The Journal of Supercomputing (2024) 80:413–434
https://doi.org/10.1007/s11227-023-05485-9

Polynomial linear discriminant analysis

Ruisheng Ran[1] · Ting Wang[1] · Zheng Li[1] · Bin Fang[2]

Accepted: 6 June 2023 / Published online: 24 June 2023

Abstract

The traditional linear discriminant analysis (LDA) is a classical dimensionality reduction method. But there are two problems with LDA. One is the small-sample-size (SSS) problem, and the other is the classification ability of LDA is not very strong. In this paper, by studying several common methods of LDA, an implicit rule of the criterion of LDA is found, and then a general framework of LDA combined with matrix function is presented. Based on this framework, the polynomials are used to reconstruct the LDA criterion, and then a new method called polynomial linear discriminant analysis (PLDA) is proposed. The proposed PLDA method has two contributions: first, it addresses the small-sample-size problem of LDA, and second, it enhances the distance of the between-class sample through polynomial function mapping, improving the classification ability. Experimental results on ORL, FERET, AR, and Coil datasets show the superiority of PLDA over existing variations of LDA.

Keywords Linear discriminant analysis · The SSS problem · Matrix function · Polynomial

✉ Bin Fang
fb@cqu.edu.cn

Ruisheng Ran
rshran@cqnu.edu.cn

Ting Wang
764874012@qq.com

Zheng Li
a1466751163@hotmail.com

[1] The College of Computer and Information Science, Chongqing Normal University, Huxi street, Chongqing 401331, Chongqing, China

[2] The College of Computer Science, Chongqing University, Huxi street, Chongqing 400044, Chongqing, China

1 Introduction

In some fields, for example, data visualization, the features are usually represented as high-dimensional data. However, the original high-dimensional data may contain redundant and noisy information, which is easy to cause errors in practical applications. The direct processing of high-dimensional data may cause higher computational costs, known as the"curse of dimensionalit". Therefore, it is necessary to reduce the dimensionality of high-dimensional data and find its low-dimensional features.

The principal component analysis (PCA) [1] and linear discriminant analysis (LDA) [2, 3] are the most widely used dimensionality reduction methods. Because of the ability of LDA to directly obtain analytical solutions based on generalized eigenvalue problems, it can avoid the local minimum problem often encountered in nonlinear algorithm construction. So, LDA is used widely in many fields. However, LDA has the small-sample-size problem in practical application [4]. In addition, there are other issues with LDA, which make its performance not enough strong, such as classification accuracy, sensitivity to outliers, nonlinear distributed samples, etc. Therefore, scholars have proposed many improvement methods to solve the above problems of LDA.

For the SSS problem of LDA, many methods have been proposed. Specially, literature [5] proposed a method combining PCA and LDA in 1997, called the Fisherface method. This method used PCA to reduce the dimension of the original data so that the within-class scatter matrix S_w is non-singular, then LDA was used to classify the dimensionally reduced subspace. The literature [6] proposed a regularization discriminant analysis (RLDA), which added a regularization term to S_w the LDA criterion, and then avoided the SSS problem. The literature [7] proposed an exponential discriminant analysis(EDA) algorithm, which used exponential function to map the scatter matrix to eliminate the singularity of the within-class scatter matrix. The literature [8] proposed a noise-free length discriminant analysis (NLDA) method, which used the hyperbolic cosine function to reconstruct the objective function and then gets the new within-class and between-class scatter matrices.

From the perspective of linear subspace, literature [9] proposed the Null-space LDA method to find the projection vector that maximizes the distance between classes in the null space of the within-class scatter matrix. Literature [10] proposed the direct LDA method (D-LDA), which discarded the null space of the between-class scatter matrix and reserved the null space of the within-class scatter matrix. In 2003, literature [10–12] proposed to use of a new variant of D-LDA, i.e., DF-LDA, which used the fractional-step LDA (F-LDA) method to safely remove the null space scheme of the between-class scatter matrix. In 2012, Sharma and Paliwal proposed a two-stage LDA (TSLDA) [13], which can be implemented in parallel by $S'^{-1}_w S_b$ and $S'^{-1}_b S_w$ to get two directional matrices and then to get the final projection matrix.

From the perspective of objective function optimization, literature [14] proposed the maximum margin criterion method (MMC) in 2006, which took the

maximization of $S_b - S_w$ as an objective function. In 2014, literature [15] proposed the angle linear discriminant embedding method (ALDE), which redefined the scatter matrix by measuring the sample dispersion with angle cosine. Some robust dimensionality reduction methods, for example [16] and L_1-LDA [17], improved the robustness of LDA by using L_1-norm or $L_{2,1}$-norm. FLDA-Lsp [18] and NPDA [19] solve the problem that L_1-LDA is not robust enough and further enhance the robustness of LDA.

The kernel-based method was also a direction for improvement, for example, kernel-based Fisher discriminant (KDA) [20], kernel-based generalized discriminant analysis (GDA) [21], and kernel principal component analysis (KPCA) [22, 23], and spectral regression kernel discriminant analysis (SRKDA). These methods mapped the samples to feature space through the nonlinear kernel functions so that the samples can be classified even when they are non-linearly distributed.

The traditional LDA only focused on the global geometric features of data. The LPP [24], CGLDA [25], LocLDA [26], and other algorithms constructed the nearest neighbor graph to preserve the local geometric information and improve the classification accuracy. The 3E-LDA method simultaneously addressed the small-sample-size, sensitivity to outliers, and absence of local geometric information problems of LDA [27]. From the perspective of semi-supervised, many researchers have proposed the TR-SDA [28] and SL-LDA [29, 30] methods. These methods combine the calculated soft labels into the scatter matrix to find the optimal projection matrix.

In this paper, we focus on improving the classification ability of LDA and addressing the small-sample-size problem of LDA. Inspired by the criterion functions of RLDA [6], EDA [7], and NLDA [8] methods, we use two scalar functions $f(x)$ and $g(x)$ to map the between-class scatter matrix S_b and the within-class scatter matrix S_w to the matrix functions $f(S_b)$ and $g(S_w)$ respectively, and then present a new framework of LDA. Under this framework and based on the polynomial approximation of any function, we can design two suitable polynomials to improve the classification ability of LDA and address the SSS problem of LDA. Thus, a new LDA method, named polynomial linear discriminant analysis (PLDA), is proposed. The main contributions of the algorithm are as follows:

1. The small-sample-size problem inherent in traditional LDA is solved.
2. The polynomial function mapping greatly enlarges the distance of the between-class samples and shows a better classification effect.

The rest of this paper is as follows: the second section discusses the theoretical background; the third section presents the PLDA method; the fourth section is the experimental results; the fifth section summarizes this research.

2 Theoretical background

2.1 Linear discriminant analysis

The idea of LDA is to project the high-dimensional samples into the optimal discriminant vector space, thereby extracting classification information and compressing the dimension of the feature space. After the projection, the between-class distance is maximized and the within-class distance is minimized simultaneously so that the separability of the patterns in this space is the best.

Let $X = \{x_1, x_2, \cdots, x_m\}$ be the n-dimensional data in R^n space, m is the number of training samples, and n is the dimensionality of training samples. Denote the category label as $c_i \in \{1, 2, \cdots, k_c\}$, where k_c is the number of categories of the samples. The purpose of LDA is to find an optimal projection matrix W such that x_j is projected to y_j of the d-dimensional data ($d < n$),

$$y_j = W^T x_j. \tag{1}$$

The LDA method uses the center point to measure distances of the within-class and between-class samples. The center point of each category is the mean vector of each category sample, let m_i be the mean vector of the ith class, where k is the number of ith class samples.

$$m_i = \frac{1}{k} \sum_{j=1}^{k} x_j. \tag{2}$$

The within-class scatter matrix S_w and between-class scatter matrix S_b of samples can be calculated as:

$$\begin{aligned} S_w &= \sum_{i=1}^{k_C} \sum_{j \in c_i} (x_j - m_i)(x_j - m_i)^T \\ S_b &= \sum_{i=1}^{k_C} (m_i - m)(m_i - m)^T \end{aligned} \tag{3}$$

where m is the mean vector of all samples.

The criterion of LDA is:

$$J(W) = \arg\max_W \frac{tr(W^T S_b W)}{tr(W^T S_w W)}. \tag{4}$$

Eq.(4) can be reduced into a generalized eigenvector problem:

$$S_b w = \lambda S_w w. \tag{5}$$

If S_w is a non-singular matrix, the optimal projection matrix W desired by LDA is composed of the eigenvectors corresponding to the maximum eigenvalues of ${S_w}^{-1} S_b$. In real life, because the number of samples is usually far less than the dimensionality

of samples, the matrix S_w is singular. Thus, the matrix S_w cannot be inverted, and Eq. (5) is unsolvable. That is the small-sample-size (SSS) problem of LDA. The common method to solve the SSS problem is "PCA+LDA" [5]. It first uses PCA to reduce the dimensionality of the original samples, so that the constructed within-class scatter matrix is non-singular. Thus, the generalized eigenvector problem is solvable. In this way, the small-sample-size problem is solved.

2.2 Regularized linear discriminant analysis

For the LDA method, when calculating the zero (or very small) eigenvalues of the within-class scatter matrix, high variance is often introduced [31]. To avoid this problem, the RLDA method [6] adds a regularization term to S_w and then replaces the original S_w matrix:

$$\tilde{S_w} = S_w + rI,$$

where r is a regularization parameter and I is an identity matrix. Then, the criterion of RLDA is:

$$J(W) = \arg\max_{W} \frac{tr(W^T S_b W)}{tr(W^T \tilde{S}_w W)}. \tag{6}$$

The goal of regularization is to reduce the high variance associated with the eigenvalues of the within-class scatter matrix. Meanwhile, the singularity of S_w is also avoided and the SSS problem is addressed.

2.3 Exponential discriminant analysis

The singularity of S_w is from the zero or small eigenvalues of the scatter matrix. The EDA method is a method for dealing with zero eigenvalues of the within-class matrix S_w, which replaces S_b and S_w with $\exp(S_b)$ and $\exp(S_w)$ respectively [7]. The criterion of the EDA method is formulated as:

$$J(W) = \arg\max_{W^T W=I} \frac{tr(W^T \exp(S_b) W)}{tr(W^T \exp(S_w) W)}. \tag{7}$$

By the properties of the exponential matrix, the EDA method makes the minimum eigenvalue of S_w is 1 and then ensures the within-class scatter matrix S_w full rank. So, the EDA method solves the singularity of the within-class scatter matrix and extracts all the discriminant information of the subspace. Where, the orthogonality is enforced on the the criterion of EDA.

2.4 Noisy-free length discriminant analysis

The LDA method and its variant version, for example, EDA, MMC, etc., assume that the data follow a single-peak Gaussian distribution. Without this assumption, the

samples in the class would be multimodal. Multimodal data means that some samples of a class form several independent clusters. Then the separability of different classes cannot be well represented by the between-class and within-class scatter matrix. In addition, noise may affect the classification effect.

To solve these problems, Murthy et al. developed noise-free relevant pattern selection, which divides the modes into boundary, non-boundary, and noise mode. A noise-free length discriminant analysis model is constructed for noiseless boundary mode and non-boundary mode [8].

In this paper, the hyperbolic cosine function is selected to map the between-class and within-class scatter matrix, so that the smaller length difference between the two classes is smaller, the larger one is larger, and there are no zero eigenvalues. The criterion of NLDA is:

$$J(W) = \arg\max_{W^T W=I} \frac{tr(W^T \cosh(S_b)W)}{tr(W^T \cosh(S_w)W)}. \tag{8}$$

It can be solved by the following generalized eigenvalue decomposition method $\cosh(S_b)w = \lambda \cosh(S_w)w$. Note that the matrix $\cosh(S_w)$ is a full rank matrix and $\cosh(S_w)^{-1}$ exists. Then the above equation can be directly solved with $\cosh(S_w)^{-1} \cosh(S_b)w = \lambda w$. Thus, there has no the small-sample-size problem.

2.5 Matrix function and its eigensystem

In this section, we give the definition and properties of the matrix function, which will be used in the paper.

Definition 1 ([32]) Let A be a square matrix and $f(x)$ be a scalar function. If one replaces the variable x in $f(x)$ with the square matrix A, the resulting matrix $f(A)$ is called the matrix function of matrix A.

Theorem 1 ([32]) *Let A be a real symmetric square matrix of n-order, $f(x)$ be a scalar function, λ_i be the eigenvalue of the matrix A and v_i be the eigenvector belonging to the eigenvalue λ_i, i.e., $Av_i = \lambda_i v_i (i = 1, 2, \cdots, n)$. For the matrix function $f(A)$, one has*:

$$f(A)v_i = f(\lambda_i)v_i (i = 1, 2, \cdots, n). \tag{9}$$

where $f(\lambda_i)$ are the eigenvalues of $f(A)$, and v_i are the eigenvectors corresponding to $f(A)$.

Theorem 2 (*Weierstrass approximation theorem* [33]) *Continuous functions on closed intervals can be uniformly approximated by polynomial series.*

3 Polynomial linear discriminant analysis

3.1 The matrix function framework for LDA

The criterion function of LDA is Eq. (4), which is presented below and denoted as Eq. (10) for convenience

$$J(W) = \arg\max_{W} \frac{tr(W^T S_b W)}{tr(W^T S_w W)}, \tag{10}$$

where W is the desired projection matrix.

By observing the criterion functions of RLDA, EDA and NLDA in Sect. 2, namely Eqs. (6–8) it can be found that they all perform some function transformation on the between-class scatter matrix S_b and the within-class scatter matrix S_w. For example, for the RLDA method, let $f(x) = x$, $g(x) = x + r$, they are used to map the matrices S_b and S_w respectively and Eq. (6) is got; for the EDA method, let $f(x) = \exp(x)$, $g(x) = \exp(x)$, they are used to map the matrices S_b and S_w respectively and Eq. (7) is got; for the NLDA method, let $f(x) = \cosh(x)$, $g(x) = \cosh(x)$, they are used to map the matrices S_b and S_w respectively and Eq. (8) is got.

More generally, we can use two scalar functions $f(x)$ and $g(x)$ to map the between-class scatter matrix S_b and within-class scatter matrix S_w in Eq. (10) into the matrix functions $f(S_b)$ and $g(S_w)$ respectively, namely:

$$S_b \to f(S_b), S_w \to g(S_w).$$

After mapping, the new criterion function of LDA can be expressed as:

$$J(W) = \arg\max_{W} \frac{tr(W^T f(S_b) W)}{tr(W^T g(S_w) W)}. \tag{11}$$

Eq. (11) can be reduced into the following generalized eigenvector problem with some linear algebra transformation:

$$f(S_b)w = \lambda g(S_w)w.$$

The optimal transformation matrix W is formed by the eigenvectors w_j corresponding to the maximum eigenvalues from the above equation. Finally, the n-dimensional original data x_i can be reduced to d-dimensional data y_i by the projection $y_i = W^T x_i$.

Thus, we give a new framework for LDA, namely Eq. (11). This criterion Eq. (11) can be regarded as a general platform, when $f(x)$ and $g(x)$ take different forms, a series of LDA methods can be obtained. For the RLDA, EDA, and NLDA methods, the criterion functions Eqs. (6–8) can be obtained when $f(x)$ and $g(x)$ in Eq. (11) taking the above forms.

3.2 Polynomial linear discriminant analysis

In Eq. (11), a new general framework for LDA is proposed. The main idea of the framework is to use two scalar functions to map the scatter matrices into the corresponding matrix functions, and then to make dimensionality reduction and feature extraction. However, in real life, there are many optional functions, but the suitable functions are difficult to find. According to Theorem 2, a polynomial can approximate any function, or any function can be represented by a polynomial, so we can use polynomials to represent the optional and various forms of functions. Then, the framework Eq. (11) is easier to use.

On the other hand, one always wants to improve the classification ability of the LDA method and avoid the SSS problem of LDA. From the above analysis, we can choose an l-order polynomial $f(x) = \sum_{k=0}^{l} a_k x^k$ to map matrix S_b, thus the gotten matrix function is $f(S_b) = a_0 + a_1 S_b + \cdots + a_l S_b^l$. Simultaneously, we use a simple linear function $g(x) = b + x(b > 0)$ to map the matrix S_w, the gotten matrix function is $g(S_w) = bI + S_w$. Thus, the criterion of LDA becomes:

$$J(W) = \arg\max_{W} \frac{tr(W^T(\sum_{k=0}^{l} a_k S_b^k)W)}{tr(W^T(bI + S_w)W)}. \tag{12}$$

Eq. (12) can be reduced to the generalized eigenvector problem:

$$\left(\sum_{k=0}^{l} a_k S_b^k\right) w = \lambda(bI + S_w)w. \tag{13}$$

Thus, a new LDA method has been presented. Because the polynomial is used to reconstruct the criterion of LDA, the new LDA method is named polynomial linear discriminant analysis (PLDA). This design can improve the classification ability of LDA and solve the SSS problem of LDA, which will be illustrated in the theoretical analysis in the next section.

The above mappings transform the matrices S_b and S_w from the original sample space to a new space.This mapping is like the kernel method. The kernel method [22] is a powerful tool for dealing with the nonlinear data, for example, the KPCA [23] and GDA [21] methods. The idea of the kernel method is to map the nonlinear data into the feature space by the kernel function, so that the nonlinear problem in the original space is transformed into a linear problem in the feature space. Similarly, in the PLDA method, we use the nonlinear polynomial functions to map the scatter matrices into the new feature space:

$$\begin{aligned} \Phi &: \mathbb{R}^n \to F \\ S_b &\mapsto \Phi_1(S_b) = \sum_{k=0}^{n} a_k S_b^k \\ S_w &\mapsto \Phi_2(S_w) = bI + S_w \end{aligned} \tag{14}$$

The difference is the kernel method, for example KPCA and GDA, map the samples from the original sample space to a feature space, while the PLDA method maps the scatter matrices into a feature space. However, note that the between-class scatter matrix S_b and the within-class scatter matrix S_w are composed of the original samples. Therefore, the kernel method and PLDA have certain similarities and PLDA should be have the characteristics of the kernel approach.

For a given dataset $X \in R^{m\times n}$, m is the number of the training samples, n is the dimensionality of the training samples. According to [34], the computational complexity of LDA is $O(mnt + t^3)$, where $t = \min(m, n)$. Compared with LDA, the time complexity of PLDA is mainly determined by calculating $f(S_b)$.

We'll see in Sect. 3.4 that PLDA chooses a seventh degree polynomial to map the S_b, i.e., $f(S_b) = a_0 + a_1 S_b + \cdots + a_7 S_b^7$. The calculation of each term of $f(S_b)$ can be divided into the calculation of even or odd power. Let $A = S_b$, $B = A^2 = A \times A$, one can get each even term recursively as follows: $A^4 = B \times B$, $A^6 = B \times A^4$... For odd power, one has: $A^3 = A \times B$, $A^5 = A \times A^4$... Thus, using recursion, each matrix power can be obtained by simple multiplication of two matrices. Then the time complexity of computing a seventh degree polynomial is $O(4n^3)$. Thus, the computation complexity of PLDA is $O\left(mnt + t^3 + 4n^3\right)$.

The PLDA, RLDA, EDA and NLDA methods all use two scalar functions to map the scatter matrix. Compared with the other three methods, the advantage of PLDA is that it can enlarge the between-class distance by the mapping of the polynomial $f(x) = \sum_{k=0}^{l} a_k x^k$ to the between-class scatter matrix S_b, while almost maintaining the within-class distance unchanged. In this way, the separation between patterns of different classes is emphasized, which helps with pattern classification. The RLDA method only adds a regularization parameter to the within-class scatter matrix to avoid the small-sample-size problem, and does not have the distance diffusion effect. Although the EDA and NLDA methods have also distance diffusion effect, they both expand the within-class and between-class distance at the same time. This expands the between-class distance, but it also increases the within-class distance, resulting in EDA and NLDA having lower classification ability than PLDA. The disadvantage of PLDA is that it has a slightly higher time complexity compared to LDA, EDA and NLDA.

3.3 Theoretical analysis

3.3.1 Avoid the SSS problem

After mapping the matrix S_w with the linear function $g(x) = b + x(b > 0)$, the gotten matrix function is $g(S_w) = bI + S_w$. Let λ_{wi} be the eigenvalues of the matrix S_w. The matrix S_w is semidefinite according to Sect. 2, and so $\lambda_{wi} \geq 0$. According to Theorem 1, $b + \lambda_{wi}$ are the eigenvalues of the matrix function $g(S_w) = bI + S_w$. Since $b > 0$, one has $b + \lambda_{wi} > 0$ and then the matrix function $g(S_w) = bI + S_w$ will not have zero eigenvalues and be nonsingular. Thus, Eq. (13) is solvable. Therefore, the PLDA method avoids the small-sample-size problem.

3.3.2 Distance diffusion mapping

The main idea of LDA is to maximize the distance of between-class samples and minimize the distance of within-class samples after projection. The two distances can be represented by traces of the scatter matrix, namely, $tr(S_b)$ and $tr(S_w)$. Where, $tr(S_b)$ represents the separability of between-class samples, and $tr(S_w)$ represents the closeness of within-class samples [7]. The trace of a matrix is the sum of all the eigenvalues of the matrix, then the above two distances can be written as:

$$tr(S_b) = \lambda_{b1} + \lambda_{b2} + \cdots + \lambda_{bn}. \tag{15}$$

$$tr(S_w) = \lambda_{w1} + \lambda_{w2} + \cdots + \lambda_{wn}. \tag{16}$$

For the PLDA method, the matrix S_b is mapped into the matrix function $f(S_b) = \sum_{k=0}^{l} a_k S_b^k$ by the polynomial $f(x) = \sum_{k=0}^{l} a_k x^k$, and the matrix S_w is mapped into the matrix function $g(S_w) = bI + S_w$ by $g(x) = b + x(b > 0)$. According to Theorem 1, $f(\lambda_{bi}) = \sum_{k=0}^{l} a_k \lambda_{bi}^k$ is the eigenvalue of the matrix function $f(S_b) = \sum_{k=0}^{l} a_k S_b^k$, and $g(\lambda_{wi}) = b + \lambda_{wi}$ is the eigenvalue of the matrix function $g(S_w) = bI + S_w$. After mapping, the new distance of the between-class samples and the within-class samples are:

$$tr(f(S_b)) = \left(\sum_{k=0}^{l} a_k \lambda_{b1}^k\right) + \left(\sum_{k=0}^{l} a_k \lambda_{b2}^k\right) + \cdots + \left(\sum_{k=0}^{l} a_k \lambda_{bn}^k\right). \tag{17}$$

$$tr(g(S_w)) = (b + \lambda_{w1}) + (b + \lambda_{w2}) + \cdots + (b + \lambda_{wn}). \tag{18}$$

According to Eqs. (17) and (18), the larger the eigenvalue λ_{bi} of the matrix S_b is, the larger the distance of between-class is; the smaller the eigenvalue λ_{wi} of the matrix S_w is, the smaller the distance of the within-class sample is. Therefore, the higher the ratio of λ_{bi} and λ_{wi}, the more discriminative the corresponding feature vector is.

In this paper, $f(x)$ is a polynomial function. After mapping with $f(x)$, the eigenvalue λ_{bi} is enlarged by the polynomial function $f(x)$, i.e., one has $f(\lambda_{bi}) > \lambda_{bi}$. And the $g(\lambda_{wi}) = b + \lambda_{wi} \approx \lambda_{wi}$. In general, the distance of between-class samples is larger than the distance of the within-class samples. Therefore, for most of the eigenvalues in Eq. (17) and Eq. (18), one has $\lambda_{bi} > \lambda_{wi}$. Then, one has:

$$\frac{f(\lambda_{bi})}{g(\lambda_{wi})} > \frac{\lambda_{bi}}{\lambda_{wi}}. \tag{19}$$

Thus, with the mapping of the polynomial functions, the PLDA method enlarges the margin of the between-class, while almost maintaining the distance of the within-class samples, which is helpful for pattern classification.

To illustrate the main idea of the PLDA method visually, we present a geometric interpretation of PLDA in Fig. 1. For convenience, two class examples are used. A class is represented by a red circle and the center of the class by a red square. The other class is represented by a blue circle, with a blue square at the center of

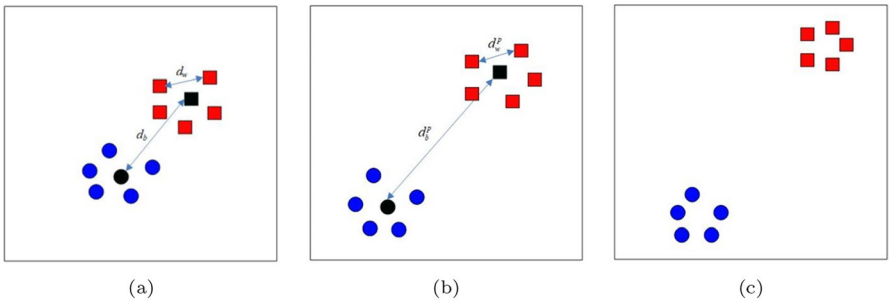


Fig. 1 Geometric interpretation of PLDA **a** The initial sample space; **b** The sample space is mapped by the polynomial matrix function; **c** The projected space

the class. Where d_w and d_b are the distances between samples in a certain class and between classes in the original sample space, while d_w^p and d_b^p are the distance between samples in a certain class and between classes in the sample space mapped by the polynomial functions. As shown in Fig. 1, the PLDA method maps the original sample to a new space using nonlinear mapping. In this new space, the distance between samples of between-class increases, and the distance between samples of within-class is almost constant. Then, the LDA method performs the linear projection and makes feature extraction on the new space.

3.4 Selection of polynomial functions

In this Section, we discuss how to select the polynomial $f(x)$ and $g(x)$. For $g(x) = b + x(b > 0)$, the constant b takes usually a little value, empirically, let $b = 0.1$, i.e., $g(x) = 0.1 + x$. The design of the polynomial $f(x)$ is discussed as follows.

It is well-known that, for any function $p(x)$, it has Taylor's expansion $p(x) = \frac{p(x_0)}{0!} + \frac{p'(x_0)}{1!}(x - x_0) + \cdots \frac{p^{(l)}(x_0)}{l!}(x - x_0)^l + o\left[(x - x_0)^l\right]$. When $x_0 = 0$, it is reduced to McLaughlin's expansion:

$$p(x) = \frac{p(0)}{0!} + \frac{p'(0)}{1!}x + \cdots + \frac{p^{(l)}(0)}{l!}x^l + o\left[x^l\right]. \tag{20}$$

In real application, the coefficients in Eq. (20) can be simplified to a certain constant c, so Eq. (20) can be rewritten as $p(x) = c\left(1 + x + \frac{x^2}{2!} \cdots + \frac{x^l}{l!}\right)$. According to the previous Section, the smaller the denominator of the polynomial, the larger the eigenvalue obtained by the polynomial mapping, so the polynomial $p(x)$ can be simplified to $p(x) = 1 + x + \frac{x^2}{2} + \cdots + \frac{x^l}{l}$.

Let's review EDA [7], which uses the exponential function $y = e^x$ to map the eigenvalues, and then to enlarge the distance of samples. In this paper, the designed polynomial should be stronger than the exponential function when the distance is enlarged.

For the polynomial $p(x)$, the larger the degree l, the larger the eigenvalue after mapping. The smaller the degree l, the smaller the influence on the eigenvalues. However, the greater the degree l, the greater the computational complexity of computing the l-order polynomial matrix function, So there is no need to choose a higher order polynomial.

Thus, we try to let $l = 3, 5, 7$, then the polynomial $p(x)$ becomes $s(x) = 1 + x + \frac{x^2}{2} + \frac{x^3}{3}$, $h(x) = 1 + x + \frac{x^2}{2} + \frac{x^3}{3} + \frac{x^4}{4} + \frac{x^5}{5}$, and $f(x) = 1 + x + \frac{x^2}{2} + \frac{x^3}{3} + \frac{x^4}{4} + \frac{x^5}{5} + \frac{x^6}{6} + \frac{x^7}{7}$. For the above three polynomials, for the same eigenvalue λ, one has $f(\lambda) > h(\lambda) > s(\lambda)$, so $f(x)$ has the strongest ability to enlarge the distance. Thus, the seventh-order polynomial $f(x) = 1 + x + \frac{x^2}{2} + \frac{x^3}{3} + \frac{x^4}{4} + \frac{x^5}{5} + \frac{x^6}{6} + \frac{x^7}{7}$ is chosen.

Now, we have $f(x) = \sum_{k=0}^{7} \frac{x^k}{k}$ and $g(x) = 0.1 + x$. when mapping the eigenvalues, we have $f(\lambda_{bi}) = \sum_{k=0}^{7} \frac{\lambda_{bi}^k}{k} \gg e^{\lambda_{bi}}$, and $g(\lambda_{wi}) = 0.1 + \lambda_{wi} \approx \lambda_{wi}$ Then, we have

$$\frac{f(\lambda_{bi})}{g(\lambda_{wi})} = \frac{\sum_{k=0}^{7} \frac{\lambda_{bi}^k}{k}}{0.1 + \lambda_{wi}} > \frac{e^{\lambda_{bi}}}{e^{\lambda_{wi}}} > \frac{\lambda_{bi}}{\lambda_{wi}}$$

So, the classification ability of PLDA should be better than that of EDA, and much better than that of LDA.

After selecting the polynomial $f(x)$ and $g(x)$, we provide an experiment to illustrate the distance diffusion effect of PLDA in four datasets: ORL, FERET, AR, and Coil-100. In the experiment, for each dataset, the original samples are normalized to avoid a large value. The between-class distance d_b and the within-class distance d_w of LDA are calculated by Eqs. (15) and (16), and the between-class distance d'_b and the within-class distance d'_w of PLDA are calculated by Eqs. (17) and (18), where the first ten largest eigenvalues are used. Table 1 shows the comparisons of the results of the LDA and PLDA methods on four datasets. As can be seen from Table 1, with the mapping $f(x) = \sum_{k=0}^{7} a_k x^k$, the between-class distance d'_b of PLDA are far larger than the between-class distance d_b of LDA. By mapping $g(x) = 0.1 + x$, however, the within-class distance d'_w of PLDA increases only slightly compared with the within-class distance d_w of LDA (and remains almost constant).

The criterion of LDA is shown in Eq. (11). If one only considers the mapping of $g(x)$ to the within-class scatter matrix S_w, the criterion of LDA will be the following form:

Table 1 The comparisons of distance of LDA and PLDA

	ORL	FERET	AR	Coil-100
d_w	4.12	24.89	28.41	93.36
d_b	4.54	66.51	33.32	154.85
d'_w	4.32	25.89	29.41	95.36
d'_b	21.53	100.49	100.40	200.48

$$J(W) = \arg\max_{W} \frac{tr(W^T S_b W)}{tr(W^T g(S_w) W)} = \arg\max_{W} \frac{tr(W^T S_b W)}{tr(W^T (0.1I + S_w) W)}$$

It is equivalent to the criterion of the RLDA.

Furthermore, if one adds the mapping of $f(x)$ to the between-class scatter matrix S_b, the criterion will be the following form:

$$J(W) = \arg\max_{W} \frac{tr(W^T f(S_b) W)}{tr(W^T g(S_w) W)} = \arg\max_{W} \frac{tr(W^T \sum_{k=0}^{7} \frac{S_b^k}{k} W)}{tr(W^T (0.1I + S_w) W)}$$

It is the criterion of the PLDA.

From the experimental results (Tables 3, 4, 5 and 6), the recognition rates of RLDA are sometimes higher than that of LDA, and sometimes even lower than that of LDA. This indicates that the mapping to the within-class matrix does not have much effect. However, compared to RLDA and LDA, in most cases, PLDA has better experimental results than RLDA and is significantly superior to LDA.

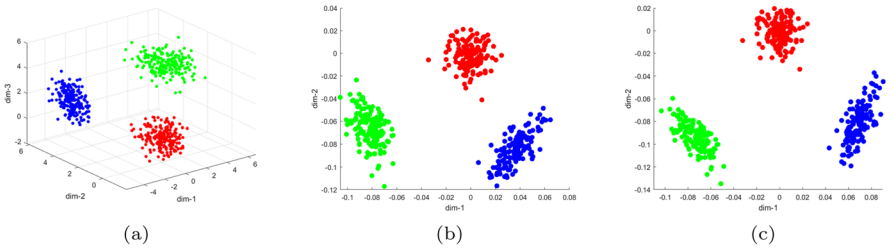
Therefore, from the above analysis, the mapping of the polynomial $f(x)$ to the between-class scatter matrix S_b is the dominant factor in improving the classification ability of PLDA.

4 Experimental results

4.1 Data visualization

To show the classification effect of PLDA more clearly, this paper carried out data visualization experiments on the synthetic data set (data with Gaussian distribution). The synthetic data with Gaussian distribution in Fig. 2a is a 3D dataset. This three-category set contains 600 points. Each class is generated using a single Gaussian function.

Figure 2b is the projection of LDA on the two-dimensional subspace, and Fig. 2c is the two-dimensional projection of PLDA. It can be seen that PLDA has good classification ability. We can also see that, compared with LDA, PLDA almost maintains



(a) (b) (c)

Fig. 2 The embedding of synthesized data **a** The synthetic data with Gaussian distribution; **b** The projection of LDA on two-dimensional subspace; **c** The projection of PLDA on two-dimensional subspace

Table 2 Details of the datasets used in the experiment

Dataset	Number of classes	Sample number of each class	Image size	Data dimension
ORL	40	10	32×32	1024
FERET	200	7	32×32	1024
Coil-100	100	72	32×32	1024
AR	120	14	40×50	2000

Fig. 3 From top to bottom are some samples of the ORL [42], FERET [43], Coin-100 [45], and AR [44] datasets

the distance of the within-class samples and enlarges the distance of the between-class samples, which is beneficial to pattern classification.

4.2 Image recognition experiments

4.2.1 Experimental setup

In this section, the experiments are performed on ORL, FERET, AR face databases, and Coil-100 object database. The details of the databases are reported in Table 2 (Fig. 3).

The ORL dataset was created by Olivetti Research Laboratories in Cambridge, UK. The images were captured at different times, under different lighting, with different facial expressions (open/closed eyes, smiling/not smiling) and facial details

Table 3 The recognition accuracy, standard deviation and optimal dimension of the ORL database

Method	3 trains	4 trains	5 trains
LDA	83.33 ± 1.76(10)	89.17 ± 0.84(10)	93.00 ± 1.32(10)
RLDA	79.17 ± 3.90(80)	82.78 ± 1.68(90)	88.17 ± 3.40(40)
NLDA	87.02 ± 1.61(40)	92.78 ± 0.89(40)	96.33 ± 1.04(40)
LDAMMC	83.57 ± 1.89(40)	88.89 ± 2.30(30)	93.17 ± 1.60(40)
KDA	84.17 ± 2.03(20)	88.75 ± 0.72(10)	93.67 ± 1.04(20)
EDA	86.07 ± 1.85(40)	91.25 ± 2.08(40)	93.83 ± 1.44(40)
TSLDA	84.05 ± 4.64(60)	88.75 ± 1.10(30)	92.33 ± 0.76(30)
ALDE	84.29 ± 2.34(40)	89.31 ± 1.93(40)	92.83 ± 1.89(40)
PLDA	**87.86 ± 1.80(40)**	**93.33 ± 1.05(40)**	**96.67 ± 1.04(40)**

The bold part represents the recognition accuracy of PLDA

(with/without glasses). All the images were taken against a dark, uniform background, with frontal faces, some slightly sideways.

The FERET dataset was published between 1993 and 1996 and was created by the FERET project. It contains grayscale face images with different lighting, each image contains only one face, and it has been widely used in the early face recognition field.

Coil-100 is a 128×128 color image dataset of different objects imaged at different angles in a 360 rotation.

AR dataset was created by Aleix Martinez and Robert Benavente at the Computer Vision Center (CVC) at the U.A.B. It contains over 4,000 color images corresponding to 126 people's faces (70 men and 56 women). Images feature frontal view faces with different facial expressions, illumination conditions, and occlusions (sunglasses and scarf). No restrictions on wear (clothes, glasses, etc.), make-up, hairstyle, etc. were imposed on participants.

On each dataset we provide a comparison of PLDA with algorithms for solving the SSS problem of LDA, including PCA+LDA [5], RLDA [6], NLDA [9], LDAMMC [14], KDA [20], EDA [7], TSLDA [13] and ALDE [15]. The LDA method below is actually the PCA+LDA method, which retains 98% of its principal component in the PCA stage. The NLDA method refers to null space LDA. For the RLDA method, we set $\tilde{S}_w = S_w + 0.1I$. For the proposed PLDA method, the polynomial functions used are: $f(x) = 1 + x + \frac{x^2}{2} + \frac{x^3}{3} + \frac{x^4}{4} + \frac{x^5}{5} + \frac{x^6}{6} + \frac{x^7}{7}$, $g(x) = 0.1 + x$. Where, $f(x)$ is an empirical function, and the other polynomial functions can be tried.

In LDA and its improved methods, the KNN algorithm is generally used for classification. In order to compare with LDA and its improved methods, the KNN algorithm is also used for classification in this paper.

4.2.2 Image recognition experiment

In the experiment, the p images are randomly selected from each category as the training sample set, and the rest images are used as the test set. For a given p, the dimension of the reduced subspace is 10 to 100 in steps of 10. In order to get stable

recognition accuracy, the above experiments were repeated ten times, and the average value of ten experiments in each dimension was taken as the recognition accuracy rate in that dimension. In the experiment, the p value is $p = 3, 4, 5$ for the ORL, FERET, and AR face datasets, and $p = 8, 10, 12$ for the Coil-100 dataset.

Tables 3, 4, 5, and 6 record the optimal recognition results of each method and the corresponding optimal subspace dimensions corresponding to the different training samples.

We also evaluate the performance of these methods when the subspace dimension varies. According to the above experimental results, for the training sample p, when the dimension of the subspace ranges from 10 to 100 (with a step size of 10), the recognition rate in each dimension can be got.

In recent years, people have proposed some latest improvements of LDA from various perspectives, including classic deep learning feature extraction algorithms, semi-supervised dimensionality reduction algorithms, etc. So, we also compare the PLDA method with these methods, and the comparison results are shown in Table 7.

Figure 4a–d below show the curve of recognition rate as a function of subspace dimension on ORL, FERET, Coil-100, and AR datasets, respectively.

Table 4 The recognition accuracy, standard deviation and optimal dimension of the FERET database

Method	3 trains	4 trains	5 trains
LDA	29.25 ± 2.38(10)	34.28 ± 0.25(10)	45.08 ± 3.17(10)
RLDA	76.17 ± 1.06(20)	85.67 ± 0.44(20)	86.33 ± 0.63(30)
NLDA	68.13 ± 2.63(30)	63.39 ± 2.44(30)	53.25 ± 3.93(40)
LDAMMC	63.04 ± 0.71(50)	75.11 ± 0.63(40)	77.83 ± 1.66(100)
KDA	41.59 ± 0.94(10)	47.94 ± 1.77(10)	55.50 ± 3.70(10)
EDA	76.88 ± 0.45(40)	82.83 ± 1.34(40)	83.50 ± 0.43(50)
TSLDA	23.54 ± 1.19(100)	23.22 ± 1.93(100)	19.50 ± 1.64(100)
ALDE	67.29 ± 1.33(40)	76.44 ± 1.36(40)	77.91 ± 1.26(40)
PLDA	**80.67 ± 0.26(50)**	**85.94 ± 0.84(40)**	**86.75 ± 0.75(100)**

The bold part represents the recognition accuracy of PLDA

Table 5 The recognition accuracy, standard deviation and optimal dimension of the Coil-100 database

Method	8 trains	10 trains	12 trains
LDA	61.09 ± 0.69(10)	65.49 ± 0.16(10)	68.09 ± 0.82(10)
RLDA	86.73 ± 0.51(20)	90.00 ± 0.47(20)	92.07 ± 0.68(30)
NLDA	71.33 ± 0.60(20)	71.53 ± 0.47(30)	71.40 ± 0.60(40)
LDAMMC	85.92 ± 0.53(30)	89.19 ± 0.21(40)	90.79 ± 0.65(40)
KDA	74.30 ± 0.70(10)	79.78 ± 0.66(10)	81.99 ± 1.43(10)
EDA	86.74 ± 0.78(30)	89.67 ± 0.35(30)	91.43 ± 0.46(30)
TSLDA	63.87 ± 0.46(40)	65.73 ± 0.07(100)	65.89 ± 0.62(50)
ALDE	85.23 ± 0.39(20)	88.67 ± 0.20(30)	90.10 ± 1.08(40)
PLDA	**86.77 ± 0.49(30)**	**90.31 ± 0.44(30)**	**92.13 ± 0.45(30)**

The bold part represents the recognition accuracy of PLDA

Table 6 The recognition accuracy, standard deviation and optimal dimension of the AR database

Method	3 trains	4 trains	5 trains
LDA	85.03 ± 2.05(10)	89.97 ± 1.35(10)	92.62 ± 0.43(10)
RLDA	71.03 ± 0.77(100)	78.42 ± 0.71(100)	82.78 ± 1.58(100)
NLDA	87.02 ± 0.76(100)	91.94 ± 0.49(100)	94.6 ± 0.42(100)
LDAMMC	83.18 ± 0.75(100)	91.05 ± 0.48(100)	94.75 ± 0.62(100)
KDA	87.32 ± 0.89(10)	91.47 ± 0.18(10)	93.80 ± 0.32(10)
EDA	86.93 ± 0.51(100)	93.08 ± 0.94(100)	95.05 ± 0.28(100)
TSLDA	87.88 ± 0.35(100)	91.61 ± 0.33(100)	93.37 ± 0.27(100)
ALDE	86.44 ± 0.59(100)	92.80 ± 0.80(100)	95.34 ± 0.79(90)
PLDA	**87.95 ± 0.55(100)**	**94.39 ± 0.78(100)**	**96.08 ± 0.28(90)**

The bold part represents the recognition accuracy of PLDA

4.3 Analysis

It can be seen from the experimental results:(1) Compared with the existing methods for solving the SSS problem of LDA, the proposed PLDA method has the best classification ability. Specially, compared with LDA, the performance of PLDA is much greater than that of LDA. This may be due to that PLDA enlarge the distance of between-class samples by polynomial mapping. (2) From Tables 4 and 5, on FERET and Coil-100, the best recognition rate of RLDA is good and close to that of PLDA. However, from Fig. 4b, c, with the increase of the subspace dimension, the recognition rate of this method shows a downward trend. In addition, the NLDA method has the same downward trend on FERET and Coil-100. The PLDA method is not affected by the size of the subspace dimension, and its recognition rate curve maintains a stable trend on all datasets. (3) Although the PLDA method has a certain similarity with the kernel method, the experimental results show that the performance of PLDA is better than that of KDA, GDA, KPCA and other kernel methods, which may also be due to the design of the polynomial function of PLDA. (4) As seen from Fig. 4, the classification ability of each method is different on different datasets. Some methods perform well on some datasets, but not on others. For example, the NLDA method performs well on the ORL dataset, but not on the FERET and Coil-100 datasets. The LDAMMC method performs well on the Coil-100 dataset, but not on the ORL and FERET datasets. The ALDE method performs well on the Coil-100 and AR datasets, but not on the other two datasets. However, the PLDA method has good and stable performance on all datasets. (5) According to Table 7, PLDA also has some advantages over the latest algorithms and deep learning methods in recent years. Therefore, PLDA not only outperforms other algorithms in the SSS problem but also performs better than other feature extraction methods. This is due to the polynomial mapping, which increases the distance between class samples.

Table 7 The comparison of recognition rates between PLDA and the latest improvement of LDA on ORL, FERET, Coil-100, and AR datasets

Methods	ORL			FERET			Coil-100		AR		
	$p = 3$	$p = 4$	$p = 5$	$p = 3$	$p = 4$	$p = 5$	$p = 8$	$p = 10$	$p = 3$	$p = 4$	$p = 6$
KPCA [35]	86.07	89.33	92.55	25.63	28.23	30.68	–	–	-	–	58.88
NWFE [35]	29.89	33.42	35.80	6.23	4.97	5.33	–	–	–	–	13.23
GMSDA [35]	88.71	92.08	94.95	77.35	80.88	82.78	–	–	–	–	95.09
3E-LDA [27]	–	88.50	94.20	–	–	–	–	–	–	–	–
NW-Fisherfaces [36]	–	68.95	73.73	–	–	–	–	–	–	68.95	78.04
RSLDA [37]	–	88.79	93.65	–	–	–	–	–	–	83.91	90.91
RNLDA [38]	91.79	95.00	96.50	–	–	–	–	–	75.09	–	–
CNN [39]	86.71	–	95.65	–	–	–	–	–	–	–	–
LRRA [40]	83.21	–	–	–	–	–	–	–	–	–	–
ASSBE [41]	–	–	–	–	–	83.75	–	–	-	–	92.58
PLDA	87.86	93.33	96.67	80.67	85.94	86.75	86.77	90.31	87.95	94.39	96.03

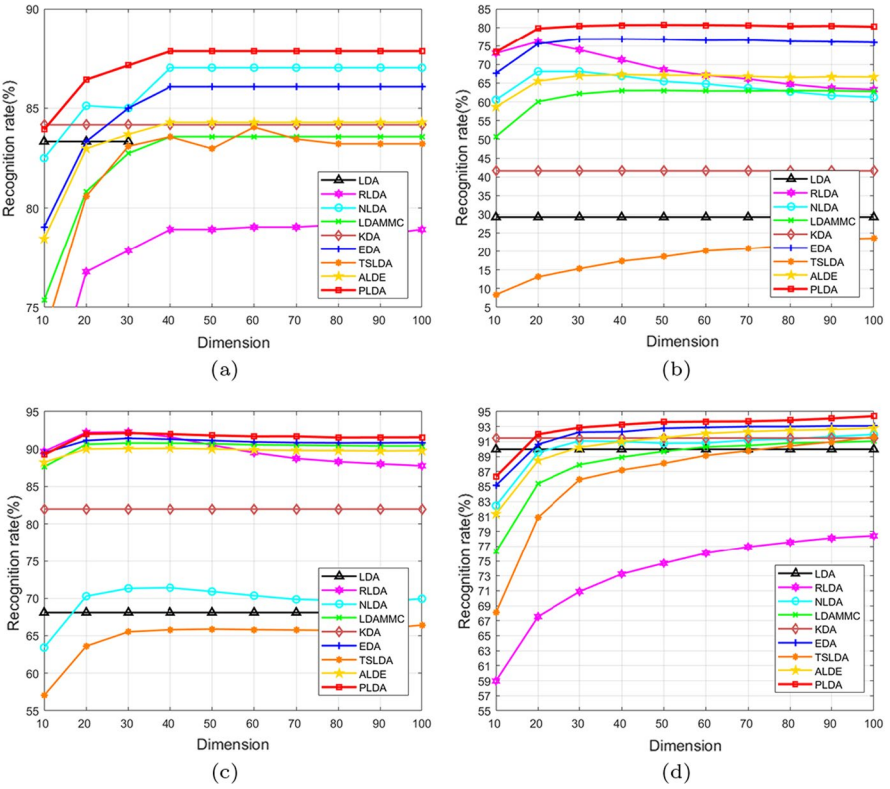


Fig. 4 The performance of the PLDA **a** comparison of performance and dimension on ORL dataset (training sample $p = 3$); **b** comparison of performance and dimension on FERET dataset (training sample $p = 3$); **c** comparison of performance and dimension on coil-100 dataset (training sample $p = 12$); **d** comparison of performance and dimension on AR dataset (training sample $p = 4$)

5 Conclusion

The traditional LDA method has the small-sample-size (SSS) problem. In this paper, by studying several common methods of LDA, we find an implicit rule of the criterion of LDA, and then we give a general criterion of LDA combined with matrix function. The new criterion uses two scalar functions $f(x)$ and $g(x)$ to map the between-class scatter matrix S_b and the within-class scatter matrix S_w into the matrix functions $f(S_b)$ and $g(S_w)$ respectively. On this basis, because any function may be approximated by the polynomial, then we use two polynomials to express the function $f(x)$ and $g(x)$. Thus, a polynomial linear discriminant analysis method (PLDA) is proposed. The contributions of PLDA has: one is it avoids the SSS problem, the other is this method enlarges the distance of the between-class samples and then helpful to pattern classification, which is worth to promote in pattern classification. In addition, this method in this paper can be regarded as

a variant of the kernel method and it is an innovation in the kernel method family of PCA and LDA. The future work is to generalize the idea of this paper in the pattern classification and then propose some new methods. However, in the case of large category problems, these methods may all have the class separation issues, which would make the got subspace overemphasize the classes that have already separated. So, in future research, we should focus on this issue as a special topic.

Acknowledgements This work was supported by National Natural Science Foundation of China under grant 61876026, Chongqing Technology Innovation and Application Development Project under Grant cstc2020jscx-msxmX0190 and cstc2019jscxmbdxX0061, Science and Technology Research Program of Chongqing Municipal Education Commission under grant KJZD-K202100505.

Author contributions RR contributed to Conceptualization, methodology, writing-review and editing; TW contributed to software, validation writing and preparation; ZL contributed to formal analysis and visualization; BF contributed to supervision; RR and BF contributed to project administration and funding acquisition.

Funding This work was supported by National Natural Science Foundation of China under grant 61876026, Chongqing Technology Innovation and Application Development Project under Grant cstc2020jscx-msxmX0190 and cstc2019jscxmbdxX0061, Science and Technology Research Program of Chongqing Municipal Education Commission under Grant KJZD-K202100505.

Availability of supporting data The datasets used or analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Conflict of interest The authors declare no conflict of interest.

Ethics approval Not applicable.

References

1. Turk M (1991) Pentland. Eigenfaces for recognition. K. Cogn. Neurosc 4, 72–86
2. Fisher RA (1936) The use of multiple measurements in taxonomic problems. Ann Eugen 7(2):179–188
3. Hart PE, Stork DG, Duda RO (2000) Pattern classification. Hoboken, New York
4. 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. https://doi.org/10.1007/s13042-013-0226-9
5. Belhumeur PN, Hespanha JP (1997) Kriegman DJ Eigenfaces vs. fisherfaces: recognition using class specific linear projection. IEEE Trans Pattern Anal Mach Intell 19(7):711–720
6. Lu J, Plataniotis KN, Venetsanopoulos AN (2005) Regularization studies of linear discriminant analysis in small sample size scenarios with application to face recognition. Pattern Recogn Lett 26(2):181–191
7. Zhang T, Fang B, Tang YY, Shang Z, Xu B (2009) Generalized discriminant analysis: A matrix exponential approach. IEEE Trans Syst Man Cybern Part B (Cybern) 40(1), pp 186–197
8. Murthy KR, Ghosh A (2017) Noisy-free length discriminant analysis with cosine hyperbolic framework for dimensionality reduction. Expert Syst Appl 81:88–107
9. Chen L-F, Liao H-YM, Ko M-T, Lin J-C, Yu G-J (2000) A new LDA-based face recognition system which can solve the small sample size problem. Pattern Recogn 33(10):1713–1726

10. Yu H, Yang J (2001) A direct LDA algorithm for high-dimensional data-with application to face recognition. Pattern Recogn 34(10):2067–2070
11. Lu J, Plataniotis KN, Venetsanopoulos AN (2003) Face recognition using LDA-based algorithms. IEEE Trans Neural Netw 14(1):195–200
12. Lotlikar R, Kothari R (2000) Fractional-step dimensionality reduction. IEEE Trans Pattern Anal Mach Intell 22(6):623–627
13. Sharma A, Paliwal KK (2012) A two-stage linear discriminant analysis for face-recognition. Pattern Recogn Lett 33(9):1157–1162
14. Li H, Jiang T, Zhang K (2003) Efficient and robust feature extraction by maximum margin criterion. Adv Neural Inf Process Syst **16**
15. Liu S, Feng L, Qiao H (2014) Scatter balance: an angle-based supervised dimensionality reduction. IEEE Trans Neural Netw Learn Syst 26(2):277–289
16. Zhao H, Wang Z, Nie F (2018) A new formulation of linear discriminant analysis for robust dimensionality reduction. IEEE Trans Knowl Data Eng 31(4):629–640
17. Zhong F, Zhang J (2013) Linear discriminant analysis based on L1-norm maximization. IEEE Trans Image Process 22(8):3018–3027
18. Ye Q, Fu L, Zhang Z, Zhao H, Naiem M (2018) Lp-and ls-norm distance based robust linear discriminant analysis. Neural Netw 105:393–404
19. Ye Q, Li Z, Fu L, Zhang Z, Yang W, Yang G (2019) Nonpeaked discriminant analysis for data representation. IEEE Trans Neural Netw Learn Syst 30(12):3818–3832
20. Mika S, Ratsch G, Weston J, Scholkopf B, Mullers K-R (1999) Fisher discriminant analysis with kernels. In: Neural Networks for Signal Processing IX: Proceedings of the 1999 IEEE Signal Processing Society Workshop (cat. No. 98th8468), pp 41–48. IEEE
21. Baudat G, Anouar F (2000) Generalized discriminant analysis using a kernel approach. Neural Comput 12(10):2385–2404
22. Schölkopf B, Smola A, Müller K-R (1998) Nonlinear component analysis as a kernel eigenvalue problem. Neural Comput 10(5):1299–1319
23. Schölkopf B, Smola A, Müller K-R (1997) Kernel principal component analysis. In: International Conference on Artificial Neural Networks, pp 583–588. Springer
24. He X, Niyogi P (2003) Locality preserving projections. Adv Neural Inf Process Syst **16**
25. Zhang D, He J, Zhao Y, Luo Z, Du M (2014) Global plus local: a complete framework for feature extraction and recognition. Pattern Recogn 47(3):1433–1442
26. Shu X, Gao Y, Lu H (2012) Efficient linear discriminant analysis with locality preserving for face recognition. Pattern Recogn 45(5):1892–1898
27. Li Y, Liu B, Yu Y, Li H, Sun J, Cui J (2021) 3E-LDA: three enhancements to linear discriminant analysis. ACM Trans Knowl Discov Data (TKDD) 15(4):1–20
28. Zhao M, Zhang Z, Chow TW (2012) Trace ratio criterion based generalized discriminative learning for semi-supervised dimensionality reduction. Pattern Recogn 45(4):1482–1499
29. Zhao M, Zhang Z, Chow TW, Li B (2014) A general soft label based linear discriminant analysis for semi-supervised dimensionality reduction. Neural Netw 55:83–97
30. Zhao M, Zhang Z, Chow TW, Li B (2014) Soft label based linear discriminant analysis for image recognition and retrieval. Comput Vis Image Underst 121:86–99
31. Veeramani K, Jaganathan S (2014) Intensified regularized discriminant analysis technique. In: International Conference on Recent Advances and Innovations in Engineering (ICRAIE-2014), pp 1–6 . IEEE
32. Golub GH, Van Loan CF (2013) Matrix computations. JHU press, Maryland
33. Rudin W (1976) Principles of mathematical analysis, vol 3. McGraw-hill, New York
34. Cai D, He X, Han J (2007) SRDA: an efficient algorithm for large-scale discriminant analysis. IEEE Trans Knowl Data Eng 20(1):1–12
35. Gao J, Li L (2019) A robust geometric mean-based subspace discriminant analysis feature extraction approach for image set classification. Optik 199:163368
36. Li D-L, Prasad M, Hsu S-C, Hong C-T, Lin C-T (2012) Face recognition using nonparametric-weighted fisherfaces. EURASIP J Adv Signal Process 2012(1):1–11
37. Hu L, Zhang W (2020) Orthogonal neighborhood preserving discriminant analysis with patch embedding for face recognition. Pattern Recogn 106:107450

38. He Z, Wu M, Zhao X, Zhang S, Tan J (2021) Representative null space LDA for discriminative dimensionality reduction. Pattern Recogn 111:107664
39. Sun K, Zhang J, Yong H, Liu J (2019) Fpcanet: fisher discrimination for principal component analysis network. Knowl-Based Syst 166:108–117
40. Kewei T, Jun Z, Changsheng Z, Lijun W, Yun Z, Wei J (2021) Unsupervised, supervised and semi-supervised dimensionality reduction by low-rank regression analysis. Chin J Electron 30(4):603–610
41. Yang R, Meng X, Feng L (2020) Scatter balance based semi-supervised dimensional reduction. In: 2020 IEEE International Conference on Smart Internet of Things (SmartIoT), pp 168–175 . IEEE
42. Olivetti & Oracle Research Laboratory (1994) The Olivenfi d Oracle Research Laboratory Face Database of Faces. https://wwwcam-orl.co.uk/facedatabase.html
43. Phillips PJ, Moon H, Rizvi SA, Rauss PJ (2000) The FERET evaluation methodology for face-recognition algorithms. IEEE Trans Pattern Anal Mach Intell 22(10):1090–1104. https://doi.org/10.1109/34.879790
44. The AR Face Database, CVC Technical Report, 24 (1998). https://portalrecerca.uab.cat/en/publications/the-ar-face-database-cvc-technical-report-24
45. Columbia object image library (Coil-100)), Technical report CUCS-006-96 (1996). https://www.cs.columbia.edu/CAVE/software/softlib/coil-100.php