Sigma}}^{\frac{1}{2}}||_F^2}$$

$$= ||\mathbf{w}^T(\overline{\mathbf{X}}_i - \overline{\mathbf{X}}_j)||_2^2 \cdot \frac{1}{||\widetilde{\boldsymbol{\Sigma}}^{\frac{1}{2}}||_F^2}\left(1 - \frac{1}{||\widetilde{\boldsymbol{\Sigma}}^{\frac{1}{2}}||_F^2}\right) \quad (21)$$

$$\leq \frac{1}{4}||\mathbf{w}^T(\overline{\mathbf{X}}_i - \overline{\mathbf{X}}_j)||_2^2$$

$$\leq \frac{1}{4}||\overline{\mathbf{X}}_i - \overline{\mathbf{X}}_j||_F^2$$

$$= \Delta'_{ij}.$$

which implies

$$-\frac{||\mathbf{w}^T(\overline{\mathbf{X}}_i - \overline{\mathbf{X}}_j)||_2^2}{||\widetilde{\boldsymbol{\Sigma}}^{\frac{1}{2}}||_F^2} \leq -||\mathbf{w}^T(\overline{\mathbf{X}}_i - \overline{\mathbf{X}}_j)||_2^2 + \Delta'_{ij} \cdot ||\widetilde{\boldsymbol{\Sigma}}^{\frac{1}{2}}||_F^2.$$

$$(22)$$

By multiplying $\frac{a}{8}\sqrt{P_i P_j}$ to both sides of (22) and summing it over all $1 \leq i < j \leq c$, we obtain the last inequality of (20).

Take $\Delta = \sum_{i<j}^c \sqrt{P_i P_j} \Delta'_{ij} = \frac{1}{4}\sum_{i<j}^c \sqrt{P_i P_j}||\overline{\mathbf{X}}_i - \overline{\mathbf{X}}_j||_F^2$,

and note that $||\widetilde{\boldsymbol{\Sigma}}^{\frac{1}{2}}||_F^2 = \sum_{i=1}^c \sum_{s=1}^{N_i} ||\mathbf{w}^T(\mathbf{X}_{is} - \overline{\mathbf{X}}_i)||_2^2$, we then obtain (12). □

References

1. Fukunaga K (2013) Introduction to statistical pattern recognition. Academic Press, New York

2. Shah JH, Sharif M, Yasmin M et al (2020) Facial expressions classification and false label reduction using LDA and threefold SVM. Pattern Recogn Lett 139:166–173

3. Ouyang AJ, Liu YM, Pei SY et al (2020) A hybrid improved kernel LDA and PNN algorithm for efficient face recognition. Neurocomputing 393:214–222

4. Miha P, Vili P (2017) Text classification method based on self training and LDA topic models. Expert Systems with Applications. 80:83–93

5. Chen Y, Zhang H, Liu R et al (2019) Experimental explorations on short text topic mining between LDA and NMF based schemes. Knowl-Based Syst 163:1–13

6. Cao G, Iosifidis A, Gabbouj M et al (2017) Multi view nonparametric discriminant analysis for image retrieval and recognition. IEEE Signal Process Lett 24(10):1537–1541

7. Liu Z, Zhang CM, Chen CX (2018) MMDF LDA An improved multi modal latent dirichlet allocation model for social image annotation. Expert Syst Appl 104:168–184

8. Wang H, Fan Y, Fang B et al (2018) Generalized linear discriminant analysis based on euclidean norm for gait recognition. Int J Mach Learn Cybern 9(4):569–576

9. Dong K, Zhao H, Tong T et al (2016) NBLDA: Negative binomial linear discriminant analysis for RNA-seq data. BMC Bioinforma 17(1):1–10

10. Ibrahim W, Abadeh MS (2019) Protein fold recognition using deep kernelized extreme learning machine and linear discriminant analysis. Neural Comput Appl 31(8):4201–4214

11. Sharma A, Paliwal KK (2015) Linear discriminant analysis for the small sample size problem: an overview. Int J Mach Learn Cybern 6(3):443–454

12. Li X, Pang Y, Yuan Y (2010) L1-norm-based 2DPCA. IEEE Trans Syst Man Cybern Part B (Cybern) 40(4):1170–1175

13. Mi JX, Zhang YN, Li Y et al (2020) Generalized two-dimensional PCA based on ℓ_2-norm minimization. Int J Mach Learn Cybern 11:2421–2438

14. Lu Y, Yuan C, Lai Z et al (2019) Horizontal and vertical nuclear norm based 2DLDA for image representation. IEEE Trans Circ Syst Video Technol 29(4):941–955

15. Zhao M, Jia ZG, Cai YF et al (2021) Advanced variations of two-dimensional principal component analysis for face recognition. Neurocomputing 452:653–664

16. Li M, Yuan B (2005) 2D-LDA: a statistical linear discriminant analysis for image matrix. Pattern Recogn Lett 26(5):527–532

17. Imani M, Ghassemian H (2015) Two dimensional linear discriminant analyses for hyperspectral data. Photogr Eng Remote Sens 81(10):777–786

18. Chen SB, Chen DR, Luo B (2015) L1-norm based two-dimensional linear discriminant analysis. J Electron Inf Technol 37(6):1372–1377

19. Li CN, Shao YH, Deng NY (2015) Robust L1-norm two-dimensional linear discriminant analysis. Neural Netw 65:92–104

20. Li CN, Shang MQ, Shao YH et al (2019) Sparse L1-norm two dimensional linear discriminant analysis via the generalized elastic net regularization. Neurocomputing 337:80–96

21. Li M, Wang J, Wang Q et al (2017) Trace ratio 2DLDA with L1-norm optimization. Neurocomputing 266(29):216–225

22. Lu Y, Yuan C, Lai Z et al (2018) Horizontal and vertical nuclear norm-based 2DLDA for image representation. IEEE Trans Circ Syst Video Technol 29(4):941–955

23. Zhang P, Deng S, Nie F et al (2019) Nuclear-norm based 2DLDA with application to face recognition. Neurocomputing 339:94–104

24. Li CN, Shao YH, Chen WJ et al (2021) Generalized two-dimensional linear discriminant analysis with regularization. Neural Netw 142:73–91

25. Li CN, Shao YH, Wang Z et al (2019) Robust bilateral Lp-norm two-dimensional linear discriminant analysis. Inf Sci 500:274–297

26. Du H, Zhao Z, Wang S et al (2017) Two-dimensional discriminant analysis based on Schatten p-norm for image feature extraction. J Vis Commun Image Represent 45:87–94

27. Lee YP (2015) Palm vein recognition based on a modified $(2d)^2$LDA. Signal Image Video Process 9(1):229–242

28. Liu X, Cao Y, Cao Y et al (2015) Novel method fusing $(2d)^2$LDA with multichannel model for face recognition. J Harbin Inst Technol 22(6):110–114

29. Wang Q, Qin Z, Nie F et al (2017) Convolutional 2DLDA for nonlinear dimensionality reduction. Int Joint Conf Artif Intell:2929–2935

30. Xiao X, Chen Y, Gong YJ et al (2019) Two-dimensional quaternion sparse discriminant analysis. IEEE Trans Image Process 29:2271–2286

31. Li CN, Shao YH, Wang Z et al (2019) Robust Bhattacharyya bound linear discriminant analysis through an adaptive algorithm. Knowl-Based Syst 183:104858

32. Nielsen F (2014) Generalized Bhattacharyya and Chernoff upper bounds on Bayes error using quasi-arithmetic means. Pattern Recogn Lett 42:25–34

33. Guo YR, Bai YQ, Li CN et al (2021) Reverse nearest neighbors Bhattacharyya bound linear discriminant analysis for multimodal classification. Eng Appl Artif Intell 97:104033

34. Yang J, Zhang D, Frangi AF et al (2004) Two-dimensional PCA: a new approach to appearance-based face representation and recognition. IEEE Trans Pattern Anal Mach Intell 26(1):131–137

35. Ayed IB, Punithakumar K, Li S (2015) Distribution matching with the Bhattacharyya similarity: A Bound Optimization Framework. IEEE Trans Pattern Anal Mach Intell 37(9):1777–1791

36. Jiang B, Zhu B (2021) Dynamic Bhattacharyya bound based approach for fault classification in industrial processes. IEEE Transactions on Industrial Informatics

37. Gyamfi KS, Brusey J, Hunt A et al (2018) Linear dimensionality reduction for classification via a sequential Bayes error minimisation with an application to flow meter diagnostics. Expert Syst Appl 91:252–262

38. Hu CF, Wang YX, Gu JW (2020) Cross domain intelligent fault classification of bearings based on tensor aligned invariant subspace learning and two dimensional convolutional neural networks. Knowl-Based Syst 209:106214

39. Hu CF, Wang YX (2019) Multidimensional denoising of rotating machine based on tensor factorization. Mech Syst Signal Process 122:273–289

40. Hu CF, He SL, Wang YX (2021) A classification method to detect faults in a rotating machinery based on kernelled support tensor machine and multilinear principal component analysis. Appl Intell 51:2609–2621

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Yan-Qin Bai is a Professor and Doctoral supervisor in the Department of Mathematics, School of Sciences at Shanghai University. She received her Ph.D. degrees majored on Operations Research and Cybernetics from Shanghai University. She held a research Fellow and postdoctoral research Fellow position in Delft University of Technology during the period of 2001-2004 and in Tilburg University, Netherlands in 2006, respectively. Her research interests include the interior-point methods for solving linear and conic optimization problems, with applications towards machine leaning, spare optimization, portfolio selection, etc. She has published over 100 research papers and a book entitled Kernel Function-Based Interior-point Algorithms for Conic Optimization published by Science Press China. She once served as the vice-president of Operations Research Society of China (ORSC), vice-president of the Branch of Mathematical Programming, and currently is the president of Operations Research Society of Shanghai, the associate editors-in-chief of Operations Research Transactions, and the editorial board member of Journal of the Operations Research Society of China (JORSC).

Chun-Na Li received her Master's degree and Ph.D degree in Department of Mathematics from Harbin Institute of Technology, China, in 2009 and 2012, respectively. Currently, she is an associate professor at Management School, Hainan University. Her research interests include optimization methods, machine learning and data mining.

Yan-Ru Guo received her Master's degree in Department of Mathematics from Zhejiang University of Technology, China, in 2015. Currently, she is pursuing a Ph.D. degree at Shanghai University. Her research interests include optimization methods and machine learning.

Lan Bai received the Ph.D. degree in mathematics from the Department of Mathematics, Jilin University, Changchun, China, in 2014. She is currently a Lecturer with the School of Mathematical Sciences, Inner Mongolia University, Hohhot, China. Her research interests include clustering techniques, feature selection, and data mining.

Yuan-Hai Shao received his B.S. degree in information and computing science in College of Mathematics from Jilin University, the masters degree in applied mathematics, and Ph.D. degree in operations research and management in College of Science from China Agricultural University, China, in 2006, 2008 and 2011, respectively. Currently, he is a professor at Management School, Hainan University. His research interests include optimization methods, machine learning and data mining. He has published over 100 refereed papers.

Affiliations

Yan-Ru Guo[1] · Yan-Qin Bai[1] · Chun-Na Li[2] [ID] · Lan Bai[3] · Yuan-Hai Shao[2]

Yan-Ru Guo
Guoyanru211@163.com

Yan-Qin Bai
yqbai@shu.edu.cn

Lan Bai
imubailan@163.com

Yuan-Hai Shao
shaoyuanhai21@163.com

[1] Department of Mathematics, Shanghai University, Shanghai 200444, People's Republic of China

[2] Management School, Hainan University, Haikou, 570228, People's Republic of China

[3] School of Mathematical Sciences, Inner Mongolia University, Hohhot, 010021, People's Republic of China