Linear Discriminant Analysis Using Rotational Invariant L_1 Norm

Xi Li[1a], Weiming Hu[2a], Hanzi Wang[3b], Zhongfei Zhang[4c]

[a]*National Laboratory of Pattern Recognition, CASIA, Beijing, China*
[b]*University of Adelaide, Australia*
[c]*State University of New York, Binghamton, NY 13902, USA*

Abstract

Linear Discriminant Analysis (LDA) is a well-known scheme for supervised subspace learning. It has been widely used in the applications of computer vision and pattern recognition. However, an intrinsic limitation of LDA is the sensitivity to the presence of outliers, due to using the Frobenius norm to measure the inter-class and intra-class distances. In this paper, we propose a novel rotational invariant L_1 norm (i.e., R_1 norm) based discriminant criterion (referred to as DCL_1), which better characterizes the intra-class compactness and the inter-class separability by using the rotational invariant L_1 norm instead of the Frobenius norm. Based on the DCL_1, three subspace learning algorithms (i.e., $1DL_1$, $2DL_1$, and TDL_1) are developed for vector-based, matrix-based, and tensor-based representations of data, respectively. They are capable of reducing the influence of outliers substantially, resulting in a robust classification. Theoretical analysis and experimental evaluations demonstrate the promise and effectiveness of the proposed DCL_1 and its algorithms.

Keywords: Linear discriminant analysis, face classification, R_1 norm.

1. Introduction

In recent years, Linear Discriminant Analysis (LDA) plays an important role in supervised learning with many successful applications of computer vision and

[1]Xi Li has moved to CNRS, TELECOM ParisTech, France. Email: xi-li@telecom-paristech.fr
[2]Email: wmhu@nlpr.ia.ac.cn
[3]Email: Hanzi.Wang@ieee.org
[4]Email: zhongfei@cs.binghamton.edu

pattern recognition. By maximizing the ratio of the inter-class distance to the intra-class distance, LDA aims to find a linear transformation to achieve the maximum class discrimination. Many variations of LDA with different properties have been proposed for discriminant subspace learning. The classical LDA [1][2] tries to find an optimal discriminant subspace (spanned by the column vectors of a projection matrix) to maximize the inter-class separability and the intra-class compactness of the data samples in a low-dimensional vector space. In general, the optimal discriminant subspace can be obtained by performing the generalized eigenvalue decomposition on the inter-class and the intra-class scatter matrices. However, an intrinsic limitation of the classical LDA is that one of the scatter matrices is required to be nonsingular. Unfortunately, the dimension of the feature space is typically much larger than the size of the training set in many applications (e.g., face recognition), resulting in the singularity of one of the scatter matrices. This is well-known as the Undersampled Problem (USP). In order to address the USP, Fukunaga [3] proposes a regularization method (RM) which adds perturbations to the diagonal entries of the scatter matrices. But the solution obtained by RM is not optimal. In recent years, many algorithms have been developed to deal with the USP, including the direct linear discriminant analysis (DLDA) [5] and the null-space linear discriminant analysis (NLDA) [4]. NLDA extracts discriminant information from the null space of the intra-class scatter matrix. In comparison, DLDA extracts the discriminant information from the null space of the intra-class scatter matrix after discarding the null space of the inter-class scatter matrix. However, NLDA and DLDA may lose discriminant information which may be useful for classification. To fully utilize all the discriminant information reflected by the intra-class and inter-class scatter matrices, Wang and Tang [6] propose a dual-space LDA approach to make full use of the discriminative information in the feature space. Another approach to address the USP is to use PCA+LDA [7][8] to extract the discriminant information (i.e., the data are pre-processed by PCA before LDA). However, PCA+LDA may lose important discriminant information in the stage of PCA.

More recent LDA algorithms work with higher-order tensor representations. Ye *et al.* [9] propose a novel LDA algorithm (i.e., 2DLDA) which works with the matrix-based data representation. Also in [9], 2DLDA+LDA is proposed for further dimension reduction by 2DLDA before LDA. Similar to [9], Li and Yuan [18] use image matrices directly instead of vectors for discriminant analysis. Xu *et al.* [19] propose a novel algorithm (i.e., Concurrent Subspaces Analysis) for dimension reduction by encoding images as 2nd or even higher order tensors. Vasilescu and Terzopoulos [15] apply multilinear subspace analysis to construct

a compact representation of facial image ensembles factorized by different faces, expressions, viewpoints, and illuminations. Lei *et al.* [14] propose a novel face recognition algorithm based on discriminant analysis with a Gabor tensor representation. He *et al.* [11] present a tensor-based algorithm (i.e., Tensor Subspace Analysis) for detecting the underlying nonlinear face manifold structure in the manner of tensor subspace learning. Yan *et al.* [10] and Tao *et al.* [13] propose their own subspace learning algorithms (i.e., DATER [10] and GTDA [13]) for discriminant analysis with tensor representations. Wang *et al.* [12] propose a convergent solution procedure for general tensor-based subspace analysis. Essentially, the aforementioned tensor-based LDA approaches perform well in uncovering the underlying data structures. As a result, they are able to handle the Undersampled Problem (USP) effectively.

However, all the aforementioned LDA approaches utilize the Frobenius norm to measure the inter-class and intra-class distances. In this case, their training processes may be dominated by outliers since the inter-class or intra-class distance is determined by the sum of squared distances. To reduce the influence of outliers, we propose a novel rotational invariant L_1 norm (referred to as R_1 norm [16][17]) based discriminant criterion called DCL_1 for robust discriminant analysis. Further, we develop three DCL_1-based discriminant algorithms (i.e., $1DL_1$, $2DL_1$, and TDL_1) for vector-based, matrix-based, and tensor-based representations of data, respectively. In contrast to the classical LDA [1], 2DLDA [9], and DATER [10], the developed $1DL_1$, $2DL_1$, and TDL_1 can reduce the influence of outliers substantially.

1.1. Related work

Pang *et al.* [20] propose a L_1-norm-based tensor analysis (TPCA-L_1) algorithm which is robust to outliers. Compared to conventional tensor analysis algorithms, TPCA-L_1 is more efficient due to its eigendecomposition-free property. Zhou and Tao [21] present a gender recognition algorithm called Manifold Elastic Net (MEN). The algorithm can obtain a sparse solution to supervised subspace learning by using L_1 manifold regularization. Especially in the cases of small training sets and lower-dimensional subspaces, it achieves better classification performances against traditional subspace learning algorithms. Pang and Yuan [22] develop an outlier-resiting graph embedding framework (referred to as LPP-L_1) for subspace learning. The framework is not only robust to outliers, but also performs well in handling the USP. Zhang *et al.* [23] propose a Discriminative Locality Alignment (DLA) algorithm for subspace learning. It takes advantage of discriminative subspace selection for distinguishing the dimension

reduction contribution of each sample, and preserves discriminative information over local patches of each sample to avoid the USP. Liu *et al.* [24] make a semi-supervised extension of linear dimension reduction algorithm called Transductive Component Analysis (TCA) and Orthogonal Transductive Component Analysis (OTCA), which leverage the intra-class smoothness and the inter-class separability by building two sorts of regularized graphs. Tao *et al.* [25] propose three criteria for subspace selection. As for the c-class classification task, these three criteria is able to effectively stop the merging of nearby classes in the projection to a subspace of the feature space if the dimension of the projected subspace is strictly lower than c–1. Tao *et al.* [26] incorporate tensor representation into existing supervised learning algorithms, and present a supervised tensor learning (STL) framework to overcome the USP. Furthermore, several convex optimization techniques and multilinear operations are used to solve the STL problem.

The remainder of the paper is organized as follows. In Sec. 2, the Frobenius and R_1 norms are briefly reviewed. In Sec. 3, a brief introduction to Linear Discriminant Analysis using the Frobenius norm is given. In Sec. 4, the details of the proposed DCL_1 and its algorithms ($1DL_1$, $2DL_1$, and TDL_1) are described. Experimental results are reported in Sec. 5. The paper is concluded in Sec. 6.

2. Frobenius and R_1 norms

Given K data samples $\mathscr{X} = \{\mathcal{X}_k\}_{k=1}^{K}$ with $\mathcal{X}_k = (x^k_{d_1 d_2 \cdots d_n})_{D_1 \times D_2 \cdots \times D_n}$, the Frobenius norm is defined as:

$$\begin{aligned} \|\mathscr{X}\| &= \sqrt{\left(\sum_{k=1}^{K}\sum_{d_1=1}^{D_1}\sum_{d_2=1}^{D_2}\cdots\sum_{d_n=1}^{D_n}\left(x^k_{d_1 d_2 \cdots d_n}\right)^2\right)} \\ &= \sqrt{\sum_{k=1}^{K}\|\mathcal{X}_k\|^2}. \end{aligned} \tag{1}$$

The rotational invariant L_1 norm (i.e., R_1 norm) is defined as:

$$\begin{aligned} \|\mathscr{X}\|_{R_1} &= \sum_{k=1}^{K}\sqrt{\left(\sum_{d_1=1}^{D_1}\sum_{d_2=1}^{D_2}\cdots\sum_{d_n=1}^{D_n}\left(x^k_{d_1 d_2 \cdots d_n}\right)^2\right)} \\ &= \sum_{k=1}^{K}\sqrt{\|\mathcal{X}_k\|^2} = \sum_{k=1}^{K}\|\mathcal{X}_k\|. \end{aligned} \tag{2}$$

When $n = 1$, the above norms are vector-based; when $n = 2$, they are matrix-based; otherwise, they are tensor-based. In the Euclidean space, the Frobenius norm has a fundamental property—rotational invariance. In comparison, the R_1 norm has the following properties: 1) triangle inequality; 2) rotational invariance, as emphasized in [16]. For convenience, we call $\|\mathcal{X}_k\|$ (s.t. $1 \leq k \leq K$) as an

element of the above norms. Clearly, the Frobenius norm is determined by the sum of the squared elements, i.e., $\sum_{k=1}^{K} \|\mathcal{X}_k\|^2$. In this case, the squared large elements dominate the sum $\sum_{k=1}^{K} \|\mathcal{X}_k\|^2$. Consequently, the Frobenius norm is sensitive to outliers. In comparison, the R_1 norm is determined by the sum of elements (i.e., $\sum_{k=1}^{K} \|\mathcal{X}_k\|$) without being squared. Thus, the R_1 norm is less sensitive to outliers than the Frobenius norm [16].

3. Linear Discriminant Analysis Using the Frobenius norm

3.1. The Classical LDA

Given the L-class training samples $\mathcal{D} = \left\{\{y_i^\ell\}_{i=1}^{N_\ell}\right\}_{\ell=1}^{L}$ with $y_i^\ell \in \mathcal{R}^{D\times 1}$ and $N = \sum_{\ell=1}^{L} N_\ell$, the classical LDA [1][2] aims to find a linear transformation $U \in \mathcal{R}^{D\times\zeta}$ which embeds the original D-dimensional vector y_i^ℓ into the ζ-dimensional vector space $\mathfrak{U}$ such that $\zeta < D$. Let $\mathbf{Tr}(\cdot)$ be the trace of its matrix argument, $S_b^{\mathfrak{U}}$ be the inter-class scatter matrix in $\mathfrak{U}$, and $S_w^{\mathfrak{U}}$ be the intra-class scatter matrix in $\mathfrak{U}$. Thus, the inter-class and intra-class distances in $\mathfrak{U}$ are respectively measured by $\mathbf{Tr}(S_b^{\mathfrak{U}})$ and $\mathbf{Tr}(S_w^{\mathfrak{U}})$, which are formulated as:

$$\begin{aligned}\mathbf{Tr}(S_b^{\mathfrak{U}}) &= \textstyle\sum_{\ell=1}^{L} N_\ell \mathbf{Tr}\left(U^T(m_\ell - m)(m_\ell - m)^T U\right) \\ &= \textstyle\sum_{\ell=1}^{L} N_\ell \|U^T(m_\ell - m)\|^2 = \|\mathscr{B}_*\|^2,\end{aligned} \tag{3}$$

$$\begin{aligned}\mathbf{Tr}(S_w^{\mathfrak{U}}) &= \textstyle\sum_{\ell=1}^{L}\sum_{i=1}^{N_\ell} \mathbf{Tr}\left(U^T(y_i^\ell - m_\ell)(y_i^\ell - m_\ell)^T U\right) \\ &= \textstyle\sum_{\ell=1}^{L}\sum_{i=1}^{N_\ell} \|U^T(y_i^\ell - m_\ell)\|^2 = \|\mathscr{W}_*\|^2,\end{aligned} \tag{4}$$

where $\mathscr{B}_* = \{b_\ell\}_{\ell=1}^{L}$ with b_ℓ being $\sqrt{N_\ell}U^T(m_\ell - m)$, $\mathscr{W}_* = \left\{\{w_{i\ell}\}_{i=1}^{N_\ell}\right\}_{\ell=1}^{L}$ with $w_{i\ell}$ being $U^T(y_i^\ell - m_\ell)$, $m_\ell = \frac{1}{N_\ell}\sum_{i=1}^{N_\ell} y_i^\ell$ is the mean of the samples belonging to the ℓth class, and $m = \frac{1}{N}\sum_{\ell=1}^{L} N_\ell m_\ell$ is the global mean of the training samples. The classical LDA aims to find an optimal transformation U by maximizing $\mathbf{Tr}(S_b^{\mathfrak{U}})$ and minimizing $\mathbf{Tr}(S_w^{\mathfrak{U}})$ simultaneously. Accordingly, we have the following optimization problem:

$$\begin{aligned}\max_U \frac{\mathbf{Tr}\left(S_b^{\mathfrak{U}}\right)}{\mathbf{Tr}(S_w^{\mathfrak{U}})} &= \max_U \frac{\sum_{\ell=1}^{L} N_\ell \|U^T(m_\ell - m)\|^2}{\sum_{\ell=1}^{L}\sum_{i=1}^{N_\ell} \|U^T(y_i^\ell - m_\ell)\|^2} \\ &= \max_U \frac{\|\mathscr{B}_*\|^2}{\|\mathscr{W}_*\|^2}\end{aligned} \tag{5}$$

Typically, the solutions to the above optimization problem can be obtained by performing the following generalized eigenvalue decomposition: $S_b^U x = \lambda S_w^U x$, s.t. $\lambda \neq 0$, where $S_b^U = \sum_{\ell=1}^{L} N_\ell(m_\ell - m)(m_\ell - m)^T$ and $S_w^U = \sum_{\ell=1}^{L}\sum_{i=1}^{N_\ell}(y_i^\ell - m_\ell)(y_i^\ell -$

$m_\ell)^T$. For convenience, let $\mathfrak{B}$ denote the matrix $\left(S_w^U\right)^{-1}S_b^U$. Indeed, the optimal transformation U is formed by the ζ eigenvectors of $\mathfrak{B}$ corresponding to its ζ largest nonzero eigenvalues.

3.2. 2DLDA

Given the L-class training samples $\mathcal{D} = \left\{\{Y_i^\ell\}_{i=1}^{N_\ell}\right\}_{\ell=1}^{L}$ with $Y_i^\ell \in \mathcal{R}^{D_1\times D_2}$ and $N = \sum_{\ell=1}^{L} N_\ell$, 2DLDA [9] is an image-as-matrix learning technique for discriminant analysis in the $(\zeta_1 \times \zeta_2)$-dimensional space $\mathfrak{U}_1 \bigotimes \mathfrak{U}_2$ ($\bigotimes$ denotes the tensor product) such that $\zeta_i < D_i$ for $1 \le i \le 2$. Suppose that $\mathfrak{U}_1$ and $\mathfrak{U}_2$ are spanned by the column vectors of $U_1 \in \mathcal{R}^{D_1\times\zeta_1}$ and $U_2 \in \mathcal{R}^{D_2\times\zeta_2}$, respectively. Thus, the low-dimensional representation of $Y_i^\ell \in \mathcal{R}^{D_1\times D_2}$ in $\mathfrak{U}_1 \bigotimes \mathfrak{U}_2$ is formulated as $U_1^T Y_i^\ell U_2 \in \mathcal{R}^{\zeta_1\times\zeta_2}$. Furthermore, we define $\mathcal{S}_b^{\mathfrak{U}}$ and $\mathcal{S}_w^{\mathfrak{U}}$ as the inter-class and intra-class scatter matrices in the low-dimensional space $\mathfrak{U}_1 \bigotimes \mathfrak{U}_2$, respectively. Like the classical LDA, 2DLDA uses the traces $\mathbf{Tr}(\mathcal{S}_b^{\mathfrak{U}})$ and $\mathbf{Tr}(\mathcal{S}_w^{\mathfrak{U}})$ in $\mathfrak{U}_1 \bigotimes \mathfrak{U}_2$ to measure the inter-class and intra-class distances, respectively. Consequently, 2DLDA aims to find the optimal transformations U_1 and U_2 from

$$\begin{aligned}\max_{U_k|_{k=1}^2} \frac{\mathbf{Tr}(\mathcal{S}_b^{\mathfrak{U}})}{\mathbf{Tr}(\mathcal{S}_w^{\mathfrak{U}})} &= \max_{U_k|_{k=1}^2} \frac{\sum_{\ell=1}^{L} N_\ell \|U_1^T(M_\ell - M)U_2\|^2}{\sum_{\ell=1}^{L}\sum_{i=1}^{N_\ell} \|U_1^T(Y_i^\ell - M_\ell)U_2\|^2} \\ &= \max_{U_k|_{k=1}^2} \frac{\|\mathscr{B}_\circ\|^2}{\|\mathscr{W}_\circ\|^2}\end{aligned}, \tag{6}$$

where $\mathscr{B}_\circ = \{B_\ell\}_{\ell=1}^{L}$ with B_ℓ being $\sqrt{N_\ell}U_1^T(M_\ell - M)U_2$, $\mathscr{W}_\circ = \left\{\{W_{i\ell}\}_{i=1}^{N_\ell}\right\}_{\ell=1}^{L}$ with $W_{i\ell}$ being $U_1^T(Y_i^\ell - M_\ell)U_2$, $M_\ell = \frac{1}{N_\ell}\sum_{i=1}^{N_\ell} Y_i^\ell$ is the mean of the samples belonging to the ℓth class, and $M = \frac{1}{N}\sum_{\ell=1}^{L} N_\ell M_\ell$ is the global mean of the training samples. However, the problem of computing the optimal U_1 and U_2 simultaneously is intractable. Consequently, 2DLDA adopts an iterative procedure to compute the optimal U_1 and U_2 asynchronously [9].

3.3. Discriminant analysis with tensor representation (DATER)

Given the L-class training samples $\mathcal{D} = \left\{\{\mathcal{Y}_i^\ell\}_{i=1}^{N_\ell}\right\}_{\ell=1}^{L}$ with $\mathcal{Y}_i^\ell \in \mathcal{R}^{D_1\times D_2\cdots\times D_n}$ and $N = \sum_{\ell=1}^{L} N_\ell$, DATER [10] is a tensor-based learning technique for discriminant analysis in the $(\zeta_1 \times \zeta_2 \cdots \times \zeta_n)$-dimensional space $\mathfrak{U}_1 \bigotimes \mathfrak{U}_2 \cdots \bigotimes \mathfrak{U}_n$ such that $\zeta_i < D_i$ for $1 \le i \le n$. Let $U_k \in \mathcal{R}^{D_k\times\zeta_k} (1 \le k \le n)$ denote the transformation matrix whose column vectors span the space $\mathfrak{U}_k$. DATER aims to pursue multiple interrelated transformation matrices (i.e., U_k for $1 \le k \le n$), which

maximize the inter-class distances while minimizing the intra-class distances under the tensor metric. More specifically, the criterion for DATER is formulated as:

$$\max_{U_k|_{k=1}^n} \frac{\sum_{\ell=1}^{L} N_\ell \|(\mathcal{M}_\ell - \mathcal{M}) \times_1 U_1 \cdots \times_n U_n\|^2}{\sum_{\ell=1}^{L} \sum_{i=1}^{N_\ell} \|(\mathcal{Y}_i^\ell - \mathcal{M}_\ell) \times_1 U_1 \cdots \times_n U_n\|^2} = \max_{U_k|_{k=1}^n} \frac{\|\mathscr{B}_\triangle\|^2}{\|\mathscr{W}_\triangle\|^2}, \tag{7}$$

where $\mathscr{B}_\triangle = \{\mathcal{B}_\ell\}_{\ell=1}^{L}$ with $\mathcal{B}_\ell$ being $\sqrt{N_\ell}(\mathcal{M}_\ell - \mathcal{M}) \times_1 U_1 \cdots \times_n U_n$, $\mathscr{W}_\triangle = \left\{\{\mathcal{W}_{i\ell}\}_{i=1}^{N_\ell}\right\}_{\ell=1}^{L}$ with $\mathcal{W}_{i\ell}$ being $(\mathcal{Y}_i^\ell - \mathcal{M}_\ell) \times_1 U_1 \cdots \times_n U_n$, $\mathcal{M}_\ell = \frac{1}{N_\ell}\sum_{i=1}^{N_\ell} \mathcal{Y}_i^\ell$ is the mean of the samples belonging to the ℓth class, and $\mathcal{M} = \frac{1}{N}\sum_{\ell=1}^{L} N_\ell \mathcal{M}_\ell$ is the global mean of the training samples. In [10], a novel algorithm called *k-mode Cluster-based Discriminant Analysis* (referred to as KCDA) is developed to iteratively learn the transformation matrices (i.e., $U_k|_{k=1}^n$) by unfolding the tensor along different tensor dimensions. More specifically, KCDA aims to optimize the problem: $\max_{U_k} \frac{\mathbf{Tr}(U_k^T S_B U_k)}{\mathbf{Tr}(U_k^T S_W U_k)}$, where S_B and S_W denote the inter-class and intra-class scatters of the k-mode unfolding matrices, respectively [10].

4. Linear Discriminant Analysis Using the R_1 norm

4.1. R_1 norm based discriminant criterion (DCL_1)

In the classical LDA, 2DLDA, and DATER, the Frobenius norm is applied to characterize the inter-class separability and intra-class compactness. Due to its sensitivity to outliers, the Frobenius norm is incompetent for robust discriminant analysis. In order to address this problem, we propose a novel R_1 norm based discriminant criterion called DCL_1, which uses the R_1 norm to replace the Frobenius norm as the cost function. As a result, the proposed DCL_1 is less sensitive to outliers. The details of DCL_1 are described as follows.

In contrast to the Frobenius norm based discriminant criteria (i.e., Eqs. (5)-(7)), the DCL_1s for vector-based, matrix-based, and tensor-based representations of data are respectively formulated as:

$$\max_{U} \mathcal{J}_\alpha = (1-\alpha)\|\mathscr{B}'_*\|_{R_1} - \alpha\|\mathscr{W}'_*\|_{R_1}, \tag{8}$$

$$\max_{U_k|_{k=1}^2} \mathcal{J}_\beta = (1-\beta)\|\mathscr{B}'_\circ\|_{R_1} - \beta\|\mathscr{W}'_\circ\|_{R_1}, \tag{9}$$

$$\max_{U_k|_{k=1}^n} \mathcal{J}_\gamma = (1-\gamma)\|\mathscr{B}'_\triangle\|_{R_1} - \gamma\|\mathscr{W}'_\triangle\|_{R_1} \tag{10}$$

where α, β, and γ are three control coefficients such that $0 < \alpha, \beta, \gamma < 1$, $\mathscr{B}'_* = \{b'_\ell\}_{\ell=1}^{L}$ with b'_ℓ being $N_\ell U^T(m_\ell - m)$, $\mathscr{W}'_* = \left\{\{w'_{i\ell}\}_{i=1}^{N_\ell}\right\}_{\ell=1}^{L}$ with $w'_{i\ell}$

being $U^T(y_i^\ell - m_\ell)$, $\mathscr{B}'_\circ = \{B'_\ell\}_{\ell=1}^L$ with B'_ℓ being $N_\ell U_1^T(M_\ell - M)U_2$, $\mathscr{W}'_\circ = \left\{\{W'_{i\ell}\}_{i=1}^{N_\ell}\right\}_{\ell=1}^L$ with $W'_{i\ell}$ being $U_1^T(Y_i^\ell - M_\ell)U_2$, $\mathscr{B}'_\triangle = \{\mathcal{B}'_\ell\}_{\ell=1}^L$ with $\mathcal{B}'_\ell$ being $N_\ell(\mathcal{M}_\ell - \mathcal{M}) \times_1 U_1 \cdots \times_n U_n$, and $\mathscr{W}'_\triangle = \left\{\{\mathcal{W}'_{i\ell}\}_{i=1}^{N_\ell}\right\}_{\ell=1}^L$ with $\mathcal{W}'_{i\ell}$ being $(\mathcal{Y}_i^\ell - \mathcal{M}_\ell) \times_1 U_1 \cdots \times_n U_n$. According to the properties of the R_1 norm, Eqs. (8)-(10) can be rewritten as:

$$\begin{aligned}\max_U \mathcal{J}_\alpha = (1-\alpha)\textstyle\sum_{\ell=1}^L N_\ell \sqrt{\|U^T(m_\ell - m)\|^2} \\ -\alpha\textstyle\sum_{\ell=1}^L \sum_{i=1}^{N_\ell} \sqrt{\|U^T(y_i^\ell - m_\ell)\|^2},\end{aligned} \tag{11}$$

$$\begin{aligned}\max_{U_k|_{k=1}^2} \mathcal{J}_\beta = (1-\beta)\textstyle\sum_{\ell=1}^L N_\ell \sqrt{\|U_1^T(M_\ell - M)U_2\|^2} \\ -\beta\textstyle\sum_{\ell=1}^L \sum_{i=1}^{N_\ell} \sqrt{\|U_1^T(Y_i^\ell - M_\ell)U_2\|^2},\end{aligned} \tag{12}$$

$$\begin{aligned}\max_{U_k|_{k=1}^n} \mathcal{J}_\gamma = (1-\gamma)\textstyle\sum_{\ell=1}^L N_\ell \sqrt{\|(\mathcal{M}_\ell - \mathcal{M}) \times_1 U_1 \cdots \times_n U_n\|^2} \\ -\gamma\textstyle\sum_{\ell=1}^L \sum_{i=1}^{N_\ell} \sqrt{\|(\mathcal{Y}_i^\ell - \mathcal{M}_\ell) \times_1 U_1 \cdots \times_n U_n\|^2},\end{aligned} \tag{13}$$

Based on Eqs. (11)-(13), three DCL_1-based subspace learning algorithms (i.e., $1DL_1$, $2DL_1$, and TDL_1) are further developed for vector-based, matrix-based, and tensor-based representations of data, respectively. The details of the three algorithms are given as follows.

4.2. $1DL_1$

$1DL_1$ aims to find the optimal transformation matrix $U \in \mathcal{R}^{D\times\zeta}$ (ζ is the final lower dimension) to maximize the objective function $\mathcal{J}_\alpha$ in Eq. (11). It is noted that $U \in \mathcal{R}^{D\times\zeta}$ is an orthogonal matrix such that $U^TU = I_\zeta$ with I_ζ being a $\zeta \times \zeta$ identity matrix. Consequently, we have the following constrained optimization problem:

$$\begin{aligned}&\max_U \mathcal{J}_\alpha, \ \text{s.t. } U^TU = I_\zeta, \\ &\Longrightarrow \max_U \mathcal{L}_\alpha = \mathcal{J}_\alpha + \tfrac{1}{2}\mathbf{Tr}\left[\Lambda(I_\zeta - U^TU)\right],\end{aligned} \tag{14}$$

where the Lagrange multiplier Λ is a diagonal matrix and $\mathcal{J}_\alpha$ is defined in Eq. (11). For simplicity, we rewrite $\mathcal{J}_\alpha$ as:

$$\begin{aligned}J_\alpha = (1-\alpha)\textstyle\sum_{\ell=1}^L N_\ell \sqrt{\mathbf{Tr}\left(U^T(m_\ell - m)(m_\ell - m)^TU\right)} \\ -\alpha\textstyle\sum_{\ell=1}^L \sum_{i=1}^{N_\ell} \sqrt{\mathbf{Tr}\left(U^T(y_i^\ell - m_\ell)(y_i^\ell - m_\ell)^TU\right)}.\end{aligned} \tag{15}$$

Algorithm: $1DL_1$

Input: $\left\{\{y_i^\ell \in \mathcal{R}^{D\times 1}\}_{i=1}^{N_\ell}\right\}_{\ell=1}^{L}$, $\mathbb{T}_{max}^*$, α, and ζ for $\zeta < D$

Output: $U \in \mathcal{R}^{D\times\zeta}$ and $\left\{\{z_i^\ell \in \mathcal{R}^{\zeta\times 1}\}_{i=1}^{N_\ell}\right\}_{\ell=1}^{L}$

1. Initialize $U = (I_\zeta, 0)^T \in \mathcal{R}^{D\times\zeta}$, where I_ζ is a $\zeta \times \zeta$ identity matrix;
2. For $t = 1$ to $\mathbb{T}_{max}^*$
3. Use the current U to compute F_α from Eq. (17);
4. Perform the eigenvalue decomposition $F_\alpha U = U\Lambda$ to update the U;
5. If U converges, break;
6. EndFor
7. $z_i^\ell \leftarrow U^T y_i^\ell$ for $i \in \{1, \ldots, N_\ell\}$ and $\ell \in \{1, \ldots, L\}$.

Figure 1: **The specific procedure of the** $1DL_1$**.**

According to the Karush-Kuhn-Tucker (**KKT**) conditions for the optimal solution, we have:

$$\frac{\partial \mathcal{L}_\alpha}{\partial U} = F_\alpha U - U\Lambda = 0 \quad \Longrightarrow \quad F_\alpha U = U\Lambda, \tag{16}$$

where

$$\begin{aligned} F_\alpha = (1-\alpha)\textstyle\sum_{\ell=1}^{L} \frac{N_\ell (m_\ell - m)(m_\ell - m)^T}{\sqrt{\mathbf{Tr}(U^T(m_\ell - m)(m_\ell - m)^T U)}} \\ - \alpha \textstyle\sum_{\ell=1}^{L} \sum_{i=1}^{N_\ell} \frac{(y_i^\ell - m_\ell)(y_i^\ell - m_\ell)^T}{\sqrt{\mathbf{Tr}\big(U^T(y_i^\ell - m_\ell)(y_i^\ell - m_\ell)^T U\big)}}. \end{aligned} \tag{17}$$

In this way, we derive an iterative algorithm for computing the optimal U. More specifically, given the current U, we can update the U by performing the eigenvalue decomposition $F_\alpha U = U\Lambda$. The specific procedure of $1DL_1$ is listed in Fig 1.

4.3. $2DL_1$

$2DL_1$ aims to find two optimal transformation matrices (i.e., $U_k|_{k=1}^2 \in \mathcal{R}^{D_k\times\zeta_k}$ with $\zeta_1 \times \zeta_2$ being the final lower dimensions) to maximize the objective function $\mathcal{J}_\beta$ in Eq. (12). It is noted that U_k ($1 \le k \le 2$) is an orthogonal matrix such that $U_k^T U_k = I_{\zeta_k}$ with I_{ζ_k} being a $\zeta_k \times \zeta_k$ identity matrix. Thus, we have the following

Algorithm: $2DL_1$

Input: $\left\{\{Y_i^\ell \in \mathcal{R}^{D_1 \times D_2}\}_{i=1}^{N_\ell}\right\}_{\ell=1}^{L}$, $\mathbb{T}_{max}^{\circ}$, β, and $\{\zeta_k\}_{k=1}^{2}$ for $\zeta_k < D_k$

Output: $\{U_k \in \mathcal{R}^{D_k \times \zeta_k}\}_{k=1}^{2}$ and $\left\{\{Z_i^\ell \in \mathcal{R}^{\zeta_1 \times \zeta_2}\}_{i=1}^{N_\ell}\right\}_{\ell=1}^{L}$

1. Initialize $\{U_k = (I_{\zeta_k}, 0)^T \in \mathcal{R}^{D_k \times \zeta_k}\}_{k=1}^{2}$, where I_{ζ_k} is a $\zeta_k \times \zeta_k$ identity matrix;
2. For $t = 1$ to $\mathbb{T}_{max}^{\circ}$
3. Use the current $\{U_k\}_{k=1}^{2}$ to compute F_{β_1} from Eq. (21);
4. Perform the eigenvalue decomposition $F_{\beta_1}U_1 = U_1\Omega_1$ to update the U_1;
5. Use the current $\{U_k\}_{k=1}^{2}$ to compute F_{β_2} from Eq. (22);
6. Perform the eigenvalue decomposition $F_{\beta_2}U_2 = U_2\Omega_2$ to update the U_2;
7. If U_k converges for $k \in \{1, 2\}$, break;
8. EndFor
9. $Z_i^\ell \leftarrow U_1^T Y_i^\ell U_2^T$ for $i \in \{1, \ldots, N_\ell\}$ and $\ell \in \{1, \ldots, L\}$.

Figure 2: **The specific procedure of the** $2DL_1$**.**

constrained optimization problem:

$$\begin{array}{l} \max\limits_{U_k|_{k=1}^2} \mathcal{J}_\beta, \text{ s.t. } U_k^T U_k = I_{\zeta_k} \text{ and } k \in \{1, 2\} \\ \Longrightarrow \max\limits_{U_k|_{k=1}^2} \mathcal{L}_\beta = \mathcal{J}_\beta + \frac{1}{2}\sum_{k=1}^{2} \mathbf{Tr}\left[\Omega_k(I_{\zeta_k} - U_k^T U_k)\right], \end{array} \tag{18}$$

where the Lagrange multiplier Ω_k ($1 \leq k \leq 2$) is a diagonal matrix and $\mathcal{J}_\beta$ is defined in Eq. (12). For simplicity, we rewrite $\mathcal{L}_\beta$ as:

$$\begin{array}{c} \mathcal{J}_\beta = (1-\beta)\sum_{\ell=1}^{L} N_\ell \sqrt{\mathbf{Tr}\left(U_1^T(M_\ell - M)U_2U_2^T(M_\ell - M)^T U_1\right)} \\ -\beta\sum_{\ell=1}^{L}\sum_{i=1}^{N_\ell}\sqrt{\mathbf{Tr}\left(U_1^T(Y_i^\ell - M_\ell)U_2U_2^T(Y_i^\ell - M_\ell)^T U_1\right)}. \end{array} \tag{19}$$

According to the Karush-Kuhn-Tucker (**KKT**) conditions for the optimal solution and the property $\mathbf{Tr}(\mathbf{AB}) = \mathbf{Tr}(\mathbf{BA})$, we have:

$$\begin{array}{c} \frac{\partial \mathcal{L}_\beta}{\partial U_1} = F_{\beta_1}U_1 - U_1\Omega_1 = 0, \quad \frac{\partial \mathcal{L}_\beta}{\partial U_2} = F_{\beta_2}U_2 - U_2\Omega_2 = 0; \\ \Longrightarrow F_{\beta_1}U_1 = U_1\Omega_1, \quad F_{\beta_2}U_2 = U_2\Omega_2; \end{array} \tag{20}$$

Algorithm: TDL_1

Input: $\left\{\{\mathcal{Y}_i^\ell \in \mathcal{R}^{D_1\times D_2\ldots\times D_n}\}_{i=1}^{N_\ell}\right\}_{\ell=1}^{L}$, $\mathbb{T}_{max}^{\triangle}$, γ, and $\{\zeta_k\}_{k=1}^n$ for $\zeta_k < D_k$

Output: $\left\{\{\mathcal{Z}_i^\ell \in \mathcal{R}^{\zeta_1\ldots\times\zeta_n}\}_{i=1}^{N_\ell}\right\}_{\ell=1}^{L}$, $\{U_k \in \mathcal{R}^{D_k\times\zeta_k}\}_{k=1}^n$

1. Initialize $\{U_k = (I_{\zeta_k}, 0)^T \in \mathcal{R}^{D_k\times\zeta_k}\}_{k=1}^n$, where I_{ζ_k} is a $\zeta_k \times \zeta_k$ identity matrix;
2. For $t = 1$ to $\mathbb{T}_{max}^{\triangle}$
3. For $k = 1$ to n
4. Use the current $U_k|_{k=1}^n$ to compute F_{γ_k} from Eq. (26);
5. Perform the eigenvalue decomposition $F_{\gamma_k}U_k{=}U_k\Gamma_k$ to update the U_k;
6. EndFor
7. If U_k converges for $k \in \{1, 2, \ldots, n\}$, break;
8. EndFor
9. $\mathcal{Z}_i^\ell \leftarrow \mathcal{Y}_i^\ell \times_1 U_1 \times_2 U_2 \cdots \times_n U_n$ for $i \in \{1, 2, \ldots, N_\ell\}$ and $\ell \in \{1, \ldots, L\}$.

Figure 3: **The specific procedure of the TDL_1.**

where

$$\begin{aligned} F_{\beta_1} = (1-\beta)\textstyle\sum_{\ell=1}^{L} \frac{N_\ell(M_\ell - M)U_2U_2^T(M_\ell - M)^T}{\mathbf{Tr}\left(U_1^T(M_\ell - M)U_2U_2^T(M_\ell - M)^TU_1\right)} \\ -\beta\textstyle\sum_{\ell=1}^{L}\sum_{i=1}^{N_\ell} \frac{(Y_i^\ell - M_\ell)U_2U_2^T(Y_i^\ell - M_\ell)^T}{\mathbf{Tr}\left(U_1^T(Y_i^\ell - M_\ell)U_2U_2^T(Y_i^\ell - M_\ell)^TU_1\right)}, \end{aligned} \tag{21}$$

$$\begin{aligned} F_{\beta_2} = (1-\beta)\textstyle\sum_{\ell=1}^{L} \frac{N_\ell(M_\ell - M)^TU_1U_1^T(M_\ell - M)}{\mathbf{Tr}\left(U_2^T(M_\ell - M)^TU_1U_1^T(M_\ell - M)U_2\right)} \\ -\beta\textstyle\sum_{\ell=1}^{L}\sum_{i=1}^{N_\ell} \frac{(Y_i^\ell - M_\ell)^TU_1U_1^T(Y_i^\ell - M_\ell)}{\mathbf{Tr}\left(U_2^T(Y_i^\ell - M_\ell)^TU_1U_1^T(Y_i^\ell - M_\ell)U_2\right)}. \end{aligned} \tag{22}$$

Consequently, we derive an iterative algorithm for computing the optimal $U_k|_{k=1}^2$. More specifically, given the current $U_k|_{k=1}^2$, we can update the U_k iteratively by performing the eigenvalue decomposition $F_{\beta_k}U_k = U_k\Omega_k$ in Eq. (20). The specific procedure of $2DL_1$ is listed in Fig 2.

4.4. TDL_1

TDL_1 aims to seek for n optimal transformation matrices (i.e., $U_k|_{k=1}^n \in \mathcal{R}^{D_k\times\zeta_k}$ with $\zeta_1 \times \zeta_2 \cdots \times \zeta_n$ being the final lower dimensions) to maximize the objective function $\mathcal{J}_\gamma$ in Eq. (13). It is noted that U_k ($1 \le k \le n$) is an orthogonal matrix such that $U_k^TU_k = I_{\zeta_k}$ with I_{ζ_k} being a $\zeta_k \times \zeta_k$ identity matrix. For convenience, define $\mathcal{G}_k^\ell$ as $(\mathcal{M}_\ell - \mathcal{M}) \times_1 U_1 \cdots \times_{k-1} U_{k-1} \times_{k+1} U_{k+1} \cdots \times_n U_n$,

define $\mathrm{G}_{(k)}^{\ell}$ as the k-mode unfolding matrix of $\mathcal{G}_k^{\ell}$, define $\mathcal{H}_k^{\ell}$ as $(\mathcal{Y}_i^{\ell} - \mathcal{M}_{\ell}) \times_1 U_1 \cdots \times_{k-1} U_{k-1} \times_{k+1} U_{k+1} \cdots \times_n U_n$, and define $\mathrm{H}_{(k)}^{\ell}$ as the k-mode unfolding matrix of $\mathcal{H}_k^{\ell}$. Thus, the objective function $\mathcal{J}_{\gamma}$ in Eq. (13) can be rewritten as:

$$\begin{aligned}
\mathcal{J}_{\gamma} &= (1-\gamma)\textstyle\sum_{\ell=1}^{L} N_{\ell}\sqrt{\|\mathcal{G}_k^{\ell} \times_k U_k\|^2} \\
&\quad -\gamma\textstyle\sum_{\ell=1}^{L}\sum_{i=1}^{N_{\ell}}\sqrt{\|\mathcal{H}_k^{\ell} \times_k U_k\|^2} \\
&= (1-\gamma)\textstyle\sum_{\ell=1}^{L} N_{\ell}\sqrt{\|(U_k)^T \mathbf{G}_{(k)}^{\ell}\|^2} \\
&\quad -\gamma\textstyle\sum_{\ell=1}^{L}\sum_{i=1}^{N_{\ell}}\sqrt{\|(U_k)^T \mathbf{H}_{(k)}^{\ell}\|^2} \\
&= (1-\gamma)\textstyle\sum_{\ell=1}^{L} N_{\ell}\sqrt{\mathbf{Tr}\left((U_k)^T\, \mathbf{G}_{(k)}^{\ell}\, (\mathbf{G}_{(k)}^{\ell})^T U_k\right)} \\
&\quad -\gamma\textstyle\sum_{\ell=1}^{L}\sum_{i=1}^{N_{\ell}}\sqrt{\mathbf{Tr}\left((U_k)^T\, \mathrm{H}_{(k)}^{\ell}(\mathrm{H}_{(k)}^{\ell})^T U_k\right)}
\end{aligned} \tag{23}$$

Please refer to [10] for the details of the aforementioned tensor operations. Consequently, we have the following constrained optimization problem:

$$\begin{aligned}
&\max_{U_k|_{k=1}^n} \mathcal{J}_{\gamma}, \text{ s.t. } U_k^T U_k = I_{\zeta_k} \text{ and } k \in \{1, 2, \ldots, n\} \\
&\Longrightarrow \max_{U_k|_{k=1}^n} \mathcal{L}_{\gamma} = \mathcal{J}_{\gamma} + \tfrac{1}{2}\textstyle\sum_{k=1}^{n} \mathbf{Tr}\left[\Gamma_k(I_{\zeta_k} - U_k^T U_k)\right],
\end{aligned} \tag{24}$$

where the Lagrange multiplier Γ_k $(1 \le k \le n)$ is a diagonal matrix. According to the Karush-Kuhn-Tucker (**KKT**) conditions for the optimal solution, we have:

$$\frac{\partial \mathcal{L}_{\gamma}}{\partial U_k} = F_{\gamma_k} U_k - U_k \Gamma_k = 0 \quad \Longrightarrow \quad F_{\gamma_k} U_k = U_k \Gamma_k \tag{25}$$

where

$$\begin{aligned}
F_{\gamma_k} &= (1-\gamma)\textstyle\sum_{\ell=1}^{L} \frac{N_{\ell}\mathbf{G}_{(k)}^{\ell}\,(\mathbf{G}_{(k)}^{\ell})^T}{\sqrt{\mathbf{Tr}\left((U_k)^T \mathbf{G}_{(k)}^{\ell}\,(\mathbf{G}_{(k)}^{\ell})^T U_k\right)}} \\
&\quad -\gamma\textstyle\sum_{\ell=1}^{L}\sum_{i=1}^{N_{\ell}} \frac{\mathbf{H}_{(k)}^{\ell}(\mathbf{H}_{(k)}^{\ell})^T}{\sqrt{\mathbf{Tr}\left((U_k)^T \mathbf{H}_{(k)}^{\ell}(\mathbf{H}_{(k)}^{\ell})^T U_k\right)}}.
\end{aligned} \tag{26}$$

As a result, an iterative strategy is adopted to compute the optimal $U_k|_{k=1}^n$. The specific procedure of TDL_1 is listed in Fig 3.

4.5. Convergence analysis

Since TDL_1 is a higher-order generalization of $1DL_1$ and $2DL_1$, we only need to make a convergence analysis of TDL_1 in theory. Based on [26], we give

the convergence proof of TDL_1 as follows. For convenience, we can reformulate the objective function of TDL_1 as $\mathcal{J}_\gamma = f(U_1, U_2, \ldots, U_n) = f(U_k|_{k=1}^n)$. Here, f is a continuous function defined as:

$$f : \mathcal{U}_1 \times \mathcal{U}_2 \times \cdots \times \mathcal{U}_n = \overset{n}{\underset{k=1}{\times}} \mathcal{U}_k \to \mathcal{R}, \tag{27}$$

where $U_k \in \mathcal{U}_k$ and $\mathcal{U}_k$ is the set which includes all possible U_k. According to Eq. (27), f has n different mappings formulated as:

$$\begin{aligned} U_k^* &\overset{\triangle}{=} \arg\max_{U_k \in \mathcal{U}_k} f(U_k|_{k=1}^n) \\ &= \arg\max_{U_k \in \mathcal{U}_k} f(U_k; U_l|_{l=1}^{k-1}, U_l|_{l=k+1}^n), \end{aligned} \tag{28}$$

where $1 \leq k \leq n$. The solution U_k^* to Eq. (28) can be computed with the given $U_l|_{l=1}^{k-1}$ in the t-th iteration and $U_l|_{l=k+1}^n$ in the $(t-1)$-th iteration of the for-loop in Step 5 in Fig. 3. Given an initial solution $U_k \in \mathcal{U}_k$ for $1 \leq k \leq n$, the optimization procedure of Eq. (28) can generate a sequence of items $\{U_{k,t}^* | 1 \leq k \leq n\}$. The sequence has the following relationship:

$$\begin{aligned} & f(U_{1,1}^*) \leq f(U_{2,1}^*) \\ & \leq \cdots \leq f(U_{n,1}^*) \leq f(U_{1,2}^*) \\ & \leq \cdots \leq f(U_{1,t}^*) \leq f(U_{2,t}^*) \\ & \leq \cdots \leq f(U_{1,\mathbb{T}}^*) \leq f(U_{2,\mathbb{T}}^*) \\ & \leq \cdots \leq f(U_{n,\mathbb{T}}^*). \end{aligned} \tag{29}$$

As $\mathbb{T} \to +\infty$, f increases monotonically. On the other hand, the upper bound of the TDL_1's objective function can be analyzed as follows:

$$\|\mathcal{J}_\gamma\| = \|f(U_k|_{k=1}^n)\| = \|(1-\gamma)\textstyle\sum_{\ell=1}^L N_\ell \mathcal{A}_\ell + \gamma \sum_{\ell=1}^L \sum_{i=1}^{N_\ell} \mathcal{B}_\ell^i\|, \tag{30}$$

where $\mathcal{A}_\ell = \|(\mathcal{M}_\ell - \mathcal{M}) \times_1 U_1 \cdots \times_n U_n\|$ and $\mathcal{B}_\ell^i = \|(\mathcal{Y}_i^\ell - \mathcal{M}_\ell) \times_1 U_1 \cdots \times_n U_n\|$. Apparently, the following relationship holds:

$$\begin{aligned} & \|f(U_k|_{k=1}^n)\| \leq (1-\gamma)\textstyle\sum_{\ell=1}^L N_\ell \|\mathcal{A}_\ell\| + \gamma \sum_{\ell=1}^L \sum_{i=1}^{N_\ell} \|\mathcal{B}_\ell^i\| \\ & \leq \left(\textstyle\prod_{k=1}^n \|U_k\|\right) \left((1-\gamma)\textstyle\sum_{\ell=1}^L N_\ell \|\mathcal{M}_\ell - \mathcal{M}\| + \gamma \sum_{\ell=1}^L \sum_{i=1}^{N_\ell} \|\mathcal{Y}_i^\ell - \mathcal{M}_\ell\|\right) \\ & \leq \left(\textstyle\prod_{k=1}^n \sqrt{\zeta_k}\right) \max\left(\textstyle\sum_{\ell=1}^L N_\ell \|\mathcal{M}_\ell - \mathcal{M}\|, \sum_{\ell=1}^L \sum_{i=1}^{N_\ell} \|\mathcal{Y}_i^\ell - \mathcal{M}_\ell\|\right). \end{aligned} \tag{31}$$

Consequently, the TDL_1's objective function has an upper bound. According to Eqs. (29) and (31), it is proved that TDL_1 converges.

Figure 4: **Exemplar training images for each person with five representative "real" images and one representative outlier image.**

5. Experiments

In order to evaluate the performances of the proposed algorithms, five datasets are used in the experiments. The first dataset is a toy set composed of ten samples categorized into two classes with an additional outlier sample. The second dataset is the 20 Newsgroups text dataset[1], which consists of 18941 documents from 20 classes. To efficiently make classification performance evaluations, we randomly split this text dataset into 20 subsets, each of which is generated by randomly selecting 20% of the original samples class by class. The other three benchmark datasets are three face recognition datasets: ORL, Yale, and PIE, respectively. More specifically, ORL[2] consists of 400 face images of 40 persons. Each person has 10 images. Yale[3] is composed of 165 images of 15 persons. Each person has 11 images. PIE is a subset of the CMU-PIE face dataset[4]. This sub-dataset contains 11560 images of 68 persons. Each person has 170 images.

Three experiments are conducted to demonstrate the superiority of the proposed $1DL_1$, $2DL_1$, and TDL_1. Specifically, the first two datasets (i.e., toy and 20 Newsgroups) are respectively used in the first two experiments. The three face datasets (i.e., ORL, Yale, and PIE) are used in the last experiment. In the experiments, each image is normalized to 32×32 pixels. For DATER and TDL_1, 40 Gabor features with five different scales and eight different directions are extracted for each image encoded as a 3rd order Gabor tensor of size $32 \times 32 \times 40$. Furthermore, α, β, and γ respectively in $1DL_1$, $2DL_1$, and TDL_1 are set to 0.20, 0.15, and 0.30, respectively. $\mathbb{T}^*_{max}$, $\mathbb{T}^\circ_{max}$, and $\mathbb{T}^\triangle_{max}$ respectively in $1DL_1$, $2DL_1$, and TDL_1 are all set to 20. In order to better evaluate the effectiveness of the various descriptors, the simple 1-Nearest-Neighbor (1NN) classifier is used for

[1] Available at http://people.csail.mit.edu/jrennie/20Newsgroups/

[2] Available at http://www.cl.cam.ac.uk/research/dtg/attarchive/ facesataglance.html

[3] Available at http://cvc.yale.edu/projects/yalefaces/yalefaces.html

[4] Available at http://www.ri.cmu.edu/projects/project_418.html

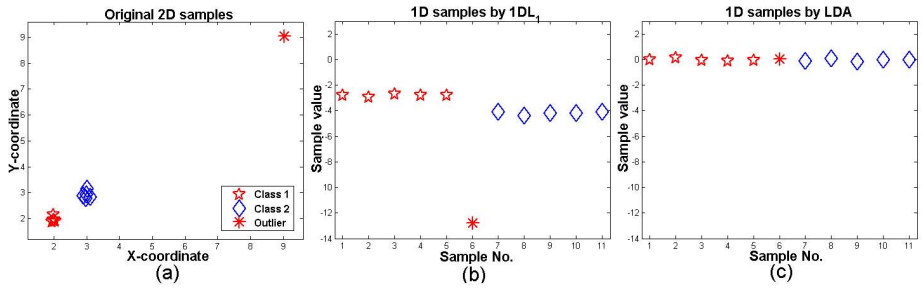

Figure 5: **Learning performances of $1DL_1$ and the classical LDA over the toy dataset. (a) shows the original dataset with an outlier. (b) and (c) display the discriminant learning results of $1DL_1$ and the classical LDA, respectively.**

final classification. Moreover, 5-fold cross validation is used for quantitative classification performance evaluations. For each person, several white noise images are generated as outlier training images. For a better illustration, some exemplar training images are shown in Fig. 4, including one representative outlier image and five representative "real" images selected from the training dataset.

The first experiment over the toy dataset (as shown in Fig. 5(a)) is to compare the dimension reduction performances between $1DL_1$ and the classical LDA. In order to demonstrate the effectiveness (i.e., insensitivity to outliers) of $1DL_1$, we intentionally include the outlier sample (plotted as "$*$" at the top-right corner of Fig. 5(a)) into the training samples of Class 1 before the binary classification. For $1DL_1$, D is equal to 2, and ζ is set as 1. For the classical LDA, ζ is also set as 1. The final learning results are plotted as two 1-dimensional signals in Figs. 5 (b) and (c) corresponding to $1DL_1$ and the classical LDA, respectively. In Figs. 5 (b) and (c), the pentacle-like and diamond-like samples correspond to the samples of Class 1 and Class 2, respectively, after the step of dimension reduction. Clearly, the inter-class scatter (i.e., 0.51) of the two-class samples except for the outlier sample in Fig. 5(b) is much larger than that (i.e., 0.01) in Fig. 5(c).

The second experiment is to evaluate the classification performances of $1DL_1$ and the classical LDA over the subsets of the 20 Newsgroups text dataset. In this dataset, each document is represented as a 26214-dimensional term frequency (TF) feature vector. Since each TF feature point lies in a high-dimensional nonlinear feature space, we need to embed it into a low-dimensional linear feature space. Motivated by this, a graph-based dimension reduction technique (i.e. Laplacian Eigenmaps [27]) is used. The edge weight of the adjacency graph is computed as the cosine similarity, i.e., $\frac{A \cdot B}{\|A\|\|B\|}$. In practice, the embedding dimension is 50. In this case, linear discriminant analysis by $1DL_1$ and the classical LDA is per-

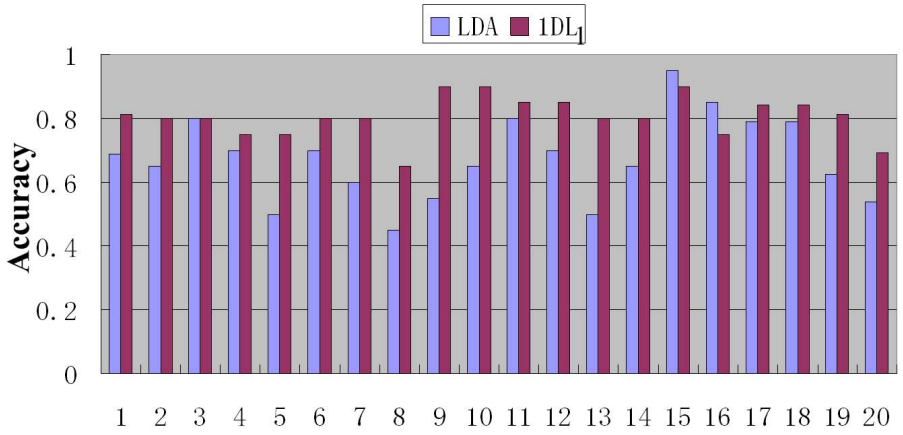

Figure 6: **Class-by-class learning performances of $1DL_1$ and the classical LDA over the 20 Newsgroups text dataset.**

formed in the 50-dimensional feature space. The final classification results are shown in Fig. 6, where x-axis corresponds to the class index and y-axis is associated with the classification accuracy. Clearly, $1DL_1$ always performs better than the classical LDA. Quantitatively, the relative gain of $1DL_1$ vs. the classical LDA is 22.91% on average.

The last experiment over the three datasets (i.e., ORL, Yale, and PIE) is to compare the classification performances of $1DL_1$, $2DL_1$, and TDL_1 with those of the classical LDA, 2DLDA, and DATER, respectively. For $1DL_1$, D is equal to 1024, and ζ is set as 26. For $2DL_1$, both D_1 and D_2 are equal to 32. ζ_1 and ζ_2 are both set as 18. For TDL_1, D_1, D_2, and D_3 are equal to 32, 32, and 40. ζ_1, ζ_2, and ζ_3 are all set as 18. For the classical LDA, 2DLDA, and DATER, the corresponding settings of ζ, $\zeta_k|_{k=1}^{2}$, and $\zeta_\ell|_{\ell=1}^{3}$ are the same as those for $1DL_1$, $2DL_1$, and TDL_1, respectively. The final learning results are shown in Figs. 7(a)-(c) corresponding to ORL, Yale, and PIE, respectively. From Figs. 7(a)-(c), it is clear that the classification accuracies of the proposed $1DL_1$, $2DL_1$, and TDL_1 are much higher than those of the classical LDA, 2DLDA, and DATER. Also, it is seen from Figs. 7(a)-(c) that the classification performances of the different data representations follow the descending order of the tensor-based, matrix-based, and vector-based ones.

Furthermore, we give a face classification example over ORL using $2DL_1$. This example aims to make performance evaluations of $2DL_1$ in the following aspects: (i) the sensitivity to β; (ii) the reconstruction effect. For (i), β is se-

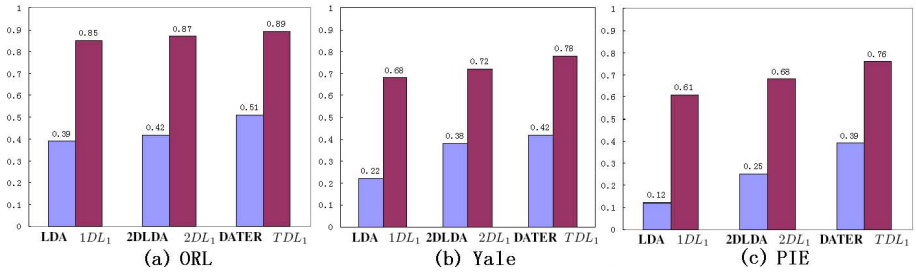

Figure 7: **Classification results of different learning methods. The x-axis corresponds to the learning method while the y-axis is associated with the average classification accuracy after 5-fold cross validation. (a), (b), and (c) correspond to the three datasets—ORL, Yale, and PIE, respectively.**

lected from nineteen different numbers within the range of $[0.05, 0.95]$. The interval between two adjacent numbers is fixed to be 0.05. For each number, $2DL_1$ is implemented to learn two projection matrices—U_1 and U_2. Using the 1-Nearest-Neighbor classifier and 5-fold cross validation technique, we obtain the average classification accuracy, as shown in Fig. 8. Clearly, it is seen that $2DL_1$ is insensitive to β. For (ii), the reconstruction image of Y is obtained by: $\mathbb{Y} = U_1U_1^TYU_2U_2^T$. For a better illustration, we compare the reconstruction performances of 2DLDA and $2DL_1$. The final reconstruction result is displayed in Fig. 9. Apparently, $2DL_1$ is robust to outliers while 2DLDA is affected by outiers.

In summary, we observe that $1DL_1$, $2DL_1$, and TDL_1 substantially outperform the classical LDA, 2DLDA, and DATER in the presence of outliers, owing to the proposed discriminant criterion DCL_1. Therefore, $1DL_1$, $2DL_1$, and TDL_1 are really promising algorithms for supervised subspace learning.

6. Conclusion

In this paper, we have proposed a novel discriminant criterion called DCL_1 that better characterizes the intra-class compactness and the inter-class separability by using the R_1 norm instead of the Frobenius norm. Based on the DCL_1, three subspace learning algorithms ($1DL_1$, $2DL_1$, and TDL_1) have been developed for the vector-based, matrix-based, tensor-based representations of data, respectively. Compared with the classical LDA [1], 2DLDA [9], and DATER [10], the developed $1DL_1$, $2DL_1$, and TDL_1 are able to reduce the influence of outliers substantially. Experimental results have demonstrated the superiority of the proposed DCL_1 and its algorithms in the existing literature.

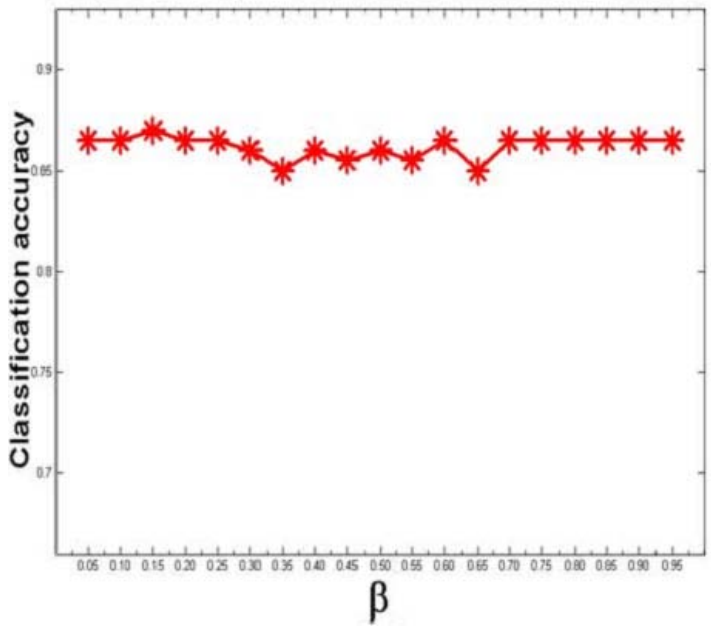

Figure 8: **Classification performance of $2DL_1$. This figure plots the average classification accuracy curve of $2DL_1$ in different cases of β.**

References

[1] R. O. Duda, P. E. Hart, and D. Stork, *Pattern Classification,* Wiley, 2000.

[2] Geoffrey J. McLachlan, *Discriminant Analysis and Statistical Pattern Recognition,* New York: Wiley, 1992.

[3] K. Fukunaga, *Introduction to Statistical Pattern Recognition (2nd),* Academic Press, 1990.

[4] F. Chen, H. Y. Liao, M. T. Ko, J. C. Lin, and G. J. Yu, "A New LDA-based Face Recognition System which Can Solve the Small Sample Size Problem," *Pattern Recognition,* Vol. 33, No. 10, pp. 1,713-1,726, 2000.

[5] H. Yu and J. Yang, "A Direct LDA Algorithm for High-dimensional Data with Application to Face Recognition," *Pattern Recognition,* Vol. 34, No. 12, pp. 2,067-2,070, 2001.

[6] X. Wang and X. Tang, "Dual-Space Linear Discriminant Analysis for Face Recognition," in *Proc. CVPR,* Vol. 2, pp. 564-569, 2004.

[7] P. Belhumeur, J. Hespanha, and D. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection," *IEEE Trans. Pattern Analysis and Machine Intelligence,* Vol. 19, pp. 711-720, 1997.

[8] D. L. Swets and J. Weng, "Using Discriminant Eigenfeatures for Image Retrieval," *IEEE*

Figure 9: **Reconstruction performances of $2DL_1$ and 2DLDA. Each red box corresponds to a person. The first row of the box shows one representative outlier image and five representative face images selected from the training dataset; the second row displays the reconstruction images by $2DL_1$; the last row exhibits the reconstruction images by 2DLDA.**

Trans. Pattern Analysis and Machine Intelligence, Vol. 18, pp. 831-836, 1996.

[9] J. Ye, R. Janardan, and Q. Li, "Two-Dimensional Linear Discriminant Analysis," *NIPS,* Vol. 2, pp. 1,569-1,576, 2004.

[10] S. Yan, D. Xu, Q. Yang, L. Zhang, X. Tang, and H. Zhang, "Discriminant analysis with tensor representation," in *Proc. CVPR,* 2005.

[11] X. He, D.Cai, and P. Niyogi, "Tensor subspace analysis," *NIPS,* 2005.

[12] H. Wang, S. Yan, T. Huang and X. Tang, "A Convergent Solution to Tensor Subspace Learning," in *Proc. IJCAI,* 2007.

[13] D. Tao, X. Li, X. Wu, and S. J. Maybank, "General Tensor Discriminant Analysis and Gabor Features for Gait Recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence,* Vol. 29, pp. 1,700-1,715, 2007.

[14] Z. Lei, R. Chu, R. He, S. Liao, and S. Z. Li, "Face Recognition by Discriminant Analysis with Gabor Tensor Representation," in *Proc. IAPR/IEEE International Conference on Biometrics,* 2007.

[15] M. A. O. Vasilescu and D. Terzopoulos, "Multilinear Subspace Analysis of Image Ensembles," in *Proc. CVPR,* Vol. 2, pp.93-99, June 2003.

[16] C. H. Q. Ding, D. Zhou, X. He, and H. Zha, "R_1-PCA: Rotational Invariant L_1-Norm Principal Component Analysis for Robust Subspace Factorization," *ICML,* pp.281-288, 2006.

[17] H. Huang and C. Ding, "Robust Tensor Factorization Using R1 Norm," in *Proc. CVPR,* 2008.

[18] M. Li and B. Yuan, "2d-lda: A novel statistical linear discriminant analysis for image matrix," *Pattern Recognition Letters,* Vol. 26, Iss. 5, pp. 527-532, 2005.

[19] D. Xu, S.C. Yan, L. Zhang, H.J. Zhang, Z.K. Liu, and H.Y. Shum, "Concurrent Subspaces Analysis," in *Proc. CVPR,* pp. 203-208, 2005.

[20] Y. Pang, X. Li, and Y. Yuan, "Robust Tensor Analysis with L1-Norm," *IEEE Trans. Circuits and Systems for Video Technology*, Vol. 20, No. 2, pp.172-178, 2010.

[21] T. Zhou, "Manifold elastic net for sparse learning," in *Proc. IEEE International Conference on Systems, Man, and Cybernetics*, pp. 3699-3704, 2009.

[22] Y. Pang and Y. Yuan, "Outlier-resisting graph embedding," *Neurocopmuting*, Vol. 4-6, pp. 968-974, 2010.

[23] T. Zhang, D. Tao, and J. Yang, "Discriminative Locality Alignment," in *Proc. ECCV*, pp. 725-738, 2008.

[24] W. Liu, D. Tao, and J. Liu, "Transductive Component Analysis," in *Proc. ICDM*, pp. 433-442, 2008.

[25] D. Tao, X. Li, X. Wu, and S. J. Maybank, "Geometric Mean for Subspace Selection," *IEEE Trans. Pattern Analysis Machine Intelligence*, Vol. 31, No.2, pp. 260-274, 2009.

[26] D. Tao, X. Li, X. Wu, W. Hu, and S. J. Maybank, "Supervised tensor learning," Knowl. Inf. Syst. 13(1): 1-42, 2007.

[27] M. Belkin and P. Niyogi, "Laplacian Eigenmaps for Dimensionality Reduction and Data Representation," *Neural Computation*, 15(6):1373-1396, 2003.