Computational Intelligence

An International Journal

Computational Intelligence, Volume 29, Number 1, 2013

SEMISUPERVISED MULTIMODAL DIMENSIONALITY REDUCTION

Zhao Zhang,[1] Tommy W.S. Chow,[1] and Ning Ye[2]

[1]*Department of Electronic Engineering, City University of Hong Kong, Kowloon, Hong Kong*

[2]*Department of Computer Science and Technology, Shandong University, Jinan, People's Republic of China*

The problem of learning from both labeled and unlabeled data is considered. In this paper, we present a novel semisupervised multimodal dimensionality reduction (SSMDR) algorithm for feature reduction and extraction. SSMDR can preserve the local and multimodal structures of labeled and unlabeled samples. As a result, data pairs in the close vicinity of the original space are projected in the nearby of the embedding space. Due to overfitting, supervised dimensionality reduction methods tend to perform inefficiently when only few labeled samples are available. In such cases, unlabeled samples play a significant role in boosting the learning performance. The proposed discriminant technique has an analytical form of the embedding transformations that can be effectively obtained by applying the eigen decomposition, or finding two close optimal sets of transforming basis vectors. By employing the standard kernel trick, SSMDR can be extended to the nonlinear dimensionality reduction scenarios. We verify the feasibility and effectiveness of SSMDR through conducting extensive simulations including data visualization and classification on the synthetic and real-world datasets. Our obtained results reveal that SSMDR offers significant advantages over some widely used techniques. Compared with other methods, the proposed SSMDR exhibits superior performance on multimodal cases.

Received 30 June 2010; Revised 8 May 2011; Accepted 26 May 2011; Published online 15 May 2012

Key words: semisupervised learning, dimensionality reduction, locality preservation, multivariate visualization, multimodality preservation, classification.

1. INTRODUCTION

Attributed to the rapid scientific and technological innovations, handling high-dimensional data, e.g., multivariate visualization and gene expressions, has become increasingly important and popular. This leads to more research on the topics of dimensionality reduction techniques. Most of previous works can be categorized as supervised, unsupervised, or semisupervised. Research has not only focused on developing new learning algorithms, but preserving intrinsic structure information has proved to be important for high-dimensional data analysis (Roweis and Saul 2000; Hinton and Salakhutdinov 2006). Unsupervised *principal component analysis* (PCA) (Mardia, Kent, and Bibby 1980), *multidimensional scaling* (MDS) (Cox and Cox 2001), and supervised *Fisher linear discriminant analysis* (FDA) (Duda, Hart, and Stor 2001) are three of the most popular dimensionality reduction techniques used for multidimensional data representation, visualization, and pattern recognition. In the area of unsupervised visualization, *visualization-induced self-organizing map* (ViSOM) (Yin 2002a, 2002b) is a widely used method. It projects high-dimensional data onto two-dimensional maps while preserving the data topology and interneuron distances, but requires the map size to be predefined. This constraint in some cases may pose certain effects on its performance, hindering effective display of data characteristics (Xu, Xu, and Chow 2010). It is well known that PCA, MDS, and ViSOM do not take the underlying class information and intrinsic local manifold information hidden in the data into account. They only perform unsupervised learning with the global structures preserved. Unlike them, FDA is a supervised globalized method which works well when the samples class labels are available, but it tends to deliver less satisfactory results when data points of the same class form several isolated clusters (Torre and Kanade 2005; Sugiyama 2007; Zhang and

Address correspondence to Zhao Zhang, Department of Electronic Engineering, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong; e-mail: itzzhang@ee.cityu.edu.hk

Ye 2011), i.e., *multimodality*. Intrinsic multimodal structure is often encountered in many real applications such as facial gender recognition. The problem of intrinsic multimodality appears when genders are classified into males and females, because face images of different subjects or individuals are captured under different external conditions. Usually, solving a multiclass problem can be performed by using a two-class method. However, this involves preprocessing procedures of merging some of the classes, which will subsequently result in the within-class multimodality problem. Thus, preserving multimodality for dimensionality reduction and multidimensional data visualization is a major issue that requires to be addressed.

To represent the data efficiently and reduce the dimensionality appropriately, it is essential to preserve the local and multimodal structures hidden in the data. The above-mentioned methods, however, fail to satisfy this requirement. In contrast, *locality preserving projection* (LPP) (He and Niyogi 2004) mines the local manifold of the data points and keeps the projections of data pairs that are in close vicinity in the original space and close in the reduced feature space. LPP is able to reduce the dimensionality of multimodal data with local structure information preserved. Recently, LPP- and FDA-based *local FDA* (LFDA) (Sugiyama 2006, 2007) and LFDA- and PCA-based *semisupervised local FDA* (SELF) (Sugiyama et al. 2008, 2010) methods were proposed for supervised and semisupervised dimensionality reduction. One of the major advantages of these methods lies in their ability to embed the multimodal real data effectively. Also, there are other graph-embedding-based nonlinear methods, such as *locally linear embedding* (LLE) (Roweis and Saul 2000; Lawrence and Sam 2003), *laplacian eigenmaps* (LE) (Belkin and Niyogi 2001), and *ISOMAP* (Tenenbaum, Silva, and Langford 2000), developed for performing dimensionality reduction and multidimensional data visualization. These unsupervised graph algorithms are popular and efficient in visualizing the synthetic data. However, ISOMAP and LLE are still constrained by certain limitations, for example, the inflexible definition of the geodesic distances for ISOMAP and the unclear interpretation of the weights-based metrics for LLE (Bengio et al. 2004). In this paper, we show that data visualizations produced by both ISOMAP and LLE tend to be relatively inaccurate on capturing the manifold structures of the multimodal datasets, e.g., *XOR* and the real *pen-based handwritten digits* dataset.

Many previous works have shown that supervised dimensionality reduction methods become inefficient when the number of labeled samples is limited. This leads to increasing interest and attention in the semisupervised techniques (Belkin, Niyogi, and Sindhwani 2006; Zhu 2006; Sugiyama et al. 2008, 2010; Zhang and Yeung 2008a). Recently, there are popular and effective semisupervised learning algorithms, such as *semisupervised discriminant analysis* (SDA) (Cai, He, and Han 2007a) and SELF (Sugiyama et al. 2008, 2010), proposed for dimensionality reduction and feature extraction. In this paper, we also refer to the use of both labeled and unlabeled samples. We propose a multimodality preserving algorithm, namely, semisupervised multimodal dimensionality reduction (SSMDR), for performing dimensionality reduction on the semisupervised scenarios. Different from the objective functions of the above-mentioned learning methods, SSMDR is capable of finding a pair of LPP transformations for two datasets. As a result, SSMDR is able to represent high-dimensional multimodal data in the best possible way, because SSMDR will not embed the multimodal data points into a single cluster. Through conducting extensive simulations, we show that SSMDR can achieve the comparative or even better results than some widely used methods.

The rest of this paper is organized as follows. In Section 2, we formulate the linear dimensionality reduction problem and briefly review some existing classical learning methods. In Section 3, we mathematically formulate the proposed projection algorithm. In Section 4, we compare SSMDR with the existing PCA, MDS, FDA, LPP, NPE, LFDA, IsoProjection, CCA, ViSOM, LE, LLE, ISOMAP, SDA, and SELF methods under several synthetic and benchmark datasets. Finally, we offer the concluding remarks in Section 5.

2. PRELIMINARIES

In this section, we present the linear dimensionality reduction problem and review two related classical methods.

2.1. Linear Dimensionality Reduction

Let $x_i \in \mathbb{R}^n (i = 1, 2, \ldots, m)$ be the vectors of m n-dimensional data and $f_i (\in \{1, 2, \ldots, c\})$ be the associated class labels, where c is the number of classes. Let $\zeta_i \in \mathbb{R}^d$ $(1 \leq d \leq n)$ be the low-dimensional representation of data x_i, where d is the selected dimensionality. Without loss of generality, we present the $n \times d$ transformation matrix by $\widehat{\xi}_x$; thus; the embedding of x_i and ζ_i is given by $\zeta_i = \widehat{\xi}_x^{\mathrm{T}} x_i$, where $^{\mathrm{T}}$ denotes the transpose of a matrix or a vector. In this paper, we focus on discussing the linear representations and also will consider extending the discussions to the nonlinear scenarios, which means that the mapping from x to ζ is nonlinear.

2.2. Fisher Linear Discriminant Analysis

FDA (Duda et al. 2001; Martinez and Kak 2001) finds the optimal vectors for discrimination. Let m_t be the number of labeled samples in the class $t \in \{1, 2, \ldots, c\}$ and $\sum_{t=1}^{c} m_t = m$. The classical FDA computes an optimal transformation matrix, mapping each column of X in the n-dimensional space to a feature vector in the d-dimensional space, satisfying $d \leq n$. Let $S^{(t)}$, $S^{(bc)}$, and $S^{(wc)}$ be the total scatter matrix, between-class scatter matrix, and within-class scatter matrix, respectively, then we have the following:

$$S^{(wc)} = \sum_{t=1}^{c} \sum_{i:f_i=t} (x_i - \widetilde{M}_{(t)})(x_i - \widetilde{M}_{(t)})^{\mathrm{T}}, \tag{1}$$

$$S^{(bc)} = S^{(t)} - S^{(wc)} = \sum_{t=1}^{c} m_t (\widetilde{M}_{(t)} - \widetilde{M})(\widetilde{M}_{(t)} - \widetilde{M})^{\mathrm{T}}, \tag{2}$$

where $\sum_{i:f_i=t}$ denotes the summation over i such that $f_i = t$, $\widetilde{M}_{(t)} = (1/m_t) \sum_{i:f_i=t} x_i$ is the average vector of samples in the tth class and $\widetilde{M} = (1/m) \sum_{t=1}^{c} \sum_{i:f_i=t} x_i$ is the global mean of all samples. Provided that $S^{(wc)}$ has full rank, then the $n \times d$ FDA transformation matrix T_{FDA} can be defined as

$$T_{FDA} = \underset{\widehat{\xi}_x \in \mathbb{R}^{n \times d}}{\arg\max} \left[\mathrm{tr} \left(\left(\widehat{\xi}_x^{\mathrm{T}} S^{(wc)} \widehat{\xi}_x \right)^{-1} \widehat{\xi}_x^{\mathrm{T}} S^{(bc)} \widehat{\xi}_x \right) \right] = \underset{\widehat{\xi}_x \in \mathbb{R}^{n \times d}}{\arg\max} \frac{\left| \widehat{\xi}_x^{\mathrm{T}} S^{(bc)} \widehat{\xi}_x \right|}{\left| \widehat{\xi}_x^{\mathrm{T}} S^{(wc)} \widehat{\xi}_x \right|} \tag{3}$$

Thus, the maximization discrimination vectors can be obtained. Note that $tr(H)$ is the trace of the matrix H.

2.3. LPPs

Unlike the global structure preservation capability of FDA, LPP (He and Niyogi 2004; He et al. 2005) aims at preserving the local manifold structure of data. It is worth noting that local information preservation is important. For a given data matrix $X = [x_1, x_2, \ldots, x_m]$, LPP finds an efficient projection transformation, T_{LPP}, for mapping a dataset into a set of

data points in $\mathbb{R}^d (d \leq n)$ with the local manifold information preserved. Let $\widetilde{A}$ denote the similarity matrix, i.e., an $m \times m$ matrix with $\widetilde{A}_{i,j}$ being the affinity value between data points x_i and x_j. Note that $\widetilde{A}_{i,j}$ is large and correspondingly $\|\zeta_i - \zeta_j\|^2$ is small if x_i and x_j are close and $\widetilde{A}_{i,j}$ is small, and correspondingly, $\|\zeta_i - \zeta_j\|^2$ is large if x_i and x_j are projected far apart. There are several different measures to define $\widetilde{A}$, for instance, the nearest neighbors method (Roweis and Saul 2000), the heat kernel method (Belkin and Niyogi 2001), and the local scaling heuristic approach (Zelnik-Manor and Perona 2005). Then, the objective function of LPP is defined as

$$\begin{aligned} T_{LPP} &= \arg\min_{\widehat{\xi}_x \in \mathbb{R}^{n \times d}} \frac{1}{2} \sum_{i,j=1}^{m} \left\| \widehat{\xi}_x^{\mathrm{T}} x_i - \widehat{\xi}_x^{\mathrm{T}} x_j \right\|^2 \widetilde{A}_{i,j} = \arg\min_{\widehat{\xi}_x \in \mathbb{R}^{n \times d}} Tr\big(\widehat{\xi}_x^{\mathrm{T}} X(\widetilde{W} - \widetilde{A}) X^{\mathrm{T}} \widehat{\xi}_x\big), \\ &\text{subject to} \quad \widehat{\xi}_x^{\mathrm{T}} X \widetilde{W} X^{\mathrm{T}} \widehat{\xi}_x = I_d, \end{aligned} \tag{4}$$

where $\widetilde{W}$ is an m-dimensional diagonal matrix with ith input element being $\widetilde{W}_{ii} = \sum_{j=1}^{m} \widetilde{A}_{i,j}$.

Equation (4) implies that LPP finds an efficient locality preserving transformation matrix such that data pairs in the close vicinity in $\mathbb{R}^n$ are still compact in the low-dimensional feature space $\mathbb{R}^d$. The matrix $\widetilde{W}$ provides an accurate measure on the data points from the perspective of geometric argument (Belkin and Niyogi 2003), but this term is usually omitted for simplicity (Ham et al. 2004; Sugiyama 2007). Let $\{\psi_r\}_{r=1}^{d}$ be the eigenvectors, ordered according to the generalized eigenvalues $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_d$ of the following generalized eigen problem: $X\widetilde{L}X^{\mathrm{T}}\psi = \lambda X\widetilde{W}X^{\mathrm{T}}\psi$, where $\widetilde{L} = \widetilde{W} - \widetilde{A}$ is the so-called Laplacian matrix (Chung 1997). Then, a solution of T_{LPP} is given as $T_{LPP} = (\psi_1 | \psi_2 | \ldots \psi_d)$. The locality preserving capability of LPP is of particular applicable for image retrieval and recognition (He et al. 2003, 2005).

3. SSMDR

We, in this section, elaborate the novel contributions of this paper. The objective is to propose an SSMDR algorithm that is able to compute a pair of projection transformations with high local information preservation and discrimination powers for two datasets $\widetilde{X}$ and $\widetilde{Y}$.

3.1. Motivation and Objective

In Figure 1, the examples of *Toy* and *XOR* datasets are shown. The two-dimensional two-class data are projected into a one-dimensional embedding space, where L_1 and L_2 are the one-dimensional embedding spaces on which the points are projected, and the black arrows are the directions of projections. Note that we take certain possible projection directions, e.g., L_1 or L_2, to represent the data points of "△"-class and "☆"-class.

For the simplest *Toy1* dataset in Figure 1(a), L_1 and L_2 are all optimal. Note that there are many methods that are able to perform well on this dataset and separate data points of different classes ("☆" and "△") nicely from each other. However, multimodality datasets, which are common in most of real-world data, are more challenging and can be problematic to the dimensionality reduction methods, because the multimodal distributions tend to be lost due to the complex intrinsic structure characteristics. Multimodal datasets usually have isolated within-class clusters, such as the following used *Toy2*, *Toy3*, and *XOR* datasets.

For the *Toy2* dataset in Figure 1(b), L_2 can embed the multimodal data respectably by separating between-class samples and preserving within-class multimodality, but L_1 gives

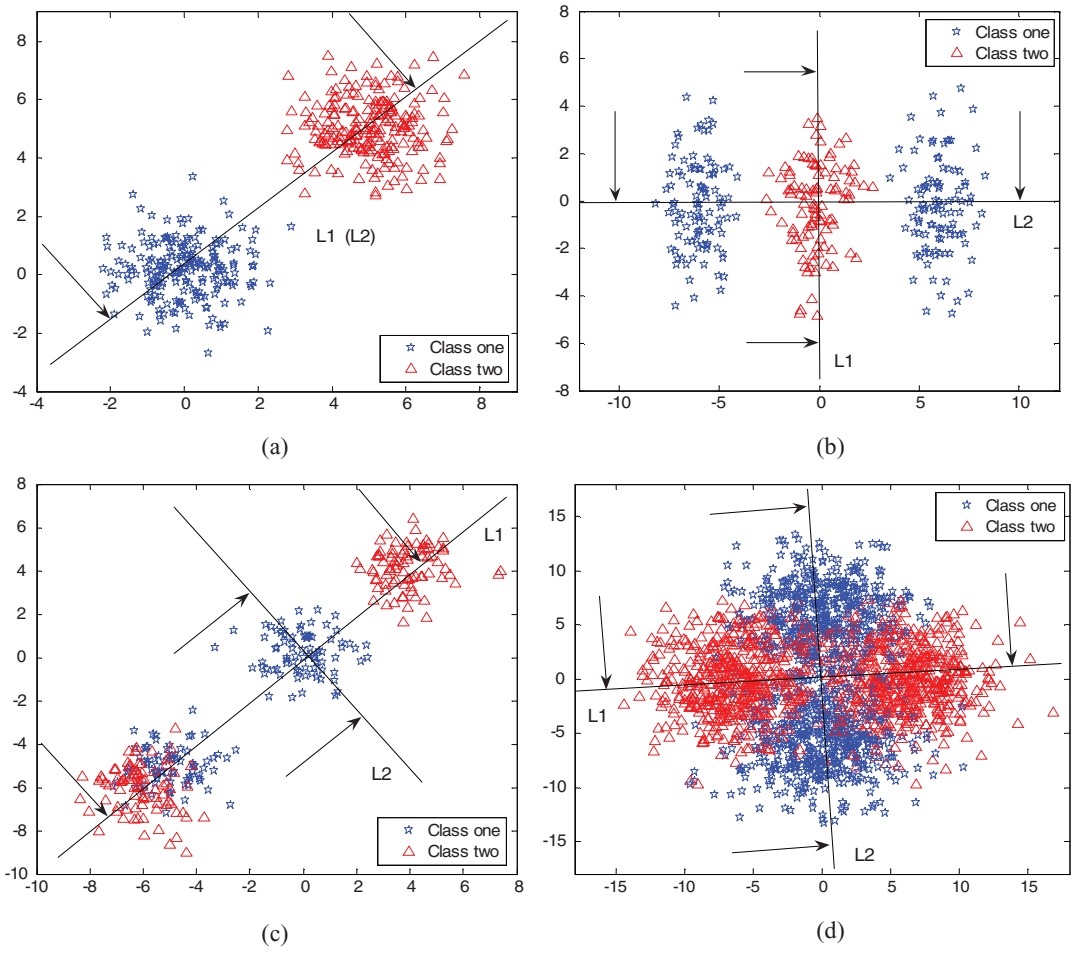


FIGURE 1. Examples of dimensionality reduction on the following four datasets: (a) on *Toy1*; (b) on *Toy2*; (c) on *Toy3*; and (d) on *XOR*.

undesired result and tends to mix the projections of the data in different classes. For the *Toy3* dataset in Figure 1(c), L_1 embeds the partial multimodal data nicely, but L_2 delivers relatively poor result that tends to mix the embeddings of the samples in different classes. The reason for collapsing the points in "☆"-class and "△"-class into a single cluster is due to the fact that the datum is originally compact in the original input space. For the *Toy2* and *Toy3* datasets, we can consistently find an optimal low-dimensional embedding space L_1 or L_2 to embed the data and nicely separate data points of different classes from each other. However, it is noted that in the *XOR* dataset shown in Figure 1(d), L_1 or L_2 are both unable to represent the data in an acceptable way. This is mainly due to the within-class multimodality of each class.

Apparently, from the one-dimensional embedding spaces constructed from these datasets, L_1 and L_2 tend to work satisfactorily, in turn, which implies that L_1 and L_2 seem to compensate the weaknesses of each other. As illustrated in Figures 1(c) and (d), neither L_1 nor L_2 is a fully optimal solution of the *Toy3* and *XOR* datasets. For the *Toy1* dataset, we may say that L_1 and L_2 deliver the same result. Thus, to represent the data points in the

best possible way, it is natural to investigate whether or not the embedding results of the multimodal data can be improved by finding both L_1 and L_2 at the same time. In this paper, we will address this issue in later section.

From the perspective of dimensionality reduction, supervised methods tend to overfit the labeled data when the labeled number is small or limited. In such cases, unlabeled samples with low cost will be useful to enhance the performance. In addition, the underlying class labels of labeled data, local manifold, and intrinsic multimodal information hidden in the data should also be considered. PCA, an unsupervised method, is able to preserve the global covariance structures of the dataset. Moreover, owing to the unsupervised nature of PCA, it can efficiently be applied to unlabeled data. In this study, we focus on proposing a new idea of integrating the local information and unlabeled data with the objective of computing the projection transformations. As a result, we are capable of deriving an effective semisupervised multimodal algorithm, called SSMDR in this paper. The proposed SSMDR algorithm utilizes the locality preserving property for the labeled data and aims at computing a pair of projection transformations for both labeled and unlabeled data in a semisupervised way.

3.2. Formulation

Given m pairs of labeled multivariate data $\{(x_i, y_i)\}_{i=1}^{m}$ accompanied with the class label $f_i(\in \{1, 2\})$, where $x_i, y_i \in \mathbb{R}^n, i = 1, 2, \ldots, m$ denotes an input vector from the data matrices $X = [x_1, x_2, \ldots, x_m]$ and $Y = [y_1, y_2, \ldots, y_m]$. The tth class has m_t data samples, i.e., $\sum_{t=1}^{2} m_t = 2m$. For unlabeled data points, we have l pairs of unlabeled samples $\{(x_i, y_i)\}_{i=m+1}^{m+l}$, that is, there are $m + l$ $(m \leq l)$ examples in total. Without loss of generality, we assume that the data points in labeled sets X and Y are ordered according to their class labels. Suppose that we are given a sample of instances $((x_1, y_1), \ldots, (x_m, y_m))$ of (X, Y), SSMDR aims to find a pair of transformations, $\widehat{\xi_x}$ and $\widehat{\xi_y}$, for them, one for each dataset. We can then use the transformation $\widehat{\xi_x}$ to represent $(x_1, x_2, \ldots, x_m)$ and analogously $\widehat{\xi_y}$ to embed $(y_1, y_2, \ldots, y_m)$ for obtaining new low-dimensional coordinates.

To improve the tightness among neighboring pairs and separate nonneighboring pairs, we consider shrinking distance metrics between the projections of similar patterns belonging to the same set X(or, Y) by minimizing $\widehat{\xi_x}^{\mathrm{T}} \sum_{i,j=1}^{m} (x_i - x_j)(x_i - x_j)^{\mathrm{T}} \widetilde{A}_{i,j}^{(wc)} \widehat{\xi_x}$ and $\widehat{\xi_y}^{\mathrm{T}} \sum_{i,j=1}^{m} (y_i - y_j)(y_i - y_j)^{\mathrm{T}} \widetilde{A}_{i,j}^{(wc)\dagger} \widehat{\xi_y}$, while maximizing the distances between the projections of pairwise patterns from different sets by maximizing $\widehat{\xi_x}^{\mathrm{T}} \sum_{i,j=1}^{m} (y_i - x_j)(y_i - x_j)^{\mathrm{T}} \widetilde{A}_{i,j}^{(bc)} \widehat{\xi_x}$ and $\widehat{\xi_y}^{\mathrm{T}} \sum_{i,j=1}^{m} (x_i - y_j)(x_i - y_j)^{\mathrm{T}} \widetilde{A}_{i,j}^{(bc)\dagger} \widehat{\xi_y}$, where the transformations $\widehat{\xi_x}^{\mathrm{T}} x$ and $\widehat{\xi_y}^{\mathrm{T}} y$ denote the low-dimensional representation of data x and y, respectively. Note that matrices $\widetilde{A}^{(wc)}$, $\widetilde{A}^{(wc)\dagger}$, $\widetilde{A}^{(bc)}$, and $\widetilde{A}^{(bc)\dagger}$ are defined for preserving the locality and are regarded as the similarity measures between data points, where $\widetilde{A}_{i,j}^{(wc)}(\widetilde{A}_{i,j}^{(wc)\dagger})$ represents the local relations of intraclass data pairs in $X(Y)$ and $\widetilde{A}_{i,j}^{(bc)}(\widetilde{A}_{i,j}^{(bc)^{\dagger}})$ represents the local relations of between-class data pairs in $X(Y)$ and $Y(X)$, which will be detailed in the next section. It is noted that the localized metric of SSMDR is employed for multimodality preservation purpose and utilized for measuring labeled samples in the semisupervised learning. This metric is similarly defined in (Ye et al. 2010) for designing binary multiplane classifier, which is a supervised globalized technique. In this present work, SSMDR optimizes the within- and between-class scatters in the similar way to the LPP criterion. It is worth noting that the within-class scatters of our proposed SSMDR can be considered as the supervised locality preserving LPP metric.

3.3. Interpretation of the Locality

In performing feature extraction or dimensionality reduction, preserving the local structure of data is important. Local nearest neighbors tend to deliver similar distributions and embeddings, so class labels of data lying on a dense area are probably the same (Zhou et al. 2004). Recently, some locality-preservation-based methods (Roweis and Saul 2000; Lawrence and Sam 2003; He and Niyogi 2004; Sugiyama 2006, 2007; Sugiyama et al. 2008, 2010) have been proposed and used in image representation and recognition. The key step of locality preservation is about constructing the neighborhood graphs and setting the weights to obtain a similarity matrix. The local neighbors of $x_i(y_i)$ can be defined by using either the Euclidean neighbor (Hinton and Salakhutdinov 2006), or the k-nearest neighborhood (Belkin and Niyogi 2003). In this work, k-nearest neighborhood search is used. We construct two within-class graphs for the samples in labeled sets X and Y by putting an edge between nodes i and j if samples $x_i(y_i)$ and $x_j(y_j)$ from the same object are "close" and have the same class labels. We can then select the simple-minded method (Belkin and Niyogi 2001), the heat kernel method (Belkin and Niyogi 2001), or the local scaling heuristic method (Zelnik-Manor and Perona 2005) to set the weights for the similarity matrix.

Suppose that there are total χ objects $O_1, O_2, \ldots, O_\chi$ in the datasets X and Y. Let $\forall_{LNN}(x_i)(\forall_{LNN}(y_i))$ denote the sample set that comprises the local neighbors of sample $x_i(y_i)$, thus $x_j(y_j)$ belongs to $\forall_{LNN}(x_i)$ $(\forall_{LNN}(y_i))$ if sample $x_j(y_j)$ is the neighbor of $x_i(y_i)$. We can then employ the heat kernel method to define the similarity matrices $\widetilde{h}^{(X)} = \{\widetilde{h}^{(X)}_{i,j}\}^m_{i,j=1}$ and $\widetilde{h}^{(Y)} = \{\widetilde{h}^{(Y)}_{i,j}\}^m_{i,j=1}$ for given sample pairs (x_i, x_j) and (y_i, y_j), where

$$\widetilde{h}^{(X)}_{i,j} = \begin{cases} \exp\left(-\|x_i - x_j\|^2/\tau^{(X)}\right), & \text{if } x_j \in \forall_{LNN}(x_i) \text{ or } x_i \in \forall_{LNN}(x_j), x_i \in O_a, x_j \in O_b, a = b, \\ 0, & \text{otherwise} \end{cases} \tag{5a}$$

$$\widetilde{h}^{(Y)}_{i,j} = \begin{cases} \exp\left(-\|y_i - y_j\|^2/\tau^{(Y)}\right), & \text{if } y_j \in \forall_{LNN}(y_i) \text{ or } y_i \in \forall_{LNN}(y_j), y_i \in O_a, y_j \in O_b, a = b, \\ 0, & \text{otherwise} \end{cases} \tag{5b}$$

where $\tau^{(X)} = \sum^m_{i,j=1} \|x_i - x_j\|^2/m(m-1)$ and $\tau^{(Y)} = \sum^m_{i,j=1} \|y_i - y_j\|^2/m(m-1)$. Note that Eqs. 5 and 6 reflect the local information around each data point, i.e. the smaller the distance $||x_i - x_j||(||y_i - y_j||)$, the closer the data points $x_i(y_i)$ and $x_j(y_j)$ of the same object. Thus the $\widetilde{h}^{(X)}_{i,j}(\widetilde{h}^{(Y)}_{i,j})$ incurs a heavy penalty, where $\widetilde{h}^{(X)}_{i,j}(\widetilde{h}^{(Y)}_{i,j})$ denote the *(i,j)*-th entry of the similarity matrix $\widetilde{h}^{(X)}(\widetilde{h}^{(Y)})$. According to the similarity matrix $\widetilde{h}^{(X)}_{i,j}(\widetilde{h}^{(Y)}_{i,j})$, we only weight the values for the sample pairs that are mutually neighbors from same object. Similarly we define an adjacency matrix $\widetilde{h}^{(YX)} = \{\widetilde{h}^{(YX)}_{i,j}\}^m_{i,j=1}$ for measuring the locality between multiple objects in Y and X sets as

$$\widetilde{h}^{(YX)}_{i,j} = \begin{cases} \exp\left(-\left\|Y_i - X_j\right\|/\tau^{(YX)}\right), & \text{if } Y_i \in \forall_{NN}(X_j) \text{ or } X_j \in \forall_{NN}(Y_i),\ X_j \in O_a, Y_i \in O_b, a \neq b, \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

where $\tau^{(YX)}$ can be similarly estimated. We can then employ the similar definition methods of [24][33] to formulate the matrices $\widetilde{A}^{(wc)}$, $\widetilde{A}^{(wc)\dagger}$ and $\widetilde{A}^{(bc)}$. Let C^*D denote the *Hadamard products* [39] between two matrices C and D with the same sizes, that is$(C^*D)_{ij} = C_{ij}D_{ij}$. Thus the weight matrices $\widetilde{A}^{(wc)}$, $\widetilde{A}^{(wc)\dagger}$, $\widetilde{A}^{(bc)}$ and $\widetilde{A}^{(bc)\dagger}$ can be defined as $\widetilde{A}^{(wc)} = \widetilde{\hbar}^{(X)*}\widetilde{\hbar}^{(X)}$, $\widetilde{A}^{(wc)\dagger} = \widetilde{\hbar}^{(Y)*}\widetilde{\hbar}^{(Y)}$, $\widetilde{A}^{(bc)} = \widetilde{\hbar}^{(XY)*}\widetilde{\hbar}^{(XY)}$ and $\widetilde{A}^{(bc)\dagger} = \widetilde{A}^{(bc)\mathrm{T}}$. In this study, we aim at keeping the local information $\widetilde{\hbar}^{(X)}_{i,j}$ and $\widetilde{\hbar}^{(Y)}_{i,j}$ unchanged, because we focus on keeping transformed data in nearby of the reduced feature space if they are in the close vicinity of the original space.

3.4. The Objective Function

In practical high-dimensional applications, small sample size (SSS) problem is frequently encountered. When the number of sample dimensions is significantly larger than the number of data or when the number of labeled data is limited, supervised dimensionality reduction methods tend to get overfitted to the labeled data (Cai, He, and Han 2007b; Sugiyama et al. 2008, 2010). To prevent overfitting, a widely used approach is to introduce a regularizer, for instance, Tikhonov regularizer (Belkin, Matveeva, and Niyogi 2004): $J^*(\theta) = \|\theta\|^2 = \theta^{\mathrm{T}}\theta$. Motivated by the graph-embeddings-based approaches, e.g., Roweis and Saul (2000), Tenenbaum et al. (2000), Belkin and Niyogi (2001), Lawrence and Sam (2003), and Zhang and Yeung (2008b), we in this study aim at constructing two regularization items for the unlabeled samples. Let $X_u = [x_{m+1}, x_{m+2}, \ldots, x_{m+l}]$ and $Y_u = [y_{m+1}, y_{m+2}, \ldots, y_{m+l}]$ denote the unlabeled sets, we define

$$\begin{aligned} j(\widehat{\xi}_x) &= \frac{1}{2l}\sum_{i,j=1}^{l} \left|\left|\widehat{\xi}_x{}^T x_{m+i} - \widehat{\xi}_x{}^T x_{m+j}\right|\right|^2 \widetilde{N}^{(X)}_{i,j} \\ &= \frac{1}{2l}\widetilde{\xi}_x{}^T \left(2\sum_{i=1}^{l}\left[\sum_{j=1}^{l}\widetilde{N}^{(X)}_{i,j}\right] x_{m+i}x_{m+i}{}^{\mathrm{T}} - 2\sum_{i,j=1}^{l} x_{m+i}\widetilde{N}^{(X)}_{i,j}x_{m+j}{}^T \right)\widetilde{\xi}_x \\ &= \frac{1}{l}\widehat{\xi}_x{}^{\mathrm{T}} X_u(\widetilde{W}^X - \widetilde{N}^X)X_u{}^{\mathrm{T}}\widehat{\xi}_x = \frac{1}{l}\widehat{\xi}_x{}^{\mathrm{T}} X_u \widetilde{L}^{(X)}_u X_u{}^{\mathrm{T}}\widehat{\xi}_x, \end{aligned} \tag{7}$$

where $\widetilde{L}^{(X)}_u = (\widetilde{W}^{(X)} - \widetilde{N}^{(X)})$ and $\widetilde{W}^{(X)}$ is a diagonal matrix with ith entry being $\widetilde{W}^{(X)}_{ii} = \sum_{j=m+1}^{m+l}\widetilde{N}^{(X)}_{i,j}$, where $\widetilde{N}^{(X)}_{i,j}$ is the (i,j)th entry of $\widetilde{N}^{(X)}$. It is noted that when each element $\widetilde{N}^{(X)}_{i,j}$ of the matrix $\widetilde{N}^{(X)}$ equals to $\frac{1}{l}$, then $J(\widehat{\xi}_x)$ is equivalent to the total scatter matrix of the PCA criterion and $\widetilde{W}^{(X)}$ will be the identity matrix. As a result, the term $J(\widehat{\xi}_x)$ will play a significant role in preserving the global covariance structures of all the data points, including labeled and unlabeled samples. The motivation for exploiting unlabeled samples is to employ them to boost the performance when the available number of labeled samples is limited. The normalization coefficient $\frac{1}{l}$ in equation (7) is used for balancing the functional value $J(\widehat{\xi}_x)$. Similarly, let matrix $\widetilde{L}^{(Y)}_u = (\widetilde{W}^{(Y)} - \widetilde{N}^{(Y)})$, where $\widetilde{W}^{(Y)}$ is a diagonal matrix whose entries are column (or row since $\widetilde{N}^{(Y)}$ is symmetric) sums of matrix $\widetilde{N}^{(Y)}$, that is, $\widetilde{W}^{(Y)}_{ii} = \sum_{j=m+1}^{m+l}\widetilde{N}^{(Y)}_{i,j}$, then similar formulation exists for the regularizer $J(\widehat{\xi}_y)$ described as

$$J(\widehat{\xi}_y) = \frac{1}{l}\widehat{\xi}_y{}^{\mathrm{T}} Y_u(\widetilde{W}^{(Y)} - \widetilde{N}^{(Y)})Y_u^{\mathrm{T}}\widehat{\xi}_y = \frac{1}{l}\widehat{\xi}_y{}^{\mathrm{T}} Y_u \widetilde{L}^{(Y)}_u Y_u^{\mathrm{T}}\widehat{\xi}_y. \tag{8}$$

By combining the terms defined for the labeled and unlabeled samples, we formulate the objective functions of the SSMDR as follows:

$$\widehat{\xi_x} = \arg\max_{\widehat{\xi_x}} \frac{(1-\mu_A)\widehat{\xi_x}^{\mathrm{T}} \widetilde{A}^{(lbc)} \widehat{\xi_x} + \mu_A J(\widehat{\xi_x})}{(1-\mu_A)\widehat{\xi_x}^{\mathrm{T}} \widetilde{A}^{(lwc)} \widehat{\xi_x} + \mu_A J^*(\widehat{\xi_x})}, \tag{9}$$

$$\widehat{\xi_y} = \arg\max_{\widehat{\xi_y}} \frac{(1-\mu_A)\widehat{\xi_y}^{\mathrm{T}} \widetilde{A}^{(lbc)\dagger} \widehat{\xi_y} + \mu_A J(\widehat{\xi_y})}{(1-\mu_A)\widehat{\xi_y}^{\mathrm{T}} \widetilde{A}^{(lwc)\dagger} \widehat{\xi_y} + \mu_A J^*(\widehat{\xi_y})}, \tag{10}$$

where $J(\widehat{\xi_x})(J(\widehat{\xi_y}))$ is a regularized term induced by the unlabeled samples in $X_u(Y_u)$ and μ_A is a control parameter. In this paper, we also employ the Tikhonov regularizers, $J^*(\widehat{\xi_x}) = \widehat{\xi_x}^{\mathrm{T}}\widehat{\xi_x}$ and $J^*(\widehat{\xi_y}) = \widehat{\xi_y}^{\mathrm{T}}\widehat{\xi_y}$, to avoid the possible singularity of the denominators, as shown in equations 9 and 10. And $\widetilde{N}^{(X)}$ and $\widetilde{N}^{(Y)}$ are set to be $l \times l$ matrices with each entry, $\frac{1}{l}$. It is noted that labeled sets X and Y and unlabeled sets X_u and Y_u are generated from $\widetilde{X}$ and $\widetilde{Y}$. In the optimizations, the unlabeled datasets X_u and Y_u are defined for measuring the total data matrix, including $\widetilde{X}$ and $\widetilde{Y}$; thus, we have $X_u \widetilde{L}_u^{(X)} X_u^{\mathrm{T}} = Y_u \widetilde{L}_u^{(Y)} Y_u^{\mathrm{T}}$ when the PCA criterion is applied.

In this work, we formulate the optimization models based on the discriminative manifold structure embedded in the supervised LPP criterion and the global covariance structures embedded in all data points. Following the objectives of PCA and LPP, the embedding transformations of SSMDR can be analytically computed by using eigen decomposition. The present methodology of this work is considered as the semisupervised extension of the optimized problems in Zhang and Ye (2011), improving the performance by incorporating unlabeled data into the problems for optimization and learning. As a result, the proposed SSMDR can be viewed as a learning method between supervised and unsupervised scenarios. It is also noted that SSMDR can characterize the within-set compactness and between-set separation by utilizing the discriminant features as the fully supervised FDA.

3.5. Definition and Typical Behavior

The projection transformation, T^1_{SSMDR}, is composed of the generalized eigenvectors associated with the first d largest generalized eigenvalues of the following generalized eigenvalue problem:

$$\widetilde{A}^{(rlbc)} \widehat{\xi_x} = \widetilde{\lambda}_x \widetilde{A}^{(rlwc)} \widehat{\xi_x}. \tag{11}$$

Similarly, the projection transformation, T^2_{SSMDR}, comprises the generalized eigenvectors associated with the first d largest generalized eigenvalues of the following generalized eigenvalue problem:

$$\widetilde{A}^{(rlbc)\dagger} \widehat{\xi_y} = \widetilde{\lambda}_y \widetilde{A}^{(rlwc)\dagger} \widehat{\xi_y}, \tag{12}$$

where $\widetilde{A}^{(rlbc)}$ and $\widetilde{A}^{(rlbc)\dagger}$ represent the extended localized between-class scatters and $\widetilde{A}^{(rlwc)}$ and $\widetilde{A}^{(rlwc)\dagger}$ are the regularized local within-class scatters, which are, respectively, defined as

$$\widetilde{A}^{(rlwc)} = (1-\mu_A)\,\widetilde{A}^{(lwc)} + \mu_A I_n,\ \widetilde{A}^{(rlbc)} = (1-\mu_A)\,\widetilde{A}^{(lbc)} + \frac{\mu_A}{l} X_u \widetilde{L}_u^{(Y)} X_u^{\mathrm{T}}, \tag{13}$$

$$\widetilde{A}^{(rlwc)\dagger} = (1-\mu_A)\,\widetilde{A}^{(lwc)\dagger} + \mu_A I_n,\ \widetilde{A}^{(rlbc)\dagger} = (1-\mu_A)\,\widetilde{A}^{(lbc)\dagger} + \frac{\mu_A}{l} Y_u \widetilde{L}_u^{(Y)} Y_u^{\mathrm{T}}, \tag{14}$$

Input: m pairs of labeled samples $\{(x_i, y_i) \mid x_i \in \mathbb{R}^n, y_i \in \mathbb{R}^n\}_{i=1}^{m}$, $\{f_i (\in \{1,2\})\}_{i=1}^{2m}$
l pairs of unlabeled samples $\{(x_i, y_i) \mid x_i \in \mathbb{R}^n, y_i \in \mathbb{R}^n\}_{i=m+1}^{m+l}$
Dimensionality of embedding space $d\,(1 \le d \le n)$
Trade-off parameter $\mu_A \in [0,1]$, $\mu_I = 10\text{e}{-5}$

Output: Two $n \times d$ transformation matrices T_{SSMDR}^1 and T_{SSMDR}^2

Step1: $\widetilde{A}^{(lwc)} = zeros(n,n)$; $\widetilde{A}^{(lbc)} = zeros(n,n)$; $\widetilde{A}^{(lwc)\dagger} = zeros(n,n)$; $\widetilde{A}^{(lbc)\dagger} = zeros(n,n)$;

Step2: Using Eqs.5 to define the adjacency matrices $\widetilde{h}^{(X)}$ and $\widetilde{h}^{(Y)}$ for intra-set samples;

Step3: Using Eq.6 to define the adjacency matrix $\widetilde{h}^{(YX)}$ for inter-set samples;

Step 4: $X \leftarrow (x_1, x_2, ..., x_m), Y \leftarrow (y_1, y_2, ..., y_m)$; *% labeled samples*

$$\widetilde{A}^{(lwc)} = X\left\{diag\left(\widetilde{A}^{(wc)}1_m\right) - \widetilde{A}^{(wc)}\right\}X^{\mathrm{T}};$$

$$\widetilde{A}^{(lbc)} = (1/2)Y\left(diag\left(\widetilde{A}^{(bc)}1_m\right)\right)Y^{\mathrm{T}} + (1/2)X\left(diag\left(1_m{}^{\mathrm{T}}\widetilde{A}^{(bc)}\right)\right)X^{\mathrm{T}} - (1/2)\left(Y\widetilde{A}^{(bc)}X^{\mathrm{T}} + X\widetilde{A}^{(bc)\mathrm{T}}Y^{\mathrm{T}}\right);$$

Step 5: $X_u \leftarrow (x_{m+1}, x_{m+2}, ..., x_{m+l})$, $Y_u \leftarrow (y_{m+1}, y_{m+2}, ..., y_{m+l})$; *% unlabeled samples*

Step 6: Compute the localized scatters $\widetilde{A}^{(lwc)\dagger}$ and $\widetilde{A}^{(lbc)\dagger}$ defined in Eqs.23 and 24;

Step 7:
$$\widetilde{A}^{(rlwc)} = (1-\mu_A)\widetilde{A}^{(lwc)} + \mu_I I_n,\ \widetilde{A}^{(rlbc)} = (1-\mu_A)\widetilde{A}^{(lbc)} + (\mu_A/l)X_u\left\{diag\left(\widetilde{N}^{(X)}1_m\right) - \widetilde{N}^{(X)}\right\}X_u{}^{\mathrm{T}};$$
$$\widetilde{A}^{(rlwc)\dagger} = (1-\mu_A)\widetilde{A}^{(lwc)\dagger} + \mu_I I_n,\ \widetilde{A}^{(rlbc)\dagger} = (1-\mu_A)\widetilde{A}^{(lbc)\dagger} + (\mu_A/l)Y_u\left\{diag\left(\widetilde{N}^{(Y)}1_m\right) - \widetilde{N}^{(Y)}\right\}Y_u{}^{\mathrm{T}};$$

Step 8: $\{\psi_r, \widehat{\lambda_{x[r]}}\}_{r=1}^{d} \leftarrow$ generalized eigenvectors and eigenvalues of $\widetilde{A}^{(rlbc)}\psi = \lambda_x \widetilde{A}^{(rlwc)}\psi$, where
$\widehat{\lambda_{x[1]}} \ge \widehat{\lambda_{x[2]}} \ge ... \ge \widehat{\lambda_{x[d]}}$ and $\psi_r{}^{\mathrm{T}}\widetilde{A}^{(rlwc)}\psi_r = 1$ for $\forall r \in \{1,2,...,d\}$; $T_{SSMDR}^1 = \left(\psi_1 \mid \psi_2 \mid ... \mid \psi_d\right)$;

Step 9: $\{\varphi_r, \widehat{\lambda_{y[r]}}\}_{r=1}^{d} \leftarrow$ generalized eigenvectors and eigenvalues of $\widetilde{A}^{(rlbc)\dagger}\varphi = \lambda_y \widetilde{A}^{(rlwc)\dagger}\varphi$, where
$\widehat{\lambda_{y[1]}} \ge \widehat{\lambda_{y[2]}} \ge ... \ge \widehat{\lambda_{y[d]}}$ and $\varphi_r{}^{\mathrm{T}}\widetilde{A}^{(rlwc)\dagger}\varphi_r = 1$ for $\forall r \in \{1,2,...,d\}$; $T_{SSMDR}^2 = \left(\varphi_1 \mid \varphi_2 \mid ... \mid \varphi_d\right)$.

FIGURE 2. SSMDR algorithm.

where $\mu_A \in [0, 1]$ is a control parameter. That is, $\widetilde{A}^{(rlwc)}(\widetilde{A}^{(rlwc)\dagger})$ changes to $\widetilde{A}^{(lwc)}(\widetilde{A}^{(lwc)\dagger})$ when $\mu_A = 0$, and $\widetilde{A}^{(rlwc)}(\widetilde{A}^{(rlwc)\dagger})$ is transformed to the identity matrix as $\mu_A = 1$. Similarly, matrix $\widetilde{A}^{(rlbc)}(\widetilde{A}^{(rlbc)\dagger})$ changes to $\widetilde{A}^{(lbc)}(\widetilde{A}^{(lbc)\dagger})$ for discrimination when $\mu_A = 0$, and $\widetilde{A}^{(rlbc)}(\widetilde{A}^{(rlbc)\dagger})$ changes to the total scatter matrix for structure preservation when $\mu_A = 1$. These properties endow the proposed SSMDR method the capability to characterize the discriminant feature with the intrinsic structure characteristics effectively preserved. Then, the twin objective functions of the proposed SSMDR algorithm can be formulated as

$$T_{SSMDR}^1 = \underset{\widehat{\xi}_x \in \mathbb{R}^{n\times d}}{\arg\max} \left[tr\left(\left(\widehat{\xi}_x^{\mathrm{T}}\widetilde{A}^{(rlwc)}\widehat{\xi}_x\right)^{-1}\widehat{\xi}_x^{\mathrm{T}}\widetilde{A}^{(rlbc)}\widehat{\xi}_x\right)\right] \tag{15}$$

and

$$T_{SSMDR}^2 = \underset{\widehat{\xi}_y \in \mathbb{R}^{n\times d}}{\arg\max} \left[tr\left(\left(\widehat{\xi}_y^{\mathrm{T}}\widetilde{A}^{(rlwc)\dagger}\widehat{\xi}_y\right)^{-1}\widehat{\xi}_y^{\mathrm{T}}\widetilde{A}^{(rlbc)\dagger}\widehat{\xi}_y\right)\right]. \tag{16}$$

In other words, SSMDR finds a pair of optimal transformations for discrimination plus preservation. SSMDR aims at structuring a low-dimensional space, under which the localized between-class spread is maximized and the regularized local within-class scatter or spread is minimized. In SSMDR, similarity matrices $\widetilde{h}^{(X)}$ and $\widetilde{h}^{(Y)}$ are, respectively, computed for the labeled samples in X and Y, and matrices $\widetilde{N}^{(X)}$ and $\widetilde{N}^{(Y)}$ are computed for unlabeled samples. The efficient implementation of the SSMDR is summarized in Figure 2, where $\{f_i\}_{i=1}^{2m}$ are the class labels for labeled samples. $zeros(n, n)$ denotes a $n \times n$ matrix with all zeros, 1_m denotes an m-dimensional vector with all ones, and $diag(A1_m)$ denotes a diagonal matrix with input elements, $A1_m$. It is noticed that the generalized eigenvalues and eigenvectors can be solved by an eigen solver, such as eigenvalue decomposition.

3.6. Computational Analysis of SSMDR

In the pattern recognition community, we often require to conduct dimensionality reduction when large amount of high-dimensional data is accompanied with the underlying class label information and the spatial information. To address this issue, we formulate the localized within-class scatter $\widetilde{A}^{(lwc)}$ and between-class scatter $\widetilde{A}^{(lbc)}$ using the following forms. In this way, one can easily describe the relations between pairs of features regarding whether sample pairs are close with each other or far apart. Thus, we can express the scatter matrices $\widetilde{A}^{(lwc)}$ and $\widetilde{A}^{(lbc)}$ in a local manner as the following formulation:

$$\widetilde{A}^{(lwc)} = \frac{1}{2}\sum_{i,j=1}^{m}(x_i - x_j)(x_i - x_j)^{\mathrm{T}}\widetilde{A}_{i,j}^{(wc)} = \frac{1}{2}\sum_{i,j=1}^{m}\left(x_i x_i^{\mathrm{T}} + x_j x_j^{\mathrm{T}} - x_i x_j^{\mathrm{T}} - x_j x_i^{\mathrm{T}}\right)\widetilde{A}_{i,j}^{(wc)}, \tag{17}$$

That is, optimizing $\widetilde{A}^{(lwc)}$ directly will not project the data points belonging to the same set into a single cluster as performing FDA, because $\widetilde{A}^{(lwc)}$ measures the pairwise distances between data points. And most importantly, the intrinsic multimodal structures can be efficiently preserved. It is noted that the metric of $\widetilde{A}^{(lwc)}$ is equivalent to the supervised LPP criterion when class labels are available. Similarly, we express $\widetilde{A}^{(lbc)}$ as

$$\widetilde{A}^{(lbc)} = \frac{1}{2}\sum_{i,j=1}^{m}(y_i - x_j)(y_i - x_j)^{\mathrm{T}}\widetilde{A}_{i,j}^{(bc)} = \frac{1}{2}\sum_{i,j=1}^{m}\left(y_i y_i^{\mathrm{T}} - y_i x_j^{\mathrm{T}} - x_j y_i^{\mathrm{T}} + x_j x_j^{\mathrm{T}}\right)\widetilde{A}_{i,j}^{(bc)}, \tag{18}$$

where $\widetilde{A}^{(wc)}$ and $\widetilde{A}^{(bc)}$ are $m \times m$ matrices. Based on the similarity matrices $\widetilde{h}^{(X)}$ and $\widetilde{h}^{(Y)}$, we weight the values for the sample pairs in the same object. We do not exert penalties for the nonneighboring sample pairs belonging to the same class. We also do not exert penalties for neighboring pairs from different objects or classes, because we aim at separate them. First, we give a matrix interpretation of the scatter $\widetilde{A}^{(lwc)}$. Then, the localized within-class scatter $\widetilde{A}^{(lwc)}$ can be formulated in a matrix form as the following:

$$\begin{aligned}
\widetilde{A}^{(lwc)} &= \frac{1}{2}\sum_{i=1}^{m} x_i\left(\sum_{j}\widetilde{A}_{i,j}^{(wc)}\right)x_i^{\mathrm{T}} + \frac{1}{2}\sum_{j=1}^{m} x_j\left(\sum_{i}\widetilde{A}_{i,j}^{(wc)}\right)x_j^{\mathrm{T}} \\
&\quad - \frac{1}{2}\sum_{i,j=1}^{m} x_i\widetilde{A}_{i,j}^{(wc)}x_j^{\mathrm{T}} - \frac{1}{2}\sum_{i,j=1}^{m} x_j\widetilde{A}_{i,j}^{(wc)}x_i^{\mathrm{T}}, \\
&= \sum_{i=1}^{m} x_i\widetilde{F}_{ii}^{(wc)}x_i^{\mathrm{T}} - \sum_{i,j=1}^{m} x_i\widetilde{A}_{i,j}^{(wc)}x_j^{\mathrm{T}} \\
&= \sum_{i=1}^{m}\widetilde{F}_{ii}^{(wc)}x_i x_i^{\mathrm{T}} - X\widetilde{A}^{(wc)}X^{\mathrm{T}} = X\widetilde{W}^{(wc)}X^{\mathrm{T}},
\end{aligned} \tag{19}$$

where $\widetilde{W}^{(wc)} = \widetilde{F}^{(wc)} - \widetilde{A}^{(wc)}$ and $\widetilde{F}^{(wc)}$ is an m-dimensional diagonal matrix with ith entry being $\widetilde{F}_{ii}^{(wc)} = \sum_j \widetilde{A}_{i,j}^{(wc)}$. $\widetilde{F}^{(wc)}$is commonly called the Laplacian matrix of $\widetilde{A}^{(wc)}$ in the spectral graph theory (Chung 1997). It is easy to verify that $\widetilde{A}^{(wc)}$ and $\widetilde{W}^{(wc)}$ are all symmetric and positive semidefinite matrices. Similarly, the localized between-class scatter matrix

$\widetilde{A}^{(lbc)}$can be expressed in a matrix interpretation as

$$\begin{aligned}\widetilde{A}^{(lbc)} &= \frac{1}{2}\sum_{i=1}^{m} y_i \left(\sum\nolimits_j \widetilde{A}_{i,j}^{(bc)}\right) y_i^{\mathrm{T}} + \frac{1}{2}\sum_{j=1}^{m} x_j \left(\sum\nolimits_i \widetilde{A}_{i,j}^{(bc)}\right) x_j^{\mathrm{T}} \\ &\quad - \frac{1}{2}\sum_{i,j=1}^{m} y_i \widetilde{A}_{i,j}^{(bc)} x_j^{\mathrm{T}} - \frac{1}{2}\sum_{i,j=1}^{m} x_j \widetilde{A}_{j,i}^{(bc)} y_i^{\mathrm{T}} \\ &= \frac{1}{2}(Y\widetilde{D}^{(bc)}Y^{\mathrm{T}} + X\widetilde{M}^{(bc)}X^{\mathrm{T}}) - \frac{1}{2}(Y\widetilde{A}^{(bc)}X^{\mathrm{T}} + X\widetilde{A}^{(bc)\mathrm{T}}Y^{\mathrm{T}})\end{aligned} \tag{20}$$

where $\widetilde{D}^{(bc)}$ and $\widetilde{M}^{(bc)}$ are m-dimensional diagonal matrices with ith (or jth) element being $\widetilde{D}_{ii}^{(wc)} = \sum_{j=1}^{m} \widetilde{A}_{i,j}^{(bc)}$ and $\widetilde{M}_{jj}^{(wc)} = \sum_{i=1}^{m} \widetilde{A}_{i,j}^{(bc)}$. Thus, the first optimization problem of SSMDR can be expressed as

$$\begin{aligned} T_{SSMDR}^{1} = \quad & \arg\max_{\hat{\xi}_{x_{[r]}} \in \mathbb{R}^{n\times 1}} \hat{\xi}_{x_{[r]}}^{\mathrm{T}} \widetilde{A}^{(rlbc)} \hat{\xi}_{x_{[r]}} \\ \text{subject to} \quad & \forall\, r \in \{1, 2, \ldots, d\} : \ \hat{\xi}_{x_{[r]}}^{\mathrm{T}} \widetilde{A}^{(rlwc)} \hat{\xi}_{x_{[r]}} - 1 = 0, \end{aligned} \tag{21}$$

or equivalently

$$T_{SSMDR}^{1} = \arg\max_{\hat{\xi}_{x_{[r]}} \in \mathbb{R}^{n\times 1}} \left[\left(\hat{\xi}_{x_{[r]}}^{T} \widetilde{A}^{(rlwc)} \hat{\xi}_{x_{[r]}}\right)^{-1} \left(\hat{\xi}_{x_{[r]}}^{T} \widetilde{A}^{(lbc)} \hat{\xi}_{x_{[r]}}\right)\right] \widetilde{A}^{(lwc)\dagger}, \tag{22}$$

where I_d is the identity matrix on $\mathbb{R}^d$. That is, SSMDR aims to find the transformation matrix T_{SSMDR}^{1} such that the localized between-class scatter in the embedding space (i.e., $T_{SSMDR}^{1^{\mathrm{T}}} \widetilde{A}^{(rlbc)} T_{SSMDR}^{1}$) is to be maximized and the regularized local within-class scatter in the embedding space (i.e., $T_{SSMDR}^{1^{\mathrm{T}}} \widetilde{A}^{(rlwc)} T_{SSMDR}^{1}$) is to be minimized. The regularized local within-class scatter $\widetilde{A}_{\Phi}^{(rlwc)\dagger}$ and the extended between-class scatter $\widetilde{A}^{(rlbc)}$ are formulated as

$$\widetilde{A}^{(rlwc)} = (1-\mu_A)\,\widetilde{A}^{(lwc)} + \mu_A I_n, \ \widetilde{A}^{(rlbc)} = (1-\mu_A)\,\widetilde{A}^{(lbc)} + \frac{\mu_A}{l} X_u \widetilde{L}_u^{(X)} X_u^{\mathrm{T}}.$$

We assume that the generalized eigenvalues $\{\widetilde{\lambda_{x_{[r]}}}\}_{r=1}^{d}$ of equation (21) or equation (22) are sorted in descending order and the generalized eigenvectors $\{\widehat{\xi}_{x_{[r]}}\}_{r=1}^{d}$ are normalized as $\widehat{\xi}_{x_{[r]}}^{\mathrm{T}} \widetilde{A}^{(rlwc)} \widehat{\xi}_{x_{[r]}} = 1$ for $r = 1, 2, \ldots, d$. Then, a solution of the transformation T_{SSMDR}^{1} is given by

$$T_{SSMDR}^{1} = (\hat{\xi}_{x_{[1]}} | \hat{\xi}_{x_{[2]}} | \ldots | \hat{\xi}_{x_{[d]}}),$$

where $\{\widehat{\xi}_{x_{[r]}}\}_{r=1}^{d}$ are the generalized eigenvectors corresponding to the first d generalized eigenvalues, i.e., $\{\widetilde{\lambda_{x_{[r]}}}\}_{r=1}^{d}$. By considering the equivalence relation between equations (21) and (22), we can see that SSMDR finds the transformation T_{SSMDR}^{1} such that nearby sample pairs in the original space R^n and the projected data are still compact in the reduced space with the normalization constraint $T_{SSMDR}^{1^{\mathrm{T}}} \widetilde{A}^{(rlwc)} T_{SSMDR}^{1} = I_d$. This can be solved by using an eigen solver. That is to say SSMDR attempts to preserve the data manifold of all the samples. All of these definitions and formulations are suitable for the transformation matrix T_{SSMDR}^{2}. We see from the Appendix A that $\widetilde{F}^{(wc)\dagger}$ is a diagonal matrix whose entries are column (or row, since $\widetilde{A}^{(wc)\dagger}$ is symmetric) sums of the matrix $\widetilde{A}^{(wc)\dagger}$, i.e., $\widetilde{F}_{ii}^{(wc)\dagger} = \sum_{j=1}^{m} \widetilde{A}_{i,j}^{(wc)\dagger}$; thus,

the localized within-class scatter $\widetilde{A}^{(lwc)\dagger}$ and localized between-class scatter $\widetilde{A}^{(lbc)\dagger}$ can be expressed as

$$\widetilde{A}^{(lwc)\dagger} = Y\left(\widetilde{F}^{(wc)\dagger} - \widetilde{A}^{(wc)\dagger}\right)Y^{\mathrm{T}} = Y\widetilde{W}^{(wc)\dagger}Y^{\mathrm{T}}, \tag{23}$$

$$\widetilde{A}^{(lbc)\dagger\dagger} = \frac{1}{2}\left(X\widetilde{D}^{(bc)\dagger}X^{\mathrm{T}} + Y\widetilde{M}^{(bc)\dagger}Y^{\mathrm{T}}\right) - \frac{1}{2}\left(X\widetilde{A}^{(bc)\dagger}Y^{T} + Y\widetilde{A}^{(bc)\dagger\mathrm{T}}Y^{\mathrm{T}}\right) \tag{24}$$

where $\widetilde{D}^{(bc)\dagger}$ and $\widetilde{M}^{(bc)\dagger}$ are m-dimensional diagonal matrix with ith (or jth) element being $\widetilde{D}^{(wc)\dagger}_{ii} = \sum_{j=1}^{m}\widetilde{A}^{(bc)\dagger}_{i,j}$ and $\widetilde{M}^{(wc)\dagger}_{jj} = \sum_{i=1}^{m}\widetilde{A}^{(bc)\dagger}_{i,j}$. It is also noted that $\widetilde{A}^{(lbc)} = \widetilde{A}^{(lbc)\dagger}$ in the computations due to the fact that $\widetilde{A}^{(bc)} = \widetilde{A}^{(bc)\dagger}$. Then, we obtain the second optimization problem for SSMDR, which is analogously defined as

$$\begin{aligned} T^{2}_{SSMDR} = \quad & \arg\max_{\hat{\xi}_{y_{[r]}}\in R^{n\times 1}} \hat{\xi}^{\mathrm{T}}_{y_{[r]}}\widetilde{A}^{(rlbc)\dagger}\hat{\xi}_{y_{[r]}} \\ \text{subject to} \quad & \forall\, r \in \{1,2,\ldots,d\}: \ \hat{\xi}^{\mathrm{T}}_{y_{[r]}}\widetilde{A}^{(rlwc)\dagger}\hat{\xi}_{y_{[r]}} - 1 = 0, \end{aligned} \tag{25}$$

where the regularized local within-class scatter or spread $\widetilde{A}^{(rlwc)\dagger}$ and the extended between-class scatter $\widetilde{A}^{(rlbc)\dagger}$ are similarly defined as the following:

$$\widetilde{A}^{(rlwc)\dagger} = (1-\mu_A)\,\widetilde{A}^{(lwc)\dagger} + \mu_A I_n, \ \widetilde{A}^{(rlbc)\dagger} = (1-\mu_A)\,\widetilde{A}^{(lbc)} + \frac{\mu_A}{l}Y_u\widetilde{L}^{(Y)}_u Y^{\mathrm{T}}_u.$$

Analogous to the computational process of T^{1}_{SSMDR}, the second transformation matrix T^{2}_{SSMDR} can be given by $T^{2}_{SSMDR} = (\hat{\xi}_{y[1]}|\hat{\xi}_{y[2]}.|\ldots|\hat{\xi}_{y[d]})$, where $\{\hat{\xi}_{y_{[r]}}\}^{d}_{r=1}$ are the generalized eigenvectors of the eigen problem in equation (25) associated with the first d largest generalized eigenvalues $\widetilde{\lambda}_{y_{[r]}}, r = 1,2,\ldots,d$, where $d \le n$.

3.7. Discussion

For efficient multidimensional data visualization and feature extraction via dimensionality reduction, there are a number of linear or nonlinear, local or global, supervised, unsupervised, or semisupervised algorithms proposed. In this work, a newly formulated objective-based multimodal SSMDR method is considered. It is interesting from some distinctive perspectives. In later sections, we will discuss some issues related to our method.

3.7.1. Comparative Analysis. Similar to the main focuses and measurement criteria of some existing discriminative dimensionality reduction methods, e.g., FDA and LFDA, the proposed SSMDR method aims to minimize the within-class compactness and maximize between-class separation as well. However, SSMDR is significantly different from them based on newly defined objectives. SSMDR exhibits certain typical behaviors and advantages over them:

(1) The classical FDA method tends to become inefficient when facing the multimodal data distributions, e.g., *XOR* dataset. Moreover, FDA is restricted by the upper bound on the dimensionality of the reduced space (Martinez and Kak 2001; He et al. 2005). So does semisupervised SDA, though SDA has a regularized term incorporating the intrinsic geometrical structure inferred from labeled and unlabeled data. Correspondingly, SSMDR is a discriminant plus preservation technique which can extricate from the bound and, most importantly, it can represent the multimodal data respectably.

(2) The popular PCA of minimizing the sample reconstruction error does not consider the local information of data. So does MDS. Although LPP, NPE, LE, LLE, SDA, LFDA, and SELF are formulated based on the local or neighborhood information preservation, but they are unable to represent the multimodal *XOR* and challenging "*incomplete tire*" in our study. Moreover, FDA and LFDA may become overfitted when the number of labeled data is small. In contrast, local information of data can be effectively kept by SSMDR and it can embed the points of *XOR* and "*incomplete tire*" effectively. Similar to SDA and SELF, SSMDR can be avoided from being overfitted to the small number of labeled data. It is worth noting that simulation results show that SELF and SDA still cannot embed the multimodal data points efficiently.

(3) The above-mentioned methods all aim at finding a set of transforming basis vectors for preservation or discrimination. Different from them, our multimodal SSMDR method aims at finding two sets of locality preserving basis vectors for learning the projections. This property is analogous to *canonical correlation analysis* (CCA). Though SSMDR and CCA can tackle the problems posed by the multimodal distributions, there are natural differences between them. These issues will be discussed in detail in next section.

3.7.2. Comparison with CCA. CCA (Hardoon, Szedmak, and Shawe-Taylor 2004) is a widely used method in pattern recognition. The objective of CCA is to find two sets of optimal basis vectors for two sets of variables, one for each set, such that the correlation between the projections of the variables onto these basis vectors was mutually maximized.

Given two multivariate sets $X \in \mathbb{R}^p$, $Y \in \mathbb{R}^q$ with $X = [x_1, x_2, \ldots, x_m]$ and $Y = [y_1, y_2, \ldots, y_m]$, CCA aims at finding two optimal projection vectors $\widehat{\xi}_x$ and $\widehat{\xi}_y$ such that the correlation coefficient between $\widehat{\xi}_x^{\mathrm{T}} X$ and $\widehat{\xi}_y^{\mathrm{T}} Y$ is maximized. That is,

$$\rho = \arg\max_{\widehat{\xi}_x, \widehat{\xi}_y} \left(\frac{\widehat{\xi}_x^{\mathrm{T}} C_{xy} \widehat{\xi}_y}{\sqrt{\widehat{\xi}_x^{\mathrm{T}} C_{xx} \widehat{\xi}_x . \widehat{\xi}_y^{\mathrm{T}} C_{yy} \widehat{\xi}_y}} \right), \quad \text{subject to} \quad \widehat{\xi}_x^{\mathrm{T}} C_{xx} \widehat{\xi}_x = 1, \widehat{\xi}_y^{\mathrm{T}} C_{yy} \widehat{\xi}_y = 1, \quad (26)$$

where C_{xy} is the between-class covariance matrix of X and Y sets, and C_{xx} and C_{yy} are the covariance matrices of X and Y sets, respectively. The maximum canonical correlation is the maximum of ρ with respect to the basis vectors $\widehat{\xi}_x$ and $\widehat{\xi}_y$, where $\widehat{\xi}_x$ and $\widehat{\xi}_y$ can be obtained by solving the following two generalized eigenvalue problems:

$$\begin{pmatrix} C_{xy} C_{yy}^{-1} C_{yx} & 0 \\ 0 & C_{yx} C_{xx}^{-1} C_{xy} \end{pmatrix} \begin{pmatrix} \widehat{\xi}_x \\ \widehat{\xi}_y \end{pmatrix} = \widetilde{\lambda} \begin{pmatrix} C_{xx} & 0 \\ 0 & C_{yy} \end{pmatrix} \begin{pmatrix} \widehat{\xi}_x \\ \widehat{\xi}_y \end{pmatrix}, \quad (27)$$

where the eigenvalue is the canonical correlation to be optimized. Though CCA and SSMDR both aim at finding a pair of projective basis vectors for dimensionality reduction and feature extraction, there are significant differences between them. One of these includes that CCA is unable to preserve the locality between points, but SSMDR incorporates the local manifold information with the scatters or spreads enabling the intrinsic local structures around data point to be preserved. Besides, CCA does not take into account the class information of data. It is naturally an unsupervised learning method, while SSMDR takes full advantage of all the underlying class labels of labeled data to improve the performance. In most practical applications, unlabeled samples are readily available, but labeled ones are expensive to obtain. SSMDR can utilize a small number of labeled data and great amount of unlabeled data to perform semisupervised dimensionality reduction. Another marked difference is that CCA and SSMDR aim at evaluating the within-class and between-class scatters using the different measurement criteria.

3.7.3. Kernel SSMDR for Nonlinear Dimensionality Reduction. In this subsection, we consider to extend SSMDR to nonlinear dimensionality reduction scenarios by employing the standard kernel approach (Schölkopf and Smola 2002). Let Φ be the mapping from $\mathbb{R}^n$ to $N^p (p > n)$, which can be implicitly defined by a kernel function, that is, the (i, j)th entry is given by

$$K_{ij} = K(x_i, x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle = \Phi(x_i)^{\mathrm{T}} \Phi(x_j), \tag{28}$$

where $\langle . \rangle$ denotes the inner product in the mapped space N. Gaussian kernel, a typical choice of the kernel function, is given by

$$K(x, x^{\mathrm{T}}) = \exp(-||x - x^{\mathrm{T}}||^2 / 2\sigma^2). \tag{29}$$

Schölkopf pointed out that every solution in the kernel space, N, could be written as an expansion in terms of the mapped training data (Schölkopf and Smola 2002). Let $\Phi(X) = (\Phi(x_1), \Phi(x_2), \ldots, \Phi(x_m))$, $\Phi(Y) = (\Phi(y_1), \Phi(y_2), \ldots, \Phi(y_m))$, then the projection vectors $\widehat{\xi_x}^{\Phi}$ and $\widehat{\xi_y}^{\Phi}$ defined in high-dimensional kernel space can be rewritten as

$$\widehat{\xi_x}^{\Phi} = \sum_{i=1}^{m} \widehat{T_{x_{[i]}}} \Phi(x_i) = \Phi(X) \widehat{T}_x, \tag{30}$$

$$\widehat{\xi_y}^{\Phi} = \sum_{i=1}^{m} \widehat{T_{y_{[i]}}} \Phi(y_i) = \Phi(Y) \widehat{T}_y. \tag{31}$$

As detailed in Section 3.5, the generalized eigenvalue problems that need to be solved in SSMDR are shown, where the identity matrix I_n on $\mathbb{R}^n$ is used to avoid the singularity and ensure the stability of the generalized eigen-decomposition problem. The localized scatters $\widetilde{A}_{\Phi}^{(rlbc)}$ and $\widetilde{A}_{\Phi}^{(rlwc)}$ in the high-dimensional kernel feature space can then be formulated as

$$\begin{aligned} \widetilde{A}_{\Phi}^{(rlbc)} = {} & \frac{(1-\mu_A)}{2} [\Phi(Y) \widetilde{D}^{(bc)} \Phi(Y)^{\mathrm{T}} + \Phi(X) \widetilde{M}^{(bc)} \Phi(X)^{\mathrm{T}} \\ & - (\Phi(Y) \widetilde{A}^{(bc)} \Phi(X)^{\mathrm{T}} + \Phi(X) \widetilde{A}^{(bc)\mathrm{T}} \Phi(Y)^{\mathrm{T}})] + \frac{\mu_A}{l} \Phi(X_u) \widetilde{L}_u^{(X)} \Phi(X_u)^{\mathrm{T}}, \end{aligned} \tag{32}$$

$$\begin{aligned} \widetilde{A}_{\Phi}^{(rlwc)} &= (1-\mu_A) \Phi(X) \left(\widetilde{F}^{(wc)} - \widetilde{A}^{(wc)} \right) \Phi(X)^{\mathrm{T}} + \mu_A I \\ &= (1-\mu_A) \Phi(X) \widetilde{W}^{(wc)} \Phi(X)^{\mathrm{T}} + \mu_A I. \end{aligned} \tag{33}$$

Substituting $\widetilde{A}_{\Phi}^{(lbc)}$, $\widetilde{A}_{\Phi}^{(rlwc)}$, and equation (30) into the generalized eigenvalue problem in equation (11), we can obtain

$$\widetilde{A}_{\Phi}^{(lbc)} \Phi(X) \widehat{T}_x = \widetilde{\delta}_x \widetilde{A}_{\Phi}^{(rlwc)} \Phi(X) \widehat{T}_x. \tag{34}$$

Multiply equation (34) by $\Phi(X)^{\mathrm{T}}$ from the left-hand side, then we can get

$$\begin{aligned} & \left[\frac{(1-\mu_A)}{2} \left(K_{xy} \widetilde{D}^{(bc)} K_{yx} + K_{xx} \widetilde{M}^{(bc)} K_{xx} - K^{(bc)} \right) + \frac{\mu_A}{l} K_{xx_u} \widetilde{L}_u^{(X)} K_{x_u x} \right] \widehat{T}_x \\ & = \widetilde{\delta}_x \left((1-\mu_A) K_{xx} \widetilde{W}^{(wc)} K_{xx} + \mu_A K_{xx} \right) \widehat{T}_x, \end{aligned} \tag{35}$$

where $K^{(bc)} = K_{xy} \widetilde{A}^{(bc)} K_{xx} + (K_{xy} \widetilde{A}^{(bc)} K_{xx})^{\mathrm{T}}$ is the kernel matrix between labeled data in X, $K_{xy} = \Phi(X)^{\mathrm{T}} \Phi(Y) = K_{yx}^{\mathrm{T}}$ is the kernel matrix between labeled samples in X and Y,

and $K_{xx_u} = \Phi(X)^{\mathrm{T}}\Phi(X_u)$ is the kernel matrix between labeled samples in X and unlabeled samples in X_u and $K_{x_ux} = K_{xx_u}^{\mathrm{T}}$. Let

$$\Upsilon_1^{\Phi} = \frac{(1-\mu_A)}{2}\left[K_{xy}\widetilde{D}^{(bc)}K_{yx} + K_{xx}\widetilde{M}^{(bc)}K_{xx} - K_{xy}\widetilde{A}^{(bc)}K_{xx} - \left(K_{xy}\widetilde{A}^{(bc)}K_{xx}\right)^{\mathrm{T}}\right] + \frac{\mu_A}{l}K_{xx_u}\widetilde{L}_u^{(X)}K_{x_ux},$$

$$\Upsilon_2^{\Phi} = (1-\mu_A)\,K_{xx}\widetilde{W}^{(wc)}K_{xx} + \mu_A K_{xx},$$

thus, equation (35) can be reformulated as the following:

$$\Upsilon_1^{\Phi}\widehat{T}_x = \widetilde{\delta}_x\Upsilon_2^{\Phi}\widehat{T}_x. \tag{36}$$

Because $\Upsilon_2^{\Phi} = (1-\mu_A)K_{xx}\widetilde{W}^{(wc)}K_{xx} + \frac{\mu_A}{l}K_{xx_u}\widetilde{L}_u^{(X)}K_{x_ux}$ is a symmetric and positive semidefinite matrix and is not always of full rank, we need to regularize it to avoid the singularity and assure the stability of the generalized eigenvalue problems by adding the term $\mu_I I$ with a small positive scalar μ_I. We therefore replace equation (36) by

$$\Upsilon_1^{\Phi}\widehat{T}_x = \widetilde{\delta}_x\left(\Upsilon_2^{\Phi} + \mu_I I\right)\widehat{T}_x. \tag{37}$$

Let $\{\widehat{T_{x_{[r]}}}\}_{r=1}^{d}$ be the generalized eigenvectors associated with the first d largest generalized eigenvalues $\widetilde{\delta_{x_{[r]}}}, r = 1, 2, \ldots, d$ of the generalized eigenvalue problem in equation (37), where they are sorted and normalized as $\widetilde{\delta_{\chi_{[1]}}} \geq \widetilde{\delta_{\chi_{[2]}}} \geq \ldots \geq \widetilde{\delta_{\chi_{[d-1]}}} \geq \widetilde{\delta_{\chi_{[d]}}}$ and $\widehat{T_{x_{[r]}}}^{\mathrm{T}}(\Upsilon_2^{\Phi} + \mu_I I)\widehat{T_{x_{[r]}}} = 1 \text{for} r = 1, 2, \ldots, d$.

The extended algorithm implies that the samples data points appear only via the forms of inner products. The projection transformation $\widehat{T}_y$ in the kernel feature space can be computed by using the similar techniques (see Appendix B for the detailed computations). Then, the solution of $\widehat{T}_y$ is given by

$$\widehat{T}_y = (\widehat{T_{y_{[1]}}}|\widehat{T_{y_{[2]}}}|\ldots|\widehat{T_{y_{[d]}}}). \tag{38}$$

It is important to note that the size of matrices to be eigen-decomposed in the kernel formulations only depends on the amount of data samples, but not on the input dimensionality. Thus, the kernelized extensions can improve the computational efficiency when the number of samples is smaller than the input dimensionality.

4. EVALUATIONS ON SEMISUPERVISED LEARNING

This section evaluates the performance of the proposed SSMDR method and other established approaches (i.e., PCA (Mardia et al. 1980), FDA (Duda et al. 2001), LPP (He and Niyogi 2004), *neighborhood preserving embedding* (NPE) (He et al. 2005), *isometric projection* (IsoProjection) (Cai et al. 2007a), LFDA (Sugiyama 2006, 2007), LE (Belkin and Niyogi 2001), *ViSOM* (Yin 2002a), LLE (Roweis and Saul 2000; Lawrence and Sam 2003), MDS (Cox and Cox 2001), *ISOMAP* (Tenenbaum et al. 2000), CCA (Hardoon et al. 2004), SDA (Cai et al. 2007b), and SELF (Sugiyama et al. 2010)). We use six datasets for visualizing the multidimensional points and classification. Note that PCA, LPP, NPE, IsoProjection, LE, ViSOM, LLE, MDS, ISOMAP, and CCA are unsupervised methods; FDA and LFDA are supervised approaches; and SDA, SELF, and SSMDR are semisupervised algorithms.

4.1. Simulation Settings

First, we introduce the setting of the parameters. For LFDA and SELF, the weigh matrix is computed by the local scaling method (Zelnik-Manor and Perona 2005) and the heat kernel (Belkin and Niyogi 2001) for LPP, LE, and SDA. The parameter and kernel width for the heat kernel and the graph weights are obtained by fivefold cross validation (Lin, Liu, and Chen 2005). For the locality or neighborhood-preservation-induced approaches, when the nearest neighbor search is applied, the amount of nearest neighbors, k, is always set to 12 when no further explanation is given. For ViSOM, the map size is 20×20, the number of iterations is 1,000, and the control parameter in the weight updating formula is set to 0.1. Besides, the control parameter between LFDA and PCA is 0.5 for SELF, and the parameter between FDA and the regularized term is 0.5 for SDA. We also test "SSMDR (0.5)" (SSMDR with $\mu_A = 0.5$) in the simulations.

Numerous simulations in (Chapelle, Schölkopf, and Zien 2006) have been conducted to evaluate the semisupervised learning methods. Our obtained results show that the performance of dimensionality reduction depends on the types of datasets. For classification, we use the one-nearest-neighbor classifier with Euclidean metric to avoid the bias caused by the choice of the learning methods. The steps can be briefly described as follows. First, we calculate the image subspaces from the sample set (labeled and unlabeled), project the points into d-dimensional feature space, and create new patterns for learning. We choose the training data (labeled and unlabeled) and test data (unlabeled) from the new sample pool, and then train a classifier model. Finally, new data points are identified by a one-nearest-neighbor classifier. For dimensionality reduction, we aim at embedding the sample points into subspaces with different reduced dimensions. In this paper, different settings over different reduced dimensions and labeled numbers are used to demonstrate the effectiveness of these methods. All the used algorithms are implemented in Matlab 7.1. We perform all simulations on a PC with Intel(R) Core(TM) i5 CPU 650 @3.20GHz 3.19 GHz 4G.

4.2. Data Preparation

In this study, two synthetic and four benchmark datasets are tested for visualization and classification. The first one is the two-class *XOR* dataset. Each class has 1,000 points and two isolated clusters. The second one is the "*incomplete tire*" dataset associated with two class labels and has 1,000 instances in each class, and each data point is represented by a three-dimensional feature. The third one is the *pen-based recognition of handwritten digits* database (available from: http://archive.ics.uci.edu/ml/machine-learning-databases/pendigits/), consisting of 10,992 images of "0"–"9", in which each image consists of 16 attributes and integers in the range from 0 to 100. The fourth one is *letter image recognition* database (available from: http://archive.ics.uci.edu/ml/ machine-learning-databases/letter-recognition/), in which the images are based on 20 different fonts and each letter was randomly distorted to produce a file of 20,000 unique stimuli. Each stimulus was converted into 16 attributes scaled to a range of integers from 0 through 255. The fifth one is the *optical recognition of handwritten digits* database which consists of 5,620 digits of "0"–"9" (available from: http://archive.ics.uci.edu/ml/ machine-learning-databases/optdigits/). The size of each image is $8\times8 = 64$ pixels, with integers ranged from 0 to 16. And the last one is the well-known *COIL-20* database (available from: http://www.cs.columbia.edu/CAVE/software/softlib/coil-20.php), in which the size of each image is 32×32 pixels, with 256 gray levels per pixel. Thus, each point of the COIL-20 database is represented by a 1,024-dimensional vector.

4.3. Multidimensional Data Visualization via Dimensionality Reduction

For data visualization, we mainly evaluate the performance of the learnt embedding spaces in terms of between-class separability, within-class compactness, and intrinsic multimodality preservation capabilities. For SSMDR, the subspace is spanned by the pair of transformations $\widehat{\xi_x} = (\hat{\xi}_{x_{[1]}}, \ldots, \hat{\xi}_{x_{[r]}})$ and $\widehat{\xi_y} = (\hat{\xi}_{y_{[1]}}, \hat{\xi}_{y_{[r]}})$, and data points will be projected onto $\widehat{\xi_x}$ and $\widehat{\xi_y}$ to generate low-dimensional feature vectors. Here, r equals to 2 that is the original data will be projected into a two-dimensional space, and the two classes of each dataset are denoted by "☆" and "△." For *XOR* and "*incomplete tire,*" we mainly evaluate the representation capability of the two-dimensional embedding space obtained by each method.

Here, we use the multimodal *XOR* shown in Figure 1(d) to demonstrate the efficiency of the embedding space found by SSMDR. For visualization, we set the k of nearest neighbors to 6. One fiftieth of the data points are used as labeled and the rest as unlabeled. The visualization results are exhibited in Figure 3, in which the horizontal axis is the first feature extracted by each method, while the vertical axis is the second feature. It is noticed that FDA can only extract one meaningful feature in the two-class problems, which is due to the limitation of the trace of the matrix (Sugiyama 2007; He et al. 2005). Thus, we randomly select the second feature. Based on the presented results, the following observations are found. First, PCA, FDA, LPP, NPE, LFDA, IsoProjection, CCA, and SELF can effectively preserve the intrinsic local and multimodal structure information hidden in the data, but they fail to separate samples of "☆"-class and "△"-class effectively. This is due to the fact that they are intrinsically compact in the original input space. Second, SSMDR achieves within-class compactness with local and multimodal structures preserved, while SDA, MDS, ViSOM, and the graph-embedding-based LE, LLE, and ISOMAP are unable to deliver the required results. They tend to mix samples of different classes with each other completely in the feature spaces. Third, CCA is able to deliver results that are comparable to our proposed SSMDR method. This is due to the fact that both CCA and SSMDR compute two sets of basis vectors for projection, which is naturally different from the other methods. We in the later section will demonstrate the outstanding characteristics of SSMDR in visualization by using multimodal *XOR* dataset. We will then conduct another two multidimensional data visualization simulations using two benchmark datasets.

Here, we test these methods on two benchmark datasets, namely, "*incomplete tire*" and *pen-based recognition of handwritten digits*, and investigate how they behave in visualization. As described in Roweis and Saul (2000) and Yang et al. (2006), the "*incomplete tire*" dataset is a highly challenging problem to dimensionality reduction (Roweis and Saul 2000), because to date there has no effective mapping function found to be able to embed the data points of "*incomplete tire*" to a linear subspace with higher between-class separation and local information preserved. For the *pen-based recognition of handwritten digits* dataset, we randomly choose 400 examples of digits "3," "5," "8," and "9" and create our digit set for visualizing the embedding results, because digits "3," "5," '8," and "9" are similar in shapes and thus tend to have similar embedding, which may impose greater difficulties to handwritten digital visualization. In our study, we merge digits "3" and "5" to a single class represented by "△," and "8," and "9" to another class represented by "☆." Figure 4 shows the original distributions of points of the two datasets. Figures 5 and 6 depict the data embedded in the embedding space learned by each method. In Figures 5 and 6, the five-pointed stars, circles, and triangles denote the three classes and the filled and unfilled symbols represent labeled and unlabeled samples, respectively.

For the synthetic "*incomplete tire*" dataset, we choose three points out of 10 of the points from each class as labeled, while the rest are treated as unlabeled. Because two class samples of the *"incomplete tire"* dataset are completely mixed with each other in the original

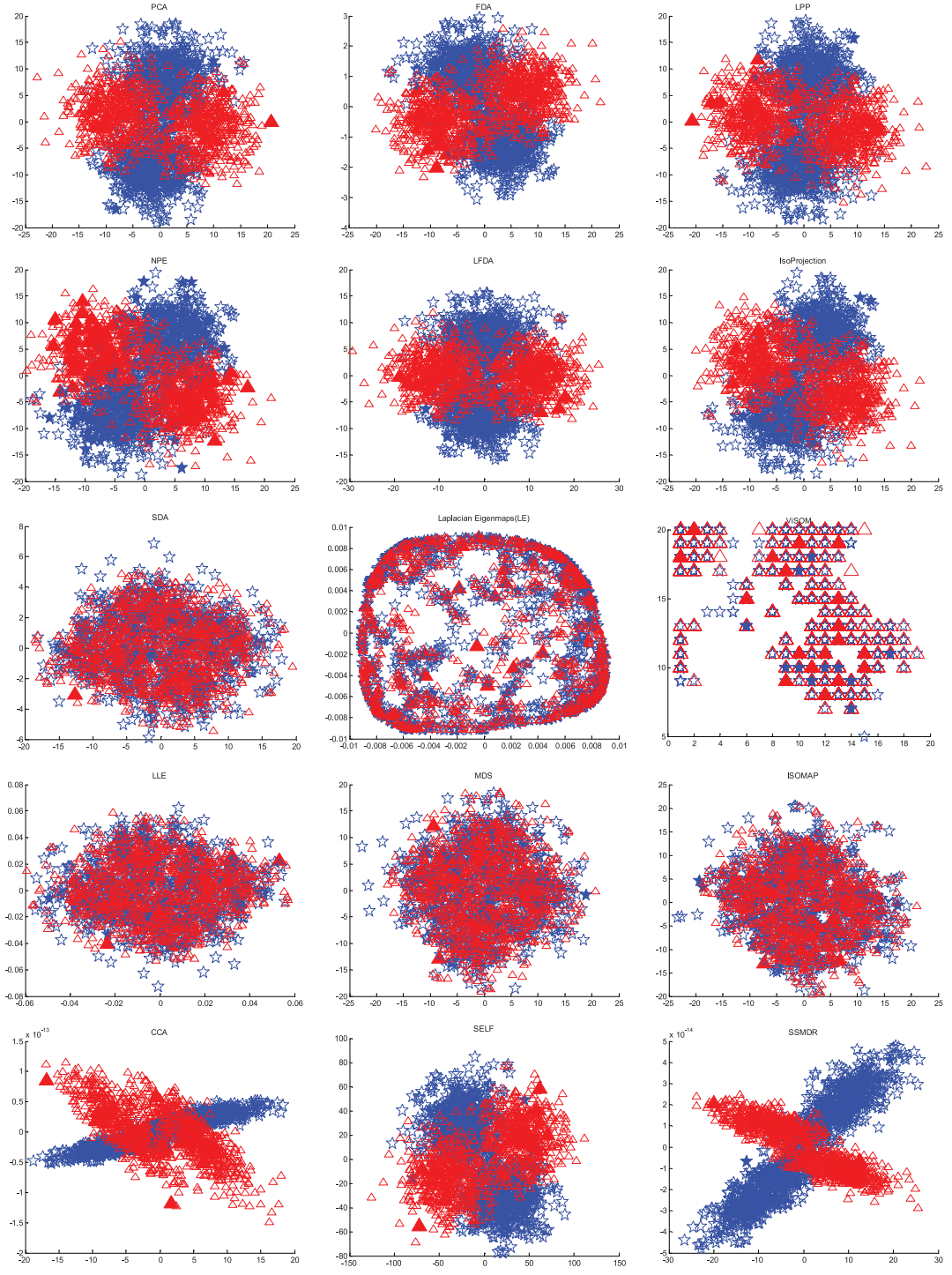


FIGURE 3. The embedding results of samples in the synthetic *XOR* dataset, where the filled or unfilled five-pointed stars and triangles are the labeled or unlabeled samples.

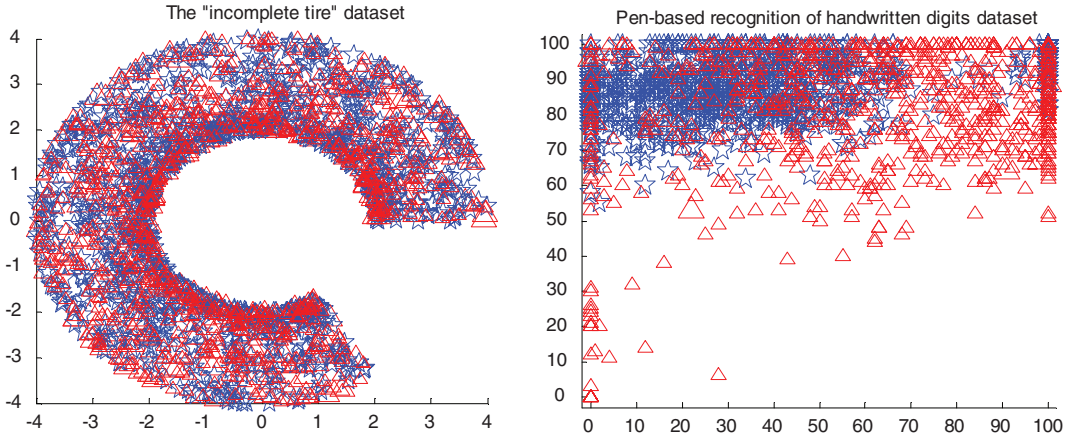


FIGURE 4. Original distributions of the points of the synthetic *"incomplete tire"* and *pen-based recognition of handwritten digits* datasets in the two-dimensional spaces.

input space, the learnt embedding spaces of unsupervised PCA, LPP, NPE, IsoProjection, LE, ViSOM, LLE, MDS, and ISOMAP; supervised FDA and LFDA; and semisupervised SDA and SELF all fail to represent the labeled and unlabeled data points, although intrinsic structure information can be effectively preserved by some of these methods. Also, these methods are all unable to achieve between-class separation. On the contrary, it can be seen from the bottom row of Figure 5 that CCA and SSMDR methods can nicely represent the labeled and unlabeled sample points.

For the *pen-based recognition of handwritten digits* dataset, a quarter of samples are used as labeled and the rest as unlabeled. We observe from the visualization results in Figure 6 that: (1) PCA, LPP, NPE, LFDA, LE, LLE, MDS, and ISOMAP clearly possess within-class multimodality preservation capability and separate one of the clusters of "☆"-class and "△"-class from each other, but data points of the other clusters are seriously mixed; (2) SDA and IsoProjection lose the multimodal structure of the "△"-class and fails to separate points belonging to different classes well; (3) ViSOM tends to mix the partial data of different classes and seems to preserve the intrinsic multimodal structure. However, a small number of points are mapped far apart from the cluster center compared to where they are expected to be; (4) CCA nicely separates the data points of "△"-class and "☆"-class, but multimodal data in each class are projected into a single cluster, which means that multimodal structure is lost; and (5) due to the multimodality, FDA not only fails to separate points, it also fails to find the intrinsic characteristics of the multimodal challenging datasets of this kind.

Based on above results and analysis, we can conclude that SSMDR is a promising method in visualizing the multimodal real data. SSMDR can overcome the difficulties posed by the challenging *"incomplete tire"* dataset. This is an essential property in recognizing the images or objects in the real-world applications.

4.4. Letter Recognition

We address a classification task by employing the real-world *letter recognition* database from the UCI ML repository (Blake and Merz 1998). In this study, we randomly choose 400 images (totally 1,600 examples) from letters "A," "B," "C," and "D" and create a new sample set by merging letters "A" and "C" to a single class (labeled as 1, denoted

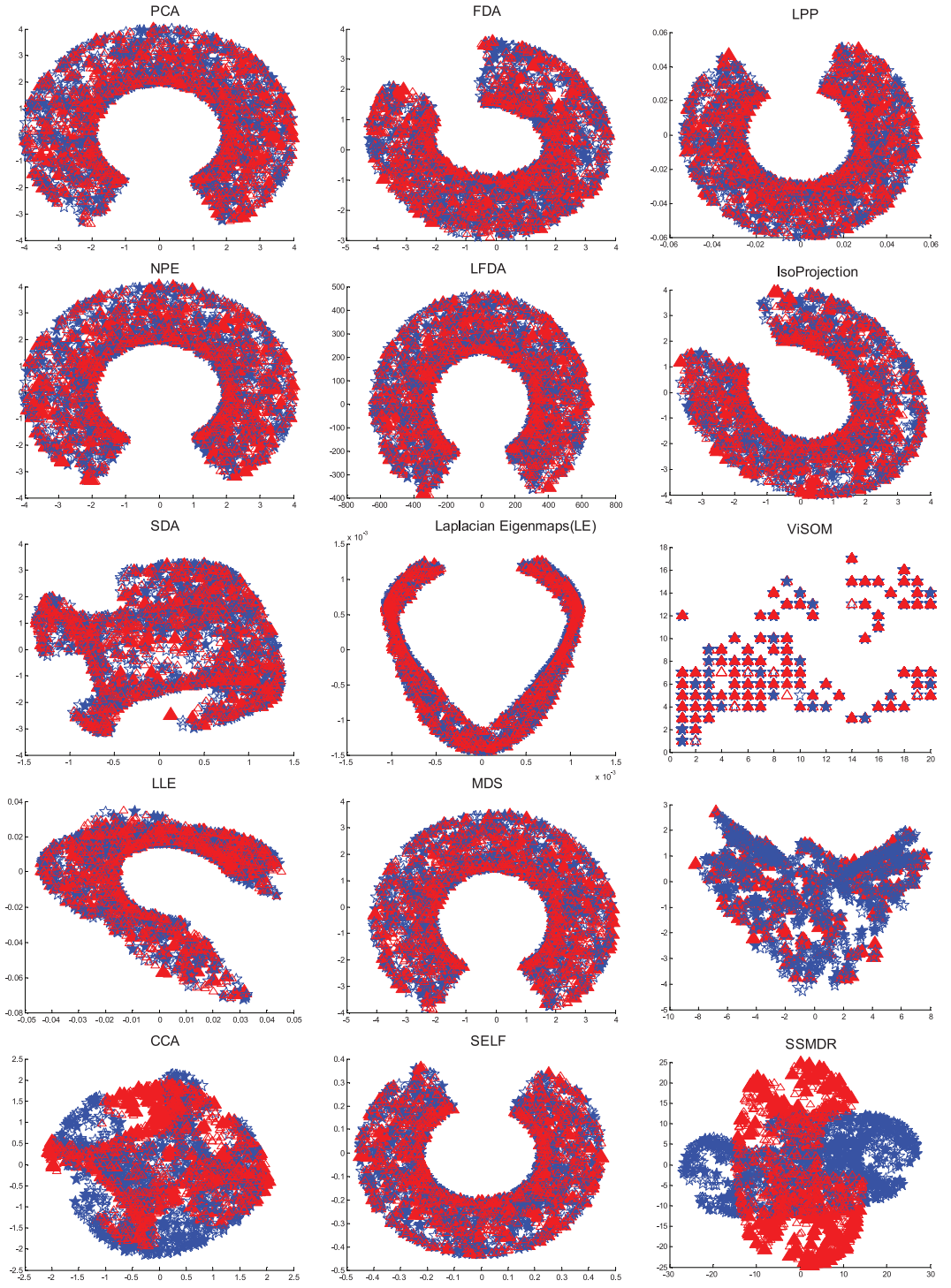


FIGURE 5. The embedding results of samples in the "*incomplete tire*" dataset, where the filled or unfilled five-pointed stars and triangles are the labeled or unlabeled samples.

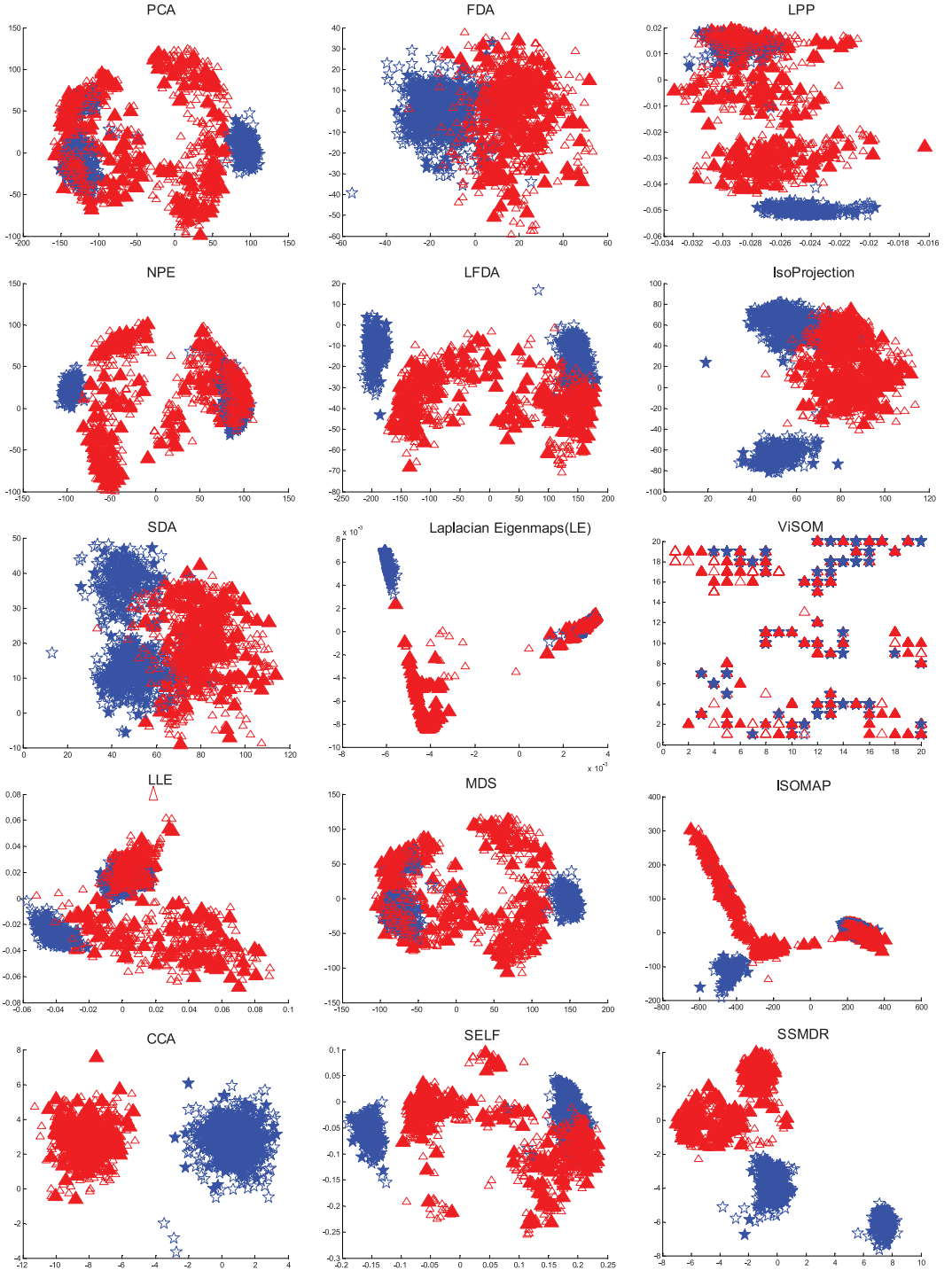
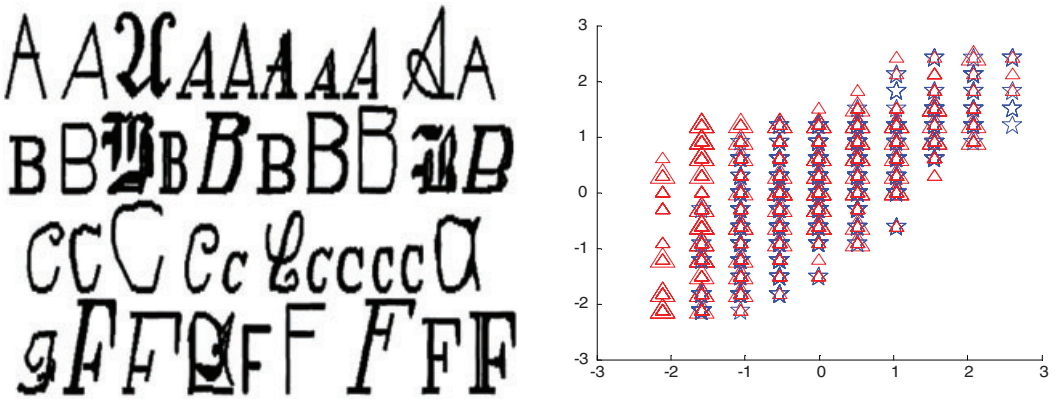


FIGURE 6. The embedding results of samples in the *pen-based recognition of handwritten digits* dataset, where the filled or unfilled five-pointed stars and triangles are the labeled or unlabeled samples.



(a) (b)

FIGURE 7. (a) Examples of character images of the *letter recognition* database. (b) The distribution of the dataset in a two-dimensional space.

by "☆") and letters "B" and "D" to another class (labeled as 2, denoted by "△"). Some preprocessed typical sample images shown in Peter and David (1991) of the letter image recognition database are displayed in Figure 7(a). And the original distribution of the data points is exhibited in Figure 7(b). Intrinsic within-class multimodality appears when samples of different letters are merged into a single class, because the character images are based on 20 different fonts of letters.

We aim at testing PCA, FDA, LPP, LFDA, SELF, SDA, and SSMDR on the dataset. We choose the average classification accuracy over repetitions as the metric. Let the mean accuracy be $\overline{\Theta} = (1/N)\sum_{i=1}^{N}\Theta_i$ and the standard deviation be $std(\Theta_i) = \sqrt{(1/(N-1))\sum_{i=1}^{N}\left(\Theta_i - \bar{\Theta}\right)^2}$, where Θ_i is the classification accuracy rate in the ith repetition and N is the number of repetitions. In all presented results, classification accuracy rates are averaged over 20 runs of different reduced dimensions or labeled numbers. The one-nearest-neighbor classifier is also applied to the original image space as baseline. Different proportions of labeled data are selected relative to the number of total samples. Here, we first fix the numbers of labeled data and vary the reduced dimensions. Four simulation settings with different degrees of supervision are considered. For semisupervised learning, we randomly choose $L\,(= 50, 75, 100, 125)$ images, accompanied with the class labels, per individual to form the labeled set ($4\times L$ in total) and the rest form the unlabeled set. For classification, labeled and unlabeled samples chosen from the labeled and unlabeled sets form the training set and the rest are regarded as test samples. For the unsupervised methods, e.g., PCA, we employ the unlabeled training set to train the learner. For the supervised methods, e.g., FDA and LFDA, only labeled samples are used for training the learner. For the semisupervised methods, we apply the training set, including both labeled and unlabeled data, to train the classifier or learner.

Figure 8 shows the mean classification accuracy rates under different degrees of supervision by a one-nearest-neighbor classifier as a function of the reduced dimensions, where *Dim* is the number of the dimensionality of the original image space, *Lab* is the number of labeled samples, *Unlab* is the number of unlabeled samples, and *Rep* is the number of repetitions. We have the following observations from the results: (1) for each configuration, the performance of FDA and SDA tends to change in a tiny small range as the number of

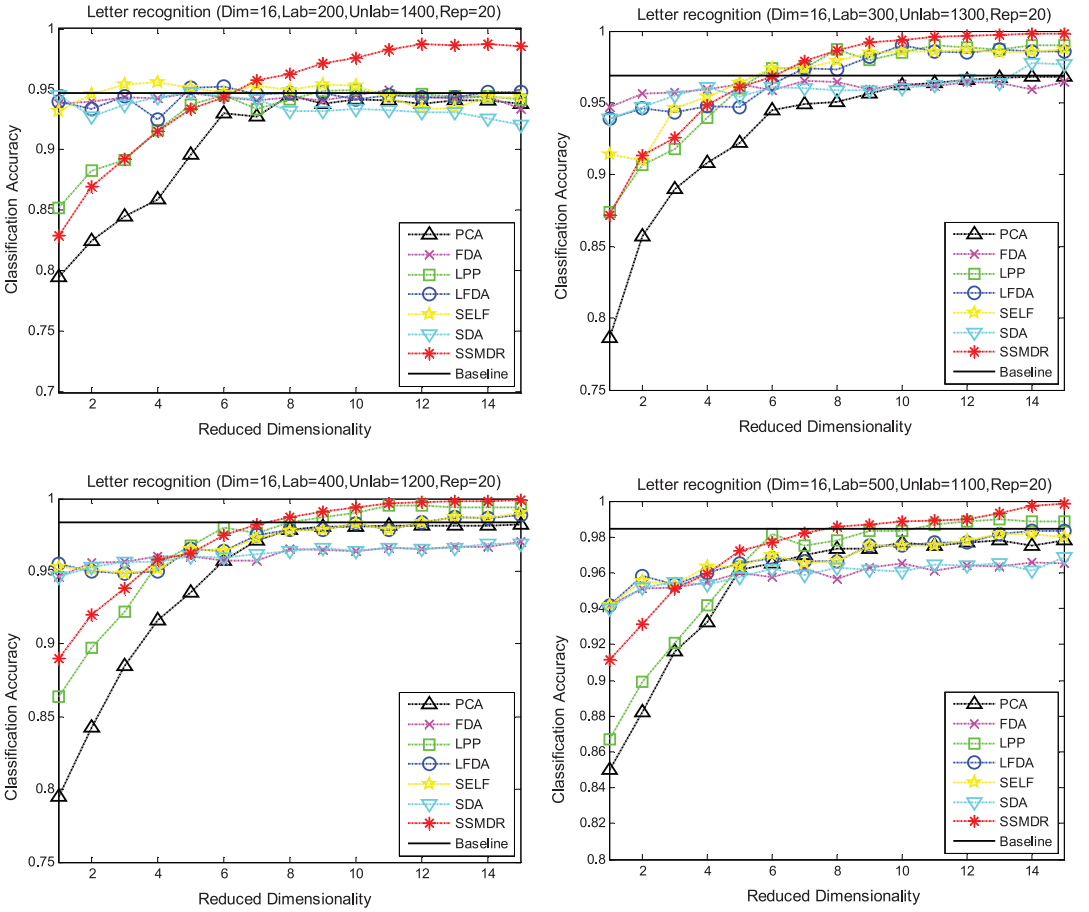


FIGURE 8. Mean classification accuracy rates for the *letter recognition* dataset by a one-nearest-neighbor classifier as a function of the reduced dimensions, where *Dim* is the number of the dimension of the original space, *Lab* is the number of labeled data, *Unlab* is the number of unlabeled data samples, and *Rep* is the number of repetitions.

reduced dimensions increases. This is due to the fact that for FDA and SDA, there are at most $c-1$ nonzero generalized eigenvalues. Thus, an upper bound on the dimensionality of the reduced space is $c-1$ (Martinez and Kak 2001; He et al. 2005), where c is the number of classes. (2) Since the baseline method uses all the dimensionality of the features for classification, the classification accuracies exceed those of FDA and SDA for *Lab*(= 400, 500). (3) PCA, LFDA, and SELF tend to perform well in a complementary way and deliver comparable results to the baseline method for the cases of *Lab*(= 200, 500). (4) SSMDR works well and tends to compensate the weaknesses of the other methods. Its classification accuracy rises steadily and at a faster rate compared with other methods when the number of reduced dimensions increases. And most importantly, when the number of dimensions reaches around five, we can obtain satisfactory results that are consistently superior to those of other methods. (5) Although LPP is able to deliver satisfactory results when L (= 75, 100, 125) images per individual, LPP is unable to deliver favorable result compared with SSMDR. LFDA and SELF deliver close results compared with LPP and SSMDR for

TABLE 1. Means and Standard Deviations of the Classification Accuracy Rates for the *Letter Recognition* Dataset over Different Numbers of Reduced Dimensions and Labeled Data.

Result / Method	Data Name: Letter (*Dim* = 16, *Lab* = 200, *Unlab* = 1,400, *Rep* = 20)				Letter (*Dim* = 16, *Lab* = 300, *Unlab* = 1,300, *Rep* = 20)			
	Mean	Std	Best	Time (s)	Mean	Std	Best	Time (s)
PCA	0.8913	0.0503	0.9450	0.0376	0.9227	0.0554	0.9704	0.0362
FDA	0.9425	0.0038	0.9497	0.1113	0.9606	0.0048	0.9652	0.1010
LPP	0.9281	0.0294	0.9500	0.0974	0.9630	0.0365	0.9900	0.0887
LFDA	0.9437	0.0068	0.9520	0.1077	0.9747	0.0152	0.9890	0.1006
SELF	0.9463	0.0076	0.9560	0.1125	0.9685	0.0201	0.9890	0.1011
SDA	0.9351	0.0078	0.9480	0.1076	0.9605	0.0044	0.9652	0.0978
Our method	0.9453	0.0489	0.9876	0.2981	0.9684	0.0382	0.9995	0.3550
Baseline	0.9420	0.0023	0.9512	0.0432	0.9531	0.0018	0.9754	0.0407
Result / Method	Data Name: Letter (*Dim* = 16, *Lab* = 400, *Unlab* = 1,200, *Rep* = 20)				Letter (*Dim* = 16, *Lab* = 500, *Unlab* = 1,100, *Rep* = 20)			
	Mean	Std	Best	Time (s)	Mean	Std	Best	Time (s)
PCA	0.9352	0.0613	0.9817	0.0334	0.9307	0.0397	0.9783	0.0378
FDA	0.9615	0.0061	0.9704	0.0946	0.9589	0.0067	0.9660	0.1042
LPP	0.9664	0.0403	0.9950	0.0848	0.9604	0.0380	0.9900	0.0923
LFDA	0.9715	0.0147	0.9887	0.0979	0.9688	0.0119	0.9833	0.1062
SELF	0.9713	0.0142	0.9900	0.0982	0.9683	0.0117	0.9817	0.1059
SDA	0.9615	0.0064	0.9690	0.0950	0.9591	0.0031	0.9687	0.1005
Our method	0.9724	0.0332	0.9998	0.3450	0.9743	0.0255	0.9984	0.3568
Baseline	0.9756	0.0020	0.9850	0.0383	0.9745	0.0014	0.9867	0.0463

Lab (= 300, 400). (6) When the number of labeled data in each class increases, the classification performance of almost all methods increases accordingly.

Table 1 presents an overall view of the means and standard deviations (Std) of the classification accuracy for the *letter recognition* dataset over the reduced dimensions. The best result and averaged running time (in seconds) of each method are also described in Table 1. For image classification and recognition, let $\theta_i = [\wp_{i1}, \wp_{i2}, \ldots, \wp_{in}]$ and $\theta_j = [\wp_{j1}, \wp_{j2}, \ldots, \wp_{jn}]$ represent two feature matrices, then the distance metric between θ_i and θ_j can be described as $Dist(\theta_i, \theta_j) = \sum_{t=1}^{n} \|\theta_{it} - \theta_{jt}\|^2$. Suppose matrices θ_1 and θ_2 are used for storing the two-class data samples labeled by 1 and 2, for any new image data θ, if $Dist(\theta, \theta_1) = \min Dist(\theta, \theta_j)$and θ_1 belongs to class 2, then θ is classified as class 2, and class 1 otherwise. The recognition accuracy, obtained by a one-nearest-neighbor classifier, is used as the measurement standards. From Table 1, it can be observed that: (1) in all cases, SSMDR consistently delivers the highest classification accuracy, e.g., the best result for the case of *Lab* (= 300) is 0.9876, and 0.9995, 0.9998, and 0.9984 for the other configurations. The subspace according to the best obtained result over repetitions is called the optimal image subspace. (2) For runtime performance, linear PCA, FDA, and LPP are

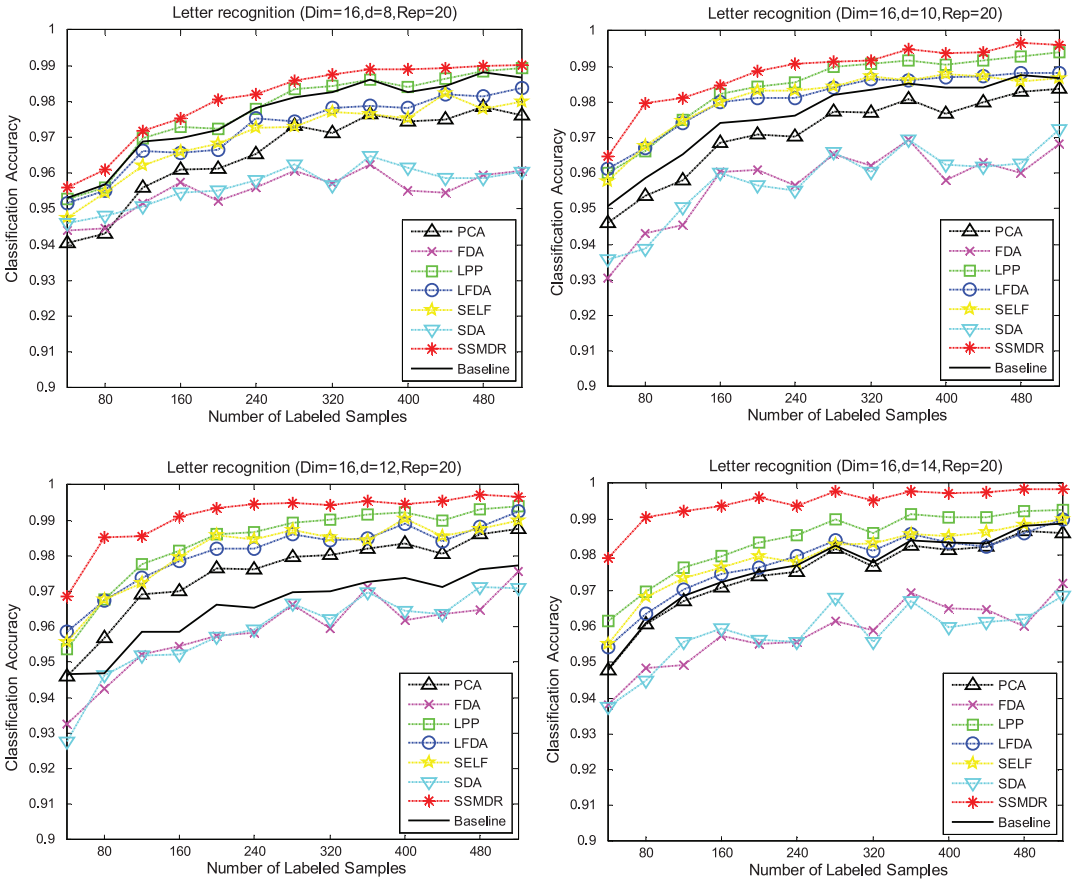


FIGURE 9. Mean classification accuracy rates for the *letter image recognition* dataset over different numbers of reduced dimensions and labeled data, where d is the number of reduced dimension.

computationally efficient. SSMDR delivers similar results compared with LFDA, SELF, and SDA. (3) As shown in Figure 8, the classification accuracies of PCA, LPP, and our proposed SSMDR rise from a relatively low level to a higher level. It is noticed that they all share the similar trends and standard deviations.

To investigate how the labeled number affects the classification accuracy, we consider four configurations with different reduced dimensions. We conduct the simulations on the *letter image recognition* dataset by varying the labeled numbers. For semisupervised learning, we randomly choose $L(= 10, 20, \ldots, 130)$ images per individual as labeled ($4\times L$ in total). Figure 9 depicts the mean classification accuracy rates by a one-nearest-neighbor classifier, from which SSMDR delivers the highest accuracies in all cases. The classification accuracies of all the methods increase gradually when the number of labeled data and reduced dimensions increases. Note that though PCA and LPP are unsupervised methods, better results are obtained, because high input dimensionality helps improving the learning performance of the classifier. LPP also works well under each configuration. Semisupervised SELF tends to deliver comparable results to LPP and SSMDR when the number of labeled data increases. Baseline still works well and delivers comparable or even better performance compared with PCA, FDA, LFDA, and SDA.

TABLE 2. Means and Standard Deviations of the Classification Accuracy Rates for the *Letter Recognition* Dataset with Different Numbers of Reduced Dimensions and Labeled Data.

Result / Method	Dataset Name: Letter ($Dim = 16$, $d = 8$, $Rep = 20$)				Letter ($Dim = 16$, $d = 10$, $Rep = 20$)			
	Mean	Std	Best	Time (s)	Mean	Std	Best	Time (s)
PCA	0.9652	0.0019	0.9785	0.0620	0.9713	0.0020	0.9837	0.0639
FDA	0.9551	0.0022	0.9625	0.6711	0.9572	0.0028	0.9695	0.6628
LPP	0.9772	0.0026	0.9892	0.3322	0.9842	0.0023	0.9940	0.3378
LFDA	0.9721	0.0023	0.9838	0.6837	0.9809	0.0015	0.9882	0.6696
SELF	0.9702	0.0024	0.9824	0.6740	0.9809	0.0084	0.9880	0.6929
SDA	0.9566	0.0028	0.9647	0.6350	0.9579	0.0031	0.9725	0.6516
Our method	0.9783	0.0029	0.9899	0.3919	0.9870	0.0024	0.9937	0.4139
Baseline	0.9762	0.0026	0.9882	0.0725	0.9764	0.0017	0.9877	0.0680
Result / Method	Dataset Name: Letter ($Dim = 16$, $d = 12$, $Rep = 20$)				Letter ($Dim = 16$, $d = 14$, $Rep = 20$)			
	Mean	Std	Best	Time (s)	Mean	Std	Best	Time (s)
PCA	0.9750	0.0014	0.9875	0.0628	0.9749	0.0016	0.9868	0.0625
FDA	0.9585	0.0029	0.9755	0.6582	0.9582	0.0019	0.9720	0.6612
LPP	0.9841	0.0017	0.9940	0.3393	0.9838	0.0016	0.9925	0.3488
LFDA	0.9809	0.0012	0.9925	0.6858	0.9779	0.0016	0.9900	0.6824
SELF	0.9812	0.0102	0.9905	0.6671	0.9795	0.0019	0.9900	0.6840
SDA	0.9587	0.0030	0.9712	0.6597	0.9579	0.0020	0.9685	0.6673
Our method	0.9914	0.0026	0.9972	0.4048	0.9945	0.0022	0.9985	0.4021
Baseline	0.9767	0.0012	0.9892	0.0705	0.9762	0.0016	0.9887	0.0713

We detail the means and standard deviations of the classification accuracy rates for the *letter recognition* dataset over different numbers of reduced dimensions and labeled data in Table 2. The best result and mean running time are also listed. Table 2 shows that SSMDR is able to deliver the highest mean accuracy and the best test result. And the performance of SSMDR is stable when the number of labeled data increases. This leads to the small errors and standard deviations that are comparable to those of LPP, SDA, and SELF. Comparing with SSMDR, other semisupervised methods deliver lower classification accuracies. Most importantly, SSMDR can deliver satisfactory results even when only a small number of labeled data are given, e.g., the accuracy is about 98% with only $L (= 10)$ labeled data in each individual. For the fixed reduced dimensions, SSMDR needs less running time compared with LFDA, SELF, and SDA when the amount of labeled data increases.

4.5. Handwritten Digital Recognition

In this subsection, we apply PCA, FDA, LPP, LFDA, SELF, SDA, and SSMDR methods to the benchmark real-world *optical recognition of handwritten digits* database from the UCI ML repository (Blake and Merz 1998) for classification and performance evaluation. In this study, we randomly choose 400 examples (1,800 examples in total) from digits "0," "1," "2," "3," "4," and "5," where even digits "0," "2," and "4" are merged to a single class (labeled

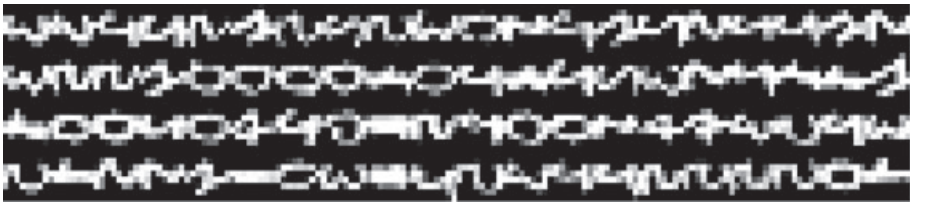


FIGURE 10. Some digit sample images of the *optical handwritten digits* database.

as 1) and odd digits "1," "3," and "5" for another class (labeled as 2), and create our sample set. Our objective is to classify the odd digits from the even digits. Figure 10 shows typical sample images of the *optical handwritten digits* database. Because data of different digits are mixed into several separate points in the original two-dimensional space, illustration for the original distribution is omitted here. Within-class multimodality appears because samples of different digits are captured from the different writing styles.

In our simulations, we first fix the labeled numbers as the optimal value and vary the amount of the reduced dimensions for performance comparison. For semisupervised learning, a random subset with L (= 20, 40, 60, 80) samples, accompanied with the class labels, per digit is selected as the labeled set ($6\times L$ in total) and the rest are regarded as unlabeled samples. The one-nearest-neighbor classifier is again used for performing classification and the classification accuracy rates are averaged over 20 runs.

Figure 11 shows the mean classification accuracy under different degrees of supervision. From the obtained results, we have the following observations: (1) For each configuration, SSMDR can almost always achieve the highest accuracy as the amount of labeled data points increases and the accuracy increases steadily till reach the best record and then keep small fluctuation. (2) PCA performs well, despite the fact that it is an unsupervised method, because the projection to the two-dimensional PCA subspace can give reasonably separate embedding. (3) LPP also works well and the performance increases slightly faster than those of LFDA, SELF, and SSMDR methods. However, SSMDR tends to outperform the other methods, including unsupervised PCA and LPP and semisupervised SDA and SELF, when the number of reduced dimension, d, increases to about 15 or higher in all cases. (4) SELF works poorly when the number of reduced dimension is relatively small and delivers the similar result compared to LFDA and FDA. (5) SDA performs better than FDA, but the performance of FDA and SDA is still constrained by the upper bound of the reduced space dimension. (6) Due to the fact that the baseline method uses all the dimension of the features, the baseline method achieves comparable result to PCA, LPP, and our method, and its performance even outperforms some other several methods.

The means and standard deviations of the classification accuracy rates for the *optical handwritten digits* dataset over the reduced dimensions are given in Table 3, in which the best test result and averaged running time (in seconds) of each method are listed. From Table 3, we can find that: (1) comparing to the other methods, SSMDR can achieve the highest classification accuracy in all cases, e.g., the best record of the case of Lab (=120) for SSMDR is 0.9904, 0.9767 for PCA, 0.8975 for FDA, 0.9635 for LPP, 0.9433 for LFDA, 0.8592 for SELF, 0.9164 for SDA, and 0.9664 for baseline. The best results of PCA, SSMDR, and the baseline method are superior to other studied methods. In addition, it is noticed that the best results of PCA and SSMDR outperform those obtained by the baseline method. The subspace according to the best record is regarded as the optimal image space. (2) From the perspective of computational time, linear PCA and FDA are still the most computationally

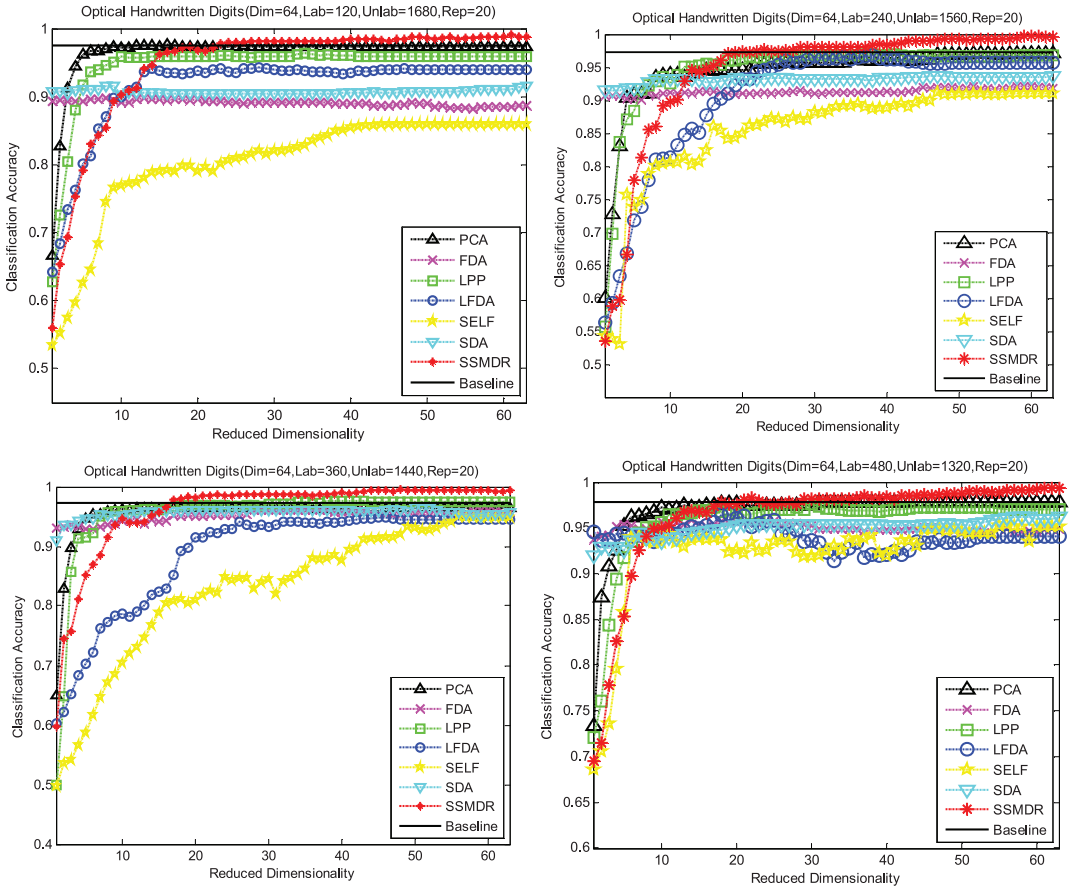


FIGURE 11. Mean classification accuracy rates for the *optical handwritten digits* dataset.

efficient among all cases. It is noted that our SSMDR is slightly slower than LPP, LFDA, SELF, and SDA for $Lab\,(=120)$, but it is worth noting that the SSMDR runtime performance tends to be close to or even better than those of other methods for $Lab\,(=360, 480)$.

Next, we aim at testing the semisupervised SELF, SDA, and SSMDR approaches using six configurations with different labeled numbers by fixing the number of reduced dimensions first. For semisupervised learning, we randomly choose $L(=5, 10, 15, \ldots, 100)$ examples per individual as labeled. Figure 12 depicts the chart of the mean classification accuracy rates and standard deviations, from which the following observations are found. First, the accuracies of SSMDR exceed those of SDA and SELF in all cases and keep steady, resulting in small standard deviations. This contributes to improving the performance of recognition systems. When the number of reduced dimensions increases, the trends of SSMDR become much steadier. Most importantly, SSMDR shows that it is able to deliver satisfactory results despite being given with small number of labeled data, e.g., the accuracy of SSMDR is around 92% for the case of $d(=20)$. Other semisupervised methods, however, fail to reach this level. Second, the accuracy rate of SDA is in the medium level lying between SELF and our proposed SSMDR. It is also noticed that SDA in general is more stable than SELF. Third, large errors frequently occur in SELF, causing large errors and standard deviations, which result in increasing the likelihood of unstable classification.

TABLE 3. Means and Standard Deviations of the Classification Accuracy Rates for the *Optical Handwritten Digits* Dataset over Different Numbers of Reduced Dimensions and Labeled Data.

Result / Method	Dataset Name: Digits (*Dim* = 16, *Lab* = 120, *Unlab* = 1,680, *Rep* = 20)				Digits (*Dim* = 16, *Lab* = 240, *Unlab* = 1,560, *Rep* = 20)			
	Mean	Std	Best	Time (s)	Mean	Std	Best	Time (s)
PCA	0.9643	0.0430	0.9686	0.0685	0.9588	0.0220	0.9707	0.0621
FDA	0.8908	0.0040	0.8975	0.0708	0.9132	0.0047	0.9208	0.2468
LPP	0.9449	0.0545	0.9635	0.0627	0.9462	0.0643	0.9689	0.2005
LFDA	0.9150	0.0624	0.9433	0.0718	0.9071	0.0972	0.9676	0.2551
SELF	0.8030	0.0798	0.8592	0.0725	0.8536	0.0868	0.9121	0.2541
SDA	0.9068	0.0033	0.9164	0.0706	0.9316	0.0054	0.9364	0.2461
Our method	0.9458	0.0864	0.9904	0.3839	0.9427	0.1015	0.9977	0.5551
Baseline	0.9664	0.0025	0.9728	0.0365	0.9614	0.0034	0.9762	0.0764
Result / Method	Dataset Name: Digits (*Dim* = 16, *Lab* = 360, *Unlab* = 1,440, *Rep* = 20)				Digits (*Dim* = 16, *Lab* = 480, *Unlab* = 1,320, *Rep* = 20)			
	Mean	Std	Best	Time (s)	Mean	Std	Best	Time (s)
PCA	0.9540	0.0434	0.9644	0.3451	0.9703	0.0344	0.9804	0.5119
FDA	0.9511	0.0092	0.9611	0.5581	0.9499	0.0045	0.9583	0.9619
LPP	0.9534	0.0732	0.9761	0.3754	0.9566	0.0443	0.9757	0.5681
LFDA	0.8878	0.0921	0.9481	0.5746	0.9396	0.0114	0.9655	0.9953
SELF	0.8289	0.1165	0.9500	0.5734	0.9224	0.0532	0.9536	0.9920
SDA	0.9564	0.0083	0.9667	0.5565	0.9500	0.0094	0.9619	0.9648
Our method	0.9604	0.0709	0.9954	0.5427	0.9619	0.0606	0.9940	0.4973
Baseline	0.9735	0.0031	0.9813	0.1035	0.9781	0.0029	0.9855	0.1181

Table 4 summarizes the means and standard deviations of the classification accuracy rates for the *handwritten digits* dataset. The best record and mean running time are also provided. As listed in Table 4, SSMDR method tends to outperform SDA and SELF in terms of classification accuracy and system stability. And SSMDR is found to be competitive in runtime performance. SELF delivers comparable result to SDA when the number of labeled data increases. Due to the fact that SELF tends to work relatively unsteady, its standard deviation is relatively larger than those of the SDA and SSMDR.

4.6. Columbia Object Image Library (COIL-20)

A classification example using the popular *Columbia object image library (COIL-20)* image database is studied. The database contains a total of 1,440 with black background for 20 different subjects. Figure 13(a) shows typical examples for the 20 subjects of the *COIL-20* database. We apply all the sample images and create two-class problem by merging the 1st, 3rd, . . . , 19th subjects to a single class (labeled as 1, denoted by "☆") and the 2nd, 4th, . . . ,

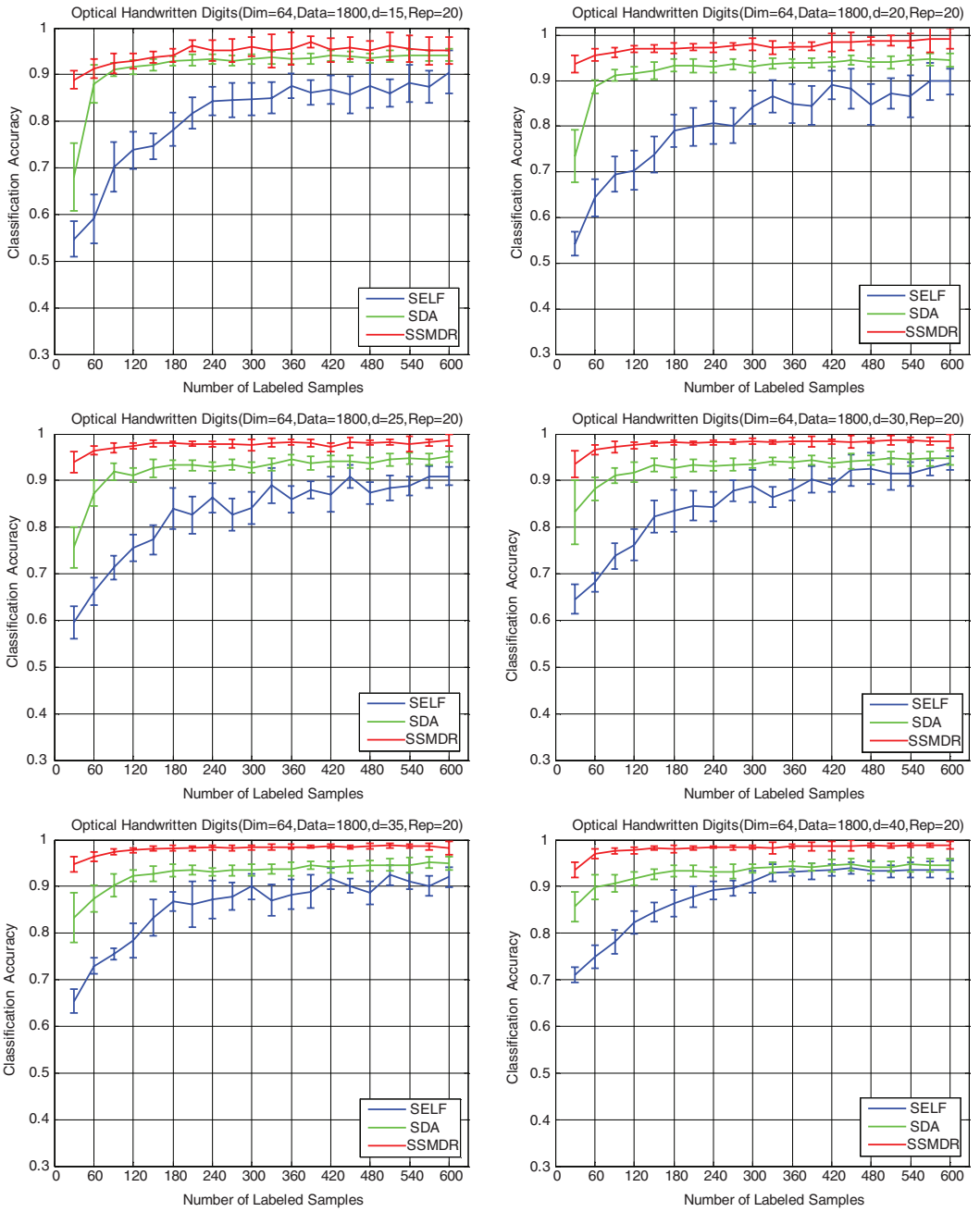


FIGURE 12. Mean classification accuracy rates and standard deviations for the *optical handwritten digits* dataset.

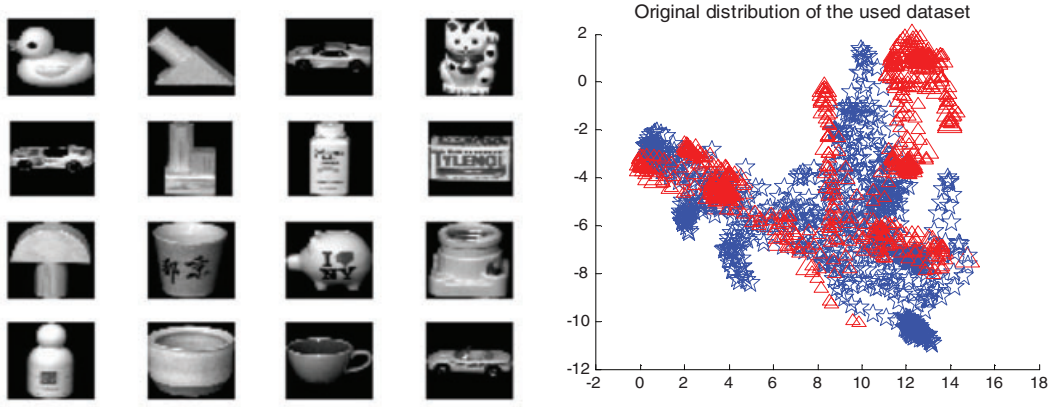
20th subject to another class (labeled as 2, denoted by "△"). The original distribution of the dataset is exhibited in Figure 13(b). The dataset is also regarded as multimodal when images of different subjects are merged into a single class, because different objects are captured under different degrees.

TABLE 4. Means and Standard Deviations of the Classification Accuracy Rates for the *Optical Handwritten Digits* Dataset over Different Numbers of Reduced Dimensions and Labeled Data.

Result / Method	Dataset Name: Digits ($Dim = 200$, $Data = 1{,}800$, $d = 15$, $Rep = 20$)				Digits ($Dim = 200$, $Data = 1{,}800$, $d = 20$, $Rep = 20$)			
	Mean	Std	Best	Time (s)	Mean	Std	Best	Time (s)
SELF	0.8087	0.0977	0.9058	0.5578	0.8038	0.0950	0.8986	0.5458
SDA	0.9177	0.0577	0.9421	0.5463	0.9224	0.0465	0.9461	0.5355
Our method	0.9460	0.0193	0.9700	0.5806	0.9751	0.0132	0.9913	0.6273

Result / Method	Dataset Name: Digits ($Dim = 200$, $Data = 1{,}800$, $d = 25$, $Rep = 20$)				Digits ($Dim = 200$, $Data = 1{,}800$, $d = 30$, $Rep = 20$)			
	Mean	Std	Best	Time (s)	Mean	Std	Best	Time (s)
SELF	0.8285	0.0867	0.9097	0.6033	0.8513	0.0833	0.9373	0.6176
SDA	0.9236	0.0428	0.9509	0.6019	0.9288	0.0274	0.9486	0.6168
Our method	0.9767	0.0099	0.9865	0.6893	0.9795	0.0116	0.9875	0.7026

Result / Method	Dataset Name: Digits ($Dim = 200$, $Data = 1{,}800$, $d = 35$, $Rep = 20$)				Digits ($Dim = 200$, $Data = 1{,}800$, $d = 40$, $Rep = 20$)			
	Mean	Std	Best	Time (s)	Mean	Std	Best	Time (s)
SELF	0.8572	0.0719	0.9247	0.6178	0.8845	0.0691	0.9396	0.6049
SDA	0.9293	0.0288	0.9515	0.6190	0.9308	0.0220	0.9474	0.6060
Our method	0.9809	0.0095	0.9878	0.6749	0.9816	0.0117	0.9890	0.6890



(a) (b)

FIGURE 13. (a) Sample examples of the 20 subjects in the *COIL-20* database. (b) The original distribution of the used dataset in a two-dimensional space.

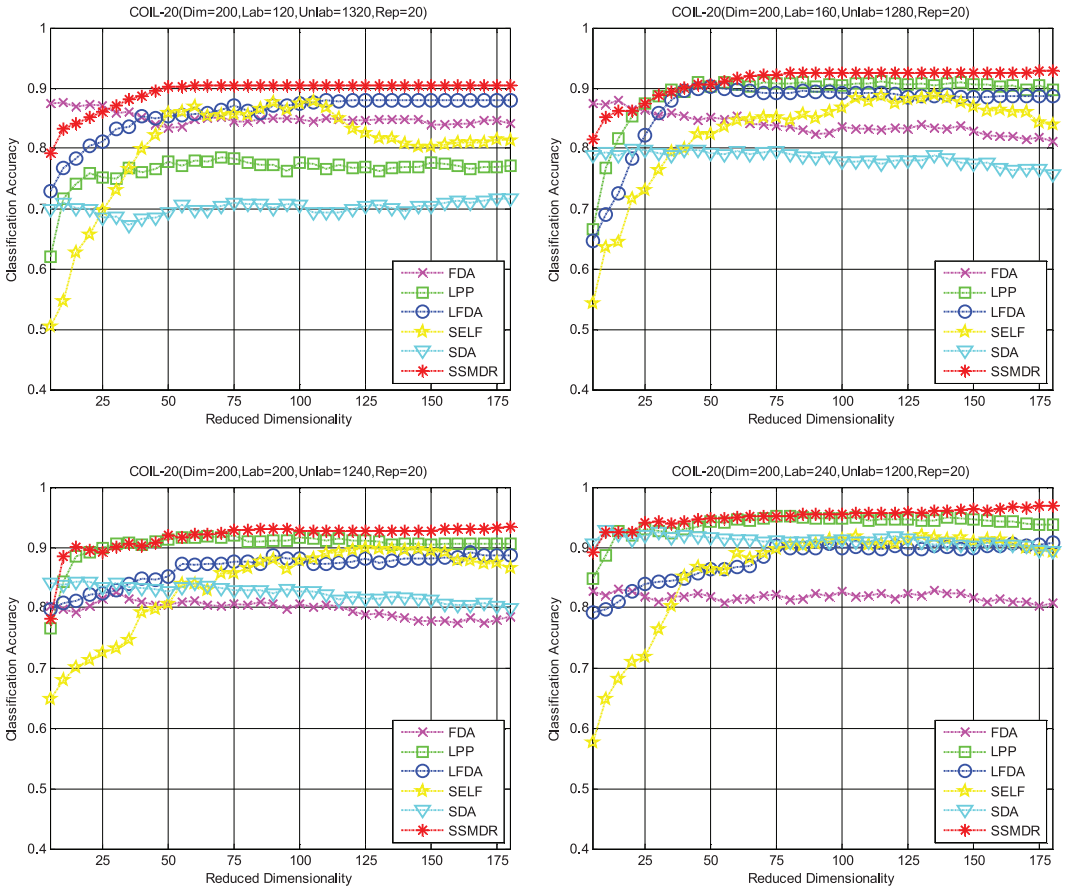


FIGURE 14. Mean classification accuracy rates for the *COIL-20* database by a one-nearest-neighbor classifier.

In this dataset, simulations are conducted in the similar way that we first fix the labeled samples and vary the number of reduced dimensions. For semisupervised learning, we randomly choose $L(= 6, 8, 10, 12)$ images per individual as labeled ($20 \times L$ in total) and the remaining as unlabeled. Prior to this study, we use PCA to preprocess the dataset by reducing the dimensionality of the original space to 200 for comparisons. We compare FDA, LPP, LFDA, SELF, and SDA with our SSMDR method. Figure 14 shows the mean classification accuracy over different degrees of supervision. From the results, it can be noticed that SSMDR delivers the best learning performance in most cases when the number of labeled data increases. The accuracy increases gradually from a relatively low level to a higher one, and the accuracy tends to stabilize around $d(= 50)$ in all cases. For the case of $Lab\,(= 120)$, LPP is unable to deliver an acceptable result, but its accuracy increases rapidly when the number of labeled data increases to high level. This is believed to be attributed to the classifier used. It is interesting to see that LPP even delivers comparable result to SSMDR. When the number of labeled data increases, SDA and SELF tend to deliver better results. The accuracy of SDA increases in a faster rate compared to SELF in most cases. As a regularized variant of FDA, SDA outperforms FDA when the number of labeled data increases to a high level. Also, as a regularized variant of LFDA, SELF works well in a complementary way

TABLE 5. Means and Standard Deviations of the Classification Accuracy Rates for the *COIL-20* Dataset over Different Numbers of Reduced Dimensions and Labeled Data.

Result / Method	Dataset Name: COIL (*Dim* = 200, *Lab* = 120, *Unlab* = 1,320, *Rep* = 20)				Dataset Name: COIL (*Dim* = 200, *Lab* = 160, *Unlab* = 1,280, *Rep* = 20)			
	Mean	Std	Best	Time (s)	Mean	Std	Best	Time (s)
FDA	0.8501	0.0108	0.8758	1.0757	0.8397	0.0172	0.8804	1.0957
LPP	0.7634	0.0275	0.7858	1.0754	0.8893	0.0476	0.9107	1.0909
LFDA	0.8563	0.0355	0.8800	1.0738	0.8679	0.0603	0.9032	1.0967
SELF	0.8000	0.0894	0.8775	1.1102	0.8259	0.0791	0.8893	1.1134
SDA	0.7009	0.0099	0.7167	1.0778	0.7840	0.0111	0.7991	1.0949
Our method	0.8921	0.0256	0.9038	1.1922	0.9111	0.0268	0.9295	1.1978

Result / Method	Dataset Name: COIL (*Dim* = 200, *Lab* = 200, *Unlab* = 1,240, *Rep* = 20)				Dataset Name: COIL (*Dim* = 200, *Lab* = 240, *Unlab* = 1,200, *Rep* = 20)			
	Mean	Std	Best	Time (s)	Mean	Std	Best	Time (s)
FDA	0.7975	0.0130	0.8250	1.1170	0.8181	0.0066	0.8302	1.1110
LPP	0.9030	0.0268	0.9200	1.1020	0.9380	0.0197	0.9510	1.0895
LFDA	0.8647	0.0257	0.8913	1.1216	0.8789	0.0334	0.9083	1.1179
SELF	0.8370	0.0719	0.9000	1.1306	0.8609	0.0873	0.9198	1.1215
SDA	0.8243	0.0128	0.8433	1.1189	0.9115	0.0075	0.9292	1.1113
Our method	0.9172	0.0267	0.9347	1.1907	0.9514	0.0152	0.9693	1.1567

and appears to be easily affected by the labeled numbers. SELF tends to achieve comparable result to LFDA.

According to the above illustrations on the COIL-20 database over different numbers of reduced dimensions, the means and standard deviations of the classification accuracy are shown in Table 5. The best test record and averaged running time are also detailed. It can be seen that SSMDR consistently delivers the best results in all cases. SSMDR also delivers the comparable or even smaller standard deviations than LFDA and SELF. The runtime performance of SSMDR is comparable to that of the semisupervised SELF method. It can also be observed that LPP delivers comparable accuracy and standard deviations to LFDA and SSMDR. SSMDR has a comparable running time to the FDA and LFDA methods in all the configurations.

Here, we prepare a classification to test the semisupervised SELF, SDA, and SSMDR approaches using six configurations over different amounts of labeled data by first setting the number of reduced dimensions to a fixed value. We randomly select $L(= 1, 2, \ldots, 15)$ examples per individual as labeled data ($20 \times L$ in total). Figure 15 illustrates the mean classification accuracy rates and standard deviations, from which we can observe that: (1) SSMDR almost always delivers the highest accuracy in all cases. It is important to note that SSMDR can deliver satisfactory results with small labeled numbers, e.g., its accuracy reaches around 88% for the case of $d(= 30)$. The other two semisupervised SDA and SELF methods, however, deliver relatively lower accuracies. Moreover, the accuracy of SSMDR

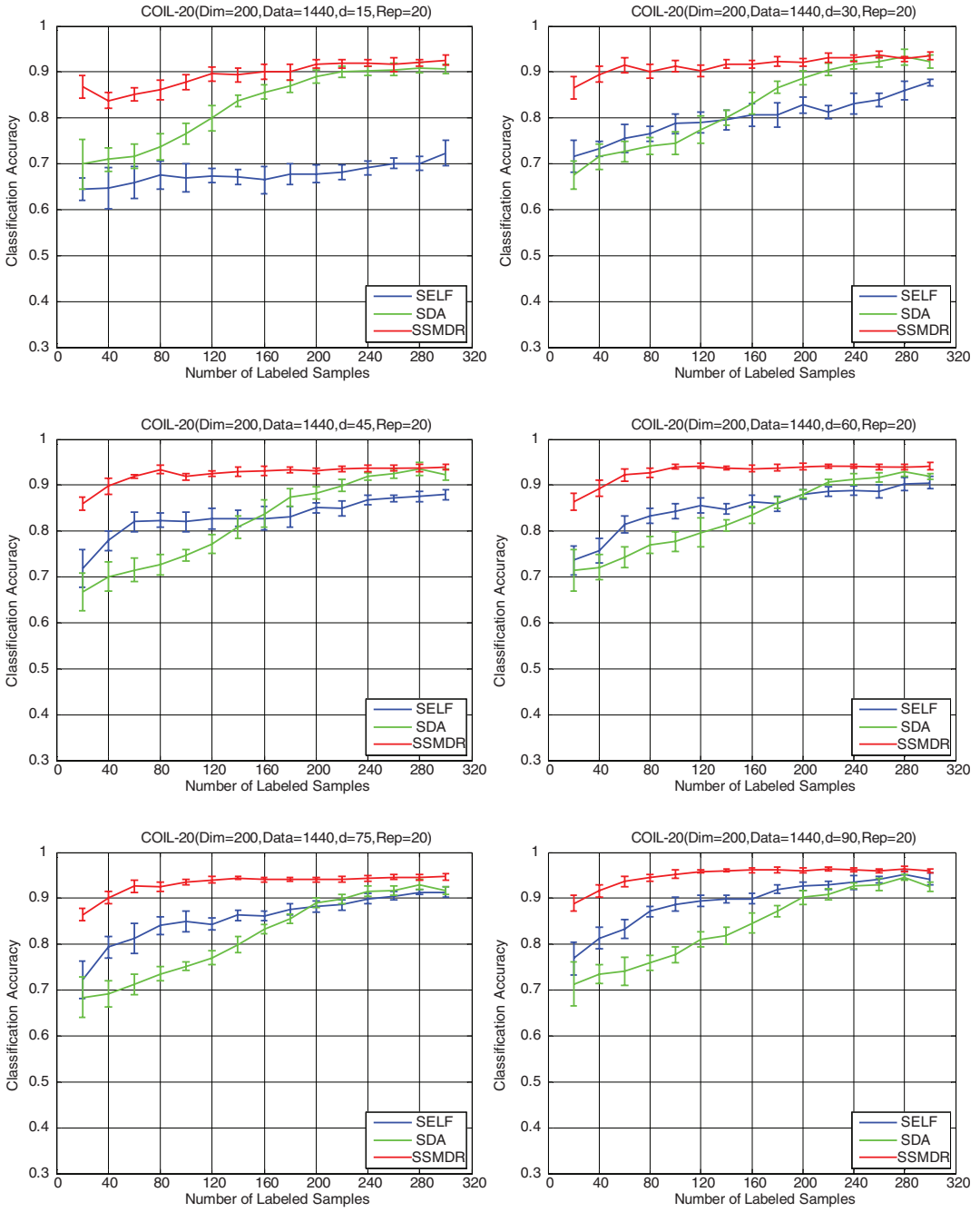


FIGURE 15. Mean classification accuracy rates and standard deviations for the *COIL-20* database over different numbers of reduced dimensions and labeled data.

keeps steady when the number of labeled data increases. This results in small standard deviations, and contributes to a stable recognition system. The accuracy of SELF increases at a faster rate than SDA when the number of labeled samples and reduced dimensionality increases. From our obtained results, the SELF method exhibits a comparable or even better

TABLE 6. Means and Standard Deviations of the Classification Accuracy Rates for the *COIL-20* Dataset over Different Numbers of Reduced Dimensions and Labeled Data.

Result / Method	Dataset Name							
	COIL ($Dim = 200$, $Data = 1{,}440$, $d = 15$, $Rep = 20$)				COIL ($Dim = 200$, $Data = 1{,}440$, $d = 30$, $Rep = 20$)			
	Mean	Std	Best	Time (s)	Mean	Std	Best	Time (s)
SELF	0.6773	0.0208	0.7235	1.3988	0.8003	0.0449	0.8779	1.3868
SDA	0.8270	0.0810	0.9099	1.3719	0.8241	0.0894	0.9325	1.3577
Our method	0.8938	0.0282	0.9263	1.7868	0.9160	0.0188	0.9383	1.6477
Result / Method	Dataset Name							
	COIL ($Dim = 200$, $Data = 1{,}440$, $d = 45$, $Rep = 20$)				COIL ($Dim = 200$, $Data = 1{,}440$, $d = 60$, $Rep = 20$)			
	Mean	Std	Best	Time (s)	Mean	Std	Best	Time (s)
SELF	0.8313	0.0412	0.8795	1.4134	0.8509	0.0494	0.9051	1.4333
SDA	0.8223	0.0938	0.9350	1.3831	0.8334	0.0765	0.9293	1.4024
Our method	0.9245	0.0208	0.9395	1.5913	0.9298	0.0220	0.9425	1.5228
Result / Method	Dataset Name							
	COIL ($Dim = 200$, $Data = 1{,}440$, $d = 75$, $Rep = 20$)				COIL ($Dim = 200$, $Data = 1{,}440$, $d = 90$, $Rep = 20$)			
	Mean	Std	Best	Time (s)	Mean	Std	Best	Time (s)
SELF	0.8594	0.0467	0.9139	1.4155	0.8941	0.0529	0.9510	1.4517
SDA	0.8204	0.0899	0.9303	1.3864	0.8395	0.0832	0.9451	1.4197
Our method	0.9325	0.0223	0.9473	1.6859	0.9503	0.0214	0.9641	1.7570

performance than that of SDA when the number of reduced dimensions increases to 60 or more. At last, it can be seen from the trends of the semisupervised SDA, its accuracy increases from relatively low levels to the high ones.

The means and standard deviations of classification accuracy rates for the *COIL-20* dataset are summarized in Table 6, in which the best result and mean running time are also listed. It can be observed that SSMDR outperforms the semisupervised SDA and SELF in terms of classification accuracy, best test record, and system stability. The running time performance of SSMDR is comparable to the other methods. As the number of reduced dimensions and labeled samples increases, SELF tends to perform better than SDA for the *COIL-20* dataset. When the number of reduced dimensions increases, the mean accuracy of SELF increases gradually from 0.6773 to 0.8003 to 0.8313 to 0.8509 to 0.8594, and finally reaches the highest record of 0.8941. The corresponding figures for SDA are 0.8270, 0.8241, 0.8223, 0.8334, 0.8204, and 0.8395. From our obtained results, the standard deviations of SDA are larger than those of the semisupervised SELF and SSMDR methods.

Based on the above results, we can conclude that SSMDR is a promising dimensionality reduction and feature extraction technique for high-dimensional image data representation.

This study shows that SSMDR is a useful technique for classifying real-world image databases.

5. CONCLUSIONS

It is widely known that most existing supervised, unsupervised, or semisupervised dimensionality reduction techniques are able to preserve the global or local structure characteristics of given data, but they are usually inapt to preserve the intrinsic multimodal structures. However, dealing with data in the class of multimodal is often required in most real-world applications. In this paper, we incorporate the local manifold information into the within- and between-set scatter matrices. Considering that labeled data are relatively expensive to obtain, to mine useful information from unlabeled data samples, terms based on the PCA criterion are defined for preserving the global covariance structures of labeled and unlabeled data. We then propose a novel semisupervised multimodal dimensionality reduction algorithm, namely, SSMDR, for efficient feature representation and extraction. SSMDR is naturally formulated on the matrix interpretations of the defined localized scatters and aims at computing a pair of optimal projection transformations for two sets of variables. As a result, SSMDR can keep within-set data pairs compact and between-set sample pairs apart. Also, the projections of interset points can be effectively separated together without losing the local and multimodal information. By defining reasonable criteria, we show that SSMDR can keep data pairs in the close vicinity of the original input space nearby in the embedding space. We also show that SSMDR can deliver excellent performance when the number of labeled data is relatively smaller than other semisupervised methods. This is an important characteristic because the amount of labeled data samples is usually small in semisupervised case. Another major advantage of SSMDR is that the embedding transformations can be obtained analytically by solving two generalized eigenvalue problems.

In this paper, we focus on linear dimensionality reduction and show that a kernelized SSMDR can be obtained by the kernel trick. However, the performance of the kernelized SSMDR heavily depends on the choice of the kernel function and its kernel parameters. Thus, how to choose the best possible kernel function and parameter still needs to be investigated. At last, SSMDR is theoretically derived for handling two-class problems, but it can still be used for dealing with multiclass by merging some of the objects or classes into a single class. Although this preprocessing approach can deliver respectable performance for many multiclass problems, our further work will lie in the area of theoretically extending SSMDR to multiclass case.

ACKNOWLEDGMENTS

The authors would like to express our sincere thanks to the anonymous reviewers' comments and suggestions that have made the paper a higher standard.

REFERENCES

BELKIN, M. and P. NIYOGI. 2001. Laplacian eigenmaps and spectral techniques for embedding and clustering. *In* Advances in Neural Information Processing System, Vol. 15. Cambridge, MA: MIT Press, pp. 585–591.

BELKIN, M. and P. NIYOGI. 2003. Laplacian eigenmaps for dimensionality reduction and data representation. Neural Computation, **15**(6):1373–1396.

BELKIN, M., I. MATVEEVA, and P. NIYOGI. 2004. Regularization and semi-supervised learning on large graphs. *In* Proceedings of the 17th Annual Conference on Learning Theory (COLT), Lecture Notes in Computer Science, Banff, Canada, pp. 624–638.

BELKIN, M., P. NIYOGI, and V. SINDHWANI. 2006. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. Journal of Machine Learning Research, **7**:2399–2434.

BENGIO, Y., J. F. PAIEMENT, P. VINCENT, O. DELALLEAU, N. L. ROUX, and M. OUIMET. 2004. Out-of-sample extensions for LLE, ISOMAP, MDS, laplacian eigenmaps, and spectral clustering. *In* Advances in Neural Information Processing Systems, Vol. 16. Cambridge, MA: MIT Press, pp. 177–184.

BLAKE C. L., and C. J. MERZ. 1998. UCI repository of machine learning databases. Available from: http://www.ics.uci.edu/&mlearn/MLRepository.html.

CAI, D., X. F. HE, and J. HAN. 2007a. Semi-supervised discriminant analysis. *In* Proceedings of the IEEE International Conference on Computer Vision (ICCV), Rio de Janeiro, Brazil, pp. 1–7.

CAI, D., X. F. HE, and J. W. HAN. 2007b. Isometric projection. *In* Proceedings of the 22nd AAAI Conference on Artificial Intelligence (AAAI), Vancouver, Canada, pp. 528–533.

CHAPELLE, O., B. SCHÖLKOPF, and A. ZIEN, editors. 2006. Semi-Supervised Learning. Cambridge, MA: MIT Press.

CHUNG, F. R. K. 1997. Spectral Graph Theory: Regional Conference Series in Mathematics, Number 92. Washington, DC: American Mathematical Society.

COX, T. F., and M. A. A. COX. 2001. Multidimensional Scaling (2nd ed.). Boca Raton: Chapman Hall.

DUDA, R. O., P. E. HART, D. G. STOR. 2001. Pattern Classification. New York: Wiley.

HAM, J., D. D. LEE, S. MIKA, and B. SCHÖLKOPF. 2004. A kernel view of the dimensionality reduction of manifolds. *In* Proceedings of the 21st International Conference on Machine Learning (ICML). New York: ACM Press, pp. 369–376.

HARDOON, D. R., S. SZEDMAK, and J. SHAWE-TAYLOR. 2004. Canonical correlation analysis: An overview with application to learning methods. Neural Computation, **16**(12):2639–2664.

HE, X. F., D. CAI, S. C. YAN, and H. J. ZHANG. 2005. Neighborhood preserving embedding. *In* Proceedings of the IEEE International Conference on Computer Vision (ICCV), Beijing, China, pp. 1208–1213.

HE, X. F., S. C. YAN, Y. X. HU, and H. J. ZHANG. 2003. Learning a locality preserving subspace for visual recognition. *In* Proceedings of the IEEE International Conference on Computer Vision (ICCV), Nice, France, p. 385.

HE, X. and P. NIYOGI. 2004. Locality preserving projections. *In* Advances in Neural Information Processing Systems, Vol. 16. *Edited by* S. Thrun, L. Saul, and B. SchÄolkopf. Cambridge, MA: MIT Press.

HE, X. F., S. C. YAN, Y. X. HU, P. NIYOGI, and H. J. ZHANG. 2005. Face recognition using Laplacianfaces. IEEE Transactions on Patten Analysis and Machine Intelligence, **27**(3):328–340.

HINTON, G. E. and R. R SALAKHUTDINOV. 2006. Reducing the dimensionality of data with neural networks. Science, **313**(5786):504–507.

LAWRENCE, K. S. and T. R. SAM. 2003. Think globally, fit locally: Unsupervised learning of low dimensional manifolds. Journal of Machine Learning Research, **4**:119–155.

LIN, Y. Y., T. L. LIU, and H. T. CHEN. 2005. Semantic manifold learning for image retrieval. *In* Proceedings of the ACM Conference on Multimedia, Singapore, pp. 249–258.

MARDIA, K. V., J. T. KENT, and J. M. BIBBY. 1980. Multivariate Analysis (Probability and Mathematical Statistics). San Diego, CA: Academic Press.

MARTINEZ, A. M., and A. C. KAK. 2001. PCA versus LDA. IEEE Transactions on Pattern Analysis and Machine Intelligence, **23**(2):228–233.

PETER, W. F. and J. S. DAVID. 1991. Letter recognition using Holland-style adaptive classifiers. Machine Learning, **6**(2):161–182.

ROWEIS, S. and L. SAUL. 2000. Nonlinear dimensionality reduction by locally linear embedding. Science, **290**:2323–2326.

SCHÖLKOPF, B. and A. SMOLA. 2002. Learning with Kernels. Cambridge, MA: MIT Press.

SCHOTT, J. R. 2005. Matrix Analysis for Statistics (Wiley Series in Probability and Statistics) (2nd ed.). Hoboken, NJ: Wiley.

SUGIYAMA, M. 2006. Local Fisher discriminant analysis for supervised dimensionality reduction. *In* Proceedings of the 23nd International Conference on Machine Learning (ICML), Pittsburgh, PA, pp. 905–912.

SUGIYAMA, M. 2007. Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis. Journal of Machine Learning Research, **8**:1027–1061.

SUGIYAMA, M., T. IDÉ, S. NAKAJIMA, and J. SESE. 2008. Semi-supervised local Fisher discriminant analysis for dimensionality reduction. *In* Advances in Knowledge Discovery and Data Mining, **Vol. 5012.** *Edited by* T. Washio, E. Suzuki, K. M. Ting, and A. Inokuchi. Berlin: Springer, pp. 333–344.

SUGIYAMA, M., T. IDÉ, S. NAKAJIMA, and J. SESE. 2010. Semi-supervised local Fisher discriminant analysis for dimensionality reduction. Machine Learning, **78**(1–2):35–61.

SUN, T. K., and S. C. CHEN. 2007. Locality preserving CCA with applications to data visualization and pose estimation. Image and Vision Computing, **25**(5):531–543.

TENENBAUM, J. B., V. D. SILVA, and J. C. LANGFORD. 2000. A global geometric framework for nonlinear dimensionality reduction. Science, **290**(5500):2319–2323.

TORRE, F. D. and T. KANADE. 2005. Multimodal oriented discriminant analysis. *In* Proceedings of the 22nd International conference on Machine learning (ICML), Bonn, Germany, pp. 177–184.

XU, L., Y. XU, and W. S. CHOW. 2010. PolSOM: A new method for multidimensional data visualization. Pattern Recognition, **43**(4):1668–1675.

YANG, X., H. FU, H. ZHA, and J. L. BARLOW. 2006. Semi-supervised nonlinear dimensionality reduction. *In* Proceedings of the International Conference on Machine Learning, Pittsburgh, PA, pp. 1065–1072.

YE, Q. L., C. X. ZHAO, N. YE, and Y. N. CHEN. 2010. Multi-weight vector projection support vector machines. Pattern Recognition Letters, **31**(13):2006–2011.

YIN, H. 2002a. ViSOM—A novel method for multivariate data projection and structure visualization. IEEE Transactions on Neural Networks, **13**(1):237–243.

YIN, H. 2002b. Data visualization and manifold mapping using the ViSOM. Neural Networks, **15**(8–9):1005–1016.

ZELNIK-MANOR, L., and P. PERONA. 2005. Self-tuning spectral clustering. *In* Advances in Neural Information Processing Systems, **Vol. 17**. *Edited by* L. K. Saul, Y. Weiss, and L. Bottou. Cambridge, MA: MIT Press, pp. 1601–1608.

ZHANG, Y., and D. Y YEUNG. 2008a. Semi-supervised discriminant analysis via CCCP. *In* Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD), Antwerp, Belgium, pp. 644–659.

ZHANG, Y., and D. Y. YEUNG. 2008b. Semi-supervised discriminant analysis using robust path-based similarity. *In* Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Anchorage, AK, pp. 1–8.

ZHANG, Z., and N. YE. 2011. Locality preserving multimodal discriminative learning for supervised feature selection. Knowledge and Information Systems, **27**(3):473–490.

ZHOU, D., O. BOUSQUET, T. LAL, J. WESTON, and B. SCHÖLKOPF. 2004. Learning with local and global consistency. Advances in Neural Information Processing Systems, Vol. 16. Cambridge, MA: MIT Press, pp. 321–328.

ZHU, X. 2006. Semi-supervised learning literature survey. Technical Report 1530, Department of Computer Sciences, University of Wisconsin at Madison, Madison, WI.

APPENDIX A: MATRIX EXPRESSIONS OF $\widetilde{A}^{(lwc)\dagger}$ AND $\widetilde{A}^{(lbc)\dagger}$

We similarly interpret the localized within-class scatter matrix $\widetilde{A}^{(lwc)\dagger}$ as the following form:

$$\widetilde{A}^{(lwc)\dagger} = \frac{1}{2}\sum_{i,j=1}^{m}(y_i - y_j)(y_i - y_j)^{\mathrm{T}}\widetilde{A}_{i,j}^{(wc)\dagger} = \frac{1}{2}\sum_{i,j=1}^{m}\left(y_i y_i^{\mathrm{T}} + y_j y_j^{\mathrm{T}} - y_i y_j^{\mathrm{T}} - y_j y_i^{\mathrm{T}}\right)\widetilde{A}_{i,j}^{(wc)\dagger},$$

then the localized within-class scatter matrix $\widetilde{A}^{(lwc)\dagger}$can be expressed in a matrix form as follows:

$$\widetilde{A}^{(lwc)\dagger} = \sum_{i=1}^{m}\widetilde{F}_{ii}^{(wc)\dagger} y_i y_i^{T}\, Y\widetilde{A}^{(wc)\dagger}Y^{T} = Y\widetilde{W}^{(wc)\dagger}Y^{T},$$

where $\widetilde{W}^{(wc)\dagger} = \widetilde{F}^{(wc)\dagger} - \widetilde{A}^{(wc)\dagger}$ and $\widetilde{F}^{(wc)\dagger}$ is the m-dimensional diagonal matrix with ith input element being $\widetilde{F}_{ii}^{(wc)\dagger} = \sum_{j=1}^{m}\widetilde{A}_{i,j}^{(wc)\dagger}$, which yields equation (14). Next, we analogously represent the between-class scatter or spread $\widetilde{A}^{(lbc)\dagger}$ in a localized manner as

$$\begin{aligned}\widetilde{A}^{(lbc)\dagger} &= \frac{1}{2}\sum_{i,j=1}^{m}(x_i - y_j)(x_i - y_j)^{\mathrm{T}}\widetilde{A}_{i,j}^{(bc)\dagger}\\ &= \frac{1}{2}\sum_{i=1}^{m}x_i\left(\sum_{j}\widetilde{A}_{i,j}^{(bc)\dagger}\right)x_i^{\mathrm{T}} + \frac{1}{2}\sum_{j=1}^{m}y_j\left(\sum_{i}\widetilde{A}_{i,j}^{(bc)\dagger}\right)y_j^{\mathrm{T}}\\ &\quad - \frac{1}{2}\sum_{i,j=1}^{m}x_i\widetilde{A}_{i,j}^{(bc)\dagger}y_j^{\mathrm{T}} - \frac{1}{2}\sum_{i,j=1}^{m}y_j\widetilde{A}_{j,i}^{(bc)\dagger}x_i^{\mathrm{T}},\end{aligned}$$

then the localized between-class scatter $\widetilde{A}^{(lbc)\dagger}$ can be expressed in a matrix interpretation as

$$\widetilde{A}^{(lbc)\dagger} = \frac{1}{2}(X\widetilde{D}^{(bc)\dagger}X^{\mathrm{T}} + Y\widetilde{M}^{(bc)\dagger}Y^{\mathrm{T}}) - \frac{1}{2}(X\widetilde{A}^{(bc)\dagger}Y^{\mathrm{T}} + Y\widetilde{A}^{(bc)\dagger\mathrm{T}}X^{\mathrm{T}}),$$

where $\widetilde{D}^{(bc)\dagger}(\widetilde{M}^{(bc)\dagger})$ is an m-dimensional diagonal matrix with the ith (or jth) element being $\widetilde{D}_{ii}^{(bc)\dagger} = \sum_{j=1}^{m}\widetilde{A}_{i,j}^{(bc)\dagger}$, $\widetilde{M}_{jj}^{(bc)\dagger} = \sum_{i=1}^{m}\widetilde{A}_{i,j}^{(bc)\dagger}$, which yields equation (24).

APPENDIX B: COMPUTATIONAL ANALYSIS OF $\widehat{T}_y$

Here, we formulate the localized scatters $\widetilde{A}_{\Phi}^{(rlbc)\dagger}$ and $\widetilde{A}_{\Phi}^{(rlwc)\dagger}$ in the kernel space in a similar manner as

$$\begin{aligned}\widetilde{A}_{\Phi}^{(rlbc)\dagger} = \frac{(1-\mu_A)}{2}\big[&\Phi(X)\widetilde{D}^{(bc)\dagger}\Phi(X)^{\mathrm{T}} + \Phi(Y)\widetilde{M}^{(bc)\dagger}\Phi(Y)^{\mathrm{T}} - (\Phi(X)\widetilde{A}^{(bc)\dagger}\Phi(Y)^{\mathrm{T}}\\ &+ \Phi(Y)\widetilde{A}^{(bc)\dagger\mathrm{T}}\Phi(X)^{\mathrm{T}})\big] + \frac{\mu_A}{l}\Phi(Y_u)\widetilde{L}_u^{(Y)}\Phi(Y_u)^{\mathrm{T}},\end{aligned}$$

By substituting $\widetilde{A}_{\Phi}^{(rlbc)\dagger}$, $\widetilde{A}_{\Phi}^{(rlwc)\dagger}$, and equation (31) into the generalized eigenvalue problem in equation (12), we obtain

$$\widetilde{A}_{\Phi}^{(rlbc)\dagger}\Phi(Y)\,\widehat{T}_y = \widetilde{\delta}_y\widetilde{A}_{\Phi}^{(rlwc)\dagger}\Phi(Y)\,\widehat{T}_y. \tag{39}$$

Let $K_{yy} = (\Phi(Y))^{\mathrm{T}}\Phi(Y) = K_{yy}^{\mathrm{T}}$ be the kernel matrix between labeled samples in Y and $K_{yy_u} = \Phi(Y)^{\mathrm{T}}\Phi(Y_u)$ be the kernel matrix between labeled samples in Y and unlabeled samples in Y_u and $K_{y_u y} = K_{yy_u}^{\mathrm{T}}$. By multiplying equation (39) by $\Phi(Y)^{\mathrm{T}}$ from the left-hand side, we can obtain

$$\begin{aligned}&\left[\frac{(1-\mu_A)}{2}(K_{yx}\widetilde{D}^{(bc)\dagger}K_{xy} + K_{yy}\widetilde{M}^{(bc)\dagger}K_{yy} - K^{(bc)\dagger}) + \frac{\mu_A}{l}K_{yy_u}\widetilde{L}_u^{(Y)}K_{y_u y}\right]\widehat{T}_y \\ &\quad = \widetilde{\delta}_y((1-\mu_A)K_{yy}\widetilde{W}^{(wc)\dagger}K_{yy} + \mu_A K_{yy})\widehat{T}_y, \end{aligned} \tag{40}$$

where $K^{(bc)\dagger} = K_{yx}\widetilde{A}^{(bc)\dagger}K_{yy} + (K_{yx}\widetilde{A}^{(bc)\dagger}K_{yy})^{\mathrm{T}}$. Here, we similarly set

$$\begin{aligned}\Upsilon_3^{\Phi} &= \frac{(1-\mu_A)}{2}[K_{yx}\widetilde{D}^{(bc)\dagger}K_{xy} + K_{yy}\widetilde{M}^{(bc)\dagger}K_{yy} - K_{yx}\widetilde{A}^{(bc)\dagger}K_{yy} - (K_{yx}\widetilde{A}^{(bc)\dagger}K_{yy})^{\mathrm{T}}] \\ &\quad + \frac{\mu_A}{l}K_{yy_u}\widetilde{L}_u^{(Y)}K_{y_u y},\end{aligned}$$

$$\Upsilon_4^{\Phi} = (1-\mu_A)\,K_{yy}\widetilde{W}^{(wc)\dagger}K_{yy} + \mu_A K_{yy}.$$

Because the matrix Υ_4^{Φ} is not always of full rank, we also need to regularize it to avoid the singularity and ensure the stability of following generalized eigen-decomposition problem by adding the generalized term $\mu_I I$ with a small positive scalar μ_I; thus, equation (40) can be replaced by

$$\Upsilon_3^{\Phi}\widehat{T}_y = \widetilde{\delta}_y\big(\Upsilon_4^{\Phi} + \mu_I I\big)\widehat{T}_y.$$

Let $\{\widehat{T_{y_{[r]}}}\}_{r=1}^{d}$ be the generalized eigenvectors associated with the first d largest generalized eigenvalues $\widetilde{\delta_{y_{[r]}}}$, $r = 1, 2, \ldots, d$, where eigenvalues and eigenvectors are sorted and normalized as $\widetilde{\delta_{y_{[1]}}} \geq \widetilde{\delta_{y_{[2]}}} \geq \cdots \geq \widetilde{\delta_{y_{[d]}}}$ and $\widehat{T_{y_{[r]}}}^{\mathrm{T}}(\Upsilon_4^{\Phi} + \mu_I I)\widehat{T_{y_{[r]}}} = 1$ for $r = 1, 2, \ldots, d$, which yields equation (38).