Locality Adaptive Discriminant Analysis Framework

Xuelong Li, *Fellow, IEEE*, Qi Wang, *Senior Member, IEEE*, Feiping Nie, and Mulin Chen

***Abstract*—Linear Discriminant Analysis (LDA) is a well-known technique for supervised dimensionality reduction, and has been extensively applied in many real-world applications. LDA assumes that the samples are Gaussian distributed, and the local data distribution is consistent with the global distribution. However, real-world data seldom satisfies this assumption. To handle the data with complex distributions, some methods emphasize the local geometrical structure, and perform discriminant analysis between neighbors. But the neighboring relationship tends to be affected by the noise in the input space. In this research, we propose a new supervised dimensionality reduction method, namely Locality Adaptive Discriminant Analysis (LADA). In order to directly process the data with matrix representation, such as images, the 2-Dimensional LADA (2DLADA) is also developed. The proposed methods have the following salient properties: (1) they find the principle projection directions without imposing any assumption on the data distribution; (2) they explore the data relationship in the desired subspace, which contains less noise; (3) they find the local data relationship automatically without the efforts for tuning parameters. The performance on dimensionality reduction shows the superiorities of the proposed methods over the state-of-the-arts.**

***Index Terms*—Dimensionality reduction, feature extraction, discriminant analysis, manifold structure**

I. Introduction

In many research areas, such as machine learning and cybernetics, data is often with high dimensionality. High-dimensional data significantly increases the computational costs, and brings large noise [1]. To mitigate this problem, dimensionality reduction techniques [2–6] are always used as the preprocessing step. Dimensionality reduction aims to learn the low-dimensional representation of the high-dimensional data, while preserving the discriminant information. *Linear discriminant analysis* (LDA) [4] is one of the most popular supervised dimensionality reduction methods. Given the class label of each sample, the goal of LDA is to learn a linear transformation, which pulls the within-class samples together and pushes the between-class samples apart. In the past few decades, LDA has been widely used in practical applications involving high-dimensional data, such as object recognition [7, 8], image retrieval [9], and image representation [10–12].

Despite its good properties, LDA has several intrinsic drawbacks. Firstly, the dimensionality reduced by LDA must be less than the class number, termed as *over-reducing* problem [13]. Supposing the class number is c, the rank of the between-class scatter matrix $\mathbf{S}_b$ is at most $c-1$. Consequently, LDA can find at most $c-1$ projection directions, which may be insufficient for retaining the valuable features, especially for binary classification tasks. Secondly, LDA suffers from the *small sample size* (SSS) problem [14]. LDA needs to calculate the inverse matrix of the within-class scatter $\mathbf{S}_w$. When the data dimensionality is very high, $\mathbf{S}_w$ becomes singular. Then LDA becomes unsolvable. Thirdly, LDA neglects the data structure in the local area. It assumes that the data obeys the Gaussian distribution and the data structure within each class is consistent with the global structure. However, this assumption is not true for the real-world tasks. Taking object clustering as an example, each object has its own pose variation, so the object images may be multi-modally distributed [15–19]. In these occasions, the performance of LDA degrades because it cannot capture the local data structure.

In the past two decades, many variants of the original LDA have been proposed, trying to improve LDA from different perspectives. To cope with the over-reducing problem, Wan et al. [13] designed the full rank between-class scatter matrix. Meanwhile, some thoughtful methods are proposed to address the SSS problem in different ways. Lu et al. [20] regularized the within-class scatter to make it reversible. Li et al. [21] put forward the Maximum Margin Criterion (MMC), which avoids calculating the inverse matrix of the within-class scatter. Ye et al. [22] developed the 2-dimensional version of LDA such that they do not need to convert the each matrix representation sample into a high-dimensional vector. The above algorithms solve the over-reducing and SSS problems excellently. But the exploration of local data relationship remains to be an open issue. A common solution to this problem is to find the k nearest neighbors of each sample, and then learn the linear transformation to pull the within-class neighbors together while making the between-class neighbors separable [23–26]. The shortcoming of this strategy is that the neighboring relationship within the input data space may be affected by the noise. Moreover, it is difficult to decide an appropriate k for various kinds of tasks.

In this paper, we present a new supervised dimensionality reduction method, *Locality Adaptive Discriminant analysis* (LADA). When processing data with matrix representation, such as images, traditional dimensionality reduction methods usually transform the data into vector form, resulting in the *curse of dimensionality*. To tackle this problem, we also develop the 2-Dimensional LADA (2DLADA), which is directly applicable for matrix data. The proposed framework can be considered as an iterative procedure: 1) the local data relationship is learned according to the samples' transformed distances; 2) the linear transformation is updated to pull the within-class similar samples together. The advantages of the proposed LADA and 2DLADA are summarized as follows:

(1) They do not resort to any assumption on the data distribution, and avoid the over-reducing and SSS problems implicitly.

The authors are with the Northwestern Polytechnical University, Xi'an 710072, Shaanxi, China. E-mails: xuelong_li@nwpu.edu.cn, crabwq@gmail.com, feipingnie@gmail.com, chenmulin@mail.nwpu.edu.cn. M. Chen is the corresponding author.

(2) They incorporate local structure learning into the discriminant analysis framework, so the local data structure can be exploited in the subspace, which contains less noise.
(3) They do not need to tune any parameter, and can be solved by the proposed optimization methods with proved convergence.

Compared to the conference version of this research [27], this paper is substantially improved by introducing more technical parts and experimental evaluations. Specifically: (1) Section III gives more details about the optimization method; (2) Section IV proposes the 2-dimensional version of LADA; (3) Section V gives more experimental results, the variances of the performance versus the training number are also provided;

The remaining parts of this paper are organized as follows. Section II reviews the classical LDA and some of its variants. Section III presents the proposed LADA and the resultant optimization algorithm. Section IV describes the 2DLADA and the optimization method. Section V gives the theoretical analysis and experimental results on real-world datasets. Section VII summarizes the conclusions.

II. Related Work

A. Linear Discriminant Analysis Revisited

Denote the data matrix as $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n] \in \mathbb{R}^{d\times n}$, where n is the number of samples and each sample $\mathbf{x}_j$ is a d-dimensional column vector. The goal of LDA [4] is to find a linear transformation matrix $\mathbf{W} \in \mathbb{R}^{d\times m}$ to embed each sample into a m-dimensional vector:

$$\mathbf{y}_j = \mathbf{W}^T\mathbf{x}_j. \tag{1}$$

When m is less than d, the compact representation of the original data can be achieved. Meanwhile, LDA considers that the optimal transformation should maximize the divergence of the between-class samples, while minimize the separation of the within-class samples. Based on the above theory, the objective function of LDA is formulated as

$$\max_{\mathbf{W}} \frac{\sum_{i=1}^{c} n_i||\mathbf{W}^T(\mu^i - \mu)||_2^2}{\sum_{i=1}^{c}\sum_{j=1}^{n_i} ||\mathbf{W}^T(\mathbf{x}_j^i - \mu^i)||_2^2}, \tag{2}$$

where c is the number of classes, n_i is the number of samples in class i, μ^i is the mean of the samples in class i, μ is the mean of all the samples, and $\mathbf{x}_j^i$ is the j-th sample in class i. Defining the between-class scatter matrix $\mathbf{S}_b \in \mathbb{R}^{d\times d}$ and the within-class scatter matrix $\mathbf{S}_w \in \mathbb{R}^{d\times d}$ as

$$\begin{aligned} \mathbf{S}_b &= \sum_{i=1}^{c} n_i(\mu^i - \mu)(\mu^i - \mu)^T, \\ \mathbf{S}_w &= \sum_{i=1}^{c}\sum_{j=1}^{n_i} (\mathbf{x}_j^i - \mu^i)(\mathbf{x}_j^i - \mu^i)^T, \end{aligned} \tag{3}$$

problem (2) can be transformed into the trace ratio form:

$$\max_{\mathbf{W}} \frac{\mathrm{tr}(\mathbf{W}^T\mathbf{S}_b\mathbf{W})}{\mathrm{tr}(\mathbf{W}^T\mathbf{S}_w\mathbf{W})}, \tag{4}$$

where $\mathrm{tr}()$ is the trace operator. Since it is difficult to solve problem (4), many researchers optimize the following ratio trace problem instead

$$\max_{\mathbf{W}} \mathrm{tr}(\frac{\mathbf{W}^T\mathbf{S}_b\mathbf{W}}{\mathbf{W}^T\mathbf{S}_w\mathbf{W}}), \tag{5}$$

and the optimal $\mathbf{W}$ is composed of the eigenvectors associated with the m largest eigenvalues of $\mathbf{S}_w^{-1}\mathbf{S}_b$. According to Jia et al. [28], the ratio trace form leads to the suboptimal solution.

Because LDA calculates the inverse matrix of $\mathbf{S}_w$, the SSS problem occurs when $\mathbf{S}_w$ is irreversible. The rank of $\mathbf{S}_b$ is at most $c-1$, which means that $\mathbf{S}_w^{-1}\mathbf{S}_b$ has at most $c-1$ non-zeros eigenvalues, leading to the over-reducing problem. In addition, As can be seen from the above formulations, LDA assumes that the divergence of the between-class samples can be reflected by the subtraction of the class mean vectors, which implies that all the classes share the same covariance. However, in real-world applications, the data samples may reside on a submanifold of the ambient space [24, 29, 30]. The ignorance of local data relationship makes LDA unsuitable for the samples with complex distributions.

B. Locality-Aware Variants of LDA

When the data distribution is more complex than Gaussian, the exploration of the local data structure is essential for a good performance. To this end, some locality-aware variants of LDA are proposed.

To capture the data structure, some methods defined the scatter matrices according to the samples' local relationship. Bressan and Vitria [25] found the k nearest neighbors of each sample and replaced the class mean with the average of the neighbors. Sugiyama [31] used Gaussian kernel to weight the scatter matrices. Nie et al. [23] denoted the scatter matrices as the covariances of the within-/between-class neighbors. Cai et al. [24] emphasized the samples with more within-class neighbors by imposing an additional constraint on the degree matrix of the kNN graph. Weinberger et al. [32] learned a Mahanalobis distance metric to find the largest margin for the k nearest neighbors. Fan et al. [26] trained a model within the neighborhood of each test sample separately, so it is time-consuming. Dong et al. [33] constructed the similarity graph with sparse representation, and utilized the learned similarity into the computation of scatter matrices. Zhang et al. [34] learned the subspace from the similar and dissimilar pairs, without the explicit label information. Nie et al. [35] found the neighbors of each sample, and learned the similarity between the neighbors. Nie et al. [36] imposed the binary and ℓ_0 constraints on the similarity graph, such that a weighted kNN graph can be obtained.

These methods build an affinity graph (Guassian graph, kNN graph) in the input space. However, the graph quality is determined by many factors, such as the scale of analysis and the data noise. Especially, when noise is large, the intrinsic similar samples may be far away from each other in the input space. So it is necessary to exploit the underlying data geometry in the desired subspace, which contains less noise and more significant statistical characteristics.

C. 2-Dimensional Variants of LDA

Traditional LDA methods work with vector data. When dealing with samples in matrix representation, they transform them into vectors by concatenating the rows together. The *matrix-to-vector alignment* produces data with very high-dimensionality, and discards the spatial information. An effective solution to this problem is to perform 2-dimensional discriminant analysis, which reduces the dimensionality along both the row and column directions.

Denote the data matrix $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n] \in \mathbb{R}^{d_1 \times d_2 n}$, where n is the sample number and each sample $\mathbf{X}_j \in \mathbb{R}^{d_1 \times d_2}$ is a matrix. Ye et al. [22] proposed to project each sample into a low dimensional matrix $\mathbf{Y} \in \mathbb{R}^{m_1 \times m_2}$ ($m_1 < d_1$ and $m_2 < d_2$)

$$\mathbf{Y}_j = \mathbf{L}^T \mathbf{X}_j \mathbf{R}, \tag{6}$$

where $\mathbf{L} \in \mathbb{R}^{d_1 \times m_1}$ and $\mathbf{R} \in \mathbb{R}^{d_2 \times m_2}$ are the transformation matrices along row and column directions respectively. Similar to Eq. (2), the objective function of the 2DLDA is rewritten as

$$\max_{\mathbf{L},\mathbf{R}} \frac{\sum_{i=1}^{c} n_i ||\mathbf{L}^T(\mu^i - \mu)\mathbf{R}||_F^2}{\sum_{i=1}^{c} \sum_{j=1}^{n_i} ||\mathbf{L}^T(\mathbf{X}_j^i - \mu^i)\mathbf{R})||_F^2}, \tag{7}$$

where the definitions of n_i, c, μ^i and μ are the same as those in Eq. (2), and $\mathbf{X}_j^i$ is the j-th sample in class i. With the above formulation, the matrix representation data can be processed directly. Sanguansat et al. [37] performed 2DLDA on the results of 2DPCA [38], and defined the scatter matrices with the prior probability of each class. To reduce the SSS problem, Yang and Dai [39] proposed the 2-Dimensional Maximum Margin Criterion (2DMMC) method, and the objective function is represented as the subtraction of the between-class and within-class scatter matrices. Wang et al. [40] introduced an additional weighted parameter on 2DMMC to balanced the between-class and within-class scatters.

The 2-dimensional methods successfully avoid the problems brought by the matrix-to-vector alignment, and have lower cost in time and space than LDA. But they still share the same assumption on the data distribution as LDA.

III. Locality-Aware Discriminant Analysis

A. Methodology

As discussed previously, the investigation of local manifold structure is crucial for handling the data with complex distributions. So we propose to learn the local data relationship with a similarity graph, and optimize the linear transformation to pull the similar points together.

Given the data samples $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n] \in \mathbb{R}^{d \times n}$ (d is the dimensionality), the objective function of LADA is formulated as

$$\begin{aligned} &\min_{\mathbf{W},\mathbf{S}} \frac{\sum_{i=1}^{c} n_i \sum_{j=1}^{n_i} \sum_{k=1}^{n_i} {s_{jk}^i}^2 \left\|\mathbf{W}^T(\mathbf{x}_j^i - \mathbf{x}_k^i)\right\|_2^2}{\frac{1}{n}\sum_{j=1}^{n}\sum_{k=1}^{n} \|\mathbf{W}^T(\mathbf{x}_j - \mathbf{x}_k)\|_2^2}, \\ &s.t. \mathbf{W}^T\mathbf{W} = \mathbf{I}, \sum_{k=1}^{n_i} s_{jk}^i = 1, s_{jk}^i \geq 0, \end{aligned} \tag{8}$$

where the n is the number of samples, $\mathbf{I} \in \mathbb{R}^{m \times m}$ is the identity matrix, s_{jk}^i is an element of the similarity graph $\mathbf{S}$, and a larger value of s_{jk}^i indicates a higher similarity between the j-th and k-th sample in class i. $\mathbf{x}_j$ is the j-th sample in the whole dataset (different from $\mathbf{x}_j^i$). The remaining definitions are the same as those in LDA.

In problem (8), the orthonormal constraint $\mathbf{W}^T\mathbf{W} = \mathbf{I}$ ensures the uniqueness of the optimal $\mathbf{W}$. To avoid the trivial solution where the optimal graph is an identity matrix, we fix s_{jj}^i as 0 and update the other elements in each iteration. When the transformation matrix $\mathbf{W}$ is obtained, s_{jk}^i will be adjusted according to the samples' transformed distance $||\mathbf{W}^T(\mathbf{x}_j^i - \mathbf{x}_k^i)||_2^2$, so the local data relationship in the learned subspace can be learned. When the data graph $\mathbf{S}$ is fixed, the optimal $\mathbf{W}$ will emphasize the within-class samples with large similarity and pull them as close as possible. By updating $\mathbf{S}$ and $\mathbf{W}$ iteratively, the proposed method incorporates local structure learning into the discriminant analysis framework.

B. Optimization Algorithm

Problem (8) involves two variables, so we propose to solve it with an alternative strategy, minimizing the objective function with respect to one variable while fixing the other one.

When S is fixed, defining the within-class scatter $\tilde{\mathbf{S}}_w \in \mathbb{R}^{d \times d}$ and total scatter $\tilde{\mathbf{S}}_t \in \mathbb{R}^{d \times d}$ as

$$\begin{aligned} \tilde{\mathbf{S}}_w &= \sum_{i=1}^{c} n_i \sum_{j=1}^{n_i} \sum_{k=1}^{n_i} {s_{jk}^i}^2 (\mathbf{x}_j^i - \mathbf{x}_k^i)(\mathbf{x}_j^i - \mathbf{x}_k^i)^T, \\ \tilde{\mathbf{S}}_t &= \frac{1}{n} \sum_{j=1}^{n} \sum_{k=1}^{n} (\mathbf{x}_j - \mathbf{x}_k)(\mathbf{x}_j - \mathbf{x}_k)^T. \end{aligned} \tag{9}$$

problem (8) is converted into

$$\min_{\mathbf{W}^T\mathbf{W}=\mathbf{I}} \frac{\mathrm{tr}(\mathbf{W}^T\tilde{\mathbf{S}}_w\mathbf{W})}{\mathrm{tr}(\mathbf{W}^T\tilde{\mathbf{S}}_t\mathbf{W})}, \tag{10}$$

which is equivalent to the following maximization problem:

$$\max_{\mathbf{W}^T\mathbf{W}=\mathbf{I}} \frac{\mathrm{tr}(\mathbf{W}^T\tilde{\mathbf{S}}_t\mathbf{W})}{\mathrm{tr}(\mathbf{W}^T\tilde{\mathbf{S}}_w\mathbf{W})}, \tag{11}$$

Supposing the rank of $\tilde{\mathbf{S}}_t$ is r, we discuss the optimization of problem (11) in two cases.

Case 1: $m > d - r$. Denoting the p-th smallest eigenvalues of $\tilde{\mathbf{S}}_w$ as β_p, it holds that $\min \mathrm{tr}(\mathbf{W}^T\tilde{\mathbf{S}}_w\mathbf{W}) = \sum_{p=1}^{m} \beta_p$ with the constraint $\mathbf{W}^T\mathbf{W} = \mathbf{I}$. Obviously, $\tilde{\mathbf{S}}_w$ is a positive semi-definite matrix, so $\sum_{p=1}^{m} \beta_p$ is larger than zero when m is larger than $d - r$. Then we have $\mathrm{tr}(\mathbf{W}^T\tilde{\mathbf{S}}_w\mathbf{W}) \geq \min \mathrm{tr}(\mathbf{W}^T\tilde{\mathbf{S}}_w\mathbf{W}) > 0$.

Denoting the objective value of problem (8) as λ, Nie et al. [23] have given the bound of λ as

$$\frac{\mathrm{tr}(\tilde{\mathbf{S}}_t)}{\mathrm{tr}(\tilde{\mathbf{S}}_w)} \leq \lambda \leq \frac{\sum_{p=1}^{m} \alpha_p}{\sum_{p=1}^{m} \beta_p}, \tag{12}$$

where α_p is the p-th largest eigenvalues of $\tilde{\mathbf{S}}_t$. In addition, denoting the sum of the first m smallest eigenvalues of $\tilde{\mathbf{S}}_t - \lambda\tilde{\mathbf{S}}_w$ as γ, Guo et al. [41] have proved that the optimal λ^* should make γ equal to zero. With $\mathrm{tr}(\mathbf{W}^T\tilde{\mathbf{S}}_w\mathbf{W}) > 0$, $\gamma > 0$

indicates that λ is smaller than the optimal value and vice visa. Together with bound of λ, the optimal value λ^* can be found by binary search. Then the optimal $\mathbf{W}^*$ is formed with eigenvectors associated with the first m smallest eigenvalues of $\tilde{\mathbf{S}}_t - \lambda^* \tilde{\mathbf{S}}_w$.

Case 2: $m \leq d - r$. In this case, we have $\min \mathrm{tr}(\mathbf{W}^T \tilde{\mathbf{S}}_w \mathbf{W}) = 0$ because the first m smallest eigenvalues of $\tilde{\mathbf{S}}_t$ are all zero. Therefore, the optimal $\mathbf{W}^*$ resides in the null space of $\tilde{\mathbf{S}}_w$ and problem (11) is equivalent to

$$\max_{\mathbf{V}^T\mathbf{V}=\mathbf{I}, \mathbf{V}\in\mathbb{R}^{(d-r)\times m}} \mathrm{tr}(\mathbf{V}^T(\mathbf{Z}^T\tilde{\mathbf{S}}_t\mathbf{Z})\mathbf{V}), \tag{13}$$

where $\mathbf{Z} \in \mathbb{R}^{d\times(d-r)}$ is formed with the eigenvectors associated with the $d-r$ zero eigenvalues of $\tilde{\mathbf{S}}_w$. The optimal $\mathbf{V}^*$ can be obtained with the first m largest eigenvectors of $\mathbf{Z}^T\tilde{\mathbf{S}}_t\mathbf{Z}$, so the optimal solution to problem (11) is $\mathbf{W}^* = \mathbf{Z}\mathbf{V}^*$.

The proposed method does not need to calculate the inverse matrix, so the SSS problem is avoided implicitly. In addition, in the proposed method, $\tilde{\mathbf{S}}_t$ is of full rank, so the over-reducing problem does not exist.

When W is fixed, problem (8) becomes

$$\begin{aligned} &\min_{\mathbf{S}} \sum_{i=1}^{c} n_i \sum_{j=1}^{n_i}\sum_{k=1}^{n_i} {s_{jk}^i}^2 \left\|\mathbf{W}^T(\mathbf{x}_j^i - \mathbf{x}_k^i)\right\|_2^2, \\ &s.t. \sum_{k=1}^{n_i} s_{jk}^i = 1, s_{jk}^i \geq 0. \end{aligned} \tag{14}$$

For each i and j, the above problem is simplified into

$$\begin{aligned} &\min_{\mathbf{u}} \sum_{k=1}^{n_i} u_k^2 ||\mathbf{W}^T(\mathbf{x}_j^i - \mathbf{x}_k^i)||_2^2, \\ &s.t. \mathbf{u}^T\mathbf{1} = 1, \mathbf{u} \geq 0, \end{aligned} \tag{15}$$

where $\mathbf{u} \in \mathbb{R}^{n_i\times 1}$ is a column vector with its k-th element u_k equal to s_{jk}^i, and $\mathbf{1} \in \mathbb{R}^{n_i\times 1}$ is a column vector with all the elements as 1. Defining a diagonal matrix $\mathbf{V}$ with its k-th element v_{kk} equal to $||\mathbf{W}^T(\mathbf{x}_j^i - \mathbf{x}_k^i)||_2^2$, problem (15) is reduced to

$$\min_{\mathbf{u}^T\mathbf{1}=1, \mathbf{u}\geq 0} \mathbf{u}^T\mathbf{V}\mathbf{u}. \tag{16}$$

Removing the constraint $\mathbf{u} \geq 0$, the Lagrangian function of problem (16) is written as

$$\mathcal{L}(\mathbf{u}, \eta) = \mathbf{u}^T\mathbf{V}\mathbf{u} - \eta(\mathbf{u}^T\mathbf{1} - 1), \tag{17}$$

Calculating the derivative of the above problem with respect to $\mathbf{u}$ and setting it to zero, we have

$$2\mathbf{V}\mathbf{u} - \eta\mathbf{1} = 0. \tag{18}$$

Together with the constraint $\mathbf{u}^T\mathbf{1} = 1$, the optimal solution $\mathbf{u}^*$ is obtained as

$$u_k^* = \frac{1}{v_{kk}} \times (\sum_{p=1}^{n_i} \frac{1}{v_{pp}})^{-1}. \tag{19}$$

Fortunately, we can see that $\mathbf{u}^*$ satisfies the constraint $\mathbf{u} \geq 0$, so $\mathbf{u}^*$ is the final solution to problem (16). Similarly, the optimal solution to problem (15) is

$${s_{jk}^i}^* = \frac{1}{||\mathbf{W}^T(\mathbf{x}_j^i - \mathbf{x}_k^i)||_2^2} \times (\sum_{p=1}^{n_i} \frac{1}{||\mathbf{W}^T(\mathbf{x}_j^i - \mathbf{x}_p^i)||_2^2})^{-1}. \tag{20}$$

By updating $\mathbf{W}$ and $\mathbf{S}$ iteratively, the local data structure in the desired subspace is exploited. Different from previous works, the proposed method learns the local structure adaptively without any additional parameter. The overall optimization algorithm for problem (8) is described in Algorithm 1.

It is easy to see that the computational cost of Algorithm 1 concentrates on the computation of $\mathbf{W}$ and $\mathbf{S}$. When updating $\mathbf{W}$, the complexity is $\mathcal{O}(n^2 + n^2d^2 + d^3)$. When updating $\mathbf{S}$, the complexity is $\mathcal{O}(n^2d^2)$. Ignoring the constant terms, the computational complexity of each iteration is $\mathcal{O}(n^2 + n^2d^2 + d^3)$, which is the same as the traditional LDA.

Algorithm 1 Algorithm of LADA

Input: Data matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n] \in \mathbb{R}^{d\times n}$, desired dimensionality m.

1: Initial the data graph $\mathbf{S}$ as $s_{jk}^i = \frac{1}{n_i}$.
2: **repeat**
3: Compute the optimal $\mathbf{W}^*$ by solving problem (10).
4: Update $\mathbf{S}$ according to Eq. (20).
5: **until** Converge

Output: Projected data $\mathbf{Y} = \mathbf{W}^*\mathbf{X}$.

IV. 2-Dimensional Locality-Aware Discriminant Analysis

A. Methodology

Given the data matrix $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n] \in \mathbb{R}^{d_1\times d_2 n}$ (d_1 and d_2 are the data dimensionalities along the row and column directions respectively), the objective function of 2D-LADA is written as

$$\begin{aligned} &\min_{\mathbf{L},\mathbf{R},\mathbf{S}} \frac{\sum_{i=1}^{c} n_i \sum_{j=1}^{n_i}\sum_{k=1}^{n_i} {s_{jk}^i}^2 \left\|\mathbf{L}^T(\mathbf{X}_j^i - \mathbf{X}_k^i)\mathbf{R}\right\|_F^2}{\frac{1}{n}\sum_{j=1}^{n}\sum_{k=1}^{n} \|\mathbf{L}^T(\mathbf{X}_j - \mathbf{X}_k)\mathbf{R}\|_F^2}, \\ &s.t. \mathbf{L}^T\mathbf{L} = \mathbf{I}, \mathbf{R}^T\mathbf{R} = \mathbf{I}, \sum_{k=1}^{n_i} s_{jk}^i = 1, s_{jk}^i \geq 0, \end{aligned} \tag{21}$$

where $\mathbf{S} \in \mathbb{R}^{n\times n}$ is the similarity graph, s_{jk}^i is an element of $\mathbf{S}$ and indicates the similarity between the j-th and k-th sample in class i, $\mathbf{X}_j$ is the j-th sample in the input dataset. The definition of $\mathbf{L}$, $\mathbf{R}$ and $\mathbf{X}_j^i$ are the same as those in Eq. (7). Different from LADA, 2DLADA projects the data along both the row and column directions with $\mathbf{L}$ and $\mathbf{R}$, so it is directly applicable to data with matrix representation.

Compared with traditional 2DLDA, the proposed 2DLADA performs local structure learning and discriminant analysis simultaneously. So 2DLADA is free from the Gaussian distribution assumption, and can handle the data with complex structures. In the following, the optimization strategy for problem (21) is presented.

B. Optimization Algorithm

Problem (21) involves three variable to be optimized, so we solve one variable while keeping the other two fixed.

When S and R are fixed, denoting two scatter matrices $\mathbf{S}_w^r$ and $\mathbf{S}_t^r$ as

$$\begin{aligned}\tilde{\mathbf{S}}_w^r &= \sum_{i=1}^{c} n_i \sum_{j=1}^{n_i}\sum_{k=1}^{n_i} {s_{jk}^i}^2 (\mathbf{X}_j^i - \mathbf{X}_k^i)\mathbf{R}\mathbf{R}^T(\mathbf{X}_j^i - \mathbf{X}_k^i)^T,\\ \tilde{\mathbf{S}}_t^r &= \frac{1}{n}\sum_{j=1}^{n}\sum_{k=1}^{n}(\mathbf{X}_j - \mathbf{X}_k)\mathbf{R}\mathbf{R}^T(\mathbf{X}_j - \mathbf{X}_k)^T,\end{aligned} \tag{22}$$

the original problem transforms into

$$\min_{\mathbf{L}^T\mathbf{L}=\mathbf{I}} \frac{\mathrm{tr}(\mathbf{L}^T\tilde{\mathbf{S}}_w^r\mathbf{L})}{\mathrm{tr}(\mathbf{L}^T\tilde{\mathbf{S}}_t^r\mathbf{L})}. \tag{23}$$

Thus, the computation of optimal $\mathbf{L}^*$ is the same as the optimization of $\mathbf{W}$ in problem (10).

When S and L are fixed, the scatter matrices $\mathbf{S}_w^l$ and $\mathbf{S}_t^l$ are defined as

$$\begin{aligned}\tilde{\mathbf{S}}_w^l &= \sum_{i=1}^{c} n_i \sum_{j=1}^{n_i}\sum_{k=1}^{n_i} {s_{jk}^i}^2 (\mathbf{X}_j^i - \mathbf{X}_k^i)^T\mathbf{L}\mathbf{L}^T(\mathbf{X}_j^i - \mathbf{X}_k^i),\\ \tilde{\mathbf{S}}_t^l &= \frac{1}{n}\sum_{j=1}^{n}\sum_{k=1}^{n}(\mathbf{X}_j - \mathbf{X}_k)^T\mathbf{L}\mathbf{L}^T(\mathbf{X}_j - \mathbf{X}_k).\end{aligned} \tag{24}$$

Since $\mathrm{tr}(\mathbf{AB})$ is equal to $\mathrm{tr}(\mathbf{BA})$, the optimal $\mathbf{R}$ can be obtained with the following problem:

$$\min_{\mathbf{R}^T\mathbf{R}=\mathbf{I}} \frac{\mathrm{tr}(\mathbf{R}^T\tilde{\mathbf{S}}_w^l\mathbf{R})}{\mathrm{tr}(\mathbf{R}^T\tilde{\mathbf{S}}_t^l\mathbf{R})}. \tag{25}$$

When L and R are fixed, the optimization of $\mathbf{S}$ yields to

$$\begin{aligned}&\min_{\mathbf{S}} \sum_{i=1}^{c} n_i \sum_{j=1}^{n_i}\sum_{k=1}^{n_i} {s_{jk}^i}^2 \left\|\mathbf{L}^T(\mathbf{X}_j^i - \mathbf{X}_k^i)\mathbf{R}\right\|_F^2,\\ &s.t. \sum_{k=1}^{n_i} s_{jk}^i = 1, s_{jk}^i \geq 0.\end{aligned} \tag{26}$$

According to the optimization of problem (14), the optimal $\mathbf{S}^*$ is computed as

$${s_{jk}^i}^* = \frac{1}{||\mathbf{L}^T(\mathbf{X}_j^i - \mathbf{X}_k^i)\mathbf{R}||_F^2} \times \Big(\sum_{p=1}^{n_i} \frac{1}{||\mathbf{L}^T(\mathbf{X}_j^i - \mathbf{X}_p^i)\mathbf{R}||_F^2}\Big)^{-1}. \tag{27}$$

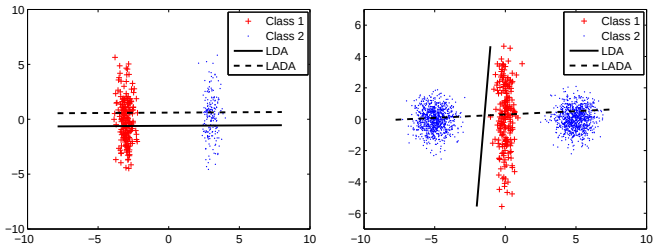
The algorithm of 2DLADA is described in Algorithm 2. Denoting d as $\max(d_1, d_2)$, the upper bound computational complexity of each iteration is $\mathcal{O}(n^2 + n^2d^2 + d^3)$, which is the same as the 2DLDA.

Algorithm 2 Algorithm of 2DLADA

Input: Data matrix $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n] \in \mathbb{R}^{d_1 \times d_2 n}$, desired dimensionality m_1 and m_2.

1: Initial the data graph $\mathbf{S}$ as $s_{jk}^i = \frac{1}{n_i}$.
2: **repeat**
3: Compute the optimal $\mathbf{L}^*$ by solving problem (23).
4: Compute the optimal $\mathbf{R}^*$ by solving problem (25).
5: Update $\mathbf{S}$ according to Eq. (27).
6: **until** Converge

Output: Projected data $\mathbf{Y} = {\mathbf{L}^*}^T\mathbf{X}\mathbf{R}^*$.



(a) Unimodal data (b) Multi-modal data

Fig. 1. Projection directions learned by LDA and LADA. LDA finds the correct direction when the class distribution is unimodal (a), but fails on muti-modal distributed data (b). LADA works well on both cases.

V. Experiments

In this section, experiments are conducted to verify the effectiveness of the proposed LADA and 2DLADA. Firstly, synthetic datasets are used to demonstrate the theoretical significance of LADA. Then, results on real-world benchmarks show the performance of LADA and 2DLADA.

A. Performance on Unimodal and Multi-Modal Datasets

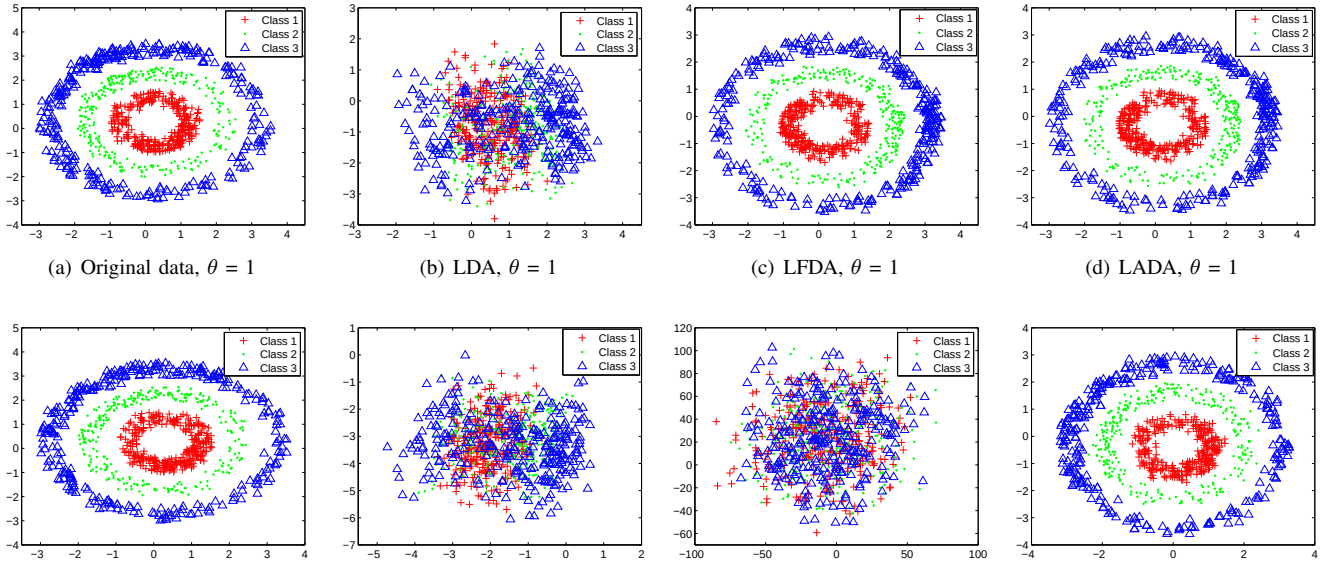
Two datasets are constructed to illustrate the significance of local structure learning. As shown in Fig. 1 (a), the first dataset is consisted of the data points generated by Gaussian distribution. Both LDA and LADA find the correct projection direction. The points in the second dataset are multi-modally distributed, which means each class has a unique distribution. As shown in Fig. 1 (b), LDA fails on the multi-modal datasets because the global data structure is unreliable. LADA still works well since it emphasizes the local relationship between the data samples. These results validate that LADA is able to handle the data with complicated distributions.

B. Performance on Three-Ring Datasets

To evaluate whether LADA is able to preserving the intrinsic manifold structure, we conduct experiments on two three-ring datasets. As visualized in Fig. 2 (a) and (e), each datasets contains the samples from three classes. In the first two dimensions, the samples are distributed in concentric circles, while the other eight dimensions are noises randomly generated in 0 and θ. θ is set as 1 and 100 for the two datasets respectively.

We compare the proposed LADA with LDA [4] and Local Fisher Discriminant Analysis (LFDA) [31]. LFDA captures the local data structure with the Gaussian kernel, and emphasizes the similar samples when calculating the scatter matrices.

The two-dimensional subspace found by different methods are shown in Fig. 2 (b)-(d) and (f)-(h). Due to the neglect of local structure, LDA cannot find the correct subspace even when the noise factor θ is 1. Benefited from the utilization of similarity graph, LFDA performs well when noise is small. However, it highly relies on the samples' distances in the input space, so it fails when the noise is large. The proposed LADA captures the samples' local relationship in the subspace, so it preserves the intrinsic geometrical structure well and shows robustness to the noise in the input space.



(a) Original data, θ = 1 (b) LDA, θ = 1 (c) LFDA, θ = 1 (d) LADA, θ = 1

(e) Original data, θ = 100 (f) LDA, θ = 100 (g) LFDA, θ = 100 (h) LADA, θ = 100

Fig. 2. (a) and (e) are the first two dimensions of the two datasets. (b)-(e) and (f)-(h) show the two-dimensional subspaces learned by LDA, LFDA and LADA on the two datasets. The performance of LDA is unsatisfying due to the neglect of local manifold structure. LFDA fails when noise is 100, because its performance depends on the data relationship in the input space. Under different noises, LADA is able to find the discriminant subspace while preserving the manifold structure.

TABLE I
DETAILED DESCRIPTION OF THE DATASETS.

	UMIST	Olivettifaces	YALE	ALLAML	TOX_171	BA
Sample number	575	400	165	72	171	1404
Class number	20	40	15	2	4	36
#Dim	112×92	64×64	80×80	7129×1	5748×1	20×16
#Dim after PCA	563	398	160	66	127	291
#Reduced dim (1D)	[5,···,70]	[5,···,70]	[5,···,70]	[5,···,65]	[5,···,70]	[5,···,70]
#Dim after resizing	25×25	25×25	25×25	—	—	—
#Reduced dim (2D)	[1×1,···,24×24]	[1×1,···,24×24]	[1×1,···,24×24]	—	—	—

C. Performance on Real-world Datasets

In this part, we show the performance of LADA and 2DLADA, and compare them with the state-of-the-art dimensionality reduction methods.

Datasets: For 1D methods, we employ six datasets for evaluation: three face image datasets, i.e., UMIST [42], Olivettifaces [43] and YALE [44], two biological datasets, i.e., ALLAML [45] and TOX_171 [45], and one handwritten dataset, i.e., Binary Alphabet (BA) [46]. Since some competitors have the SSS problem, we take principle Component Analysis (PCA) [2] as preprocessing for a fair comparison, keeping 99.5% of the covariance.

For 2D methods, experiments are conducted on UMIST, Olivettifaces and Yale. The 2D methods do not have the SSS problem, so they do not need PCA. Instead, we resize all the images into a resolution of 25×25 pixels for efficiency. The descriptions of these datasets are given in Table I. The UMIST dataset contains 575 images of 20 persons. The images of each person cover a ranges of poses from profile to frontal views. We use the pre-cropped version in the experiments. Olivettifaces dataset consists of the face images of 40 persons. For each person, the images were taken at different poses and facial expressions (smilling/not smilling). YALE dataset contains 165 images of 15 individuals. The images are captured under different lighting effects, facial expressions (happy, sad) and facial details (with/without glasses).

Competitors: the proposed LADA is compared with four state-of-the-art algorithms, including LDA, Maximum Margin Criterion (MMC) [21], Non-parametric Discriminant Analysis (NDA) [25] and Local Fisher Discriminant Analysis (LFDA) [31]. For NDA and LFDA, the scaling parameter k is set as 5 empirically. The classification result without dimensionality reduction is taken as the baseline, termed as RAW.

For the 2DLADA, we employ five 2D dimensionality reduction methods for comparison. They are 2-Dimensional PCA (2DPCA) [38], Robust 2DPCA (R2DPCA) [47], 2-Dimensional Locality Preserving Projections (2DLPP) [48], 2-Dimensional LDA (2DLDA) [22] and 2-Dimensional MMC (2DMMC) [39].

Evaluation mechanism: for each class, we randomly choose N images for training and the remaining images are used for

TABLE II
THE AVERAGE CLASSIFICATION ACCURACIES (MEAN ± STD. DEV. %) ON UMIST DATASET. THE BEST REDUCED DIMENSIONALITY IS SHOWN IN THE BRACKETS. THE BEST RESULTS ARE IN BOLD FACE.

1D Methods	$N = 6$	$N = 7$	$N = 8$	$N = 9$	$N = 10$
RAW	87.88±2.87 (—)	90.67±2.16 (—)	92.64±2.03 (—)	94.66±1.75 (—)	95.23±1.91 (—)
LDA	65.02±5.83 (19)	66.84±7.06 (19)	67.39±6.26 (19)	68.57±7.16 (19)	76.19±6.47 (19)
MMC	87.70±2.78 (65)	90.30±2.08 (70)	96.04±1.99 (15)	94.54±1.90 (70)	95.25±2.04 (65)
NDA	91.04±2.67 (20)	94.11±1.34 (55)	95.86±1.77 (70)	96.86±1.18 (35)	97.33±1.35 (35)
LFDA	88.34±2.33 (70)	90.71±2.31 (70)	91.93±2.30 (70)	92.91±1.77 (70)	93.73±1.76 (70)
LADA	**94.44±1.81** (10)	**95.70±1.09** (10)	**96.67±1.74** (20)	**97.19±1.28** (10)	**97.79±1.34** (15)
2D Methods	$N = 6$	$N = 7$	$N = 8$	$N = 9$	$N = 10$
RAW	88.13±1.97 (—)	90.90±1.70 (—)	93.64±1.81 (—)	95.14±1.56 (—)	93.97±1.78 (—)
2DPCA	91.43±2.04 (3)	93.14±0.88 (3)	95.45±1.62 (3)	97.78±1.58 (3)	96.40±1.68 (3)
R2DPCA	91.16±2.48 (5)	93.07±1.26 (4)	94.74±1.51 (5)	96.50±1.69 (4)	94.93±1.40 (4)
2DLPP	91.17±2.07 (2)	94.23±0.87 (2)	94.31±1.37 (2)	97.31±0.63 (2)	96.51±1.49 (2)
2DLDA	86.33±2.12 (10)	89.74±1.88 (12)	91.57±1.56 (9)	94.38±0.80 (12)	95.31±1.88 (10)
2DMMC	91.56±2.04 (6)	93.79±1.36 (6)	95.13±1.52 (6)	96.76±1.82 (6)	95.20±1.17 (6)
2DLADA	**93.10±1.98** (14)	**95.96±1.47** (16)	**96.48±1.58** (13)	**98.03±1.05** (17)	**97.76±1.07** (15)

TABLE III
THE AVERAGE CLASSIFICATION ACCURACIES (MEAN ± STD. DEV. %) ON OLIVETTIFACES DATASET. THE BEST REDUCED DIMENSIONALITY IS SHOWN IN THE BRACKETS. THE BEST RESULTS ARE IN BOLD FACE.

1D Methods	$N = 6$	$N = 7$	$N = 8$	$N = 9$
RAW	90.13±1.88 (—)	92.33±2.33 (—)	93.63±1.89 (—)	94.13±3.73 (—)
LDA	64.21±3.17 (39)	61.58±3.07 (39)	60.75±6.18 (39)	57.45±5.74 (39)
MMC	89.44±2.14 (65)	92.17±2.25 (65)	94.25±1.42 (40)	93.25±2.32 (35)
NDA	93.06±2.04 (65)	94.13±3.03 (45)	92.50±5.19 (70)	89.54±4.18 (70)
LFDA	87.50±1.80 (70)	88.75±4.72 (70)	89.37±2.89 (70)	83.25±2.80 (70)
LADA	**95.38±1.84** (50)	**94.67±1.92** (60)	**95.63±1.73** (70)	**95.14±2.65** (60)
2D Methods	$N = 6$	$N = 7$	$N = 8$	$N = 9$
RAW	90.94±2.60 (—)	93.04±2.43 (—)	94.82±2.48 (—)	96.07±2.84 (—)
2DPCA	91.72±2.39 (8)	93.51±2.33 (9)	95.60±2.13 (12)	96.55±2.94 (12)
R2DPCA	91.03±2.65 (19)	93.04±2.58 (10)	95.12±2.07 (11)	96.55±2.83 (11)
2DLPP	91.47±2.61 (17)	93.08±2.43 (10)	95.12±2.11 (8)	96.07±2.51 (13)
2DLDA	90.47±2.68 (14)	91.49±2.62 (14)	93.99±2.03 (16)	95.60±2.66 (15)
2DMMC	91.28±2.47 (19)	93.19±2.33 (22)	95.30±2.28 (19)	96.43±2.74 (13)
2DLADA	**94.19±2.51** (16)	**95.25±2.52** (20)	**97.08±1.51** (19)	**97.62±1.96** (18)

TABLE IV
THE AVERAGE CLASSIFICATION ACCURACIES (MEAN ± STD. DEV. %) ON YALE DATASET. THE BEST REDUCED DIMENSIONALITY IS SHOWN IN THE BRACKETS. THE BEST RESULTS ARE IN BOLD FACE.

1D Methods	$N = 6$	$N = 7$	$N = 8$	$N = 9$	$N = 10$
RAW	76.40±3.24 (—)	74.25±3.43 (—)	74.00±6.89 (—)	78.50±7.41 (—)	76.33±9.30 (—)
LDA	55.47±5.31 (14)	57.33±4.31 (14)	48.33±7.44 (14)	53.83±8.39 (14)	52.33±6.31 (14)
MMC	76.67±3.33 (65)	74.92±3.31 (45)	90.11±5.60 (15)	79.33±7.57 (25)	77.00±7.77 (45)
NDA	83.33±3.78 (15)	84.58±4.41 (15)	84.78±6.26 (35)	86.00±6.72 (20)	76.67±6.45 (40)
LFDA	84.93±6.44 (50)	86.00±3.35 (40)	81.89±6.93 (60)	83.17±6.95 (60)	87.67±7.68 (40)
LADA	**91.20±3.02** (45)	**92.82±2.86** (70)	**94.89±3.03** (70)	**95.83±3.75** (70)	**96.00±3.67** (70)
2D Methods	$N = 6$	$N = 7$	$N = 8$	$N = 9$	$N = 10$
RAW	76.84±3.59 (—)	77.06±4.32 (—)	77.93±4.87 (—)	76.67±7.30 (—)	76.22±7.84 (—)
2DPCA	76.98±3.44 (10)	77.17±4.24 (9)	78.30±5.06 (6)	78.33±5.43 (2)	77.56±6.50 (2)
R2DPCA	76.89±3.25 (13)	77.06±4.32 (20)	78.07±5.03 (13)	77.00±7.37 (11)	76.22±7.83 (7)
2DLPP	76.49±4.27 (2)	78.06±4.06 (2)	79.33±5.01 (2)	79.11±6.61 (2)	78.89±8.81 (2)
2DLDA	81.34±3.94 (14)	83.44±4.57 (13)	84.81±4.66 (14)	85.67±6.95 (14)	85.56±4.32 (12)
2DMMC	77.16±3.75 (18)	77.39±4.44 (18)	78.81±4.89 (14)	77.78±7.32 (11)	77.33±6.97 (17)
2DLADA	**85.60±3.81** (23)	**87.00±4.27** (19)	**88.96±4.24** (22)	**89.11±4.86** (17)	**90.22±4.04** (17)

TABLE V
THE AVERAGE CLASSIFICATION ACCURACIES (MEAN ± STD. DEV. %) ON ALLAML DATASET. THE BEST REDUCED DIMENSIONALITY IS SHOWN IN THE BRACKETS. THE BEST RESULTS ARE IN BOLD FACE.

1D Methods	$N = 6$	$N = 7$	$N = 8$	$N = 9$	$N = 10$
RAW	78.52±7.61 (—)	82.24±6.07 (—)	81.43±4.39 (—)	81.48±5.31 (—)	84.81±6.17 (—)
LDA	56.51±7.18 (1)	56.72±8.28 (1)	54.64±6.42 (1)	53.89±9.49 (1)	59.42±9.24 (1)
MMC	78.55±5.61 (15)	82.24±6.07 (15)	88.57±4.46 (5)	82.22±3.12 (5)	85.14±6.01 (15)
NDA	75.53±6.53 (25)	80.52±9.19 (15)	76.79±9.77 (5)	83.52±8.66 (5)	77.88±8.86 (5)
LFDA	85.33±4.96 (50)	85.34±7.33 (40)	88.93±4.36 (45)	86.11±4.74 (30)	87.31±3.12 (40)
LADA	87.67±5.62 (5)	88.10±5.64 (15)	89.82±2.99 (35)	90.41±3.63 (25)	90.38±4.71 (55)

TABLE VI
THE AVERAGE CLASSIFICATION ACCURACIES (MEAN ± STD. DEV. %) ON TOX_171 DATASET. THE BEST REDUCED DIMENSIONALITY IS SHOWN IN THE BRACKETS. THE BEST RESULTS ARE IN BOLD FACE.

1D Methods	$N = 6$	$N = 7$	$N = 8$	$N = 9$	$N = 10$
RAW	52.18±4.72 (—)	52.94±3.75 (—)	56.62±3.39 (—)	58.15±4.76 (—)	57.79±3.26 (—)
LDA	35.10±9.42 (3)	30.14±7.16 (3)	32.45±5.12 (3)	36.07±4.84 (3)	36.49±8.49 (3)
MMC	52.65±5.33 (20)	52.94±3.58 (25)	59.78±3.11 (5)	58.15±4.76 (35)	58.24±2.86 (35)
NDA	54.35±4.20 (5)	49.44±5.24 (45)	51.22±5.53 (35)	55.41±5.15 (45)	49.24±4.79 (50)
LFDA	57.01±4.67 (70)	56.78±3.28 (60)	59.78±5.21 (70)	61.48±4.23 (65)	60.53±4.03 (40)
LADA	**60.27±4.13** (10)	**59.58±4.26** (10)	**62.37±3.91** (20)	**67.11±5.28** (15)	**69.08±5.18** (25)

TABLE VII
THE AVERAGE CLASSIFICATION ACCURACIES (MEAN ± STD. DEV. %) ON BA DATASET. THE BEST REDUCED DIMENSIONALITY IS SHOWN IN THE BRACKETS. THE BEST RESULTS ARE IN BOLD FACE.

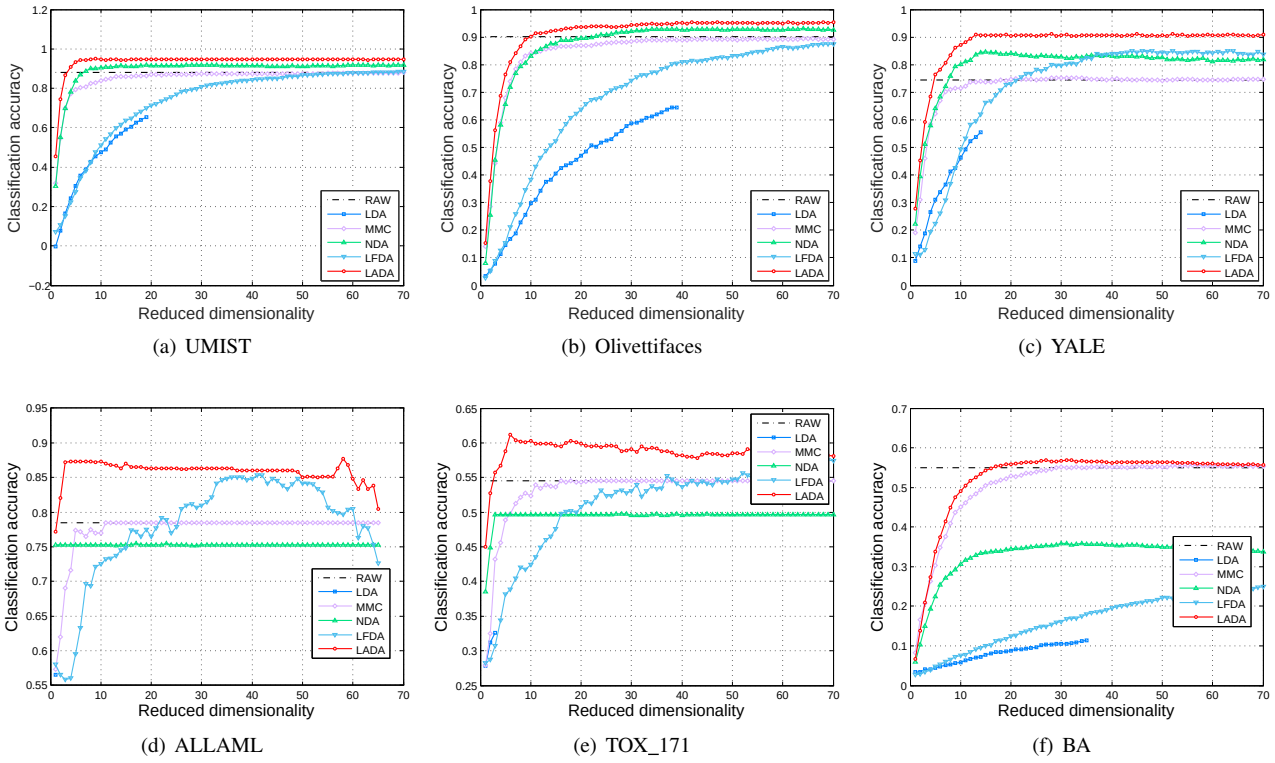
1D Methods	$N = 6$	$N = 7$	$N = 8$	$N = 9$	$N = 10$
RAW	54.13±1.42 (—)	56.41±0.91 (—)	57.61±0.85 (—)	58.18±1.85 (—)	59.73±0.90 (—)
LDA	13.01±1.26 (35)	10.32±1.99 (35)	9.72±2.84 (35)	10.12±1.27 (35)	15.65±1.17 (35)
MMC	55.18±1.50 (40)	57.27±1.03 (50)	59.86±1.51 (70)	60.20±0.96 (70)	61.81±1.38 (65)
NDA	35.40±1.63 (35)	29.18±1.85 (60)	19.94±1.17 (60)	7.57±1.48 (70)	16.39±0.63 (70)
LFDA	24.71±1.87 (70)	20.99±2.21 (70)	27.21±1.92 (70)	34.32±1.79 (70)	39.19±1.25 (70)
LADA	**56.06±1.50** (35)	**58.28±0.87** (25)	**60.25±0.48** (25)	**60.94±1.7** (40)	**62.22±0.82** (25)

testing. In UMIST, YALE, ALLAML, TOX_171 and BA, N is set as 6, 7, 8, 9 and 10. In Olivettifaces, N is set as 6, 7, 8 and 9. After performing dimensionality reduction, nearest neighbor classifier is used to classify the obtained low-dimensional data. We repeat the random split for 30 times, and report the average classification accuracy and standard deviation.

For the 1D methods, the optimal reduced dimensionality is found by grid search in the range of $[5, 10, \cdots, 70]$. Since the data dimensionality of ALLAML is 66, we search the optimal reduced dimensionality in the range of $[5, 10, \cdots, 65]$. For the 2D methods, the optimal reduced dimensionality is searched in the range of $[1\times1, 2\times2, \cdots, 24\times24]$. LDA and 2DLDA have the over-reducing problem, so their maximum dimensionalities are set as $c-1$ and $(c-1)\times(c-1)$ respectively (c is the class number).

Experimental results: the classification results on different are exhibited in Table II, III, IV, V, VI and VII respectively. Each method is with its optimal reduced dimensionality. The classification accuracy increases with the number of training samples. Among the 1D methods, LADA achieves the best performance on all the datasets. The results of LDA are unsatisfying because it can only find $c-1$ projection directions, which is insufficient for sustaining the discriminant information. MMC overcomes the over-reducing problem, and performs better than LDA. But it also assumes that the samples are Gaussian distributed. NDA and LFDA use the kNN and Gaussian graph respectively to capture the local manifold. NDA outperforms LFDA on most occasions, because the kNN graph preserves the classification boundary better than Gaussian graph. However, the performance of NDA is still inferior to the proposed LADA since it relies on the samples distances in the input data space. LADA learns the local data structure adaptively in the learned subspace, so it is robust to the data noise and shows the best classification results.

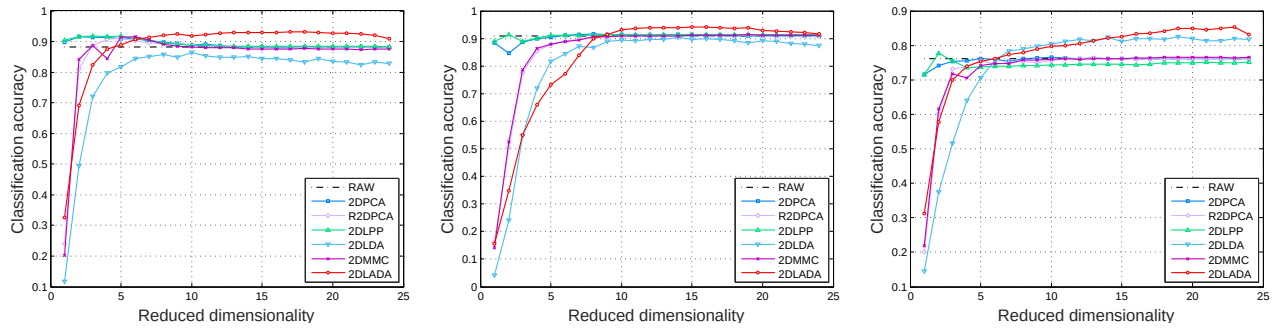
As can be seen in Table II, III and IV, most of the 2D methods outperform the baseline, which implies the fact that the discriminant features reside in a low dimensional subspace. This phenomenon is not very manifest for the 1D methods due to the utilization of PCA. Since the input data is with relatively low dimensionality, the over-reducing problem does not affect 2DLDA too much. The classification accuracies of 2DPCA, R2DPCA, 2DLPP, 2DLDA and 2DMMC are very close. Compared with the competitors, 2DLADA is able to perceive the geometrical structure, so it shows the best results. Note that, insteading of employing PCA to eliminate the null space, we just use the resized raw data for the 2D methods.



(a) UMIST (b) Olivettifaces (c) YALE

(d) ALLAML (e) TOX_171 (f) BA

Fig. 3. Average classification accuracies of 1D methods versus the reduced dimensionality on the datasets.



(a) UMIST (b) Olivettifaces (c) YALE

Fig. 4. Average classification accuracies of 2D methods versus the reduced dimensionality on (a) UMIST, (b) Olivettifaces and (c) YALE datasets.

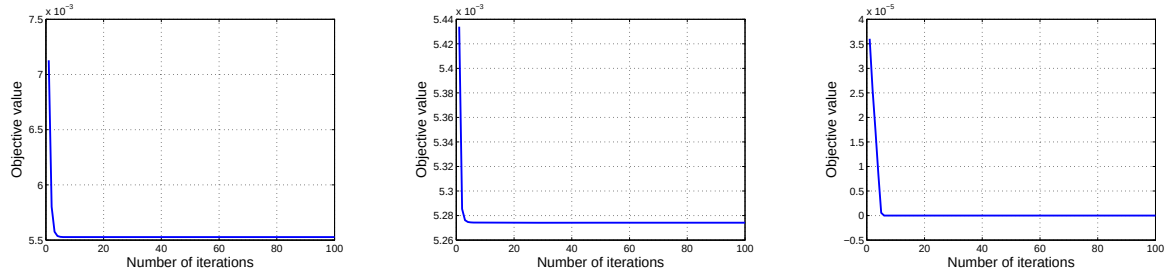
Therefore, the performance of 2DLADA is inferior to LADA on some occasions.

Fig. 3 and 4 visualize the changes of average classification accuracy with respect to the reduced dimensionality. Here N is set as 6. It can be seen that LADA achieves the highest accuracy constantly. The performance 2DLADA is slightly worse than the competitors when the dimensionality is low, but it shows promising results when the dimensionality exceeds 10. As evident from the curves, the performance of each method becomes relatively stable when the dimensionality increases to a certain value, which indicates that the input data contains redundant features and it is necessary to perform dimensionality reduction.

VI. Convergence and Parameter Sensitivity

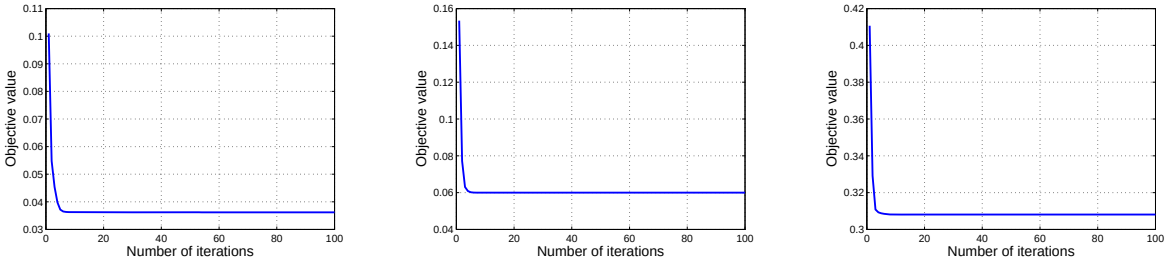
In this section, we first discuss the convergence behaviors of LADA and 2DLADA. Taking LADA as an example, the original objective function is decomposed into two sub-problems in the optimization. When updating $\mathbf{W}$, the global minimum solution is obtained by binary search. When updating $\mathbf{S}$, the final solution satisfies the KKT condition. So the objective value decreases monotonously in the optimization of each variable, and reaches to a local optimal value after a few iterative steps. Fig. 5 and 6 plot the convergence curves of LADA and 2DLADA, the optimization methods converge very fast, which ensures the efficiency.

In addition, we also study the parameter sensitivity of the locality-aware competitors. We run NDA, LFDA and LADA on two random split of YALE dataset (different training and



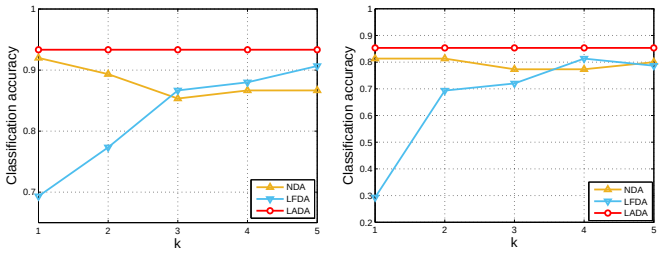
(a) UMIST (b) Olivettifaces (c) YALE

Fig. 5. Convergence curves of LADA on (a) UMIST, (b) Olivettifaces and (c) YALE datasets.



(a) UMIST (b) Olivettifaces (c) YALE

Fig. 6. Convergence curves of 2DLADA on (a) UMIST, (b) Olivettifaces and (c) YALE datasets.



(a) Split 1 of YALE (b) Split 2 of YALE

Fig. 7. Classification accuracy of NDA, LFDA and LADA on two random splits of YALE dataset. The results of NDA and LFDA change dramatically with the scaling parameter k.

testing sets), and plot the classification accuracy curves with respect to the scaling parameter k. As shown in Fig. 7, the performance of NDA and LFDA change dramatically with the value of k. The decision of k is important for the locality-aware LDA methods. However, even on the same dataset, the optimal k is different for two random splits. Thus, it is impractical to decide an appropriate k for various real-world tasks. The proposed LADA does not have this problem because it is totally parameter-free.

VII. Conclusions

In this paper, we propose a new supervised dimensionality reduction framework, including two algorithms, Locality Adaptive Discriminant Analysis (LADA) and 2-Dimensional LADA (2DLADA). LADA works with the data in vector representation, while 2DLADA processes the matrix data directly. The proposed methods integrate local structure learning and discriminant analysis jointly, so they are able to exploit the underlying data structure within the desired subspace. Since they do not rely on the data relationship in the input space, the influence of noise is alleviated. In addition, the over-reducing and SSS problems are avoided implicitly in our methods. Experimental results show that the proposed methods achieve state-of-the-art performance on face image classification. The parameter-free property improves the applicability of LADA and 2DLADA.

In the future work, it is desirable to extend the proposed framework to large-scale unlabeled data processing tasks. Besides, we also plan to learn the pair-wise relationship with the Graph Convolutional Network (GCN), which has been demonstrated to be effective on graph learning.

Acknowledgments

This work was supported by The National Key Research and Development Program of China under Grant 2018YFB1107400, and The National Natural Science Foundation of China under U1864204, 61773316, U1801262, and 61871470.

References

[1] X. Lan, M. Ye, R. Shao, B. Zhong, P. C. Yuen, and H. Zhou, "Learning modality-consistency feature templates: A robust rgb-infrared tracking system," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 12, pp. 9887–9897, 2019.

[2] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 2, no. 1-3, pp. 37–52, 1987.

[3] Y. Liu, X. Gao, Q. Gao, L. Shao, and J. Han, "Adaptive robust principal component analysis," *Neural Networks*, vol. 119, pp. 85–92, 2019.

[4] M. Friedman and A. Kandel, *Introduction to Pattern Recognition - Statistical, Structural, Neural and Fuzzy Logic Approaches*, ser. Series in Machine Perception and Artificial Intelligence. WorldScientific, 1999, vol. 32.

[5] X. Liu, L. Wang, J. Zhang, J. Yin, and H. Liu, "Global and local structure preservation for feature selection," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 6, pp. 1083–1095, 2014.

[6] J. Xi, X. Yuan, M. Wang, A. Li, X. Li, and Q. Huang, "Inferring subgroup-specific driver genes from heterogeneous cancer samples via subspace learning with subgroup indication," vol. 36, no. 6, pp. 1855–1863, 2020.

[7] T. Luo, C. Hou, F. Nie, and D. Yi, "Dimension reduction for non-gaussian data by adaptive discriminative analysis," *IEEE Transactions on Cybernetics*, vol. 49, no. 3, pp. 933–946, 2019.

[8] Z. Li, Z. Zhang, J. Qin, Z. Zhang, and L. Shao, "Discriminative fisher embedding dictionary learning algorithm for object recognition," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 3, pp. 786–800, 2020.

[9] M. Zhao, Z. Zhang, T. Chow, and B. Li, "Soft label based linear discriminant analysis for image recognition and retrieval," *Computer Vision and Image Understanding*, vol. 121, pp. 86–99, 2014.

[10] X. Lan, M. Ye, S. Zhang, and P. C. Yuen, "Robust collaborative discriminative learning for rgb-infrared tracking," in *AAAI Conference on Artificial Intelligence*, 2018, pp. 7008–7015.

[11] Q. Wang, Z. Qin, F. Nie, and X. Li, "C2DNDA: A deep framework for nonlinear dimensionality reduction," *IEEE Transactions on Industrial Electronics*, vol. 68, no. 2, pp. 1684–1694, 2021.

[12] Q. Wang, Z. Qin, F. Nie, and Y. Yuan, "Convolutional 2d LDA for nonlinear dimensionality reduction," in *International Joint Conference on Artificial Intelligence*, 2017, pp. 2929–2935.

[13] H. Wan, G. Guo, H. Wang, and X. Wei, "A new linear discriminant analysis method to address the over-reducing problem," in *International Conference on Pattern Recognition and Machine Intelligence*, 2015, pp. 65–72.

[14] R. Huang, Q. Liu, H. Lu, and S. Ma, "Solving the small sample size problem of lda," in *Pattern Recognition*, vol. 3, 2002, pp. 29–32.

[15] X. Liu, M. Li, C. Tang, J. Xia, J. Xiong, L. Liu, M. Kloft, and E. Zhu, "Efficient and effective regularized incomplete multi-view clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, DOI:10.1109/TPAMI.2020.2974828, 2020.

[16] X. Liu, X. Zhu, M. Li, L. Wang, C. Tang, J. Yin, D. Shen, H. Wang, and W. Gao, "Late fusion incomplete multi-view clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 10, pp. 2410–2423, 2019.

[17] X. Li, M. Chen, F. Nie, and Q. Wang, "A multiview-based parameter free framework for group detection," in *AAAI Conference on Artificial Intelligence*, 2017, pp. 4147–4153.

[18] J. Wang, A. Suzuki, L. Xu, F. Tian, L. Yang, and K. Yamanishi, "Orderly subspace clustering," in *AAAI Conference on Artificial Intelligence*, 2019, pp. 5264–5272.

[19] J. Wang, L. Xu, F. Tian, A. Suzuki, C. Zhang, and K. Yamanishi, "Attributed subspace clustering," in *International Joint Conference on Artificial Intelligence*, S. Kraus, Ed., 2019, pp. 3719–3725.

[20] J. Lu, K. Plataniotis, and A. Venetsanopoulos, "Regularization studies of linear discriminant analysis in small sample size scenarios with application to face recognition," *Pattern Recognition Letters*, vol. 26, no. 2, pp. 181–191, 2005.

[21] H. Li, T. Jiang, and K. Zhang, "Efficient and robust feature extraction by maximum margin criterion," *IEEE Transactions on Neural Networks*, vol. 17, no. 1, pp. 157–165, 2006.

[22] J. Ye, R. Janardan, and Q. Li, "Two-dimensional linear discriminant analysis," in *Advances in Neural Information Processing Systems*, 2004, pp. 1569–1576.

[23] F. Nie, S. Xiang, and C. Zhang, "Neighborhood minmax projections." in *International Joint Conference on Artificial Intelligence*, 2007, pp. 993–998.

[24] D. Cai, X. He, K. Zhou, J. Han, and H. Bao, "Locality sensitive discriminant analysis." in *International Joint Conference on Artificial Intelligence*, 2007, pp. 708–713.

[25] M. Bressan and J. Vitria, "Nonparametric discriminant analysis and nearest neighbor classification," *Pattern Recognition Letters*, vol. 24, no. 15, pp. 2743–2749, 2003.

[26] Z. Fan, Y. Xu, and D. Zhang, "Local linear discriminant analysis framework using sample neighbors," *IEEE Transactions on Neural Networks*, vol. 22, no. 7, pp. 1119–1132, 2011.

[27] X. Li, M. Chen, F. Nie, and Q. Wang, "Locality adaptive discriminant analysis," in *International Joint Conference on Artificial Intelligence*, 2017, pp. 2201–2207.

[28] Y. Jia, F. Nie, and C. Zhang, "Trace ratio problem revisited," *IEEE Transactions on Neural Networks*, vol. 20, no. 4, pp. 729–735, 2009.

[29] X. Li, M. Chen, and Q. Wang, "Self-tuned discrimination-aware method for unsupervised feature selection," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 8, pp. 2275–2284, 2019.

[30] M. Chen, Q. Wang, and X. Li, "Discriminant analysis with graph learning for hyperspectral image classification," *Remote Sensing*, vol. 10, no. 6, p. 836, 2018.

[31] M. Sugiyama, "Local fisher discriminant analysis for supervised dimensionality reduction," in *International conference on Machine learning*, 2006, pp. 905–912.

[32] K. Weinberger and L. Saul, "Distance metric learning for large margin nearest neighbor classification," *Journal of Machine Learning Research*, vol. 10, pp. 207–244, 2009.

[33] Y. Dong, B. Du, L. Zhang, and L. Zhang, "Dimensionality reduction and classification of hyperspectral images

using ensemble discriminative local metric learning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 5, pp. 2509–2524, 2017.
[34] L. Zhang, H. P. H. Shum, and L. Shao, "Discriminative semantic subspace analysis for relevance feedback," *IEEE Transactions on Image Processing*, vol. 25, no. 3, pp. 1275–1287, 2016.
[35] F. Nie, Z. Wang, R. Wang, Z. Wang, and X. Li, "Adaptive local linear discriminant analysis," *ACM Transactions on Knowledge Discovery from Data*, vol. 14, no. 1, pp. 9:1–9:19, 2020.
[36] F. Nie, Z. Wang, R. Wang, and X. Li, "Submanifold-preserving discriminant analysis with an auto-optimized graph," *IEEE Transactions on Cybernetics*, vol. 50, no. 8, pp. 3682–3695, 2020.
[37] P. Sanguansat, W. Asdornwised, S. Jitapunkul, and S. Marukatat, "Two-dimensional linear discriminant analysis of principle component vectors for face recognition," *IEICE Transactions*, vol. 89-D, no. 7, pp. 2164–2170, 2006.
[38] J. Yang, D. Zhang, A. Frangi, and J. Yang, "Two-dimensional PCA: A new approach to appearance-based face representation and recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 1, pp. 131–137, 2004.
[39] W. Yang and D. Dai, "Two-dimensional maximum margin feature extraction for face recognition," *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, vol. 39, no. 4, pp. 1002–1012, 2009.
[40] H. Wang, S. Chen, and Z. Hu, "Image recognition using weighted two-dimensional maximum margin criterion," in *International Conference on Natural Computation*, 2007, pp. 582–586.
[41] Y. Guo, S. Li, J. Yang, T. Shu, and L. Wu, "A generalized foley-sammon transform based on generalized fisher discriminant criterion and its application to face recognition," *Pattern Recognition Letters*, vol. 24, no. 1, pp. 147–158, 2003.
[42] D. Graham and N. Allinson, "Characterising virtual eigensignatures for general purpose face recognition," *Face Recognition Form Theory to Applications*, vol. 163, no. 2, pp. 446–456, 1998.
[43] F. Samaria and A. Harter, "Parameterisation of a stochastic model for human face identification," in *IEEE Workshop on Applications of Computer Vision*, 1994, pp. 138–142.
[44] A. Georghiades, P. Belhumeur, and D. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE transactions on pattern analysis and machine intelligence*, vol. 23, no. 6, pp. 643–660, 2001.
[45] J. Li, K. Cheng, S. Wang, F. Morstatter, T. Robert, J. Tang, and H. Liu, "Feature selection: A data perspective," *arXiv:1601.07996*, 2016.
[46] P. Belhumeur, J. Hespanha, and D. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711–720, 1997.
[47] R. Zhang, F. Nie, and X. Li, "Auto-weighted two-dimensional principal component analysis with robust outliers," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017, pp. 6065–6069.
[48] D. Hu, G. Feng, and Z. Zhou, "Two-dimensional locality preserving projections (2DLPP) with its application to palmprint recognition," *Pattern Recognition*, vol. 40, no. 1, pp. 339–342, 2007.

Xuelong Li (M'02-SM'07-F'12) is currently a Full Professor with the School of Computer Science and with the Center for OPTical Imagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an 710072, China.

Qi Wang (M'15-SM'15) received the B.E. degree in automation and the Ph.D. degree in pattern recognition and intelligent systems from the University of Science and Technology of China, Hefei, China, in 2005 and 2010, respectively. He is currently a Professor with the School of Computer Science and with the Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision and pattern recognition.

Feiping Nie received the Ph.D. degree in computer science from Tsinghua University, Beijing, China, in 2009.

He is currently a Professor with the Center for OPTical Imagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an, China. He has authored over 100 papers in prestigious journals and conferences like the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, the IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, the International Conference on Machine Learning, the Conference on Neural Information Processing Systems, and the Conference on Knowledge Discovery and Data Mining. His current research interests include machine learning and its applications fields, such as pattern recognition, data mining, computer vision, image processing, and information retrieval.

Dr. Nie serves as an Associate Editor or a PC Member for several prestigious journals and conferences in the related fields.

Mulin Chen received the B.E. degree in software engineering and the Ph.D. degree in computer application technology from Northwestern Polytechnical University, Xi'an, China, in 2014 and 2019 respectively. He is currently a researcher with the Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an, China. His current research interests include computer vision and machine learning.