Applied Intelligence (2024) 54:9508–9523
https://doi.org/10.1007/s10489-024-05632-6

Generalized robust linear discriminant analysis for jointly sparse learning

Yufei Zhu[1] · Zhihui Lai[1] · Can Gao[1] · Heng Kong[2]

Accepted: 19 June 2024 / Published online: 17 July 2024

Abstract
Linear discriminant analysis (LDA) is a well-known supervised method that can perform dimensionality reduction and feature extraction effectively. However, traditional LDA-based methods need to be turned into the trace ratio form to compute the closed-form solution, in which the within-class scatter matrix should be nonsingular. In this article, we design a new model named generalized robust linear discriminant analysis (GRLDA) method to tackle this disadvantage and improve the robustness. GRLDA uses $L_{2,1}$-norm on both loss functions to reduce the influence of outliers and on regularization term to obtain joint sparsity simultaneously. The intrinsic graph and the penalty graph are constructed to characterize the intraclass similarity and interclass separability, respectively. A novel optimization method is proposed to solve the proposed model, in which a quadratic problem on the Stiefel manifold is involved to avoid the inverse computation on a singular matrix. We also analyze the computational complexity rigorously. Finally, the experimental results on face, object, and medical images exhibit the superiority of GRLDA.

Keywords Generalized Robust Linear Discriminant Analysis (GRLDA) · Feature extraction · Convex optimization problem

1 Introduction

With the development of pattern recognition, dimensionality reduction (DR) algorithms have become more and more accessible [1]. Generally speaking, DR algorithms can reduce model's computational complexity and run time, alleviating the impact of noisy information [2, 3] and redundant features [4, 5]. As one of the hottest topics in pattern recognition recently, researchers have paid greater attention to DR algorithms. DR algorithms can be used in many practical applications, such as human gait recognition and object identification [6], which aim to find an optimal projection matrix to maintain the most important features [7, 8], and [9].

DR algorithms are classified as linear DR algorithms [10] and non-linear DR algorithms [11] based on whether the mapping functions are linear or nonlinear. The most typical linear DR algorithms include principal component analysis (PCA) and locality preserving projection (LPP). The core idea of PCA is to obtain a set of orthogonal bases and the variance among the reduced data needs to be maximized [12] and the reconstruction error needs to be minimized [13]. In LPP [14], we first assign larger weights to the data points at closer distances. The projection matrix is obtained by minimizing the sum of products between the distances among point pairs and their corresponding weights. Non-linear DR algorithms include locally linear embedding (LLE) [15], isomap [16], and laplacian eigenmaps (LE) [17]. LLE tries to keep the linear relationship between samples in the neighborhood, while the purpose of Isomap is to keep the distance between near-neighbor samples different. And LE is aimed to construct relationships between data from a local approximation perspective.

LDA [13] is a well-known method in the supervised DR field. However, traditional LDA-based algorithms have many disadvantages. For example, on the one hand, LDA is required to be changed to trace ratio form and then utilizes the generalized eigenvalue decomposition (GEVD) method to obtain the closed-form solution [18]. This may result in errors between the optimal solution and the estimated solution since the trace ratio problem is not a convex optimization problem [19]. To solve the trace ratio problem, trace ratio linear discriminant analysis (TRLDA) [20] and

✉ Heng Kong
generaldoc@126.com

Zhihui Lai
lai_zhi_hui@163.com

[1] College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China

[2] Department of Breast and Thyroid Surgery, BaoAn Central Hospital of Shenzhen, Shenzhen 518102, China

ratio sum linear discriminant analysis (RSLDA) [21] were proposed, which effectively overcome this problem. The TRLDA method is a novel formulation of LDA, which can be turned into a quadratic problem with regard to the Stiefel manifold [20], but RSLDA is dedicated to maximizing the ratio of the between-class scatter to the within-class scatter in each dimension [21], so it could avoid selecting features with small variance. On the other hand, LDA aims to preserve the global Euclidean structure and does not preserve the local geometric features of the original high-dimensional data. However, locality is considered to be a more important characteristic than global structure and affects the performance of databases in the real-world applications [22]. As such, a large body of research has been done on various LDA algorithms that concentrate on locality relationships. Marginal Fisher Analysis (MFA) [23], which seeks to learn a more discriminating projection given neighbor information, is one of the most often used algorithms.

However, the above algorithms share a common drawback, i.e., Frobenius norm is used as the basic metric might enlarge the effect of outliers in certain senses. Therefore, some scholars proposed alternative approaches using L_1 -norm to replace the Frobenius norm. Some classical algorithms, such as R1-PCA [24] and LDA-R1 [25], were proposed to improve the robustness. Pang et al. proposed L_1-norm tensor analysis [26], which can enhance the robustness of the model for tensor feature extraction. Recently, sparse learning has become more and more popular [23–25]. One of the most well-known methods, Sparse Principal Component Analysis [27], transforms PCA into a regression-type optimization problem and exerts a quadratic and a lasso regularization terms. Lai et al. proposed a novel sparse method called sparse bilinear discriminant analysis (SBDA [6]) to obtain the sparse subspace for gait recognition. However, Nie et al. point out that L_1-norm enlarges the gaps among data points and degrades the subsequent classification performance [28]. He pointed out that L -norm integrates the advantage for distance measurement of L_2 -norm and sparsity for enhancing the robustness of L_1 -norm [28].

Although the aforementioned algorithms, such as new formulation of TRLDA and RSLDA, are able to release trace ratio problem, they are still sensitive to outliers. Despite the fact that the effect of outliers can be suppressed by R1-PCA and LDA-R1, they cannot guarantee joint sparsity. Nie et al. [29] first pointed out that the mean value calculated by L_2 -norm is not optimal, and they proposed a novel robust PCA with an optimal re-weighted mean. Then, Zhao et al. [30] proposed a robust LDA measured by $L_{2,1}$ -norm which can alleviate the influence of outliers. Even though extensive algorithms based on $L_{2,1}$ -norm are proposed on different occasions, they do not use the local information of the original data effectively. Lai et al. [31] proposed a rotational invariant framework using $L_{2,1}$ -norm as the basic metric, including rotational invariant LDA (RILDA) and rotational invariant MFA (RIMFA) which use $L_{2,1}$ -norm as the measurement on scatter matrices to reduce the impact of outliers. Then, in [32], locally joint sparse marginal embedding (LJSME) was proposed by Mo et al. which can break through the small sample size (SSS) problem and enhance the ability to preserve the locality relationships with the joint sparsity simultaneously. Lin et al. [33] proposed the generalized robust multiview discriminant analysis (GRMDA) method for addressing the SSS problem of LDA in multiple-view scenarios. Drawing inspiration from the robust discriminative ability of LDA and the benefits of feature extraction with $L_{2,p}$ -norm regularization, Li et al. [34] proposed STR-LDA for classification tasks. Singular value decomposition served as inspiration for the development of JSOLDA [35], a brand-new subspace learning technique that addresses the challenge of obtaining orthogonal sparse solutions in OLDA. However, the last three methods are essentially the trace ratio optimization problem [19], and finally, eigenvalue decomposition is used to obtain the optimal solution. Therefore, to improve the performance of LDA-based methods for classification, a more robust and effective method is essential.

In this article, we propose a new LDA algorithm called generalized robust linear discriminant analysis (GRLDA) for feature extraction. This method is capable of releasing the SSS problem in the LDA-based methods and meanwhile ensuring locality preservation in a robust and effective form to obtain discriminant projections. Moreover, it can also guarantee the joint sparsity of the projection matrix. The main contributions or novelty of GRLDA are highlighted as follows.

1) Unlike the trace ratio problem, a new robust LDA in the form of trace and square root of trace is proposed, which can be converted into a convex optimization problem so as to obtain the local optimal solution. We prove the equivalence of our proposed method and the trace ratio method. Furthermore, it is possible to circumvent the drawback that the trace ratio problem can only have an approximate solution.
2) Several robust factors are integrated into the proposed GRLDA. Firstly, we construct two weighted graphs, the intrinsic graph and the penalty graph. But different from MFA, two scatter matrices are measured by $L_{2,1}$ -norm so as to preserve the locality relationship with higher reconstruction ability and enhance the robustness to outliers. Meanwhile, we impose the $L_{2,1}$-norm-based regularization term to ensure that the learned projections are jointly sparse and thus improve the performance of feature selection.
3) To calculate the best answer to the corresponding optimization problem, we design an iterative method. Additionally, the computational complexity is examined.

Numerous tests have shown that the suggested GRLDA can outperform most algorithms and that the designed method has a rapid convergence speed.

The remainder of this paper is briefly outlined as follows. In Section 2, we first present some notations and then we will discuss our motivation and the objective function, meanwhile its optimal solution is also shown. The computational complexity is also presented. In Section 3, a series of experiments have been conducted. Finally, we draw a conclusion for this article in Section 6.

2 The proposed method (GRLDA)

In this section, some notations and definitions will be given and then the motivation and the objective function of the model are presented. We will also show how to compute the optimal solution using an iterative algorithm.

2.1 Notations and definitions

Let $||A||_P$ and $tr(A)$ represent L_P -norm and the trace of the matrix A, respectively. Given a sample matrix $X = [x_1, x_2, ..., x_N] \in R^{d \times N}$ including all the training samples $\{x_i\}_{i=1}^{N} \in R^d$ in its columns, where N is the number of samples and d is the original dimension. Let $U \in R^{d \times m}$ denote the projection matrix, where m is the subspace dimension.

Given any matrix $Q = [q_{ij}] \in R^{m \times n}$, the Frobenius norm of the matrix Q is denoted as:

$$||Q||_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} q_{ij}^2} = \sqrt{\sum_{i=1}^{m} ||Q^i||_2^2} \tag{1}$$

the $L_{2,1}$ -norm of the matrix Q is denoted as:

$$||Q||_{2,1} = \sum_{i=1}^{m} \sqrt{\sum_{j=1}^{n} q_{ij}^2} = \sum_{i=1}^{m} ||Q^i||_2 \tag{2}$$

Many theoretical analyses and experiments have shown that imposing $L_{2,1}$ -norm as the basic metric on the objective function can ensure joint sparsity. And for any given rotational matrix A,$||QA||_{2,1} = ||Q||_{2,1}$. In [36], Nie et al. pointed out that this characteristic is rotational invariant.

2.2 Motivation and objective function

Numerous techniques have been developed to mitigate the adverse impact of traditional LDA on outliers. However, there are still many drawbacks. Firstly, they cannot preserve the local structure of data [37]. However, this characteristic makes differences in reconstructing the locality relationship in a low-dimensional space. Secondly, those extensions of LDA that keep the locality relationships are required to be transformed into another form and then adopt GEVD so that the within-class scatter matrix is required to be nonsingular. Therefore, we cannot obtain the closed-form solution if the sample size is very small. Moreover, the GEVD of between-class scatter matrix and within-class scatter matrix can only approximate the true value. Last but not least, despite the fact that some algorithms take the SSS and trace ratio problems into account, they do not consider joint sparsity [38]. Therefore, in this paper, we propose a generalized robust linear discriminant analysis (GRLDA) for feature extraction and dimensionality reduction. This approach addresses the joint sparsity utilizing $L_{2,1}$ -norm as the basic metric on the regularization term as well as the primary component to increase the robustness. It also inherits the property of RIMFA in addition to taking into account the benefits of TRLDA.

The local within-class and local between-class scatter using $L_{2,1}$ -norm as the basic metric can be computed as follows:

$$\sum_{i=1}^{N}\sum_{j=1}^{N} ||U(x_i - x_j)||_2 W_{ij}^w = tr(U^T X_{Gw}^T D_{Gw} X_{Gw} U) \tag{3}$$

$$\sum_{i=1}^{N}\sum_{j=1}^{N} ||U(x_i - x_j)||_2 W_{ij}^b = tr(U^T X_{Gb}^T D_{Gb} X_{Gb} U) \tag{4}$$

where W_{ij}^w and W_{ij}^b are defined the same as MFA [23] and the new sample matrix X_{Gw} and X_{Gb} are defined as:

$$\begin{aligned} X_{Gw} = [&W_{11}^w(x_1 - x_1), ..., W_{1N}^w(x_1 - x_N), ..., \\ &W_{N1}^w(x_N - x_1), ..., W_{NN}^w(x_N - x_N)]^T \end{aligned} \tag{5}$$

$$\begin{aligned} X_{Gb} = [&W_{11}^b(x_1 - x_1), ..., W_{1N}^b(x_1 - x_N), ..., \\ &W_{N1}^b(x_N - x_1), ..., W_{NN}^b(x_N - x_N)]^T \end{aligned} \tag{6}$$

and the two diagonal matrices are defined as:

$$\begin{aligned} D_{Gw} = \mathrm{diag}(&\frac{1}{2||W_{11}^w(x_1 - x_1)^T U||_2}, ..., \frac{1}{2||W_{1N}^w(x_1 - x_N)^T U||_2} \\ &, ..., \frac{1}{2||W_{N1}^w(x_N - x_1)^T U||_2}, ..., \frac{1}{2||W_{NN}^w(x_N - x_N)^T U||_2}) \end{aligned} \tag{7}$$

$$\begin{aligned} D_{Gb} = \mathrm{diag}(&\frac{1}{2||W_{11}^b(x_1 - x_1)^T U||_2}, ..., \frac{1}{2||W_{1N}^b(x_1 - x_N)^T U||_2} \\ &, ..., \frac{1}{2||W_{N1}^b(x_N - x_1)^T U||_2}, ..., \frac{1}{2||W_{NN}^b(x_N - x_N)^T U||_2}) \end{aligned} \tag{8}$$

We denote $S_{Gw} = X_{Gw}^T D_{Gw} X_{Gw}$ and $S_{Gw} = X_{Gb}^T D_{Gb} X_{Gb}$, and according to the definition of $L_{2,1}$ -norm, a diagonal matrix D with the i-th diagonal element is defined as:

$$D_{ii} = \frac{1}{2||u^i||_2} \tag{9}$$

where u^i denotes the i-th row of the matrix U.

RIMFA aims to learn the projection matrix U by maximizing the interclass separability and minimizing the intraclass compactness simultaneously. The objective function of RIMFA is as follows:

$$\begin{aligned} &\max_U \frac{tr(U^T S_{Gb} U)}{tr(U^T S_{Gw} U)} \\ &\text{s.t.} U^T U = I \end{aligned} \tag{10}$$

The optimal U of RIMFA could be obtained by the eigenvalue decomposition of the matrix $S_{Gw}^{-1} S_{Gb}$. However, when the number of training samples is very small, the inversion of within-class scatter matrix does not exist. The most commonly used method is to add a regularization term to the within-class scatter matrix. Therefore, the eigenvalue decomposition of S_{Gw} and S_{Gb} can only approximate the true value. Since the optimal solution of RIMFA cannot be obtained directly, it is natural to construct a novel function with respect to U and a learnable variable s, which is equivalent to the problem (10):

$$\begin{aligned} &\min_{s,U} s^2 tr(U^T S_{Gw} U) - 2s\sqrt{tr(U^T S_{Gb} U)} \\ &\qquad \text{s.t.} U^T U = I \end{aligned} \tag{11}$$

We can know that the Eq. (11) is equivalent to the Eq. (10) from the following theorem.

Theorem 1 *The optimization problem in* (11) *is equivalent to the trace ratio problem* (10).

Proof The optimal solution of s can be obtained by taking the derivative of (11) with respect to s and setting it to be 0, we get.

$$\begin{aligned} &2s tr(U^T S_{Gw} U) - 2\sqrt{tr(U^T S_{Gb} U)} = 0 \\ &\qquad \Rightarrow s = \frac{\sqrt{tr(U^T S_{Gb} U)}}{tr(U^T S_{Gw} U)} \end{aligned} \tag{12}$$

Substitute the optimal s into the problem (11), we get

$$\begin{aligned} &\min_U s^2 tr(U^T S_{Gw} U) - 2s\sqrt{tr(U^T S_{Gb} U)} \\ &= \min_U \frac{tr(U^T S_{Gb} U)}{tr(U^T S_{Gw} U)} - 2\frac{tr(U^T S_{Gb} U)}{tr(U^T S_{Gw} U)} \\ &= \min_U - \frac{tr(U^T S_{Gb} U)}{tr(U^T S_{Gw} U)} \\ &\Rightarrow \max_U \frac{tr(U^T S_{Gb} U)}{tr(U^T S_{Gw} U)} \end{aligned} \tag{13}$$

which completes the proof. □

With the above preparations and ensuring the joint sparsity of the projection matrix, the objective function of GRLDA is finally defined as follows:

$$\begin{aligned} &\min_{s,U} s^2 tr(U^T S_{Gw} U) - 2s\sqrt{tr(U^T S_{Gb} U)} + \alpha||U||_{2,1} \\ &\qquad \text{s.t.} U^T U = I \end{aligned} \tag{14}$$

2.3 The optimization

In problem (14), there are 2 iterative variables, i.e., projection matrix U, balance parameter s. In this paper, we firstly fix U to iterate s, then fix s to compute U.

1) Fix U to update s

We can easily get the updated s by setting the derivative with respect to (*w.r.t.*) s to 0, then we have:

$$s = \frac{\sqrt{tr(U^T S_{Gb} U)}}{tr(U^T S_{Gw} U)} \tag{15}$$

2) Fix s to update U

For updating the matrix U, it is worthwhile for us to introduce a theorem proposed by Nie [29], which is a solution to the maximization problem as follows:

$$\max_{m \in \Omega} \sum\nolimits_i f_i(h_i(m)) \tag{16}$$

where $f_i(h_i(m)) \leq 0$ is required to be an arbitrary convex function $w.r.t. h_i(m)$ under the arbitrary constraint of $m \in \Omega$. Before introducing the theorem, we firstly give the following lemma.

Lemma1 [39] *Assume m is a matrix, vector or scalar, f(m) is a scalar output function while h(m) is arbitrary whether it is a matrix, vector or scalar. Then we can get the following equality*:

$$\begin{aligned} \frac{\partial f(h(m))}{\partial m} &= \sum\nolimits_i \sum\nolimits_j \frac{\partial h_{ij}(m)}{\partial m} \frac{\partial f(h(m))}{\partial h_{ij}(m)} \\ &= \frac{tr(\partial h(m)) \cdot \left(f\prime(h(m))^T\right)}{\partial m} \end{aligned} \tag{17}$$

which is also called Chain-Rule.

The Lagrange function of the problem (16) *is given as follows*:

$$J_1(m, \mu) = \sum\nolimits_i f_i(h_i(m)) - \Gamma(m, \mu) \tag{18}$$

Take the derivative of the Eq. (18) *and use the lemma* 1, *we have the following derivation*:

$$\begin{aligned} &\frac{\partial J_1(m,\mu)}{\partial m} = \sum\nolimits_i \frac{\partial f_i(h_i(m))}{\partial m} - \frac{\partial \Gamma(m,\mu)}{\partial m} \\ &= \sum\nolimits_i \frac{tr(\partial h_i(m) G_i^T)}{\partial m} - \frac{\partial \Gamma(m,\mu)}{\partial m} \\ &= \frac{\partial(\sum_i tr(G_i^T h_i(m)) - \Gamma(m,\mu))}{\partial m} \end{aligned} \tag{19}$$

For simplicity, we suppose:

$$J_2(m,\mu)=\sum\nolimits_i tr(G_i^T h_i(m))-\Gamma(m,\mu) \tag{20}$$

Namely, (20) *is the Lagrange function of the following optimization problem* [20]:

$$\max_{m\in\Omega}\sum\nolimits_i tr(G_i^T h_i(m)) \tag{21}$$

With the aforementioned preparations, we present the following theorem to compute the problem (14).

Theorem 2: *The solution to the optimization problem* (16) *can be obtained by the following iterative procedure* [20].

1) Initialize m that satisfies $m\in\Omega$;
2) Compute $G_i=f'_i(h_i(m))$, where $f'_i(h_i(m))$ is an arbitrary super gradient of the function $f_i(m)$ at the point $h_i(m)$;
3) Obtain the optimal $m*$ of the problem (21);
4) Update $m\leftarrow m*$.
5) Repeat 2) – 4) until convergence.

The proof of the theorem 2 is in the appendix. According to the theorem 2, we are required to convert the problem (14) into a convex function *w.r.t.U* as follows:

$$\begin{aligned}\max_U tr(U^T\widetilde{S}_{Gw}U)+2s\sqrt{tr(U^TS_{Gb}U)}+tr(U^T\widetilde{D}U)\\ \text{s.t.}U^TU=I\end{aligned} \tag{22}$$

Where $a,\widetilde{D}=\alpha(\gamma I-D)$, β and γ need to be large enough to make $\widetilde{S}_{Gw}$ and $\widetilde{D}$ be positive semi-definite (PSD). From (22), we can obviously know the first and third term are convex terms. It is necessary for us to prove the second term is a convex term but firstly we need to introduce Lemma 2.

Lemma 2 [21] *If $f(s)$ is a linear output function, and $F(s)$ is a convex function, then $F(f(s))$ is still convex.*

Proof If $f(s)$ is a linear output function, and $F(s)$ is a convex function, then according to the definition of convex function, we have:

$$\begin{aligned}F(f(\lambda_1s_1+\lambda_2s_2))&=F(\lambda_1f(s_1)+\lambda_2f(s_2))\\&\le\lambda_1F(f(s_1))+\lambda_2F(f(s_2))\end{aligned} \tag{23}$$

which proves that $F(f(s))$ is still a convex function. □

We can easily know that S_{Gb} is a PSD matrix so a corresponding matrix $P\in R^{n\times d}$ can be found which satisfies $S_{Gb}=P^TP$, then we get:

$$\begin{aligned}&2s\sqrt{tr(U^TS_{Gb}U)}\\&=2s\sqrt{tr(U^TP^TPU)}\\&=2s\sqrt{tr((PU)^TPU)}\\&=2s\sqrt{tr(R^TR)}\\&=2s||R||_F\end{aligned} \tag{24}$$

where $R=PU\in R^{n\times m}$. It is easy to know that $||\cdot||_F^2$ is a convex function, and in each iteration, P is a constant matrix, so $R(U)$ is a linear function *w.r.t. U*. According to the lemma 2, we can conclude that the second term of problem (22) is convex.

For simplicity, we denote $\widetilde{S}_{Gb}=sS_{Gb}U/\sqrt{tr(U^TS_{Gb}U)}$, so the objective function can be rewritten as:

$$\begin{aligned}&\max_{U^TU=I} tr(U^T\widetilde{S}_{Gw}U)+2tr(U^T\widetilde{S}_{Gb})+tr(U^T\widetilde{D}U)\\&=\max_{U^TU=I}\sum\nolimits_{i=1}^m u_i^T\widetilde{S}_{Gw}u_i+2\sum\nolimits_{i=1}^m u_i^T(\widetilde{S}_{Gb})_i+\sum\nolimits_{i=1}^m u_i^T\widetilde{D}u_i\\&=\max_{U^TU=I}\sum\nolimits_{i=1}^m(u_i^T\widetilde{S}_{Gw}u_i+2u_i^T(\widetilde{S}_{Gb})_i+u_i^T\widetilde{D}u_i)\end{aligned} \tag{25}$$

where $f_i(u_i)=u_i^T\widetilde{S}_{Gw}u_i+2u_i^T(\widetilde{S}_{Gb})_i+u_i^T\widetilde{D}u_i$, $h_i(u_i)=u_i$, u_i denotes the i−th column of the matrix U. It is easy to know that $f_i(u_i)$ is a convex function. According to the theorem 1, the convex optimization problem (25) can be rewritten as:

$$\begin{aligned}&{}_{U^TU=I}{}^{max}\sum\nolimits_{i=1}^m\left(f_i\prime\left(h_i\left(u_i\right)\right)\right)^Th_i\left(u_i\right)\\&={}_{U^TU=I}{}^{max}\sum\nolimits_{i=1}^m\left(2\widetilde{S}_{GW}u_i+2\left(\widetilde{S}_{Gb}\right)_i+2\widetilde{D}_{u_i}\right)^Tu_i\\&={}_{U^TU=I}{}^{max}\sum\nolimits_{i=1}^m c_i^Tu_i\\&={}_{U^TU=I}{}^{max}tr\left(U^TC\right)\end{aligned} \tag{26}$$

where a and $C=\left[c_1,c_2,\ldots,c_m\right]$. The optimal $U*$ can be gained by the following method [21].

Firstly, we adopt the Singular Value Decomposition on C, and we get $\widetilde{U}\in R^{d\times d}$ and $V\in R^{m\times m}$ which satisfy $C=\widetilde{U}\Sigma V^T$. Then we will have:

$$\begin{aligned}&tr(U^TC)=tr(U^T\widetilde{U}\Sigma V^T)=tr(\Sigma V^TU^T\widetilde{U})\\&=tr(\Sigma Z)=\sum\nolimits_{i=1}^m\sigma_{ii}z_{ii}\end{aligned} \tag{27}$$

where $Z=V^TU^T\widetilde{U}\in R^{m\times d}$. We can easily know that $ZZ^T=I_m\in R^{m\times m}$,$\sigma_{ii}$ and z_{ii} are diagonal elements of Σ and Z, so $z_{ii},(i=1,2,3,\ldots,m)$ is no more than 1. That means that the function $tr(\Sigma Z)$ can be maximized when all of the diagonal elements of Z are equal to 1. According to $Z=V^TU^T\widetilde{U}\in R^{m\times d}$, the optimal U can be obtained by $U^*=\widetilde{U}Z^TV^T=\widetilde{U}[I_m;0_{(d-m)\times m}]V^T$. The optimal U^* of problem (22) can be obtained via Algorithm 1. The whole algorithm flow is shown in Table 1

Table 1 GRLDA algorithm

```
Input: sample matrix $X \in R^{d\times N}$, the number of iterations $T_1$, parameter $\alpha$, the subspace dimension $m$.
Output: projection matrix $U$.
Step 1: Initialize $U$ as column orthogonal matrix with size $d\times m$ and $t=1$.
Step 2: Construct similarity matrix $X_{Gw}$, $X_{Gb}$ and $D$
Step 3: while not converge or $t \le T_1$ do
        - Construct matrix $S_{Gw}$, $S_{Gb}$;
        - Update $s_t = \sqrt{tr(U_{t-1}^T S_{Gb}^{t-1} U_{t-1})} \Big/ \sqrt{tr(U_{t-1}^T S_{Gw}^{t-1} U_{t-1})}$;
        - Compute $\tilde{D}_t = \alpha(\gamma I_d - D_{t-1})$, $\tilde{S}_{Gw}^t = s_t^2(\beta I_d - S_{Gw}^{t-1})$, $\tilde{S}_{Gb}^t = s_t S_{Gb}^{t-1} U_{t-1} \Big/ \sqrt{tr(U_{t-1}^T S_{Gb}^{t-1} U_{t-1})}$;
        - Update $U$ via Algorithm 1;
        - Update matrix $D_{Gw}$, $D_{Gb}$, $D$ via (7)-(9);
        - $t = t+1$;
    end while
Step 4: Output the projection matrix $U$ for further feature analysis
```

Algorithm 1 For solving the problem (22)

```
Input: PSD matrices $\tilde{S}_{Gb}$, $\tilde{S}_{Gw}$, $\tilde{D}$ and the number of iterations $T_2$.
Output: The projection matrix $U$.
Initialize $U$ with a column-orthogonal matrix, and $t=1$;
While not converge or $t \le T_2$ do
Compute matrix $C$ via (26);
Adopt the full SVD on $C$: $C = \tilde{U}\Sigma V^T$;
Update matrix $U = \tilde{U}[I_m; 0_{(d-m)\times m}]V^T$;
$t = t+1$;
End while
```

.2.4 Computational complexity

From the proposed GRLDA algorithm, we can know that the outer loop is to update s, D_{Gw}, D_{Gb}, and the inner loop is to update the projection matrix U. The inner loop consists of three parts: computing the matrix C, adopting the full SVD of C, updating the projection matrix U. The computational complexity of both computing the matrix $C \in R^{d\times m}$ and performing the full SVD on the matrix C is $O(md^2)$. The outer loop consists of computing s,$\widetilde{S}_{Gw}$, $\widetilde{S}_{Gw}$ and updating D_{Gw},D_{Gb},D, whose complexities are $O(md^2)$ at most. Therefore, the total computational complexity of the proposed algorithm is $O(T_1 T_2 md^2)$, where T_1 and T_2 are the number of outer loop and inner loop iterations. Extensive experiments demonstrate that T_1 is no more than 3 on lots of databases.

3 Experiments

In this section, we compare the proposed GRLDA with some state-of-the-art algorithms including classical dimensionality reduction algorithms and some recent algorithms to illustrate the performance on some well-known databases. The classical dimensionality reduction methods include principle component analysis (PCA) [40], linear discriminant analysis (LDA) [10] and marginal fisher analysis (MFA) [23]. The recent methods include $L_{2,1}$-norm-based algorithms (i.e. rotational invariant LDA and MFA (RILDA, RIMFA) [31], the $L_{2,1}$-norm regularized locally joint sparse marginal embedding (LJSME) [32]), trace and square root of trace [18] based algorithms (i.e. trace ratio LDA (TRLDA) [20], ratio sum LDA (RSLDA) [21], sparse trace ratio LDA (STR-LDA) [34]) and jointly sparse orthogonal LDA (JSOLDA) [35]. The original data must be pre-processed because the dimensionality of each image is extremely high and there are very few training samples [41]. As a result, the scatter matrices may contain null space. In order to mitigate the effects of null space and preserve the primary energy, we employ PCA to minimize the dimensionality. Every image had its dimensionality lowered to 200. Next, the primary feature on the COIL-20, FERET, ORL, Extended Yale B, AR, Yale, BreastMNIST, and PneumoniaMNIST databases is extracted using GRLDA. Lastly, additional categorization is performed using the closest neighbor classifier (NN).

3.1 Explorations on Parameter Setting

In the experiment of GRLDA, we first try to find the optimal range of the parameter value from which we can obtain the best performance. We explore the recognition rate with the variable α variously from $[10^{-9}, 10^{-8}, ..., 10^{8}, 10^{9}]$ on 5 databases. The recognition rates versus the variable α are illustrated in Fig. 1a. From the graph shown in Fig. 1a, it can be found that the best parameter area is $[10^{2}, ..., 10^{9}]$. The potential reason lies in the fact that when the parameter α is larger, the projection obtained through the proposed GRLDA exhibits a greater number of rows with all zeros, which indicates the learned projection matrix is more

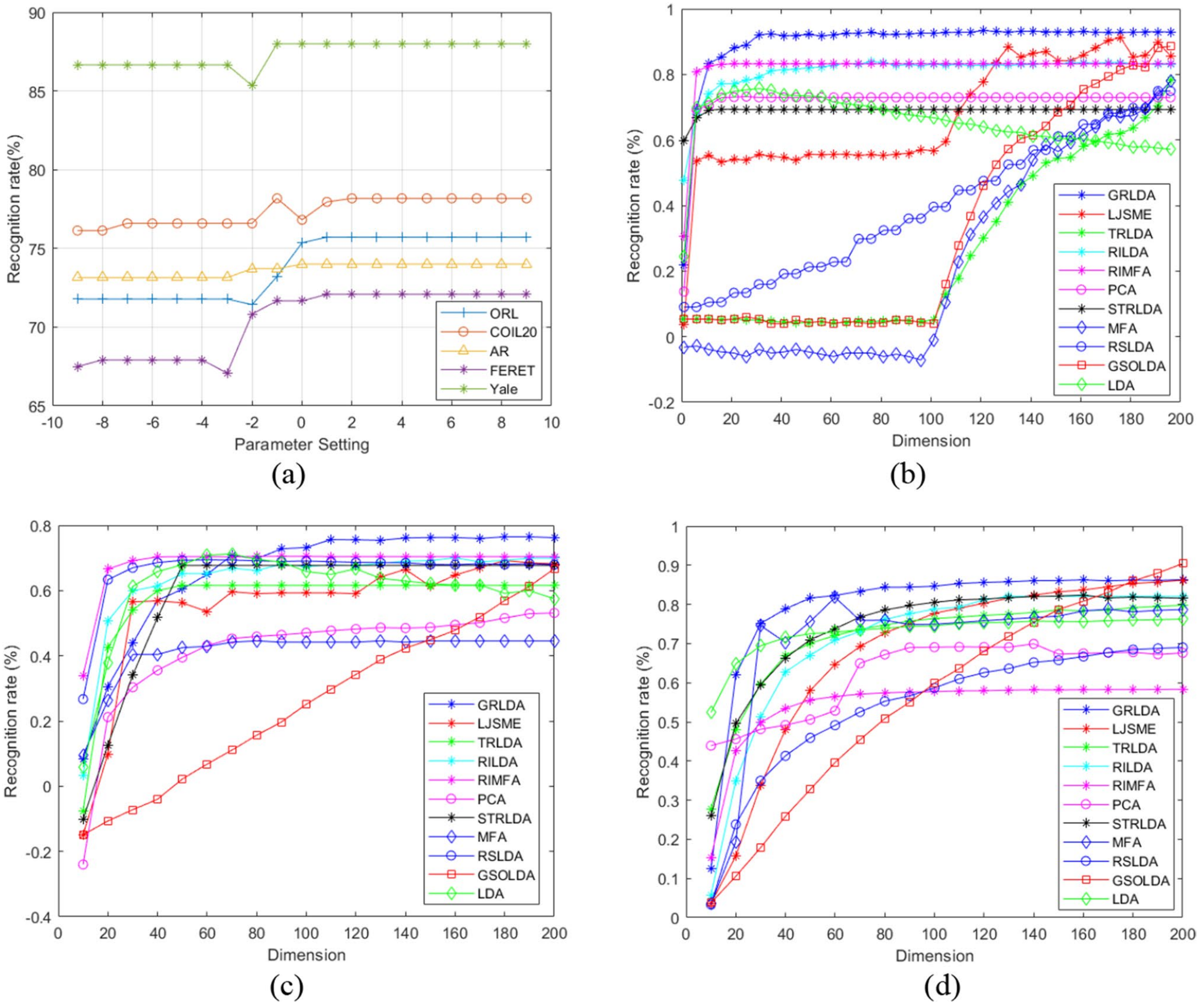


Fig. 1 **a** The recognition rates versus the variation of parameter α of GRLDA on different databases. The recognition rates versus the subspace dimension of different algorithms on **b** COIL-20 dataset, on **c** FERET dataset, on **d** ORL dataset

jointly sparse, ultimately enabling the proposed GRLDA to achieve adaptive feature selection. The recognition rates do not decline significantly like other methods (i.e. LJSME, JSOLDA) when the parameter α falls within other ranges, so we can conclude that our model is very stable even though the range of parameter α is very small. For simplicity, we set $\alpha \in [10^2, 10^9]$ in experiments for each database.

3.2 Experiments on COIL-20 Database

The COIL-20 object image database contains a total of 1440 images of 20 different objects, where the image size is 128 × 128. The 10 images of the first object with the different pose are shown in Fig. 2a. We randomly select the $t(t=5, 10, 15, 20)$ images as gallery set [32], and the rest of images are used as probe set [32]. We conduct the experiment to test the performance of GRLDA under the circumstances where images are rotated with 360 degrees. The experiment is conducted a total of 10 times.

The average recognition rates of the feature extraction algorithms (i.e., the proposed GRLDA and PCA, LDA, MFA, RILDA, RIMFA, LJSME, TRLDA, RSLDA, STRLDA, GSOLDA) when selecting 200 projections are shown in Table 2 and the recognition rates versus subspace dimension using above algorithms are illustrated in Fig. 1b when the training samples are 10. Table 2 illustrates that GRLDA outperforms other algorithms when the number of training samples are 10, 15 and 20. Even though when the number of training samples are 5, the recognition of RSLDA is slightly more than GRLDA but GRLDA also performs very well and obtains the second place. Figure 1b indicates that the proposed GRLDA

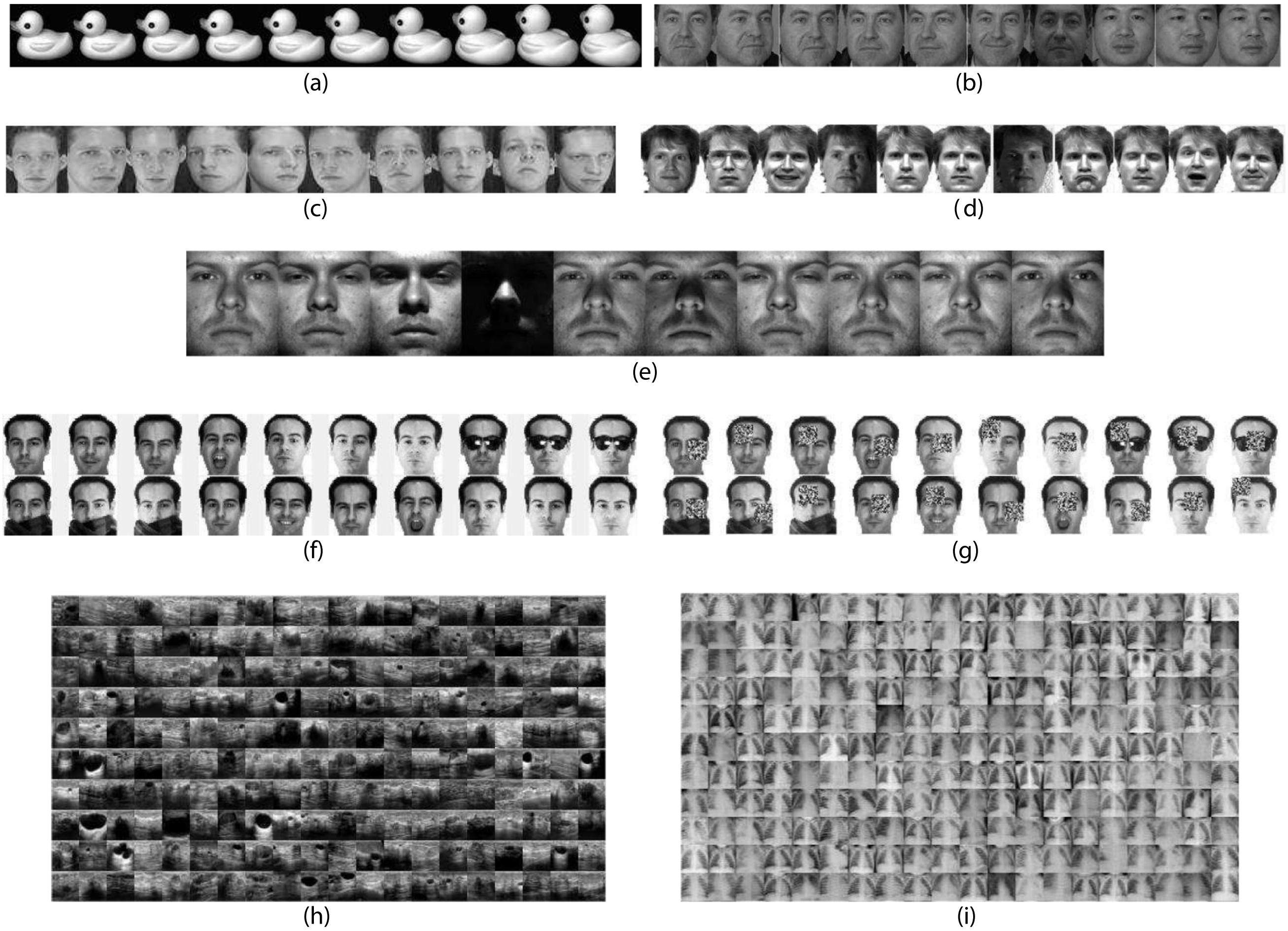


Fig. 2 The samples of COIL-20 images in (**a**), FERET images in (**b**), ORL images in (**c**), Yale images in (**d**), Extended Yale B images in (**e**), AR original images in (**f**), AR images with block size 10*10 in (**g**), BreastMNIST images in (**h**), PneumoniaMNIST images in (**i**)

performs better than other methods at a low dimension and maintains a relatively stable value. Despite the fact that the recognition rate of GRLDA decreases slightly as the subspace dimension rises, it still performs better than other methods.

3.3 A. Experiments on FERET Database

The FERET dataset contains 200 people and every person has 7 images. It is a well-known dataset to test the robustness of many classical methods. Based on the eye area, the original image of each face is automatically cut [42]. The

Table 2 The average recognition accuracy (%), standard deviation, training samples of different methods on COIL-20 face database

Training samples	PCA	LDA	MFA	RILDA	LJSME	TRLDA	RSLDA	RIMFA	STR-LDA	GSOLDA	GRLDA
5	80.17 ±1.95	61.60 ±3.57	74.06 ±1.82	70.67 ±2.55	81.94 ±1.45	82.47 ±1.31	**82.86** **±1.55**	80.07 ±1.60	76.70 ±1.65	81.85 ±1.68	82.70 ±1.25
10	83.84 ±1.42	64.81 ±2.34	79.93 ±3.40	73.49 ±1.81	85.85 ±2.22	84.64 ±2.02	86.00 ±3.70	84.09 ±1.29	82.86 ±2.04	86.45 ±1.52	**86.73** **±1.12**
15	85.32 ±1.54	68.37 ±2.30	82.14 ±2.00	74.27 ±3.97	87.54 ±1.27	86.98 ±0.70	88.64 ±1.67	86.46 ±1.34	84.31 ±1.37	88.67 ±0.26	**89.62** **±1.43**
20	85.64 ±2.31	70.60 ±3.06	84.08 ±0.89	75.23 ±2.69	82.20 ±1.86	88.65 ±1.38	89.39 ±1.64	88.58 ±1.54	87.34 ±1.55	89.96 ±1.39	**92.64** **±1.20**

cropped image size is 40 × 40. The sample images of the first person are illustrated in Fig. 2b.

In this experiment, we randomly use $L(L = 3, 4, 5)$ images of each individual as gallery set [32], and we use the remained images as probe set [32]. Table 3 illustrates the recognition rates of different algorithms. Table 3 indicates that the proposed GRLDA fully displays its robustness to outliers against some classical methods such as LDA, MFA etc. Despite the fact that GRLDA is slightly less than RILDA when the training samples are 5, it becomes evident that GRLDA's superiority increases as the number of training samples decreases. We select 6 training samples for example, Fig. 1c depicts the trends of the recognition rates versus the subspace dimension. The figure indicates that GRLDA and STR-LDA are equally effective when the training samples are small and they can both obtain a high recognition rate. Nevertheless, as the dimension rises, GRLDA performs better than STR-LDA and they reach the corresponding peak when the subspace dimension is at 110 and 70, respectively.

3.4 Experiments on ORL Database

The ORL dataset contains 400 images of 40 individuals, where the image size is 56×46. In the experiment, we randomly select $l(l=3, 4, 5)$ images as gallery set [32], and the rest of images are used as probe set [32]. Meanwhile, according to Fig. 1a, we set the parameter α to be10^3. Table 4 illustrates the average recognition rates of different algorithms. Figure 1d shows that the recognition rates versus the subspace dimension when the training samples are 3. It is obviously that GRLDA performs better than other methods again (Tables 5, 6, 7 and 8).

3.5 Experiments on Extended Yale B Database

The Extended Yale B database is a high-dimensional dataset used for face recognition research. It contains 2414 frontal-facial images with dimensions of 192 × 168 pixels. The dataset includes images of 39 individuals, with

Table 3 The average recognition accuracy (%), standard deviation, training samples of different methods on FERET face database

Training samples	PCA	LDA	MFA	RILDA	LJSME	TRLDA	RSLDA	RIMFA	STR-LDA	GSOLDA	GRLDA
3	45.66 ±2.10	33.18 ±1.73	36.48 ±2.65	36.98 ±3.06	45.58 ±2.56	39.93 ±4.26	36.28 ±2.65	45.09 ±3.27	39.74 ±1.56	45.78 ±1.78	**46.79 ±1.86**
4	52.52 ±2.58	48.69 ±2.01	44.41 ±1.88	50.46 ±2.87	52.39 ±3.22	39.87 ±3.17	48.41 ±2.62	54.62 ±3.31	53.26 ±1.55	54.98 ±1.97	**55.19 ±1.47**
5	56.69 ±3.18	54.46 ±2.88	49.94 ±3.04	**59.56 ±3.21**	57.53 ±3.54	57.78 ±3.64	58.22 ±3.07	48.12 ±3.84	59.12 ±2.68	58.67 ±3.16	59.00 ±3.87

Table 4 The average recognition accuracy (%), standard deviation, training samples of different methods on ORL face database

Training samples	PCA	LDA	MFA	RILDA	LJSME	TRLDA	RSLDA	RIMFA	STR-LDA	GSOLDA	GRLDA
3	88.09 ±1.98	84.95 ±2.86	81.53 ±2.73	83.53 ±2.05	88.19 ±1.34	87.91 ±1.49	87.35 ±2.29	88.16 ±1.34	86.79 ±2.14	87.86 ±2.57	**88.27 ±2.32**
4	90.46 ±2.35	86.10 ±2.49	87.17 ±2.42	84.40 ±2.19	91.45 ±2.79	91.31 ±2.43	91.64 ±1.83	91.31 ±2.43	90.59 ±2.56	90.21 ±1.37	**91.67 ±2.13**
5	93.25 ±1.33	86.98 ±2.05	88.43 ±2.89	84.85 ±3.36	94.23 ±1.60	93.70 ±1.90	94.10 ±1.74	90.53 ±2.37	94.69 ±1.75	94.09 ±3.07	**95.28 ±1.77**

Table 5 The average recognition accuracy (%), standard deviation, training samples of different methods on YALE face database

Training samples	PCA	LDA	MFA	RILDA	LJSME	TRLDA	RSLDA	RIMFA	STR-LDA	GSOLDA	GRLDA
4	81.90 ±2.00	79.90 ±3.48	74.06 ±3.48	70.33 ±6.22	82.58 ±2.56	82.10 ±2.50	82.61 ±2.03	77.90 ±4.54	81.99 ±3.25	81.78 ±2.46	**82.86 ±2.11**
5	83.44 ±2.61	80.17 ±6.67	74.56 ±3.33	73.17 ±4.77	82.39 ±2.95	82.94 ±2.88	82.22 ±2.17	76.89 ±6.44	83.56 ±1.97	82.56 ±1.46	**84.05 ±1.82**
6	81.87 ±2.90	85.26 ±5.42	77.93 ±2.51	78.20 ±4.81	80.47 ±3.11	81.20 ±3.04	81.87 ±3.51	79.87 ±7.29	83.46 ±2.79	84.31 ±3.42	**84.33 ±2.53**

Table 6 Details of MedMNIST dataset

Name	Data modality	Number of features	Tasks/labels	Train/Validation/Test
Breast MNIST	Breast Ultrasound	784(28×28)	Binary-Class (2)	546 / 78 / 156
Pneumonia MNIST	Chest X-Ray	784(28×28)	Binary-Class (2)	4,708 / 524 / 624

Table 7 The average validating accuracy (%), standard deviation of different methods on two datasets

Dataset	PCA	LDA	MFA	RILDA	LJSME	TRLDA	RSLDA	RIMFA	STR-LDA	GSOLDA	GRLDA
Breast MNIST	85.97 ±1.03	84.49 ±2.56	85.02 ±3.42	83.29 ±5.64	83.21 ±5.43	83.21 ±2.24	88.42 ±1.21	87.93 ±3.74	82.64 ±2.17	87.51 ±2.32	**89.79 ±1.68**
Pneumonia MNIST	90.32 ±5.47	90.04 ±4.65	83.57 ±2.27	82.47 ±4.89	93.49 ±2.64	92.65 ±2.64	91.87 ±2.64	83.54 ±6.67	93.34 ±1.31	**94.85 ±1.64**	94.33 ±1.67

Table 8 The average testing accuracy (%), standard deviation of different methods on two datasets

Dataset	PCA	LDA	MFA	RILDA	LJSME	TRLDA	RSLDA	RIMFA	STR-LDA	GSOLDA	GRLDA
Breast MNIST	71.90 ±2.00	76.97 ±2.34	74.56 ±3.94	77.65 ±5.32	83.79 ±2.54	81.45 ±2.31	80.45 ±2.08	75.48 ±4.59	83.47 ±2.19	83.19 ±2.16	**83.50 ±2.59**
Pneumonia MNIST	82.57 ±1.54	85.26 ±6.64	75.97 ±3.64	73.87 ±4.67	84.56 ±2.34	83.57 ±2.64	82.68 ±1.16	74.76 ±6.87	88.54 ±1.62	87.85 ±1.39	**88.67 ±2.36**

an average of about 64 images per person. The images were taken under diverse lighting circumstances and with a range of facial expressions. Figure 2e displays the 10 photos of the initial individual. Like the FERET dataset, the Extended Yale B dataset includes images of faces taken in challenging conditions, leading to the presence of outliers. Table 9 presents the test results of GRLDA and other cutting-edge approaches, with training samples of 5, 10, 15, and 20, respectively. Experimental results demonstrate that our proposed GRLDA approach achieves a recognition rate of approximately 90% on this high-dimensional dataset when the training samples are 20. This performance is notably superior to the present state-of-the-art method STR-LDA. Therefore, our proposed GRLDA is also capable to effectively handle high-dimensional datasets.

3.6 Experiments on AR Database

The AR dataset consists of 3120 images of 120 different individuals, where each image size is 50×40. In this experiment, we use a subsection of AR dataset, namely the first 20 images are selected to verify the performance of GRLDA in the case when images are varying with different illustration, expressions and occlusions [43]. We randomly select $t(t = 3, 4, 5)$ images as gallery set [32], and the rest of images are used as probe set [32].

3.6.1 Robustness Evaluation

To test the robustness of the proposed GRLDA, we randomly add a block noise to each image. The sample images are shown in Fig. 2f and g respectively. Table 10 lists the highest recognition rates, the dimensions, and the standard deviations of different algorithms with different block size. Figure 3a shows that when the training samples are 5, the recognition rates versus the subspace dimension of the original images, and images with block size 15×15, 10×10, 5×5, respectively. Results in Table 10 illustrates the stronger robustness of the proposed GRLDA to the corruption of image than other methods.

3.6.2 Face Reconstruction and Learned Projections

In this subsection, we further conduct a series of experiments to explore the reconstructed face images and the projection visualization of the proposed GRLDA and some state-of-the-art methods, i.e., LDA, MFA, RSLDA. We take two kinds of experiments into account. On the one hand, we use the first 5 original images of each individual to train and obtain the learned projections of LDA, MFA, RSLDA, GRLDA. Figure 4a illustrates one original image of the first person. The reconstructed images by LDA, MFA, RSLDA, GRLDA are shown in Fig. 4b-e. In each method, we use the first 50 projections to reconstruct the face image. To

Table 9 The average recognition accuracy (%), standard deviation, training samples of different methods on Extended YALE B face database

Training samples	PCA	LDA	MFA	RILDA	LJSME	TRLDA	RSLDA	RIMFA	STR-LDA	GSOLDA	GRLDA
5	56.12 ±1.07	61.45 ±3.80	63.87 ±3.54	67.93 ±6.21	68.50 ±2.97	62.00 ±3.70	64.17 ±2.19	70.49 ±3.27	70.27 ±3.41	**71.67** **±1.03**	71.40 ±1.87
10	60.24 ±2.05	65.37 ±1.87	65.02 ±3.96	70.54 ±4.02	72.58 ±2.40	68.10 ±2.50	72.61 ±2.07	76.88 ±4.76	74.79 ±3.45	75.70 ±2.10	**77.69** **±2.14**
15	64.50 ±2.43	68.79 ±4.21	70.25 ±2.35	74.43 ±1.89	79.37 ±2.64	72.46 ±2.23	74.70 ±3.04	82.37 ±5.65	81.92 ±2.74	77.64 ±1.51	**83.87** **±3.11**
20	66.64 ±1.29	71.34 ±2.84	73.21 ±1.28	78.27 ±3.57	83.12 ±3.27	73.49 ±2.50	76.73 ±2.18	84.67 ±4.32	85.03 ±3.24	81.58 ±2.78	**89.76** **±2.31**

Table 10 The top recognition accuracy (%), standard deviation, training samples and dimensions of different methods on AR face database with block corruption

Block size	Training samples	PCA	LDA	MFA	RILDA	LJSME	TRLDA	RSLDA	RIMFA	GSO LDA	STR-LDA	GRLDA
15 × 15	3	44.04 ±2.33 (115)	71.70 ±1.60 (140)	61.51 ±3.55 (150)	60.40 ±3.00 (200)	62.43 ±1.76 (190)	43.50 ±2.67 (150)	58.08 ±3.71 (180)	72.70 ±2.27 (100)	73.43 ±1.54 (200)	66.46 ±3.58 (100)	**76.52** **±1.77** **(200)**
	4	63.25 ±1.72 (115)	78.90 ±1.17 (125)	72.41 ±2.31 (150)	67.17 ±2.09 (200)	62.60 ±2.05 (185)	49.86 ±2.77 (150)	63.79 ±4.68 (180)	80.44 ±1.70 (100)	83.79 ±1.02 (200)	73.57 ±2.43 (100)	**83.97** **±2.28** **(200)**
	5	67.93 ±1.12 (115)	71.67 ±2.53 (140)	80.05 ±1.86 (150)	67.84 ±1.78 (200)	66.68 ±1.51 (190)	67.01 ±1.63 (150)	66.48 ±3.15 (170)	78.19 ±1.95 (100)	85.95 ±1.21 (200)	77.24 ±2.31 (100)	**85.96** **±1.48** **(200)**
10 × 10	3	61.70 ±2.45 (115)	83.37 ±1.09 (140)	68.47 ±2.95 (150)	82.21 ±1.21 (200)	61.17 ±1.12 (175)	62.95 ±1.57 (150)	63.64 ±7.35 (180)	78.17 ±2.40 (100)	84.34 ±1.52 (200)	67.57 ±1.74 (100)	**87.57** **±1.96** **(200)**
	4	67.38 ±1.82 (115)	84.75 ±1.46 (140)	79.75 ±1.39 (150)	85.42 ±1.29 (200)	67.80 ±1.96 (190)	67.30 ±1.54 (150)	66.59 ±1.70 (180)	82.02 ±3.12 (110)	90.06 ±3.24 (200)	79.15 ±2.23 (100)	**91.10** **±1.17** **(200)**
	5	71.64 ±2.01 (115)	78.43 ±2.57 (140)	85.68 ±2.40 (150)	79.20 ±1.78 (200)	72.22 ±1.90 (190)	71.48 ±1.90 (150)	71.49 ±1.99 (180)	81.69 ±1.81 (100)	**91.24** **±1.38** **(200)**	81.09 ±2.21 (100)	91.04 ±1.70 (200)
5 × 5	3	61.82 ±1.90 (115)	85.40 ±1.48 (140)	73.80 ±2.31 (150)	85.97 ±2.13 (200)	62.64 ±2.22 (190)	62.74 ±2.39 (150)	64.28 ±6.25 (180)	78.10 ±2.93 (110)	83.56 ±1.60 (200)	78.36 ±3.24 (100)	**88.54** **±1.59** **(200)**
	4	67.76 ±2.55 (115)	87.79 ±1.24 (140)	83.23 ±1.39 (150)	88.44 ±0.94 (200)	67.86 ±1.84 (190)	68.85 ±1.64 (150)	69.41 ±3.98 (180)	81.41 ±4.98 (100)	88.29 ±1.79 (200)	81.30 ±2.41 (100)	**89.29** **±1.56** **(200)**
	5	72.26 ±2.38 (120)	82.26 ±1.95 (140)	88.82 ±1.69 (150)	82.70 ±1.11 (200)	71.85 ±2.18 (190)	72.48 ±1.79 (150)	71.62 ±2.41 (180)	84.10 ±2.67 (100)	90.24 ±1.88 (200)	83.23 ±2.37 (100)	**90.28** **±1.54** **(200)**

explore the learned subspace, we also show the first learned projection of LDA, MFA, RSLDA and GRLDA, respectively in Fig. 4f-i. On the other hand, we use the first 5 images corrupted by the block size 15 × 15 to obtain the projection matrix and conduct the same operation as above mentioned. The results are illustrated in Fig. 5 and some conclusions can be drawn from Figs. 4 and 5:

From Fig. 4h, we know that RSLDA performs well in selecting the most discriminative feature in the original space. However, it cannot effectively avoid the impact of outliers, which is obviously highlighted in Fig. 5h. From Fig. 4b to 4e, it is obvious to see that the reconstructing ability of LDA is the least but GRLDA not only has a relatively good reconstructing ability, but also can find discriminative projections for feature extraction and selection. As we know that GRLDA can reduce the impact of outliers by the $L_{2,1}$ -norm as the basic measurement and the regularized term, and both Fig. 5e and i indicate this theoretical explanation since the images are less influenced by noised block.

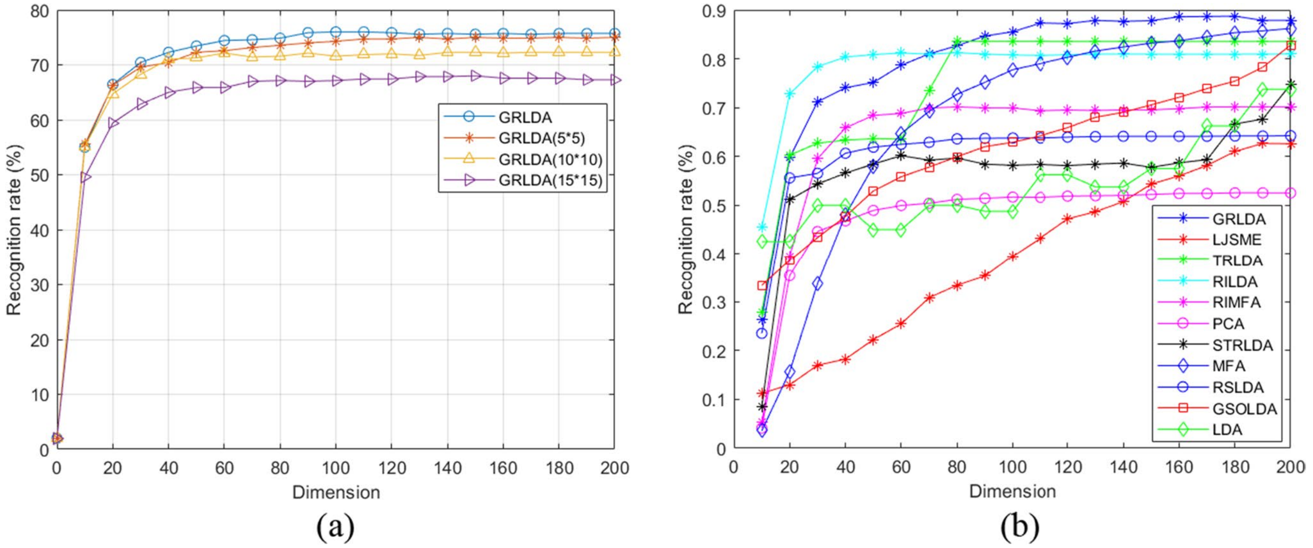


Fig. 3 **a** The recognition rates versus the subspace dimension by GRLDA with different block size. **b** The recognition rates versus the subspace dimension by different methods on the Yale dataset

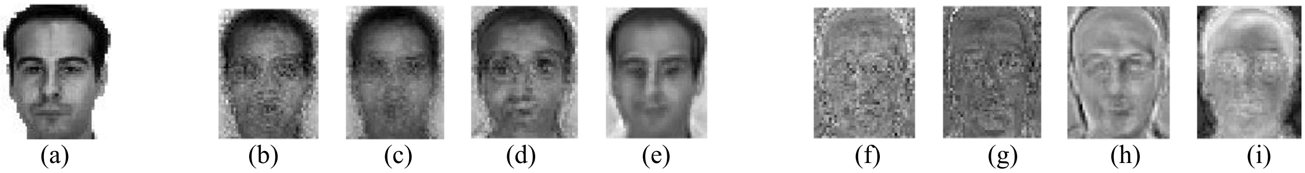


Fig. 4 Original images in (**a**), reconstructed images by LDA (**b**), MFA (**c**), RSLDA (**d**), GRLDA (**e**); the first projection obtained by LDA (**f**), MFA (**g**), RSLDA (**h**), GRLDA (**i**)

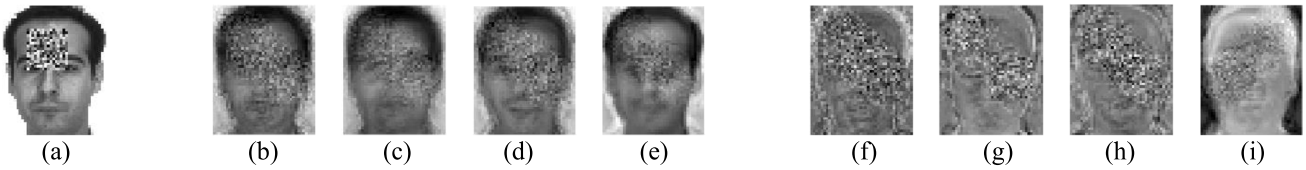


Fig. 5 Images with random block size 15*15 in (**a**), reconstructed images by LDA (**b**), MFA (**c**), RSLDA (**d**), GRLDA (**e**); the first projection obtained by LDA (**f**), MFA (**g**), RSLDA (**h**), GRLDA (**i**)

3.7 Experiments on Yale Database

The Yale dataset involves a total of 15 people, each with 11 face images, with size of 100×80. This dataset contains variations in illumination, facial expression and with or without glasses [43], Fig. 2d shows the sample images of the first person. In this experiment, we randomly choose $t(t=5$, 6, 7) images of each individual as gallery set [32], and the rest of images are used as probe set [32]. And the corresponding parameter α is set in 10^3 according to Fig. 1a.

The average recognition rates of different algorithms are shown in Table 5. When the first 6 images of each people were used as gallery set, the testing recognition rates versus the subspace dimension are illustrated in Fig. 3b. Both of them indicates that the proposed GRLDA is robust to small samples and can reach a high recognition in a very low dimension and maintain a good stability.

3.8 Experiments on MedMNIST Database

The MedMNIST dataset comprises a vast collection of standardized biomedical images, consisting of 708,069 2D medical images across 12 different categories and 9,998

3D medical images spanning 6 different types. The dimensions of 2D images are 28 × 28, while the dimensions of 3D images are 28 × 28 × 28. Additionally, the background information of the images is eliminated, making them very suitable for testing machine learning methods. This dataset is utilized for lightweight image classification tasks, encompassing binary classification, multi-classification, ordinary regression, and multi-category tasks. For this study, we specifically chose two datasets, namely BreastMNIST and PneumoniaMNIST, to evaluate the performance of our approach. The sample images of the two subdatasets are shown in Fig. 2h and i. Table 6 provides a more comprehensive overview of the two subdatasets. The experimental results presented in Tables 7 and 8 clearly indicate that GRLDA achieves the highest level of classification accuracy among all the methods tested on breast cancers and pneumonia test datasets. Furthermore, GRLDA also attains the second-highest accuracy on the PneumoniaMNIST validation set.

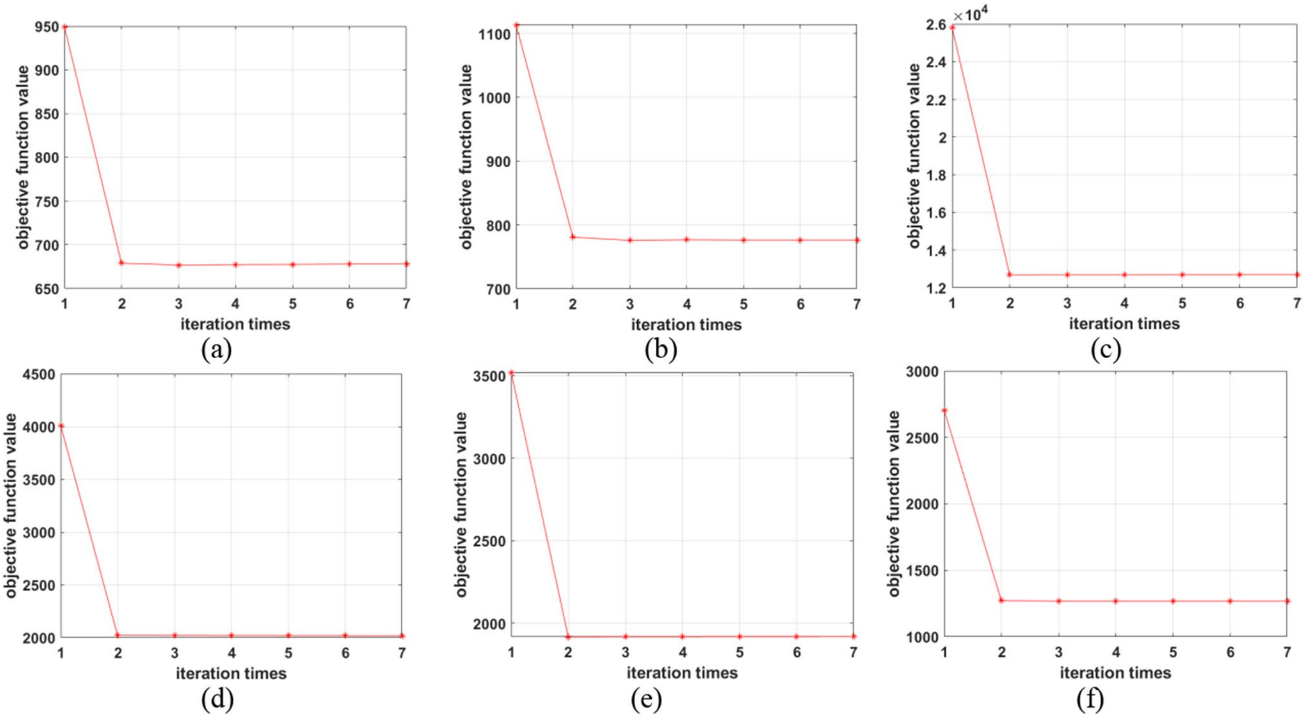
3.9 Convergence Study

It is interesting and necessary for us to explore how fast GRLDA converges. Figure 6 depicts the objective function value of the proposed GRLDA versus the number of iteration times on different databases. It clearly indicates that the proposed GRLDA can converge within 3 iterations.

3.10 Experimental Results and Discussions

The experimental results in terms of the recognition accuracy of the proposed GRLDA and the classical algorithms (i.e., PCA, LDA and MFA), and some methods based on rotational invariant $L_{2,1}$-norm (i.e., RILDA and RIMFA [31], LJSME [32]) and other new methods (i.e. TRLDA [20], RSLDA [21], STR-LDA [34], and GSOLDA [35]) are listed in tables and figures above, we can draw some interesting conclusions:

1. In most cases, the proposed GRLDA and GSOLDA [35] outperform other algorithms. The main reason is that the loss functions of the two algorithms are based on $L_{2,1}$-norm so that they are more robust to outliers than methods using L_2-norm as the basic metric. However, GRLDA not only computes the intraclass and interclass scatter matrices using $L_{2,1}$-norm, but also defines a new formulation measured by trace and square root of trace, which can obtain the local optimal solution.
2. The performances of rotational invariant algorithms and some classical algorithms approach GRLDA when the training samples are big. However, when the quantity of training samples falls, their performances significantly deteriorate. Furthermore, when the training samples change, GRLDA can retain greater stability. The possible explanation is that, in order to avoid computing the inverse of the intraclass (or interclass) scatter matrix and



(a) (b) (c)

(d) (e) (f)

Fig. 6 Convergence curves of GRLDA on (**a**) Yale, (**b**) FERET, (**c**) COIL-20, (**d**) AR, (**e**) ORL, (**f**) PIE

to obtain the optimal solution, the optimization problem is converted into a quadratic problem on the Stiefel manifold and singular value decomposition is used.

3. Based on the performance statistics of various approaches as the subspace dimension increases, it can be observed that LDA, MFA, and RSLDA are susceptible to variations in the subspace dimension in some datasets. However, in the majority of databases, the suggested GRLDA can successfully overcome this drawback and reach the peak at a very low dimension.
4. In most cases, the proposed GRLDA outperforms RSLDA, which is aimed to maximize the separability of the data point in each dimension of the subspace so that it can obtain the features with more discriminative ability. Figure 4h indicates that. However, when we compare Fig. 4d and e, we can observe that the reconstructing ability of GRLDA is better than RSLDA. The potential reason is that GRLDA preserves the local geometric structure, which is helpful to learn discriminative information and reconstruction ability simultaneously.

4 Conclusion

In this paper, we propose a more robust linear discriminant analysis method incorporating multiple factors and using $L_{2,1}$ -norm as the basic metric on both loss function and regularization term. GRLDA tends to preserve the local geometric structures in the learned subspace with joint sparsity to obtain more discriminative features. Two sub problems can be broken down into an iterative approach that is designed to compute the optimal answer. The ideal solution can be found simply in the first section. The suggested goal function is transformed into a quadratic problem on the Stiefel manifold in the second section, and the best solution is found by applying SVD. This technique effectively avoids computing the inverse of a singular matrix so that the number of samples can be very small. Moreover, we rigorously analyze the computational complexity. Extensive experiments on face, object, and medical databases indicate that the speed of the convergence is very fast and the performance of the proposed GRLDA is superior to most state-of-the-art algorithms.

Appendix

First, we introduce a lemma as follows:

Lemma 3

Assuming that *f(x)* is a convex function of *x* where *x* can be a scalar, vector or matrix variable, then we obtain:

$$\mathrm{f}(\mathrm{x}_1) - \mathrm{f}(\mathrm{x}_2) \geq \mathrm{Tr}((\mathrm{f}'(\mathrm{x}_2))^{\mathrm{T}}(\mathrm{x}_1 - \mathrm{x}_2)) \tag{28}$$

where $f'(x_2)$ is the super-gradient of *f(x)* at x_2.

Proof of the theorem 2

It is easy to know that $\mathrm{f}_\mathrm{i}(\mathrm{h}_\mathrm{i}(\mathrm{m}))$ is an arbitrary convex function *w.r.t.* $\mathrm{h}_\mathrm{i}(\mathrm{m})$ under the arbitrary constraint of $\mathrm{m} \in \Omega$. We assume that $\mathrm{f}_\mathrm{i}(\mathrm{h}_\mathrm{i}(\mathrm{m})) \leq 0$. In the *t*-th iteration, we denote $\mathrm{G}_\mathrm{i}^\mathrm{t} = \mathrm{f}\prime_\mathrm{i}(\mathrm{h}_\mathrm{i}(\mathrm{m}^{\mathrm{t}-1}))$. For each *i*, according to the lemma 3, we have:

$$\begin{aligned}&\mathrm{f}_\mathrm{i}(\mathrm{h}_\mathrm{i}(\mathrm{m}^\mathrm{t})) - \mathrm{f}_\mathrm{i}(\mathrm{h}_\mathrm{i}(\mathrm{m}^{\mathrm{t}-1}))\\ &\geq Tr((^{\mathrm{G}_\mathrm{i}^\mathrm{t}}\mathrm{h}_\mathrm{i}(\mathrm{m}^\mathrm{t})) - Tr((^{\mathrm{G}_\mathrm{i}^\mathrm{t}}\mathrm{h}_\mathrm{i}(\mathrm{m}^{\mathrm{t}-1}))\end{aligned} \tag{29}$$

According to (21), the following can be derived:

$$\mathrm{Tr}((\mathrm{G}_\mathrm{i}^\mathrm{t})^\mathrm{T}\mathrm{h}_\mathrm{i}(\mathrm{m}^\mathrm{t})) \geq \mathrm{Tr}((\mathrm{G}_\mathrm{i}^\mathrm{t})^\mathrm{T}\mathrm{h}_\mathrm{i}(\mathrm{m}^{\mathrm{t}-1})) \tag{30}$$

Summing (29) and (30), we have:

$$\sum_\mathrm{i} \mathrm{f}_\mathrm{i}(\mathrm{h}_\mathrm{i}(\mathrm{m}^\mathrm{t})) \geq \sum_\mathrm{i} \mathrm{f}_\mathrm{i}(\mathrm{h}_\mathrm{i}(\mathrm{m}^{\mathrm{t}-1})) \tag{31}$$

Summing (31) and $\mathrm{f}_\mathrm{i}(\mathrm{h}_\mathrm{i}(\mathrm{m})) \leq 0$, the value of the objective function (16) will monotonically increase until convergence.

Authors contribution All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Yufei Zhu, Zhihui Lai, Can Gao, and Heng Kong. The first draft of the manuscript was written by Yufei Zhu and Zhihui Lai, and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Data availability The datasets analyzed during the current study are available in the following databases.

(1) **COIL20 database** (http://www.cs.columbia.edu/CAVE/software/softlib/coil-20.php),

(2) **FERET database** (http://www.nist.gov/itl/iad/ig/colorferet.cfm),

(3) **ORL database** (http://www.uk.research.att.com/facedatabase.html),

(4) **AR database** (http://www2.ece.ohio-state.edu/~aleix/ARdatabase.html).

(5) **MedMNIST database** (https://medmnist.com/).

(6) **Extended Yale B Database** (https://www.kaggle.com/datasets/tbourton/extyalebcroppedpng).

Declarations

Competing interests This work was supported in part by the Natural Science Foundation of China under Grant 62272319, and in part by the Natural Science Foundation of Guangdong Province (Grant 2023A1515010677, 2024A1515011637) and Science and Technology Planning Project of Shenzhen Municipality under Grant JCYJ20210324094413037 and JCYJ20220818095803007.

Ethical and informed consent for data used I confirm that I have obtained informed consent from all participants whose data I used in my research.

References

1. Tao D, Tang X, Li X (2006) Direct kernel biased discriminant analysis: a new content based image retrieval relevance feedback algorithm. IEEE Trans Multimed 8(4):716–727
2. Passalis N, Tefas A (2018) Dimensionality reduction using similarity-induced embeddings. IEEE Trans Neural Netw Learn Syst 29(8):3429–3441
3. Vashishtha G, Kumar R (2023) Unsupervised learning model of sparse filtering enhanced using wasserstein distance for intelligent fault diagnosis. J Vib Eng Technol 11(7):2985–3002
4. Lou Q, Deng Z, Choi K-S, Shen H, Wang J, Wang S (2021) Robust multi-label relief feature selection based on fuzzy margin co-optimization, IEEE Trans Emerg Topics Comput Intell, early access
5. Vashishtha G, Chauhan S, Kumar A, KumarAn R (2022) Ameliorated African Vulture Optimization Algorithm to Diagnose the Rolling Bearing Defects. Meas Sci Technol 33(7):075013
6. Lai Z, Xu Y, Jin Z, Zhang D (2014) Human gait recognition via sparse discriminant projection learning. IEEE Trans Circuits Syst Video Technol 24(10):1651–1662
7. Yang L, Song S, Gong Y (2019) Nonparametric dimension reduction via maximizing pairwise separation probability. IEEE Trans Neural Netw Learn Syst 30(10):3205–3210
8. Bhadra T, Maulik U (2022) Unsupervised Feature Selection Using Iterative Shrinking and Expansion Algorithm. IEEE Trans Emerg Topics Comput Intell 5(5):1453–1462
9. Vashishtha G, Kumar R (2022) Feature Selection Based on Gaussian Ant Lion Optimizer for Fault Identification in Centrifugal Pump, Recent Advances in Machines and Mechanisms: Select Proceedings of the iNaCoMM 2021. Singapore: Springer Nature Singapore. 295–310
10. Cunningham JP, Ghahramani Z (2015) Linear dimensionality reduction: Survey, insights, and generalizations. J Mach Learn Research 16(1):2859–2900
11. Sumithra V, Surendran S (2015) A review of various linear and non linear dimensionality reduction techniques. Int J Comput Sci Inf Technol 6(3):2354–2360
12. Kwak N (2008) Principal component analysis based on L1-norm maximization. IEEE Trans Pattern Anal Mach Intell 30(9):1672–1680
13. Martinez AM, Kak AC (2001) PCA versus LDA. IEEE Trans Pattern Anal Mach Intell 23(2):228–233
14. He X (2004) Locality preserving projections. Proc Adv Neural Inf Process Syst 16(1):186–197
15. Roweis ST, Saul LK (2000) Nonlinear dimensionality reduction by locally linear embedding, science, 2000, vol. 290, no. 5500, pp. 2323–2326
16. Balasubramanian M, Schwartz EL (2002) The isomap algorithm and topological stability. Science 295(5552):7–7
17. Belkin M, Niyogi P (2003) Laplacian eigenmaps for dimensionality reduction and data representation. Neural Comput 15(6):1373–1396
18. Wen J, Fang X, Cui J, Fei L, Yan K, Chen Y, Xu Y (2018) Robust sparse linear discriminant analysis. IEEE Trans Circuits Syst Video Technol 29(2):390–403
19. Wang H, Yan S, Xu D, Tang X, Huang T (2007) Trace ratio vs. ratio trace for dimensionality reduction, in Proc IEEE Comput Soc Conf Comput Vision Pattern Recognit, pp. 1–8
20. Wang J, Wang L, Nie F, Li X (2021) A novel formulation of trace ratio linear discriminant analysis, IEEE Trans Neural Netw Learn Syst, pp. 1–11
21. Wang J, Wang H, Nie F, Li X (2022) Ratio sum vs. sum ratio for linear discriminant analysis. IEEE Trans Pattern Anal Mach Intell 44(12):10171–10185
22. Pang Y, Yuan Y (2010) Outlier-resisting graph embedding. Neurocomputing 73(4–6):968–974
23. Yan S, Xu D, Zhang B, Zhang H-J, Yang Q, Lin S (2006) Graph embedding and extensions: a general framework for dimensionality reduction. IEEE Trans Pattern Anal Mach Intell 29(1):40–51
24. Ding C, Zhou D, He X, Zha H (2006) R1-PCA: Rotational invariant L1-norm principal component analysis for robust subspace factorization, Proc 23rd Int Conf Mach Learn: 281–288
25. Li X, Hu W, Wang H, Zhang Z (2010) Linear discriminant analysis using rotational invariant L1-norm. Neurocomputing 73(13–15):2571–2579
26. Pang Y, Li X, Yuan Y (2010) Robust tensor analysis with L1-norm. IEEE Trans Circuits Syst Video Technol 20(2):172–178
27. Zou H, Hastie T, Tibshirani R (2004) Sparse principal component analysis. J Comput Graph Stat 15(2):265–286
28. Nie F, Wang Z, Wang R, Wang Z, Li X (2019) Towards robust discriminative projections learning via non-greedy L2,1-norm minmax. IEEE Trans Pattern Anal Mach Intell 43(6):2086–2100
29. Nie F, Yuan J, Huang H (2014) Optimal mean robust principal component analysis, Int Conf Mach Learn, PMLR
30. Zhao H, Wang Z, Nie F (2019) A new formulation of linear discriminant analysis for robust dimensionality reduction. IEEE Trans Knowl Data Eng 31(4):629–640
31. Lai Z, Xu Y, Yang J, Shen L, Zhang D (2016) Rotational invariant dimensionality reduction algorithms. IEEE Trans Cybern 47(11):3733–3746
32. Mo D, Lai Z, Wong WK (2019) Locally joint sparse marginal embedding for feature extraction. IEEE Trans Multimed 21(12):3038–3052
33. Lin Y, Lai Z, Zhou J, Wen J, Kong H (2023) Multiview Jointly Sparse Discriminant Common Subspace Learning. Pattern Recognit 138:109342
34. Li Z, Nie F, Wu D, Wang Z, Li X (2023) Sparse Trace Ratio LDA for Supervised Feature Selection, IEEE Trans Cybern, early access. https://doi.org/10.1109/TCYB.2023.3264907
35. Mo D, Lai Z, Zhou J, Hu Q (2023) Scatter matrix decomposition for jointly sparse learning. Pattern Recognit 140:109485
36. Nie F, Huang H, Cai X, Ding C (2010) Efficient and robust feature selection via joint L2,1-norms minimization, Proc Adv Neural Inf Process Syst, pp. 1813–1821
37. Wong WK, Zhao HT (2012) Supervised optimal locality preserving projection. Pattern Recognit 45(1):186–197
38. Yang Y, Shen HT, Ma Z, Huang Z, Zhou X (2011) L2,1-norm regularized discriminative feature selection for unsupervised, IJCAI Int Joint Conf Artif Intell
39. Nie F, Wu D, Wang R, Li X (2007) Truncated robust principle component analysis with a general optimization framework. IEEE Trans Pattern Anal Mach Intell 40(1):339–342
40. Turk M, Pentland A (1991) Eigenfaces for recognition. J Cogn Neurosci 3(1):71–86

41. Ye J, Xiong T, Madigan D (2006) Computational and theoretical analysis of null space and orthogonal linear discriminant analysis J Machine Learn Res, vol. 7, no. 7
42. Mo D, Lai Z, Wang X, Wong WK (2019) Jointly sparse locality regression for image feature extraction. IEEE Trans Multimed 22(11):2873–2888
43. Lai Z, Mo D, Wen J, Shen L, Wong WK (2018) Generalized robust regression for jointly sparse subspace learning. IEEE Trans Circuits Syst Video Technol 29(3):756–772
44. Wang K, He R, Wang L, Wang W, Tan T (2015) Joint feature selection and subspace learning for cross-modal retrieval. IEEE Trans Pattern Anal Mach Intell 38(10):2010–2023
45. Huang J, Li G, Huang Q, Wu X (2017) Joint feature selection and classification for multilabel learning. IEEE Trans Cybern 47(3):876–889
46. Shi X, Yang Y, Guo Z, Lai Z (2014) Face recognition by sparse discriminant analysis via joint L2,1-normminimization. Pattern Recognit 47:2447–2453

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

Yufei Zhu is currently with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China. His research interests include image processing and pattern recognition.

Zhihui Lai received the B.S degree in mathematics from South China Normal University M.S degree from Jinan University, and the Ph D degree in pattern recognition and intelligence system from Nanjing University of Science and Technology (NUST), China in 2002 2007 and 2011, respectively. He has been a Research Associate, Postdoctoral Fellow and Research Fellow at The Hong Kong Polytechnic University, Hong Kong He has published over 200 scientific articles. His research interests include face recognition, image processing and content based image retrieval, pattern recognition, compressive sense, human vision modelization and applications in the fields of intelligent robot research. Now he is an associate editor of International Journal of Machine Learning and Cybernetics. For more information including all papers and related codes, the readers are referred to the website (http://www.scholat.com/laizhihui).

Can Gao received the Ph.D. degree in pattern recognition and intelligent systems from Tongji University, Shanghai, China, in 2013. From 2010 to 2011, he was a Visiting Scholar with the University of Alberta, Edmonton, AB, Canada. From 2015 to 2018, he was a Research Associate, a Postdoctoral Fellow, and a Research Fellow with The Hong Kong Polytechnic University, Hong Kong. He is currently an Assistant Professor with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China. He has authored or coauthored more than 60 academic papers and is an Associate Editor for the Journal of Intelligent & Fuzzy Systems. His current major research interests include semi-supervised learning, soft computing, and anomaly detection.

Heng Kong received the M.D. and B.S. degree from Chongqing Medical University, M.S. degree from Guangzhou Medical University, and Ph.D. degree from Southern Medical University, China, in 2000, 2005 and 2008, respectively. She works as a visiting scholar in Cancer Center of Georgia Reagent University at Augusta in USA in 2014 2016. She is a director in department of thyroid and breast, BaoAn Central Hospital of Shenzhen, China. She is also doing basic and clinic research associated breast cancer. Her research interests include gene therapy, immunotherapy, early diagnosis and prognosis analysis of breast cancer, as well as the tumor image processing and recognition using machine learning and artificial intelligent methods.