Dimension Reduction for Non-Gaussian Data by Adaptive Discriminative Analysis

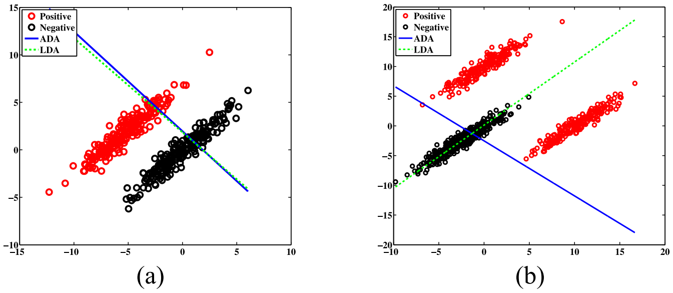
Tingjin Luo, Chenping Hou, *Member, IEEE*, Feiping Nie, and Dongyun Yi

***Abstract*—High-dimensional non-Gaussian data are ubiquitous in many real applications. Face recognition is a typical example of such scenarios. The sampled face images of each person in the original data space are more closely located to each other than to those of the same individuals due to the changes of various conditions like illumination, pose variation, and facial expression. They are often non-Gaussian and differentiating the importance of each data point has been recognized as an effective approach to process the high-dimensional non-Gaussian data. In this paper, to embed non-Gaussian data well, we propose a novel unified framework named adaptive discriminative analysis (ADA), which combines the sample's importance measurement and subspace learning in a unified framework. Therefore, our ADA can preserve the within-class local structure and learn the discriminative transformation functions simultaneously by minimizing the distances of the projected samples within the same classes while maximizing the between-class separability. Meanwhile, an efficient method is developed to solve our formulated problem. Comprehensive analyses, including convergence behavior and parameter determination, together with the relationship to other related approaches, are as well presented. Systematical experiments are conducted to understand the work of our proposed ADA. Promising experimental results on various types of real-world benchmark data sets are provided to examine the effectiveness of our algorithm. Furthermore, we have also evaluated our method in face recognition. They all validate the effectiveness of our method on processing the high-dimensional non-Gaussian data.**

***Index Terms*—Adaptive discriminative analysis (ADA), dimensionality reduction, face recognition, high-dimensional non-Gaussian data, linear discriminant analysis (LDA).**

I. Introduction

HIGH-DIMENSIONAL non-Gaussian data [1]–[4] are ubiquitous in many real applications, especially in the fields of face recognition [5]–[9]. Due to the effect of the changes of uncontrolled conditions like complex backgrounds, illumination, pose variation, occlusion, and facial expressions, in the original feature space the sampled face images of different persons are located more closely than those of the same individual. If we assume such type of data are Gaussian distributed, the domain of Gaussian variable violates the boundary property. Therefore, face recognition [8], [10], [11] is a typical example of such scenarios. Moreover, face recognition plays an important role in security systems and intelligent surveillance system. It is still a hot topic in the area of machine learning, computer vision and pattern recognition. However, directly processing such high-dimensional non-Gaussian data not only degrades its performance but also is time consuming [12]–[16] in learning tasks. To solve this problem, many dimensionality reduction methods have been introduced for data processing, such as non-negative matrix factorization-based methods [17], [18], k-dimensional coding schemes-based methods [19], and so on. While dimensionality reduction can be applied to both supervised and unsupervised learning, we focus on the problem of supervised learning for non-Gaussian data, where the label information is available.



(a) (b)

Fig. 1. Dashed line is one projection of LDA, and the solid line is one projection of ADA, for two artificial data sets. (a) Data of each class is sampled from a Gaussian distribution. (b) One class is sampled from a Gaussian distribution, while the other class is sampled from a Gaussian mixture model with two kernels. Note that LDA fails to find the optimal direction in the case with non-Gaussian distributed data.

Although non-Gaussian data arises in many fields, the original data is always assumed as Gaussian distributed in the traditional methods, such as linear discriminative analysis (LDA) [3], [20] and its variants unrelated LDA (ULDA) [21], orthogonal LDA (OLDA) [22], [23], max–min distance analysis [24], and weighted Harmonic mean of trace ratios for multiclass discriminant analysis (WHMDA) [10], [25]. Placing the Gaussian assumption on the non-Gaussian data will discard a great deal of useful structural information [3], [26]. As seen from the toy example

Manuscript received January 9, 2017; revised August 20, 2017; accepted December 29, 2017. Date of publication January 24, 2018; date of current version February 14, 2019. This work was supported by the National Natural Science Foundation of China under Grant 61473302 and Grant 61503396. This paper was recommended by Associate Editor B. Zhang. *(Corresponding author: Chenping Hou.)*

T. Luo, C. Hou, and D. Yi are with the College of Science, National University of Defense Technology, Changsha 410073, China (e-mail: luotingjin03@nudt.edu.cn; hcpnudt@hotmail.com; dongyun.yi@gmail.com).

F. Nie is with the Center for Optical Imagery Analysis and Learning, Northwestern Polytechnical University, Xi'an 710072, China (e-mail: feipingnie@gmail.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCYB.2018.2789524

in Fig. 1(left), the methods based on Gaussian assumption will fail to solve the non-Gaussian data. To address this problem, many dimensionality reduction algorithms [11], [27]–[29] have been proposed in the literature.

According to the ways to incorporate in evaluating and measuring the contribution of samples, the dimensionality reduction methods for non-Gaussian data can be organized into two categories.

1) Locality-based methods [27]–[29], which differentiate the importance of samples using the local structure chosen by k-nearest neighbors (kNNs) method.
2) Adaptivity-based methods [11], [26], which measure the contribution of instances by the training data adaptively.

To process the non-Gaussian data well, Qiu and Wu [29] proposed stepwise nearest neighbor discriminant analysis (SNNDA) which further extended nonparametric discriminant analysis method [30] with a stepwise process based on margin maximum criterion [31] and the local structural information. However, the stepwise process is time-consuming, and in the construction of the scatter matrices, the values of the weight matrix are empirical and only the nearest neighbor is used, which is sensitive to noise. Nie *et al.* [27] proposed neighborhood minmax projections (NMMPs) by using the pairwise points where the two points are neighbors of each other. Fan *et al.* [28] proposed local LDA (LLDA) using sample neighbors selected by kNN and adding the affinity matrix to weight the importance of samples.

To measure the importance of samples adaptively, Roweis and Saul [12] and Belkin and Niyogi [32] proposed locally linear embedding and Laplacian eigenmaps (LEs) to distinguish the contribution of samples adaptively and preserve the local structural information by heat kernel, respectively. However, this nonlinear property makes them computationally expensive. Moreover, they yield mappings that are defined only on the training data points and it remains unclear how to naturally evaluate the maps on new-coming testing points. Thus, He and Niyogi [26] proposed local preserve projections (LPPs), which is a linear approximation of the nonlinear LE. However, LPP is an unsupervised dimensionality reduction method and does not use the label information. LPP may not obtain good performance for classification task. Thus, Sugiyama [11] proposed local Fisher discriminant analysis (LFDA) to preserve within-class local structure by combining the ideas of LPP [26] and LDA. LFDA adopts the two-step strategy, which first computed the importance of each sample in the same class by kNN-based heat kernel [26] and then extracted the new feature representation.

From the above analysis, we can easily find that, differentiating the importance of each data point is an effective strategy to process the high-dimensional non-Gaussian data. The idea of differentiating the importance of samples has been applied into many machine learning methods. For example, support vector machine (SVM) [3], [33] computes the optimal classification hyperplane and boosting methods [34], [35] obtains the weak learners by adaptively weighting the importance of each sample. Furthermore, compared with processing Gaussian distributed data, the locality and adaptivity are two effective approaches to discriminate the importance of each sample for the non-Gaussian data. Nevertheless, these existing locality-based methods require many parameters to construct the model and cannot weight the importance of instances automatically, while these adaptivity-based methods measure the importance of samples adaptively only in the original feature space. Therefore, we propose a novel adaptive approach to measure the importance of samples adaptively and simultaneously preserve the local structural information.

TABLE I
NOTATIONS AND DEFINITIONS

Notations	Descriptions
d	Dimensionality of the original data
n	Data size
c	Number of classes
m	Reduced dimensionality
$\boldsymbol{x}_i \in \mathbb{R}^d$	The i-th data point
$\mathcal{N}_k(\boldsymbol{x})$	The k nearest neighbors of $\boldsymbol{x}$
$y_i \in \mathbb{N}$	The label of i-th data point
$\boldsymbol{x}_i^k \in \mathbb{R}^d$	The i-th data point in the k-th class
$\mathbf{X}_i \in \mathbb{R}^{d\times n_i}$	Data matrix in the i-th class
$\mathbf{X} \in \mathbb{R}^{d\times n}$	Data matrix
$\boldsymbol{y} \in \mathbb{N}^n$	The labels of $\mathbf{X}$
$\boldsymbol{u} \in \mathbb{R}^d$	The total sample mean vector
$\boldsymbol{u}_i \in \mathbb{R}^d$	The mean vector of the i-th class
$\mathbf{S}_t \in \mathbb{R}^{d\times d}$	Total scatter matrix
$\mathbf{S}_w \in \mathbb{R}^{d\times d}$	Within-class scatter matrix
$\mathbf{S}_b \in \mathbb{R}^{d\times d}$	Between-class scatter matrix
$\boldsymbol{W} \in \mathbb{R}^{d\times m}$	Transformation matrix

Based on the adaptive approach, we introduce a novel framework, referred to as adaptive discriminative analysis (ADA), for high-dimensional non-Gaussian data. To embed non-Gaussian distributed data well, it combines the sample's importance measurement and subspace learning into a unified framework. Some popular methods such as LDA and LFDA, can be viewed as special cases within the proposed framework. By using this approach, ADA not only discriminate the contribution of each instance in the same class, but also weights the importance of instance in the learned low-dimensional feature space automatically. ADA learns the optimal transformation matrix to preserve more discriminative information by maximizing the similarity of pairwise within-class samples. As illustrated in Fig. 1, ADA can discover the most discriminative direction in both cases, while LDA fails to find the optimal discriminative directions for non-Gaussian data. We also provide an efficient method to solve the proposed problem in an alternate way. Meanwhile, comprehensive analyses, including convergence behavior and the relationships to other related approaches, are provided. Compared with traditional methods, our algorithm is demonstrated to have better performance on systematic data and various types of real-world benchmark data sets. Furthermore, we have also evaluated our method in a real application scenario: face recognition. These results all verify the effectiveness of ADA. Besides, we highlight the contributions of this paper as follows.

1) Propose an unified framework for supervised dimensionality reduction of high-dimensional data, which combines the measurement of samples' importance and the subspace learning.
2) Provide an unified view to analyze many traditional methods by our proposed framework. Furthermore, it

encourages us to develop a new algorithm for high-dimensional non-Gaussian data.

3) Develop an efficient method to solve our formulated problem and rigorously analyze the performance of our method in aspects of the convergence behavior and connection with related methods.
4) Evaluate the effectiveness of ADA by extensive experimental results on synthetic data and various kinds of real-world data sets. Moreover, We apply ADA to face recognition and demonstrate its promising performance in real applications.
5) Only one parameter needs to be tuned in our model.

The rest of this paper is organized as follows. Section II summarizes the related works. We formulate the proposed ADA and provide an effective method of solving this problem in Section III. We discuss convergence analysis and the relationships to prior related works in Section IV. Section V provides promising comparison results on various kinds of data sets. We evaluate our method on face recognition in Section VI, followed by the conclusions and future works in Section VII. For convenience, the important notations used in this paper are summarized in Table I.

II. Related Work

A. Review of Linear Discriminant Analysis

Let $\mathbf{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n] \in \mathbb{R}^{d \times n}$ be a data matrix consisting of n data points in $\mathbb{R}^d$. In classical LDA [3], the within-class and between-class scatter matrices are defined as

$$\mathbf{S}_w = \frac{1}{n} \sum_{i=1}^{c} \sum_{\boldsymbol{x} \in \mathcal{X}_i} (\boldsymbol{x} - \boldsymbol{u}_i)(\boldsymbol{x} - \boldsymbol{u}_i)^T$$

$$\mathbf{S}_b = \frac{1}{n} \sum_{i=1}^{c} n_i (\boldsymbol{u}_i - \boldsymbol{u})(\boldsymbol{u}_i - \boldsymbol{u})^T \tag{1}$$

where $\mathcal{X}_i$ represents the feature set of ith class, n_i and $\boldsymbol{u}_i$ are the sample size and the centroid of ith class, respectively, and $\boldsymbol{u}$ is the global mean vector of $\mathbf{X}$. It follows from the definitions that $\text{tr}(\mathbf{S}_w)$ measures the within-class cohesion and $\text{tr}(\mathbf{S}_b)$ measures the between-class separation, where $\text{tr}(\cdot)$ of a square matrix is the summation of its diagonal entries. The objective function of classical LDA [23] is formulated as

$$\max_{\boldsymbol{W} \in \mathbb{R}^{d \times m}} \text{tr}\Big(\big(\boldsymbol{W}^T \mathbf{S}_t \boldsymbol{W}\big)^{-1} \boldsymbol{W}^T \mathbf{S}_b \boldsymbol{W}\Big). \tag{2}$$

The problem in (2) has a closed form solution, i.e., the m eigenvectors of $\mathbf{S}_t^{-1}\mathbf{S}_b$ corresponding to the m largest nonzero eigenvalues, provided that the within-class scatter matrix $\mathbf{S}_w$ is nonsingular [3], [23]. Note that classical LDA does not handle singular scatter matrices, which limits its applicability to the high-dimensional under-sampled problems. Several methods, including ULDA [21] and OLDA [22], [23], were proposed to deal with such singularity problem. The key properties of ULDA is to obtain the uncorrelated feature and reduce the redundancy in the lower-dimensional space. ULDA aims to find the optimal $\mathbf{S}_t$-orthogonal discriminant vectors.[1] ULDA aims to compute the optimal $\boldsymbol{W}$, such that

$$\max_{\boldsymbol{W}^T \mathbf{S}_t \boldsymbol{W} = I} \text{tr}\Big(\big(\boldsymbol{W}^T \mathbf{S}_w \boldsymbol{W}\big)^{-1} \boldsymbol{W}^T \mathbf{S}_b \boldsymbol{W}\Big) \tag{3}$$

where $\mathbf{S}_t = \mathbf{S}_w + \mathbf{S}_b$ is the total scatter matrix. OLDA requires that the discriminant vectors are orthogonal to each other. The optimal transformation in OLDA can be computed by solving the following optimization problem:

$$\max_{\boldsymbol{W}^T \boldsymbol{W} = I} \text{tr}\Big(\big(\boldsymbol{W}^T \mathbf{S}_t \boldsymbol{W}\big)^{+} \boldsymbol{W}^T \mathbf{S}_b \boldsymbol{W}\Big) \tag{4}$$

where M^+ denotes the pseudo-inverse of matrix M. Finally, the optimal transformation of OLDA can be computed by diagonalizing the three scatter matrices simultaneously.

However, ULDA and OLDA still fail to solve the high-dimensional non-Gaussian problems, since they are same as classical LDA under the assumption of Gaussian distribution. LLDA and LFDA were proposed to deal with this problem.

B. Local Linear Discriminant Analysis

Fan *et al.* [28] proposed an LLDA framework for the non-Gaussian distributed classification problem. LLDA captures the local structural information using training sample neighbors chosen by kNN. The within-class and between-class scatter matrices are defined as follows:

$$\begin{cases} \tilde{\mathbf{S}}_w = \sum_{i=1}^{\tilde{c}} \tilde{n}_i (\tilde{\boldsymbol{u}}_i - \tilde{\boldsymbol{u}})(\tilde{\boldsymbol{u}}_i - \tilde{\boldsymbol{u}})^T \\ \tilde{\mathbf{S}}_b = \sum_{i=1}^{\tilde{c}} \sum_{j=1}^{\tilde{n}_i} \Big(\tilde{\boldsymbol{x}}_j^i - \tilde{\boldsymbol{u}}_i\Big)\Big(\tilde{\boldsymbol{x}}_j^i - \tilde{\boldsymbol{u}}_i\Big)^T \end{cases} \tag{5}$$

where $\tilde{c}$, $\tilde{n}_i$, $\tilde{\boldsymbol{x}}_j^i$, $\tilde{\boldsymbol{u}}_i$, and $\tilde{\boldsymbol{u}}$ are the number of classes, the number of the ith class, the jth sample, the mean vector of the ith class, and the mean vector of the total k determined nearest neighbors, respectively. Then, based on the local Fisher criterion, the objective function of LLDA is expressed as

$$\max_{\boldsymbol{W}} \text{tr}\bigg(\Big(\boldsymbol{W}^T \tilde{\mathbf{S}}_w \boldsymbol{W}\Big)^{-1} \boldsymbol{W}^T \tilde{\mathbf{S}}_b \boldsymbol{W}\bigg). \tag{6}$$

C. Local Fisher Discriminant Analysis

LFDA [11] incorporates the merits of LPP and LDA for supervised dimensionality reduction. The main idea is that both $\mathbf{S}_w$ and $\mathbf{S}_b$ are weighted by the affinity matrix $\hat{\mathbf{A}}$ of the training data. Each element of $\hat{\mathbf{A}}$ is calculated using a local scaling method [36], i.e., choosing kNNs and assigning individual scaling for samples

$$\hat{A}_{ij} = \begin{cases} \exp\Big(-\delta \left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\|_2^2\Big), & \boldsymbol{x}_j \in \mathcal{N}_k(\boldsymbol{x}_i) \\ 0, & \text{otherwise} \end{cases}$$

where $\mathcal{N}_k(\cdot)$ represents the kNNs. With the use of affinity matrix, $\hat{\mathbf{S}}_w$ and $\hat{\mathbf{S}}_b$ can be computed by

$$\begin{cases} \hat{\mathbf{S}}_b = \sum_{i,j=1}^{n} \hat{A}_{ij}^b \big(\boldsymbol{x}_i - \boldsymbol{x}_j\big)\big(\boldsymbol{x}_i - \boldsymbol{x}_j\big)^T \\ \hat{\mathbf{S}}_w = \sum_{i,j=1}^{n} \hat{A}_{ij}^w \big(\boldsymbol{x}_i - \boldsymbol{x}_j\big)\big(\boldsymbol{x}_i - \boldsymbol{x}_j\big)^T \end{cases} \tag{7}$$

where $\hat{A}_{ij}^b = \hat{A}_{ij}/k$ if $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ belong to the same class, otherwise $\hat{A}_{ij}^b = 0$ and $\hat{A}_{ij}^w = \hat{A}_{ij}((1/n) - (1/k))$ if $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ belong to the same class, otherwise $\hat{A}_{ij}^w = 0$.

[1] For any vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ are $\mathbf{S}_t$-orthogonal, if $\boldsymbol{x}^T \mathbf{S}_t \boldsymbol{y} = 0$.

Finally, the optimal transformation matrix $\boldsymbol{W}$ can be obtained by solving the following objective function:

$$\max_{\boldsymbol{W}} \ \operatorname{tr}\left(\left(\boldsymbol{W}^T\hat{\mathbf{S}}_w\boldsymbol{W}\right)^{-1}\boldsymbol{W}^T\hat{\mathbf{S}}_b\boldsymbol{W}\right). \tag{8}$$

III. Adaptive Discriminative Analysis for Supervised Dimensionality Reduction

A. Motivation and Formulation of Our Framework

Given the training data matrix $\mathbf{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n] \in \mathbb{R}^{d\times n}$ comprising examples of the input vectors along with their corresponding labels $\boldsymbol{y} = [y_1, \ldots, y_n]$. Denote $\mathbf{X}_i = [\boldsymbol{x}_1^i, \ldots, \boldsymbol{x}_{n_i}^i] \in \mathbb{R}^{d\times n_i}(1 \le i \le c)$ as data matrix belonging to the ith class. Thus the training data matrix $\mathbf{X}$ can be represented as $[\mathbf{X}_1, \ldots, \mathbf{X}_c]$. $\mathbf{1} = [1, 1, \ldots, 1]^T \in \mathbb{R}^{n\times 1}$ and $\boldsymbol{W} = [\boldsymbol{w}_1, \ldots, \boldsymbol{w}_m] \in \mathbb{R}^{d\times m}$ is the transformation matrix.

As seen in Fig. 1, classic LDA [3], [20] and its variants [21], [23] perform poorly if samples in some class are non-Gaussian distributed and form several separate clusters (i.e., multimodal). The undesired behavior is by the globality property of the within-class scatter and between-class scatter matrices in these methods. In other words, Gaussian assumption with equal co-variance structure for the data is too strong in LDA. It also leads to ignore the difference of within-class and between-class samples. Intuitively, the contribution of each sample in the same class should be not equal for the non-Gaussian distributed data. For instance, the maximum-margin hyperplane of SVM [3] is determined by the support vectors (SVs), which are the samples nearest to the hyperplane. Moreover, the hyperplane of SVM can be computed by its SVs efficiently. Boosting methods [34] obtain the weak learners by discriminating the importance of each sample gradually and then improve the classification performance.

On the other hand, by analyzing the properties of manifold learning methods [12], [26], [32], it can be found that the global structure of nonlinear manifolds can be represented by a locally linear structure and differentiating the importance of each sample is an effective approach to process the non-Gaussian data. If data points are close in the original high-dimensional space, these manifold learning methods can make samples of different classes overlapped by persevering local structure. Motivated by this idea, LLDA and LFDA adaptively measure the importance of each instance by kNN-based methods. However, LLDA and LFDA only measured the importance of samples in the original feature space. Moreover, they empirically adopted kNN-based method to compute the sample's importance. LLDA and LFDA are not only sensitive to noise or outliers, but also required more priori information to determine the number of nearest neighbors.

Considering the above analysis of LLDA and LFDA, the sample's importance measure and subspace learning interact with each other. In the different feature spaces, the distributions of data points are very different and the contribution of each sample in the same class will be different. The optimal transformation matrix $\boldsymbol{W}$ will alter with the change of sample's importance. Therefore, the solutions of LLDA and LFDA are not optimal, only measuring sample's importance in the original feature space. Their performances can be improved. To solve these problems, we propose a unified framework to jointly measure the sample's importance and extract the discriminative features for high-dimensional non-Gaussian data. The objective function of this framework can be summarized as

$$\max \frac{1}{2n}\sum_{i=1}^{c}\frac{1}{n_i}\sum_{j=1}^{n_i}\sum_{k=1}^{n_i}\exp\left(-\delta\cdot l\left(\boldsymbol{W}, \boldsymbol{x}_j^i, \boldsymbol{x}_k^i\right)\right) \tag{9}$$

where $\delta > 0$ is the scale parameter and $l(\boldsymbol{W}, \boldsymbol{x}_j^i, \boldsymbol{x}_k^i)$ is a metric which measures the distance between pairwise points $\boldsymbol{x}_j^i$ and $\boldsymbol{x}_k^i$ in the projected space, such as Euclidean distance $l(\boldsymbol{W}, \boldsymbol{x}_j^i, \boldsymbol{x}_k^i) = \|\boldsymbol{W}^T(\boldsymbol{x}_j^i - \boldsymbol{x}_k^i)\|_2^2$ and ℓ_1-norm distance $l(\boldsymbol{W}, \boldsymbol{x}_j^i, \boldsymbol{x}_k^i) = \|\boldsymbol{W}^T(\boldsymbol{x}_j^i - \boldsymbol{x}_k^i)\|_1$.

It is worth noting that $l(\boldsymbol{W}, \boldsymbol{x}_j^i, \boldsymbol{x}_k^i)$ is able to preserve the local structural information and the term $\exp(-\delta\cdot l(\boldsymbol{W}, \boldsymbol{x}_j^i, \boldsymbol{x}_k^i))$ adaptively weights the similarity between the pairwise points $\boldsymbol{x}_j^i$ and $\boldsymbol{x}_k^i$ in the low-dimensional feature subspace. When $\boldsymbol{W} = I$ is an identity matrix, $\exp(-\delta\cdot l(\boldsymbol{W}, \boldsymbol{x}_j^i, \boldsymbol{x}_k^i))$ is equivalent to compute the similarity between the pairwise points $\boldsymbol{x}_j^i$ and $\boldsymbol{x}_k^i$ in original feature space, that is, the samples' importance measure of LLDA and LFDA is a spacial case of our framework. Moreover, the formulation in (10) not only measures the within-class similarity in the learning subspace, but it will also help to choose the neighbors of current sample in the same class automatically. For example, when the distance between $\boldsymbol{x}_j^i$ and $\boldsymbol{x}_k^i$ is very large, the value $\exp(-\delta\cdot l(\boldsymbol{W}, \boldsymbol{x}_j^i, \boldsymbol{x}_k^i))$ will be very close to zero, that is $\boldsymbol{x}_k^i$ is not the neighbor of $\boldsymbol{x}_j^i$. Another benefit of this adaptive approach is to improve the robustness to noisy data.

To embed non-Gaussian data well, we develop an novel semisupervised dimensionality reduction method named ADA based on our proposed framework (9). We adopt Euclidean distance to measure the similarity between pairwise points. The objective function of ADA can be simplified as

$$\max_{\boldsymbol{W}} \ \frac{1}{2n}\sum_{i=1}^{c}\frac{1}{n_i}\sum_{j=1}^{n_i}\sum_{k=1}^{n_i}\exp\left(-\delta\left\|\boldsymbol{W}^T\left(\boldsymbol{x}_j^i - \boldsymbol{x}_k^i\right)\right\|^2\right). \tag{10}$$

Denote the total scatter matrix by $\mathbf{S}_t = (1/n)\mathbf{X}^T H\mathbf{X}$, where $H = I - (1/n)\mathbf{1}\mathbf{1}^T$ is the centering matrix. To avoid arbitrary scaling and the trivial solution of all zeros, we constrain the subspace with $\boldsymbol{W}^T\mathbf{S}_t\boldsymbol{W} = I$ such that the data on transformed subspace are statistically uncorrelated, as in ULDA [21]. Then, the problem of ADA can be formulated as

$$\begin{aligned}\max_{\boldsymbol{W}\in\mathbb{R}^{d\times m}} \ & \frac{1}{2n}\sum_{i=1}^{c}\frac{1}{n_i}\sum_{j=1}^{n_i}\sum_{k=1}^{n_i}\exp\left(-\delta\left\|\boldsymbol{W}^T\left(\boldsymbol{x}_j^i - \boldsymbol{x}_k^i\right)\right\|^2\right)\\ \text{s.t.} \ \ & \boldsymbol{W}^T\mathbf{S}_t\boldsymbol{W} = I.\end{aligned} \tag{11}$$

Compared with classic LDA, our ADA model in (11) is able to process the non-Gaussian data and preserve the local structural information by differentiating the importance of each data point. Different from LLDA and LFDA which find kNNs in the original feature space, ADA finds the soft neighbors in the optimal subspace to preserve more discriminative information. Therefore, our ADA will be robust to noisy data and obtain a better solution to process high-dimensional non-Gaussian

Algorithm 1 Algorithm to Solve the Problem (11) by Iteratively Efficient Way

```
Input: Training data $\mathbf{X}$ and its corresponding labels $\boldsymbol{y}$.
Output: Optimal transformation matrix $\boldsymbol{W}$.
Let $t = 1$. Initialize $\boldsymbol{W}_t \in \mathbb{R}^{d\times m}$ such that $\boldsymbol{W}^T\mathbf{S}_t\boldsymbol{W} = I$;
while not converge do
    1. Calculate each $\mathbf{A}^i, i = 1, 2, ..., c$ as Eq. (13) and Laplacian matrix $\mathbf{L} = (\mathbf{D} - \mathbf{A})/n$ by $\boldsymbol{W}_t$;
    2. Update transformation matrix $\boldsymbol{W}_{t+1}$. The columns of the updated $\boldsymbol{W}_{t+1}$ are the first $m$ eigenvectors of $\mathbf{S}_t^+\mathbf{X}L\mathbf{X}^T$ corresponding to the first $m$ smallest eigenvalues; $t = t + 1$;
    3. Check convergence condition (22): $Div(t) < 10^{-6}$.
end while
```

data. To solve this difficult problem, in the next section, we will propose an efficient algorithm to optimize it.

B. Optimization

Based on the iterative optimization strategy, we propose an algorithm as described in Algorithm 1 to solve the problem (11) of ADA. In each iteration, we first compute the similarity of each sample by the current solution $\boldsymbol{W}_t$, and then learn the optimal transformation matrix $\boldsymbol{W}_{t+1}$ by the updated similarity information. The iteration procedure is repeated until converges. From Algorithm 1, we can see our algorithm can be easily implemented without using other optimization toolbox.

Denote the set $\mathbb{C} = \{\boldsymbol{W} \in \mathbb{R}^{d\times m} | \boldsymbol{W}^T\mathbf{S}_t\boldsymbol{W} = I\}$ and

$$\phi(\boldsymbol{W}) = \frac{1}{2n}\sum_{i=1}^{c}\frac{1}{n_i}\sum_{j=1}^{n_i}\sum_{k=1}^{n_i}\exp\left(-\delta\left\|\boldsymbol{W}^T\left(\boldsymbol{x}_j^i - \boldsymbol{x}_k^i\right)\right\|^2\right).$$

By the concept of functions, we know that $\phi(\boldsymbol{W})$ is a convex and smoothed function. Therefore, in the $t + 1$th iteration, we can use the following quadric model to approximate the original problem (11):

$$\boldsymbol{W}_{t+1} = \arg\max_{\boldsymbol{W}\in\mathbb{C}}\frac{1}{2n}\sum_{i=1}^{c}\sum_{j=1}^{n_i}\sum_{k=1}^{n_i} -A_{jk}^i\left\|\boldsymbol{W}^T\left(\boldsymbol{x}_j^i - \boldsymbol{x}_k^i\right)\right\|^2 \quad (12)$$

where

$$A_{jk}^i = \frac{\delta}{n_i}\exp\left(-\delta\left\|\boldsymbol{W}_t^T\left(\boldsymbol{x}_j^i - \boldsymbol{x}_k^i\right)\right\|^2\right). \quad (13)$$

Note that at $\boldsymbol{W}_t$, the objective function of (12) has the same gradient as the problem (11). Denote $\mathbf{A} = \text{diag}((\mathbf{A}^1, \mathbf{A}^2, \ldots, \mathbf{A}^c)$ and a Laplacian matrix

$$\mathbf{L} = (\mathbf{D} - \mathbf{A})/n$$

where $\mathbf{D}$ is a diagonal matrix with the jth element as $D_{jj} = \sum_{k=1}^{n} A_{jk}$. The problem (12) can be written as the following matrix form:

$$\boldsymbol{W}_{t+1} = \arg\min_{\boldsymbol{W}^T\mathbf{S}_t\boldsymbol{W}=I}\text{Tr}\left(\boldsymbol{W}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\boldsymbol{W}\right). \quad (14)$$

The Lagrangian function of the problem (14) is

$$\mathcal{L}(\boldsymbol{W}) = \text{Tr}\left(\boldsymbol{W}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\boldsymbol{W}\right) - \text{Tr}\left(\Lambda\left(\boldsymbol{W}^T\mathbf{S}_t\boldsymbol{W} - I\right)\right).$$

Taking the derivative of $\mathcal{L}(\boldsymbol{W})$ with respect to $\boldsymbol{W}$, we have

$$\frac{\partial\mathcal{L}(\boldsymbol{W})}{\partial\boldsymbol{W}} = \mathbf{X}\mathbf{L}\mathbf{X}^T\boldsymbol{W} - \mathbf{S}_t\boldsymbol{W}\Lambda. \quad (15)$$

Setting the derivative in (15) to zero, we can obtain

$$\mathbf{X}\mathbf{L}\mathbf{X}^T\boldsymbol{W} = \mathbf{S}_t\boldsymbol{W}\Lambda. \quad (16)$$

By the matrix theory, the problem in (16) is the generalized eigenvalue problem. It is easy to show that the matrices $\mathbf{X}\mathbf{L}\mathbf{X}^T$ and $\mathbf{S}_t$ are symmetric and positive semidefinite. Thus, the optimal solution of $\boldsymbol{W}$ is the first m eigenvectors of $\mathbf{S}_t^+\mathbf{X}\mathbf{L}\mathbf{X}^T$ corresponding to the first m smallest eigenvalues, where $\mathbf{S}_t^+$ is the pseudo-inverse matrix of $\mathbf{S}_t$. The transformation matrix $\boldsymbol{W}$ that minimize the objective function in (14) are given by the minimum eigenvalue solutions to the generalized eigenvalue problem in (16). For small sample size cases, ADA adopted the similar strategy with OLDA [22], [23] and learned the optimal transformation matrix by the pseudo-inverse matrix of $\mathbf{S}_t$, instead of using the inverse matrix $\mathbf{S}_t^{-1}$ directly.

We iteratively update $\mathbf{L}$ and $\boldsymbol{W}$ by (15), i.e., KKT condition holds. Note that $\mathbf{L}$ is not a variable to optimize. In the iterative steps, we solve the problem (16) to update $\boldsymbol{W}$, and then recalculate (16), where $\mathbf{L}$ is only an intermediate value to help calculate optimal $\boldsymbol{W}$. The algorithm will converge to a local optimum, which is proved in the next section.

IV. Discussion

A. Convergence Analysis

In this section, we prove that the objective function of (11) is nondecreasing under the updating rules of $\boldsymbol{W}$ and S in Algorithm 1. First, the following lemma is introduced.

Lemma 1: Let $f(x) = e^{-\delta x}$, $\delta > 0$ and $x > 0$, and the series expansion of $f(x)$ at x_0 is denoted as

$$T(x, x_0) = e^{-\delta x_0} - \delta e^{-\delta x_0}(x - x_0) - \frac{\delta}{2}e^{-\delta x_0}(x - x_0)^2.$$

For any $x > 0$, the following inequality holds:

$$g(x) = f(x) - T(x, x_0) \geq 0.$$

Proof: According to differential and integral theory, we have

$$g'(x) = -\delta e^{-\delta x} + \delta e^{-\delta x_0} + \delta e^{-\delta x_0}(x - x_0).$$

Obviously, if $x > x_0$, $g'(x) > 0$ and else $g'(x) < 0$ $(x < x_0)$, so $x = x_0$ is the only minimum point. Therefore, for any $x > 0$, we can get $g(x) \geq g(x_0) = 0$. ■

Theorem 1: The procedure in Algorithm 1 will monotonically increase the objective of the problem (14) in the each iteration, and converge to a local optimum.

Proof: Let $f_{jk}^i(\boldsymbol{W}) = \exp(-\delta\|\boldsymbol{W}^T(\boldsymbol{x}_j^i - \boldsymbol{x}_k^i)\|^2)$. According to Lemma 1, we have

$$\begin{aligned} T_{jk}^i(\boldsymbol{W}, \boldsymbol{W}_0) = {} & f_{jk}^i(\boldsymbol{W}_0) - 2tf_{jk}^i(\boldsymbol{W}_0)\text{tr}\Big((\boldsymbol{W} - \boldsymbol{W}_0)^T P_{jk}^i\boldsymbol{W}_0\Big) \\ & - \delta f_{jk}^i(\boldsymbol{W}_0)\text{tr}\Big((\boldsymbol{W} - \boldsymbol{W}_0)^T P_{jk}^i(\boldsymbol{W} - \boldsymbol{W}_0)\Big) \end{aligned}$$

and

$$f_{jk}^i(\boldsymbol{W}) \geq T_{jk}^i(\boldsymbol{W}, \boldsymbol{W}_0), f_{jk}^i(\boldsymbol{W}_0) = T_{jk}^i(\boldsymbol{W}_0, \boldsymbol{W}_0) \tag{17}$$

where $f_{jk}^i(\boldsymbol{W}_0) = \exp(-\delta\|\boldsymbol{W}_0^T(\boldsymbol{x}_j^i - \boldsymbol{x}_k^i)\|^2)$ and $P_{jk}^i = (\boldsymbol{x}_j^i - \boldsymbol{x}_k^i)(\boldsymbol{x}_j^i - \boldsymbol{x}_k^i)^T$. It is easily note that

$$\begin{aligned} T_{jk}^i(\boldsymbol{W}, \boldsymbol{W}_0) = & -\delta f_{jk}^i(\boldsymbol{W}_0)\mathrm{tr}\Big(\boldsymbol{W}^T P_{jk}^i \boldsymbol{W}\Big) + f_{jk}^i(\boldsymbol{W}_0) \\ & - \delta f_{jk}^i(\boldsymbol{W}_0)\mathrm{tr}\Big(\boldsymbol{W}_0^T P_{jk}^i \boldsymbol{W}_0\Big) \end{aligned} \tag{18}$$

where $f_{jk}^i(\boldsymbol{W}_0) - \delta f_{jk}^i(\boldsymbol{W}_0)\mathrm{tr}(\boldsymbol{W}_0^T P_{jk}^i \boldsymbol{W}_0)$ is constant relative to $\boldsymbol{W}$. Thus in the $(t+1)$th iteration, the problem (14) is equivalent to the following problem:

$$\boldsymbol{W}_{t+1} = \arg \max_{\boldsymbol{W}^T\mathbf{S}_t\boldsymbol{W}=I} \frac{1}{2n}\sum_{i=1}^{c}\frac{1}{n_i}\sum_{j=1}^{n_i}\sum_{k=1}^{n_i} T_{jk}^i(\boldsymbol{W}, \boldsymbol{W}_t).$$

The objective of LPP [26] can be formulated as

$$\min_{\boldsymbol{W}^T\mathbf{X}\bar{\mathbf{D}}\mathbf{X}^T\boldsymbol{W}=I} \mathrm{Tr}\big(\boldsymbol{W}^T\mathbf{X}\bar{\mathbf{L}}\mathbf{X}^T\boldsymbol{W}\big) \tag{19}$$

where $\bar{L} = \bar{\mathbf{D}} - \bar{\mathbf{A}}$ is the Laplacian matrix, $\bar{A}_{ij} = \exp(-\delta\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2)$, and $\bar{\mathbf{D}}$ is a diagonal matrix with the ith diagonal element $\bar{D}_{ii} = \sum_{j=1}^n \bar{A}_{ij}$. By the analysis in [26], we know that the problem (19) can obtain the optimal solution by the minimum eigenvalue solution to the generalized eigenvalue problem and has closed form solution. Note that the problem (14) has the similar formulation with LPP after some matrix transformation. Thus, in $(t + 1)$th iteration, we obtain the optimal transformation matrix $\boldsymbol{W}$ of the problem (14)

$$\boldsymbol{W}_{t+1} = \arg \min_{\boldsymbol{W}^T\mathbf{S}_t\boldsymbol{W}=I} \mathrm{tr}\big(\boldsymbol{W}^T\mathbf{XLX}\boldsymbol{W}\big)$$

which indicates that

$$\sum_{i=1}^{c}\sum_{j,k=1}^{n_i}\frac{T_{jk}^i(\boldsymbol{W}_{t+1}, \boldsymbol{W}_t)}{2nn_i} \geq \sum_{i=1}^{c}\sum_{j,k=1}^{n_i}\frac{T_{jk}^i(\boldsymbol{W}_t, \boldsymbol{W}_t)}{2nn_i}. \tag{20}$$

By Lemma 1 and (17), the following inequality holds:

$$\begin{aligned} \sum_{i=1}^{c}\sum_{j=1}^{n_i}\sum_{k=1}^{n_i}\frac{f_{jk}^i(\boldsymbol{W}_{t+1})}{2nn_i} &\geq \sum_{i=1}^{c}\sum_{j=1}^{n_i}\sum_{k=1}^{n_i}\frac{T_{jk}^i(\boldsymbol{W}_{t+1}, \boldsymbol{W}_t)}{2nn_i} \\ \sum_{i=1}^{c}\sum_{j=1}^{n_i}\sum_{k=1}^{n_i}\frac{f_{jk}^i(\boldsymbol{W}_t)}{2nn_i} &= \sum_{i=1}^{c}\sum_{j=1}^{n_i}\sum_{k=1}^{n_i}\frac{T_{jk}^i(\boldsymbol{W}_t, \boldsymbol{W}_t)}{2nn_i}. \end{aligned} \tag{21}$$

Combining (20) and (21), we have

$$\sum_{i=1}^{c}\sum_{j=1}^{n_i}\sum_{k=1}^{n_i}\frac{f_{jk}^i(\boldsymbol{W}_{t+1})}{2nn_i} \geq \sum_{i=1}^{c}\sum_{j=1}^{n_i}\sum_{k=1}^{n_i}\frac{f_{jk}^i(\boldsymbol{W}_t)}{2nn_i}.$$

That is to say

$$\phi(\boldsymbol{W}_{t+1}) \geq \phi(\boldsymbol{W}_t).$$

Thus, Algorithm 1 will monotonically increase the objective of the problem in (11) in each iteration t. Note that the objective function has upper bounds, so the above iteration will converge. Therefore, Algorithm 1 will monotonically increase the objective value in the each iteration until converges to a local optimum. ■

As we use the transformation matrix $\boldsymbol{W}$ to get the new feature representation, we also need to make clear the convergence behavior of it. Following [37], we measure the divergence between two sequential $\boldsymbol{W}$s at the kth iteration by the following metric:

$$\mathrm{Div}(k) = \sum_{i=1}^{m}\big|\,\|\boldsymbol{w}_{k+1}^i\|_2 - \|\boldsymbol{w}_k^i\|_2\big|. \tag{22}$$

The metric defined above acts as an indicator to show whether the final results would be changed drastically.

B. Connection to Related Approaches

In this section, we discuss the connections between ADA and LDA, LPP, LLDA, and LFDA. Before going into the details, let us first introduce the following lemma.

Lemma 2: $\mathbf{S}_t = (1/n)\sum_{i=1}^n(\boldsymbol{x}_i - \boldsymbol{u})(\boldsymbol{x}_i - \boldsymbol{u})^T$ can be rewritten as $\mathbf{S}_t = (1/2n^2)\sum_{i,j=1}^n(\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^T$, where $\boldsymbol{u} = (1/n)\sum_{i=1}^n \boldsymbol{x}_i$ [38]. Meanwhile, the matrix form of $\mathbf{S}_t$ is equivalent to $\mathbf{S}_t = (1/n)\mathbf{X}H\mathbf{X}^T$, where $H = I - (1/n)\mathbf{11}^T$.

Proof: Substitute $\boldsymbol{u} = (1/n)\sum_{i=1}^n \boldsymbol{x}_i$ into $\mathbf{S}_t$, we can get by

$$\begin{aligned} \mathbf{S}_t &= \frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{x}_i - \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\right)\left(\boldsymbol{x}_i - \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\right)^T \\ &= \frac{1}{n}\sum_{i=1}^{n}\left[\boldsymbol{x}_i\boldsymbol{x}_i^T - \frac{2}{n}\boldsymbol{x}_i\sum_{j=1}^{n}\boldsymbol{x}_j^T + \frac{1}{n^2}\sum_{i=1}^{n}\boldsymbol{x}_i\sum_{j=1}^{n}\boldsymbol{x}_j^T\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^T - \frac{1}{n^2}\sum_{i=1}^{n}\boldsymbol{x}_i\sum_{j=1}^{n}\boldsymbol{x}_j^T = \frac{1}{n}\mathbf{X}H\mathbf{X}^T. \end{aligned} \tag{23}$$

On the other hand

$$\begin{aligned} &\frac{1}{2n^2}\sum_{i,j=1}^{n}(\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^T \\ &\quad = \frac{1}{2n^2}\sum_{i,j=1}^{n}\Big(\boldsymbol{x}_i\boldsymbol{x}_i^T + \boldsymbol{x}_j\boldsymbol{x}_j^T - 2\boldsymbol{x}_i\boldsymbol{x}_j^T\Big) \\ &\quad = \frac{1}{2n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^T + \frac{1}{2n}\sum_{j=1}^{n}\boldsymbol{x}_j\boldsymbol{x}_j^T - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_j^T \\ &\quad = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^T - \frac{1}{n^2}\sum_{i=1}^{n}\boldsymbol{x}_i\sum_{j=1}^{n}\boldsymbol{x}_j^T. \end{aligned} \tag{24}$$

Combining (23) and (24), the following equality holds:

$$\mathbf{S}_t = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i - \boldsymbol{u})(\boldsymbol{x}_i - \boldsymbol{u})^T = \frac{1}{2n^2}\sum_{i=1}^{n}(\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^T.$$

Therefore, the data covariance matrix $\mathbf{S}_t$ can be rewritten as the pairwise form $\mathbf{S}_t = (1/n)\mathbf{X}H\mathbf{X}^T$. ■

Proposition 1: When $\delta \to 0$, LDA in (2) can be regarded as a special case of ADA in (11).

Proof: As seen from (11) and (12), when $\delta \to 0$, $A_{jk}^i \to 1$ and then the loss function is equivalent to

$$\begin{aligned}\phi(\boldsymbol{W}) &= -\frac{1}{2n}\sum_{i=1}^{c}\frac{1}{n_i}\sum_{j=1}^{n_i}\sum_{k=1}^{n_i}\left\|\boldsymbol{W}^T\left(\boldsymbol{x}_j^i-\boldsymbol{x}_k^i\right)\right\|_2^2\\ &= -\frac{1}{2n}\sum_{i=1}^{c}\frac{1}{n_i}\sum_{j,k=1}^{n_i}\left\|\boldsymbol{W}^T\left[\left(\boldsymbol{x}_j^i-\boldsymbol{u}_i\right)-\left(\boldsymbol{x}_k^i-\boldsymbol{u}_i\right)\right]\right\|_2^2\\ &= -\frac{1}{2n}\sum_{i=1}^{c}\frac{1}{n_i}\sum_{j,k=1}^{n_i}-2\left(\boldsymbol{x}_j^i-\boldsymbol{u}_i\right)^T\boldsymbol{W}\boldsymbol{W}^T\left(\boldsymbol{x}_k^i-\boldsymbol{u}_i\right)\\ &\quad+\left\|\boldsymbol{W}^T\left(\boldsymbol{x}_j^i-\boldsymbol{u}_i\right)\right\|_2^2+\left\|\boldsymbol{W}^T\left(\boldsymbol{x}_k^i-\boldsymbol{u}_i\right)\right\|_2^2 \qquad (25)\end{aligned}$$

where $\boldsymbol{u}_i$ is the centroid of ith class. Note that $\sum_{j=1}^{n_i}\sum_{k=1}^{n_i}(\boldsymbol{x}_j^i-\boldsymbol{u}_i)^T\boldsymbol{W}\boldsymbol{W}^T(\boldsymbol{x}_k^i-\boldsymbol{u}_i)=0$, the problem (25) is equivalent to

$$\phi(\boldsymbol{W}) = -\frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n_i}\left\|\boldsymbol{W}^T\left(\boldsymbol{x}_j^i-\boldsymbol{u}_i\right)\right\|^2. \qquad (26)$$

By the definition of within-class scatter matrix in (1), we can easily obtain

$$\phi(\boldsymbol{W}) = -\mathrm{tr}\left(\boldsymbol{W}^T\mathbf{S}_w\boldsymbol{W}\right).$$

Therefore, the problem (11) is equivalent to

$$\min_{\boldsymbol{W}^T\mathbf{S}_t\boldsymbol{W}=I}\ \mathrm{tr}\left(\boldsymbol{W}^T\mathbf{S}_w\boldsymbol{W}\right) \qquad (27)$$

that is the equivalent form of traditional LDA in (2) [23]. ■

Similarly, another dimensionality reduction method, i.e., LPP can also be analyzed within this framework. LPP first builds a graph by using neighborhoods, which preserves local structural information, and then computes the transformation matrix by general eigen decomposition. In other words, LPP can be regarded as solving the following problem:

$$\min_{\boldsymbol{W}^T\mathbf{X}\bar{\mathbf{D}}\mathbf{X}^T\boldsymbol{W}=I}\sum_{i,j=1}^{n}\left\|\boldsymbol{W}^T\boldsymbol{x}_i-\boldsymbol{W}^T\boldsymbol{x}_i\right\|_2^2\bar{A}_{ij} \qquad (28)$$

where $\bar{A}_{ij}=\exp(-\delta\|\boldsymbol{x}_i-\boldsymbol{x}_j\|_2^2)$ and $\bar{\mathbf{D}}$ is a diagonal matrix with the ith diagonal element $\bar{D}_{ii}=\sum_{j=1}^{n}\bar{A}_{ij}$. By simple algebra formulation, the objective function in (28) can be reformulated as the matrix form in (19).

Proposition 2: If we compute the weighted matrix $\mathbf{A}$ in (13) by employing Gaussian function in high-dimensional feature space, that is $\mathbf{A}=\bar{\mathbf{A}}$, LPP can be regarded as a special case of ADA when $\mathbf{S}_t=\mathbf{X}\bar{\mathbf{D}}\mathbf{X}^T$.

Proof: When we fix $\mathbf{A}=\bar{\mathbf{A}}$ in each iteration, ADA will obtain the optimal solution by one step. Meanwhile, if $\mathbf{S}_t=\mathbf{X}\bar{\mathbf{D}}\mathbf{X}^T$, the problem (11) is converted to the problem (28) to solve $\boldsymbol{W}$. In other words, when $\mathbf{S}_t=\mathbf{X}\bar{\mathbf{D}}\mathbf{X}^T$ and $\mathbf{A}=\bar{\mathbf{A}}$, the solution of ADA by one iteration can be regarded as solving the problem in (28), which is the formulation of LPP. ■

Proposition 3: If we compute the within-class and between-class scatter matrices using the sample neighbors (chosen by kNN) by (5), LLDA in (6) can be regarded as a special case of ADA in (11) when $\delta \to 0$.

When the number k of the chosen samples is equal to n, the formulation of LLDA in (6) is equivalent to the formulation of LDA (2). Thus, by the results of Proposition 1, if we compute the weight matrix $\mathbf{A}$ in (13) by the sample neighbors chosen by kNN, which is the same as LLDA, the formulation in (11) will be equivalent to the problem (6).

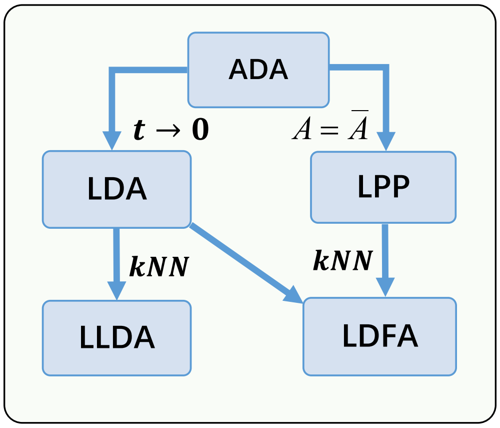


Fig. 2. Relationship of our ADA and other related methods. Different dimensionality reduction approaches can be regarded as special cases of ADA.

TABLE II
DATA SETS DESCRIPTIONS

Dataset	Size	Dim	#Class	Type
Breast	286	9	2	UCI
Wine	178	13	3	UCI
IRIS	150	4	3	UCI
DNA	3186	150	3	Biodata
Splice	3175	60	2	Biodata
Coil20	1440	1024	20	Object
USPS	9298	256	10	Handwritten
MNIST	70000	784	10	Handwritten

Proposition 4: If we compute the weighted matrix $\mathbf{A}$ in (13) by $\hat{\mathbf{A}}^w$, that is $\mathbf{A}=\hat{\mathbf{A}}^w$, LFDA can be regarded as a special case of ADA when $\mathbf{S}_t=\hat{\mathbf{S}}_w+\hat{\mathbf{S}}_b$ and $\delta \to 0$.

By Proposition 2, the proof of this result is the same as that in Proposition 3. We would like to exclude it.

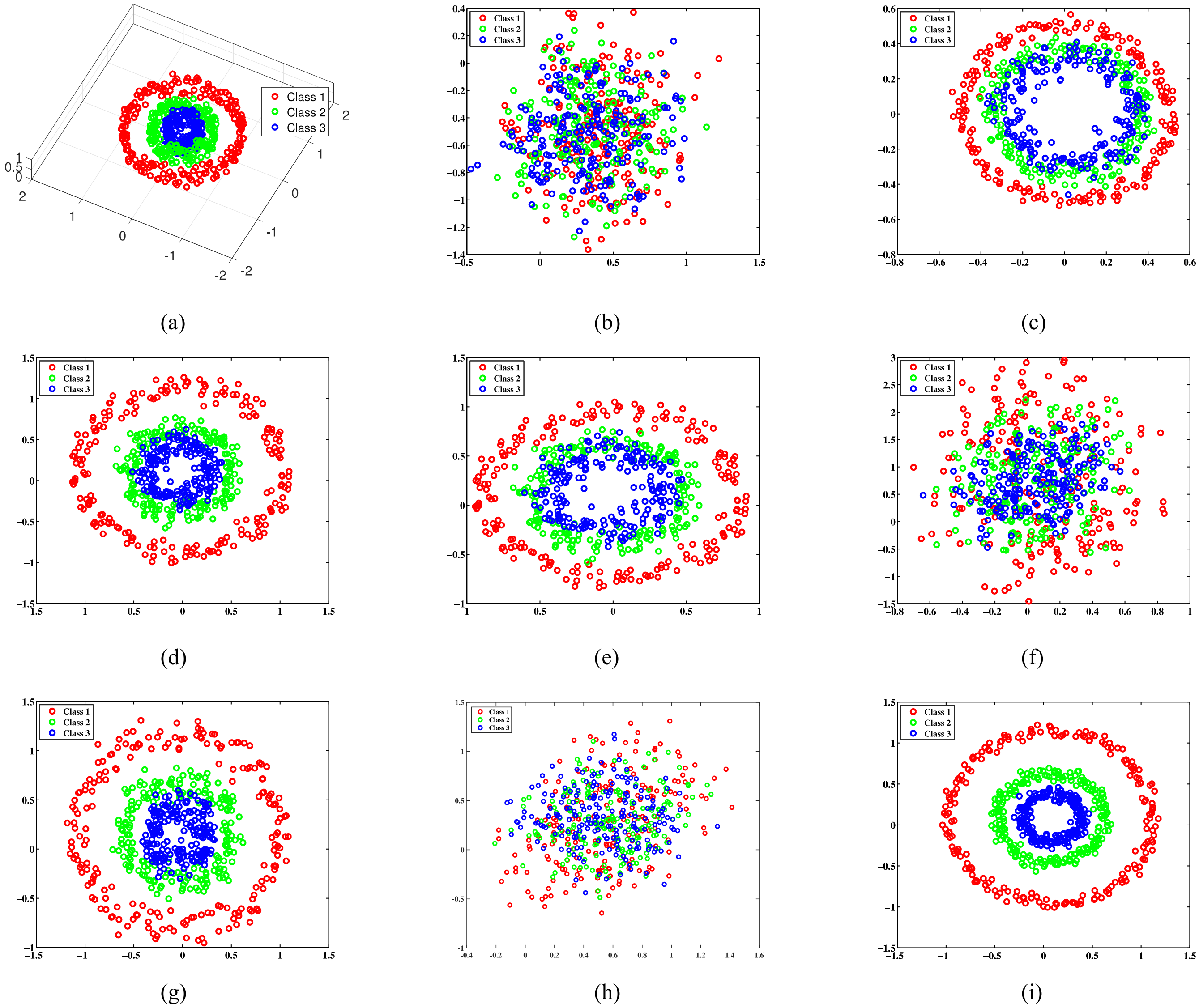
To sum up, the relationships are listed in Fig. 2. As indicated by the results in the previous propositions, different dimensionality reduction approaches can be regarded as special cases of our proposed ADA and it can be regarded as a unified framework in viewing them.

V. Experimental Results

A. Data Description and Experimental setups

In our experiments, eight public data sets with various statistical characters are collected to present the performance of different dimensionality reduction methods. These data sets include three image data sets including Coil20,[2] MNIST,[3] and USPS,[4] three UCI machine learning repository data sets, breast cancer (Breast),[5] IRIS,[6] and Wine,[7] and two biological

[2] http://www.cs.columbia.edu/CAVE/research/coil-20.html
[3] http://yann.lecun.com/exdb/mnist/
[4] http://www.cad.zju.edu.cn/home/dengcai/Data/USPS
[5] https://archive.ics.uci.edu/ml/datasets/Breast+Cancer
[6] https://archive.ics.uci.edu/ml/datasets/Iris
[7] https://archive.ics.uci.edu/ml/datasets/Wine



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Fig. 3. 2-S subspace learned by LDA, LPP, NMMP, SNNDA, LLDA, LFDA, WHMDA, and ADA, for an artificial 8-D data set. It illustrates that ADA can find a subspace preserving manifold structure with more discriminability. (a) Original data. (b) LDA. (c) LPP. (d) NMMP. (e) SNNDA. (f) LLDA. (g) LFDA. (h) WHMDA. (i) ADA.

data sets, DNA[8] and Splice.[9] All data sets are standardized to be zero-mean and normalized by standard deviation. We summarize the statistics of the data sets in Table II.

We compare the proposed ADA method against the following related methods. Unsupervised dimensionality reduction methods include: 1) principle component analysis (PCA) [39]; 2) locality preserving projections (LPP) [26]. Supervised dimensionality reduction methods consist of: 3) LDA [3], [20]; 4) ULDA [21]; 5) OLDA [22], [23]; 6) NMMPs [27]; 7) LFDA [11]; 8) LLDA [28]; 9) SNNDA [29]; and 10) WHMDA [10]. For all compared methods except PCA, we reduce the dimensionality of input data to be $c - 1$. For PCA, we reduce the dimensionality of input data such that 90% of data variance is preserved. We implement PCA, LPP, LDA, ULDA, OLDA, LFDA, and SNNDA using the codes published by the authors.[10]

To test the quality of the reduced features and analyze the effect of classifiers, we choose the classification accuracy as evaluation metric. Once the projection matrix is obtained by the dimensionality reduction methods, nearest neighbor classifier (NNC) method and the linear SVM are used to classify the unlabeled data points in the projected space. NNC and linear SVM are the traditional representative of nonlinear and linear classifier, respectively. In NNC, we use the most widely used Euclidean distance. We implement SVM by LIBSVM[11] package and implement the "one-against-one" approach for multiclass cases (for more details see [33], [40]). Following [3] and [41], the SVM classifier is individually performed on each

[8] https://www.csie.ntu.edu.tw/ cjlin/libsvmtools/datasets/multiclass.html#dna

[9] https://www.csie.ntu.edu.tw/ cjlin/libsvmtools/datasets/binary.html#splice

[10] http://www.cad.zju.edu.cn/home/dengcai/Data

[11] http://www.csie.ntu.edu.tw/ cjlin/libsvm/

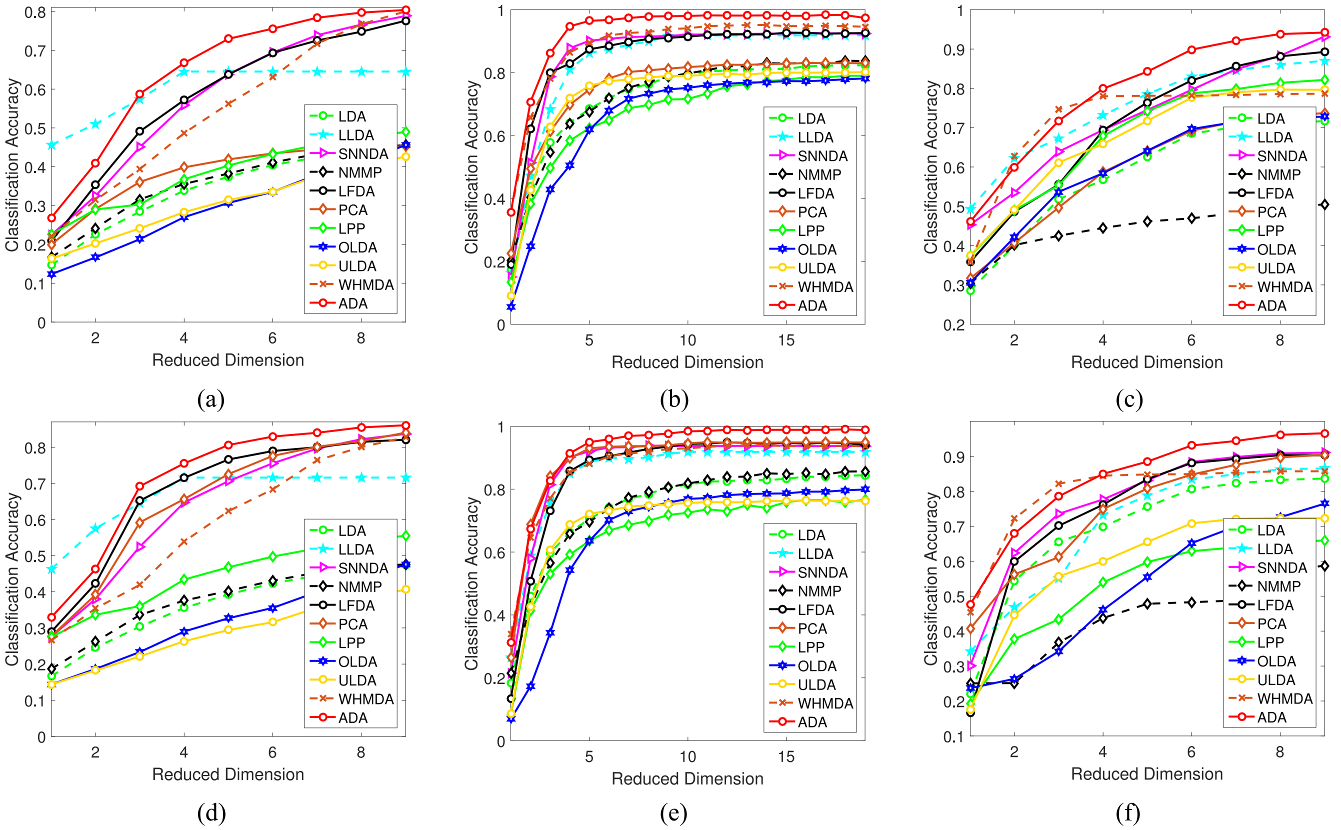


Fig. 4. Classification accuracy of compared methods on three benchmark data sets by NNC and linear SVM when the dimension of the learned subspace varies. *x*-axis is the number of reduced features and *y*-axis is the classification accuracy. Top: the classification results by NNC method on MNIST, COIL20, and USPS. Bottom: the classification results by linear SVM classifier on MNIST, COIL20, and USPS.

data set with reduced features, using the linear kernel with $C = 1$ by all compared methods.

B. Synthetic Data

We explore a toy example to present the performance of our ADA. We artificially generate a 3-D data set. The data set consists of three classes [shown in Fig. 3(a) by different colors]. In the first two dimensions, the classes are distributed in concentric circles, while the third dimensions are Gaussian noise with large variance. Fig. 3 shows the 2-D subspace learned by LPP, LDA, LLDA, NMMP, LFDA, SNNDA, WHMDA, and ADA, respectively. The distributions of synthetic data is not exact Gaussian distribution, that traditional methods like LDA fails to solve it. The results illustrate that ADA outperforms all other methods, which demonstrate the effectiveness of our method in learning the discriminative subspace. This is because ADA choose neighbors and simultaneously measure the importance of samples by using adaptive method. Thus, ADA can find more robust and discriminative subspace by this adaptive approach. Meanwhile, it also verifies that locality and adaptivity are the effective approaches to process the non-Gaussian data.

C. Comparison Between ADA and Other Algorithms

In this section, we evaluate our method in a typical supervised task, i.e., classification by two groups of experiments. As in [40], we randomly sample 50% data points as the training set and the remaining are used for testing. The process is repeated for 50 times and results in 50 different partitions. First, we use the training data as the input of dimensionality reduction methods and learn the optimal transformation matrices and new mapped features. Then, the classifiers, i.e., NNC and SVM are employed for classification, where the new subspace features are determined as training samples and the new projected features of original unlabeled data by learned transformation matrix are testing examples.

One group is to test the performance with different size of projected features. With different number of new projected features, we have conducted experiments on three multiclass data sets, i.e., MNIST, COIL20, and USPS. As in [3], other parameters are tuned by cross validation if necessary. The mean classification accuracy with different numbers of reduced dimension is shown in Fig. 4. Similar with [42] and [43], another group of experiments is to further investigate the impact of classifiers on the performance of the proposed ADA. Due to the maximal dimension of the projected subspace is $c - 1$ for LDA and its variants, in this group of experiments, the dimensionality of mapped features is ranged from 1 to $c-1$ for LDA, OLDA, ULDA, and WHMDA, while the dimensionality of mapped features of other methods is ranged from 1 to $d - 1$. Meanwhile, six benchmark data sets Breast, DNA, Splice, IRIS, Wine, and USPS are chosen to test the impact of dimensionality reduction methods on classifiers. After repeating 50 times of experiments, we calculate the mean accuracy

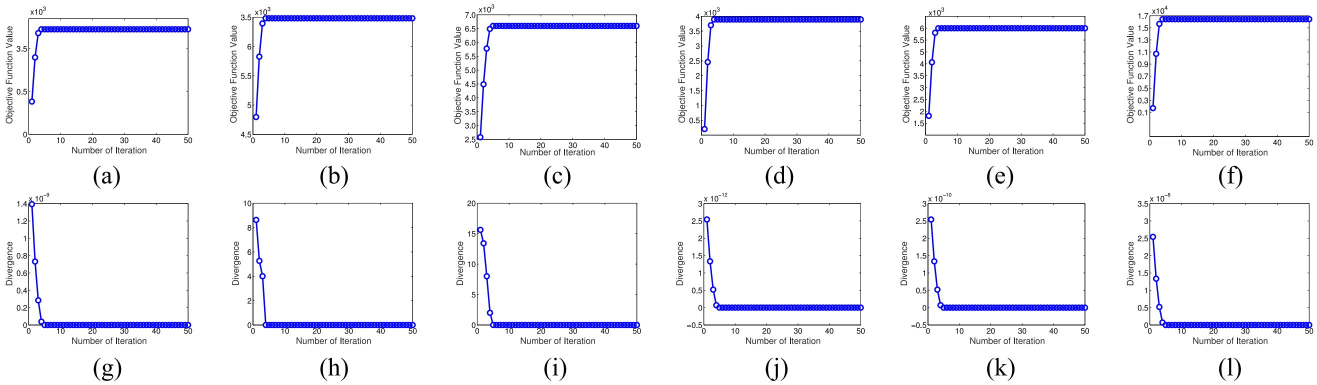


Fig. 5. Convergence behavior of ADA on DNA (left) and Wine (right), respectively. Top line is the objective value of ADA. Bottom line is divergence between tow consecutive $\boldsymbol{W}$ measured by (22). (a) and (g) DNA, $\delta = 0.0001$. (b) and (h) DNA, $\delta = 0.001$. (c) and (i) DNA, $\delta = 0.01$. (d) and (j) Wine, $\delta = 0.0001$. (e) and (k) Wine, $\delta = 0.001$. (f) and (l) Wine, $\delta = 0.01$.

and standard deviation values for each methods and classifiers with different dimensions of mapped features. The best mean classification results and their corresponding dimensions of mapped features are listed in Table III.

Fig. 4 and Table III show that ADA performs much better than other methods in most cases. The results from both NNC and linear SVM indicate our ADA can learn more discriminative subspace than other methods very efficiently. Moreover, the performance of ADA does not depend on the type of classifiers (nonlinear or linear). Another point should be highlighted here. No matter the number of training samples is larger than the dimensionality of the data, like DNA, Splice, USPS, and MNIST, or the small sample size cases in the data sets, such as Coil20, Japanese female facial expressions (JAFFE), Yale face (Yale), UMIST, and Caltech, ADA all performs well. ADA is not only applied for large sample size problem, but also used for small sample size cases.

D. Convergence Analysis

To validate the efficiency of our proposed algorithm to solve the problem of ADA, we present the convergence behavior curves of Algorithm 1 when $\delta = \{0.0001, 0.01, 0.1\}$. Two kinds of results are provided. The first concerns objective function value and the other is the divergence between two consecutive $\boldsymbol{W}$s as shown in (22). We show the results on two data sets DNA and Wine, since the algorithm has similar convergence behavior on the other data sets. The convergence curves are displayed in Fig. 5.

As seen from Fig. 5, the objectives of ADA with $\delta = \{0.0001, 0.001, 0.1\}$ are nondecreasing during the iterations, and they all converge to a fixed value. Additionally, in all cases, the divergence between two sequential $\boldsymbol{W}$s converges to zero, which indicates that the final results will not be changed drastically. Furthermore, ADA converges within 15 iterations on this two data sets for the three δ values. Therefore, our proposed ADA scales well in practice because of the fast convergence speed.

E. Parameter Determination

There is only one parameter, i.e., δ, in our proposed ADA model. We would like to provide some results of ADA with different parameters. Since parameter determination is still an open problem, we determine δ by a heuristic way. More concretely, we determine it by grid search at first and then change them within certain ranges. The classification accuracy results of NNC on the testing set with different δ on Wine, DNA, Splice, Coil20, USPS, and MNIST are shown in Fig. 6.

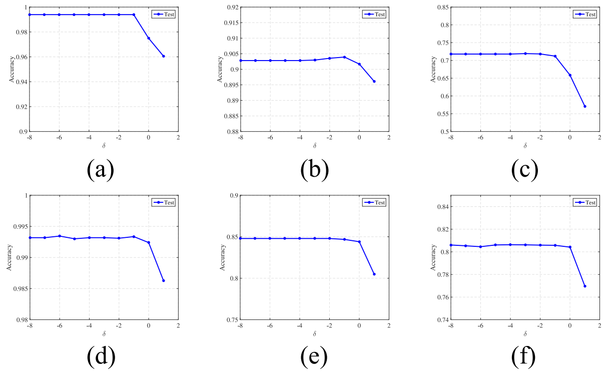


Fig. 6. Classification accuracy of NNC on (a) Wine, (b) DNA, (c) Splice, (d) Coil20, (e) USPS, and (f) MNIST with different δ value. x-axis represents the log-scale value of δ and y-axis is the classification accuracy of NNC.

As seen from Fig. 6, parameter determination takes influence on the performance of ADA. Different values of δ may result in different learned feature subspace. Then, the classification accuracy results of NNC change.

VI. Application to Face Recognition

As a major open challenge in computer vision and machine learning, face recognition [7], [8], [44] is the most important task in a number of application domains, including access control, visual surveillance system, and duplication of government issued identity documents (e.g., passport and driver license), to name a few. Thus face recognition has been extensively studied over the past two decades. In this section, we will evaluate ADA in this challenging application scenario.

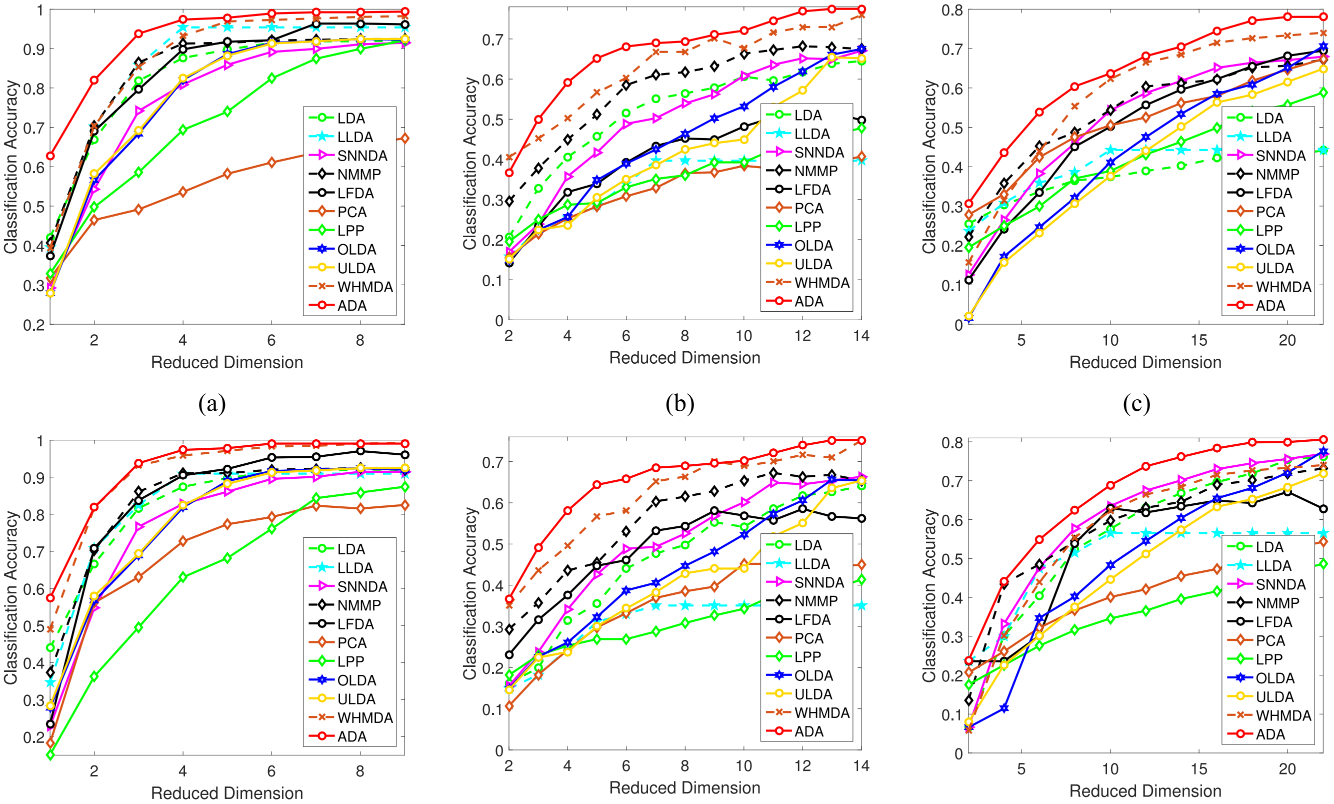
There are six diverse public face databases collected to illustrate the performance of different dimensionality reduction approaches. These data sets include JAFFE,[12]

[12] http://www.kasrl.org/jaffe.html

TABLE III

BEST MEAN CLASSIFICATION RESULTS AND THE CORRESPONDING DIMENSIONALITY OF 11 METHODS ON DATA SETS: BREAST, DNA, SPLICE, IRIS, WINE, AND USPS. THE BOLD NUMBERS ARE THE HIGHEST IN STATISTICAL VIEW. [MEAN±STD(DIMENSION)]

Classifier	Methods	Breast	DNA	Splice	IRIS	Wine	USPS
NNC	LDA	78.47±7.73(1)	48.92±1.70(2)	63.76±0.66(1)	94.33±2.67(2)	61.25±13.69(2)	83.84±0.47(9)
	OLDA	78.47±7.73(1)	46.10±1.87(2)	63.76±0.66(1)	94.18±2.38(2)	56.75±6.89(2)	83.10±0.54(9)
	ULDA	78.47±7.73(1)	50.58±1.97(2)	63.76±0.66(1)	81.34±5.36(2)	58.25±7.94(2)	83.31±0.39(9)
	PCA	92.58±2.23(3)	87.85±1.25(6)	88.55±1.00(5)	96.40±1.23(3)	88.86±2.19(7)	94.16±0.39(17)
	LPP	89.46±1.46(2)	86.15±1.70(7)	89.73±1.58(7)	92.81±1.07(3)	86.52±1.66(5)	93.15±0.48(20)
	NMMP	90.71±2.95(3)	90.46±1.34(4)	89.08±2.41(4)	96.73±0.75(3)	90.29±2.43(6)	92.40±2.51(16)
	SNNDA	89.93±4.30(4)	90.63±4.13(3)	88.18±2.29(4)	94.52±0.71(3)	89.50±1.37(6)	92.81±2.46(15)
	LLDA	92.77±1.29(2)	89.31±1.98(3)	91.20±1.77(5)	96.26±0.49(3)	90.99±2.10(5)	93.13±1.38(19)
	LFDA	93.49±1.75(2)	90.31±1.70(4)	90.81±0.66(4)	96.39±0.67(3)	92.14±3.69(7)	93.97±0.47(18)
	WHMDA	95.72±0.68(1)	91.23±0.48(2)	76.08±0.91(1)	92.78±0.96(2)	66.29±2.39(2)	91.28±0.69(9)
	ADA	**96.98±0.75(3)**	**94.58±0.65(5)**	**94.07±0.83(4)**	**97.99±1.07(3)**	**91.14±1.43(5)**	**96.51±0.09(15)**
SVM	LDA	67.67±10.85(1)	42.10±1.31(2)	66.98±1.48(1)	93.43±3.35(2)	61.88±14.52(2)	85.07±0.44(9)
	OLDA	67.67±10.85(1)	42.33±1.21(2)	66.98±1.48(1)	94.48±2.62(2)	57.00±6.98(2)	84.05±0.61(9)
	ULDA	78.20±7.61(1)	41.80±1.66(2)	67.21±1.75(1)	77.01±10.69(2)	67.12±8.67(2)	84.21±0.48(9)
	PCA	93.86±0.95(3)	89.35±2.00(6)	89.47±0.50(5)	95.62±0.76(3)	90.29±1.89(7)	94.20±0.27(17)
	LPP	90.88±1.27(2)	88.54±0.97(7)	90.05±1.29(7)	94.02±0.40(3)	88.00±1.99(5)	93.80±0.25(20)
	NMMP	90.52±1.85(3)	86.42±1.43(4)	89.81±2.01(4)	96.63±0.62(3)	93.57±2.54(6)	92.45±2.55(16)
	SNNDA	89.30±4.15(4)	88.84±3.86(3)	91.19±2.33(4)	96.16±0.36(3)	92.88±1.84(6)	93.53±2.05(15)
	LLDA	94.11±1.93(2)	87.73±1.28(3)	90.95±1.74(5)	97.53±0.63(3)	92.71±1.01(5)	94.78±2.16(19)
	LFDA	95.18±1.85(2)	90.77±1.31(4)	92.46±1.48(4)	96.99±0.75(3)	93.43±1.52(7)	94.86±0.44(18)
	WHMDA	95.89±0.53 (1)	93.45±0.32(2)	81.25±0.59(1)	97.33±0.61(2)	73.14±1.56(2)	91.60±0.68(9)
	ADA	**97.77±0.60(3)**	**96.50±0.58(5)**	**91.82±0.49(4)**	**97.93±1.30(3)**	**96.86±1.05(5)**	**96.16±0.37(15)**



(a) (b) (c)

(d) (e) (f)

Fig. 7. Recognition rates of different methods on (a) and (d) JAFFE, (b) and (e) Yale, and (c) and (f) Caltech databases with different numbers of reduced features under changes of illumination and facial expression. Top line is the results of NNC. Bottom line is the results of linear SVM.

Yale,[13] Caltech,[14] CMU pose, illumination, and expression (CMU-PIE),[15] UMIST,[16] and Extended Yale face database B (YaleB).[17] The brief details of them are presented as follows.

1) The JAFFE data set contains 217 images of ten Japanese female models at a resolution of 256×256. It mainly

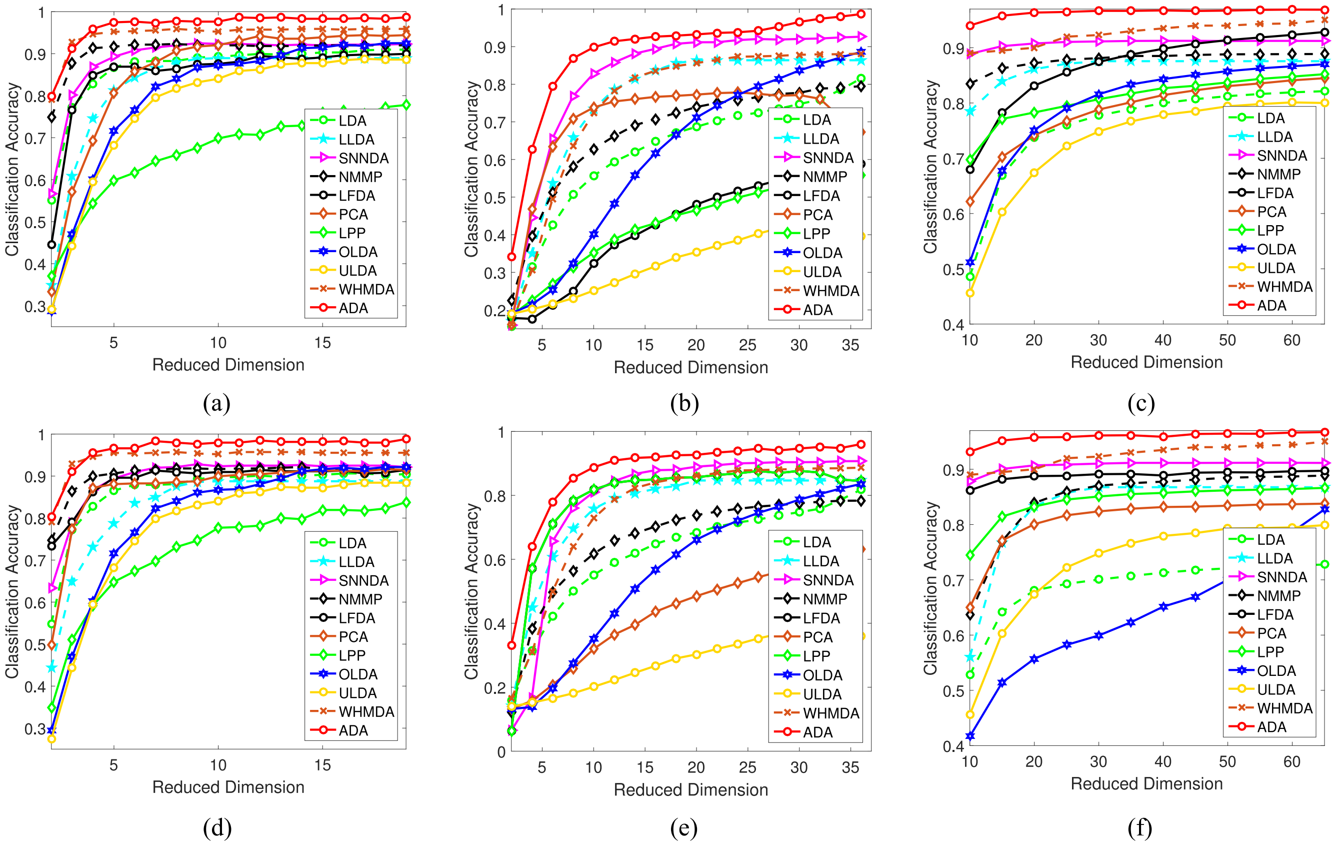
[13] http://cvc.yale.edu/projects/yalefaces/yalefaces.html

[14] http://www.vision.caltech.edu/Image_Datasets/faces

[15] http://vasc.ri.cmu.edu/idb/html/face/

[16] http://www.sheffield.ac.uk/eee/research/iel/research/face

[17] http://vision.ucsd.edu/ iskwak/ExtYaleDatabase/ExtYaleB.html



(a) (b) (c)

(d) (e) (f)

Fig. 8. Recognition rates of different methods on (a) and (d) UMIST, (b) and (e) YaleB, and (c) and (f) CMU-PIE databases with different numbers of reduced features under pose variation. Top line is the results of NNC. Bottom line is the results of linear SVM.

includes seven expressions, i.e., happiness, sadness, fear, anger, surprise, disgust, and neutrality. The expressions were posed without instruction by Japanese participants in Japan.

2) Yale data set contains 165 images of 15 persons at a resolution of 64×64. There exist various lighting conditions, gender, facial expressions and configurations among these images.
3) The YaleB [5] contains 16 128 images of 38 human subjects under nine poses and 64 illumination conditions. The cropped and normalized 192×168 face images were captured under various laboratory-controlled lighting and pose conditions.
4) CMU-PIE database contains 41 368 source images of 68 individuals with various conditions (13 poses, 43 illuminations, three expressions, and three talking situations) for each individual. For our experiment, we choose a subdatabase from the camera C05, which contains 68 subjects with 49 images per person. It contains images with varying poses and illumination conditions.
5) The Sheffield face database (previously UMIST) consists of 564 images of 20 individuals in a resolution of 220×220. Each covering a range of poses from profile to frontal views. Subjects cover a range of race, sex, and appearance.
6) Caltech face data set is collected by Markus Weber at California Institute of Technology. It contains 450 images of 27 individuals in a resolution of 896×592. The original images are captured under different lighting, expressions, and backgrounds conditions.

Face recognition is very difficult in real world since there exists varying poses, facial expressions, occlusion and illumination conditions. Before going into the details, we would like to introduce the procedure briefly. There are mainly three steps for our experiments.

1) We use the well-known feature descriptors to characterize each face image. We do not use the raw features of original faces since the images in this task are complicated and the raw features are not good enough to describe them. Thus, as in traditional face recognition methods [6]–[9], we use 640-D dense SIFT feature, 32-D Gabor feature, 512-D GIST feature, and 256-D LBP feature to represent each face image. For SIFT feature, we first adopt two-level pyramid model to segment original faces into five patches and then represent each patch as 128-D SIFT features. Finally, each face is described as 1440-D feature vector.
2) We adopt different dimensionality reduction models to learn the projection matrix $\boldsymbol{W}$.
3) After feature embedding, all face images are represented by lower-dimensional vectors. As in previous experiments, we randomly split all data sets into training and testing parts. Finally, NNC and linear SVM are employed for recognition.

We compare ADA algorithm with several widely used related methods: PCA, LPP, LDA, ULDA, OLDA, SNNDA, NMMP, LFDA, and LLDA. In our experiments, we implement grid search to estimate the values of δ. The parameter δ ranges in $[0.0000001, 0.000001, \ldots, 0.1, 1]$. Then fivefold cross validation is used to decide the optimal parameter. The parameters in other methods are also tuned by grid search. Finally, we report the best recognition rate for each algorithm.

A. Experiments With Changes of Illumination and Expression

To verify the performance of our proposed method with changes of illumination and facial expression, we choose JAFFE, Yale, and Caltech databases in this experiment. Since different databases have different scale, we randomly select 9, 7, and 10 face images of each person from JAFFE, Yale, and Caltech as training set and the remaining face images as testing. We set the number of reduced feature dimension between 1 and $c - 1$ for all databases. Then we repeat this procedure for 50 independent runs by NNC and linear SVM classifiers. The mean recognition rates are presented in Fig. 7.

From Fig. 7, we can see that ADA outperforms other compared methods with changes of illumination and facial expression conditions, no matter which kind of classifiers we have employed. Interestingly, the improvement is more significant with the increase of the number of reduced features. Specially, the improvement of ADA over other methods is 3%–12% under variation of illumination and expression.

B. Experiments With Pose Variation

In this group of experiments, we further study the impact of pose variation on face recognition of our ADA. We evaluate the performance of all compared methods on UMIST, YaleB, and CMU-PIE databases with pose variation. For each database, we randomly select 60% samples for training and the rest for testing. Based on the procedures mentioned above, every method is tested 50 times. The mean recognition rates are shown in Fig. 8.

As seen from the numerical results in Fig. 8, our ADA achieves the highest recognition rates in each case against all compared methods. The recognition rates of ADA is 5%–9% higher than others. In other words, ADA is more effective than other methods for face recognition with pose variation.

In summary, ADA can obtain the most discriminative features from different visual features. This is because ADA combines the sample's importance measurement and discriminant analysis into the unified framework to enhance the performance of classification.

VII. Conclusion

In this paper, we aim to provide insights into the relationship between sample's importance measure and subspace learning, as well as to facilitate the design of new algorithms for non-Gaussian data. The framework was proposed to provide a unified perspective for the understanding this relationship. Moreover, this framework can be used as a general platform to develop new algorithms. Meanwhile, we have developed a new supervised dimensionality reduction method named ADA based on this framework. A byproduct is a series of theoretical analysis and some interesting optimization strategies. Plenty of experimental results on different kinds of data sets have been shown that ADA can extract more discriminative subspace features. Furthermore, ADA has been applied to face recognition. One of our future works is the selection of optimal parameter δ, which is an unsolved and open problem in many learning algorithms. Another future work is to propose more efficient optimization method to solve our proposed problem in (11).

References

[1] S. Boutemedjet, N. Bouguila, and D. Ziou, "A hybrid feature extraction selection approach for high-dimensional non-Gaussian data clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 8, pp. 1429–1443, Aug. 2009.

[2] J. B. Tenenbaum, V. De Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[3] I. Jolliffe, "Principal component analysis," in *International Encyclopedia of Statistical Science*, M. Lovric, Ed. Berlin, Germany: Springer, 2011.

[4] D. Tao, X. Li, X. Wu, and S. J. Maybank, "General tensor discriminant analysis and Gabor features for gait recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 10, pp. 1700–1715, Oct. 2007.

[5] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 643–660, Jun. 2001.

[6] D. A. Forsyth and J. Ponce, *Computer Vision: A Modern Approach*. Englewood Cliffs, NJ, USA: Prentice-Hall, 2002.

[7] X. Tan and B. Triggs, "Enhanced local texture feature sets for face recognition under difficult lighting conditions," *IEEE Trans. Image Process.*, vol. 19, no. 6, pp. 1635–1650, Jun. 2010.

[8] A. R. Rivera, J. R. Castillo, and O. O. Chae, "Local directional number pattern for face analysis: Face and expression recognition," *IEEE Trans. Image Process.*, vol. 22, no. 5, pp. 1740–1752, May 2013.

[9] M. Yang, L. Zhang, S. C. Shiu, and D. Zhang, "Gabor feature based robust representation and classification for face recognition with Gabor occlusion dictionary," *Pattern Recognit.*, vol. 46, no. 7, pp. 1865–1878, 2013.

[10] Z. Li, F. Nie, X. Chang, and Y. Yang, "Beyond trace ratio: Weighted harmonic mean of trace ratios for multiclass discriminant analysis," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 10, pp. 2100–2110, Oct. 2017.

[11] M. Sugiyama, "Local Fisher discriminant analysis for supervised dimensionality reduction," in *Proc. ACM ICML*, 2006, pp. 905–912.

[12] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[13] L. Van Der Maaten, E. Postma, and J. Van den Herik, "Dimensionality reduction: A comparative," *J. Mach. Learn. Res.*, vol. 10, pp. 66–71, 2009.

[14] T. Luo, C. Hou, D. Yi, and J. Zhang, "Discriminative orthogonal elastic preserving projections for classification," *Neurocomputing*, vol. 179, pp. 54–68, Feb. 2016.

[15] D. Tao, X. Li, X. Wu, and S. J. Maybank, "Geometric mean for subspace selection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 260–274, Feb. 2009.

[16] D. Tao, X. Tang, X. Li, and X. Wu, "Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 7, pp. 1088–1099, Jul. 2006.

[17] N. Guan, D. Tao, Z. Luo, and B. Yuan, "Manifold regularized discriminative nonnegative matrix factorization with fast gradient descent," *IEEE Trans. Image Process.*, vol. 20, no. 7, pp. 2030–2048, Jul. 2011.

[18] T. Liu and D. Tao, "On the performance of Manhattan nonnegative matrix factorization," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 9, pp. 1851–1863, Sep. 2016.

[19] T. Liu, D. Tao, and D. Xu, "Dimensionality-dependent generalization bounds for k-dimensional coding schemes," *Neural Comput.*, vol. 28, no. 10, pp. 2213–2249, 2016.

[20] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Ann. Eugenics*, vol. 7, no. 2, pp. 179–188, 1936.

[21] J. Ye, R. Janardan, Q. Li, and H. Park, "Feature reduction via generalized uncorrelated linear discriminant analysis," *IEEE Trans. Knowl. Data Eng.*, vol. 18, no. 10, pp. 1312–1322, Oct. 2006.

[22] J. Ye, "Characterization of a family of algorithms for generalized discriminant analysis on undersampled problems," *J. Mach. Learn. Res.*, vol. 6, pp. 483–502, Dec. 2005.

[23] J. Ye and T. Xiong, "Computational and theoretical analysis of null space and orthogonal linear discriminant analysis," *J. Mach. Learn. Res.*, vol. 7, pp. 1183–1204, Jul. 2006.

[24] W. Bian and D. Tao, "Max-min distance analysis by using sequential SDP relaxation for dimension reduction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 1037–1050, May 2011.

[25] S. Zheng, F. Nie, C. Ding, and H. Huang, "A harmonic mean linear discriminant analysis for robust image classification," in *Proc. IEEE 28th Int. Conf. Tools Artif. Intell. (ICTAI)*, 2016, pp. 402–409.

[26] X. He and P. Niyogi, "Locality preserving projections," in *Proc. NIPS*, 2004, pp. 153–160.

[27] F. Nie, S. Xiang, and C. Zhang, "Neighborhood MinMax projections," in *Proc. IJCAI*, 2007, pp. 993–998.

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

[29] X. Qiu and L. Wu, "Stepwise nearest neighbor discriminant analysis," in *Proc. IJCAI*, 2005, pp. 829–834.

[30] K. Fukunaga and J. Mantock, "Nonparametric discriminant analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-5, no. 6, pp. 671–678, Nov. 1983.

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

[32] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," in *Proc. NIPS*, vol. 14, 2001, pp. 585–591.

[33] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, p. 27, 2011.

[34] G. Rätsch, T. Onoda, and K.-R. Müller, "Soft margins for AdaBoost," *Mach. Learn.*, vol. 42, no. 3, pp. 287–320, 2001.

[35] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Stat.*, vol. 29, no. 5, pp. 1189–1232, 2001.

[36] L. Zelnik-Manor and P. Perona, "Self-tuning spectral clustering," in *Proc. NIPS*, 2004, pp. 1601–1608.

[37] C. Hou, F. Nie, X. Li, D. Yi, and Y. Wu, "Joint embedding learning and sparse regression: A framework for unsupervised feature selection," *IEEE Trans. Cybern.*, vol. 44, no. 6, pp. 793–804, Jun. 2014.

[38] M. Loog, R. P. W. Duin, and R. Haeb-Umbach, "Multiclass linear dimension reduction by weighted pairwise Fisher criteria," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 7, pp. 762–766, Jul. 2001.

[39] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed. San Diego, CA, USA: Academic Press, 1990.

[40] C.-W. Hsu and C.-J. Lin, "A comparison of methods for multiclass support vector machines," *IEEE Trans. Neural Netw.*, vol. 13, no. 2, pp. 415–425, Mar. 2002.

[41] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *J. Mach. Learn. Res.*, vol. 7, pp. 2399–2434, Dec. 2006.

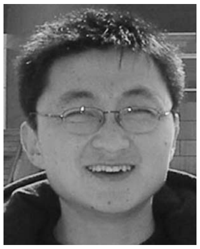
[42] B.-C. Kuo and D. A. Landgrebe, "Nonparametric weighted feature extraction for classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 5, pp. 1096–1105, May 2004.

[43] J.-M. Yang, P.-T. Yu, and B.-C. Kuo, "A nonparametric feature extraction and its application to nearest neighbor classification for hyperspectral image data," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 3, pp. 1279–1293, Mar. 2010.

[44] A. K. Jain, B. Klare, and U. Park, "Face recognition: Some challenges in forensics," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit. Workshops*, 2011, pp. 726–733.

Tingjin Luo received the B.S. and master's degrees from the College of Information System and Management, National University of Defense Technology, Changsha, China, in 2011 and 2013, respectively, where he is currently a Ph.D. degree with the College of Science.

He was a visiting Ph.D. student with the University of Michigan, Ann Arbor, MI, USA, from 2015 to 2017. His current research interests include machine learning, multimedia analysis, optimization, and computer vision.



Chenping Hou (M'12) received the B.S. and Ph.D. degrees in applied mathematics from the National University of Defense Technology, Changsha, China, in 2004 and 2009, respectively.

He is currently an Associate Professor with the College of Science, National University of Defense Technology. He has authored several papers in journals and conferences, such as the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, the IEEE TRANSACTIONS ON CYBERNETICS, the IEEE TRANSACTIONS ON IMAGE PROCESSING, *Pattern Recognition*, the International Joint Conferences on Artificial Intelligence (IJCAI), and the Association for the Advancement of Artificial Intelligence (AAAI) Conference on AI. His current research interests include pattern recognition, machine learning, data mining, and computer vision.

Dr. Hou has served a PC Member for the Proceedings of the Neural Information Processing Systems Conference, IJCAI, AAAI, and the IEEE International Conference on Acoustics, Speech and Signal Processing.

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

He has published over 100 papers in the top journals and conferences with over 2000 Google scholar citations. His current research interests include machine learning and its application fields, such as pattern recognition, data mining, computer vision, image processing, and information retrieval.

Dr. Nie currently serves as an Associate Editor or a PC Member for several prestigious journals and conferences.

Dongyun Yi received the B.S. degree from Nankai University, Tianjin, China, and the M.S. and Ph.D. degrees from the National University of Defense Technology, Changsha, China.

He was a Visiting Researcher with the University of Warwick, Coventry, U.K., in 2008. He is a Professor with the College of Science, National University of Defense Technology. His current research interests include statistics, systems science, and data mining.