Trace Ratio Optimization-Based Semi-Supervised Nonlinear Dimensionality Reduction for Marginal Manifold Visualization

Zhao Zhang, *Student Member*, *IEEE*, Tommy W.S. Chow, *Senior Member*, *IEEE*, and Mingbo Zhao, *Student Member*, *IEEE*

Abstract—Visualizing similarity data of different objects by exhibiting more separate organizations with local and multimodal characteristics preserved is important in multivariate data analysis. Laplacian Eigenmaps (LAE) and Locally Linear Embedding (LLE) aim at preserving the embeddings of all similarity pairs in the close vicinity of the reduced output space, but they are unable to identify and separate interclass neighbors. This paper considers the semi-supervised manifold learning problems. We apply the pairwise Cannot-Link and Must-Link constraints induced by the neighborhood graph to specify the types of neighboring pairs. More flexible regulation on supervised information is provided. Two novel multimodal nonlinear techniques, which we call trace ratio (TR) criterion-based semi-supervised LAE ($\mathrm{S^2LAE}$) and LLE ($\mathrm{S^2LLE}$), are then proposed for marginal manifold visualization. We also present the kernelized $\mathrm{S^2LAE}$ and $\mathrm{S^2LLE}$. We verify the feasibility of $\mathrm{S^2LAE}$ and $\mathrm{S^2LLE}$ through extensive simulations over benchmark real-world MIT CBCL, CMU PIE, MNIST, and USPS data sets. Manifold visualizations show that $\mathrm{S^2LAE}$ and $\mathrm{S^2LLE}$ are able to deliver large margins between different clusters or classes with multimodal distributions preserved. Clustering evaluations show they can achieve comparable to or even better results than some widely used methods.

Index Terms—Semi-supervised manifold learning, trace ratio optimization, nonlinear dimensionality reduction, multimodality preservation, pairwise constraints, marginal manifold visualization

1 Introduction

Visualizing image data via dimensionality reduction (DR) has become increasingly important since many emerging applications are closely related with high-dimensional data, such as human gene distributions. These data sets often contain numerous samples, each of which contains huge number of features. The major issue of DR is to find a projection matrix to transform the high-dimensional data into the low-dimensional representations appropriately with the intrinsic local or global geometry structures being effectively preserved [1], [2]. When DR is appropriately conducted, the compact meaningful representation of the original data can be utilized for various subsequent tasks, such as visualization. Linear *Principal Component Analysis (PCA)* [3] and *Linear Discriminant Analysis (LDA)* [4] are two representative DR methods.

Intrinsic multimodal and nonlinear structures are often come across in real-life applications. For example, in facial gender recognition, intraclass multimodal and nonlinear structures appear when genders are classified to males and females, because images of different persons are usually captured under different conditions (e.g., lightings and poses). Similarly, merging even or odd digits to a single class usually involves multimodal nonlinear structures. Obviously, it is advantageous to employ the nonlinear DR techniques to the real data sets [5], [6]. Data points in a dense area deliver similar manifolds [7]. To represent given data well, it is vital to consider the local information of data. This leads to the appearance of many locality or neighborhood preserving methods [8], e.g., *Laplacian Eigenmaps (LAE)* [10], *Locally Linear Embedding (LLE)* [1] and *ISOMAP* [36]. Nonlinear LAE, LLE, and ISOMAP deliver optimal embeddings directly without exhibiting the projection axes and are developed for data visualization. These unsupervised methods are efficient in visualizing synthetic data sets [5] and are powerful to handle nonlinear data. Note that LAE and LLE keep all neighboring pairs close in the reduced space. On the contrary, neighbors of different objects or classes also deliver similar embeddings, so transformed data of interclass neighbors are likely to be congregated. The major reason for mixing the interclass neighbors is that they do not consider any form of supervised information, such as class labels or pairwise constraints (PC). In the real word, unlabeled data are readily available but labeled ones are usually expensive to obtain. By incorporating the class information into the manifold learning, *Supervised LLE* (SLLE) [27], *Supervised ISOMAP* (S-ISOMAP) [26] and *Semi-Supervised Maximum Margin Projection* (MMP) [35] are then proposed. Note that SLLE and S-ISOMAP are supervised and may be overfitted to the training data when only a limited number of labeled data are available. Recently, by utilizing the prior information obtained from the mapping of certain points, another type of semi-supervised manifold learning algorithms including *Semi-Supervised Local Tangent Space Alignment* (SS-LTSA) [16], [17],

• *The authors are with the Department of Electronic Engineering, City University of Hong Kong, Room G6409, 6/F, Academic Building, 83 Tat Chee Avenue, Kowloon, Hong Kong. E-mail: itzzhang@ee.cityu.edu.hk, eetchow@cityu.edu.hk, mzhao4@student.cityu.edu.hk.*

Manuscript received 5 May 2011; revised 16 Nov. 2011; accepted 2 Feb. 2012; published online 2 Mar. 2012.
Recommended for acceptance by A. Tung.
For information on obtaining reprints of this article, please send e-mail to: tkde@computer.org, and reference IEEECS Log Number TKDE-2011-05-0252.
Digital Object Identifier no. 10.1109/TKDE.2012.47.

Semi-Supervised ISOMAP (SS-ISOMAP) [16], and *Semi-Supervised LLE* (SS-LLE) [16], are proposed. SS-LTSA is a semi-supervised extension of LTSA [11].

Domain knowledge in the form of pairwise *Cannot-Link (CL)* and *Must-Link (ML)* constraints are widely used in many areas, for instance [15], [18], [19], [22]. PC are created depending on whether point pairs are in the same class or not. Compared with the class labels, PC can be achieved with minimal human effort and can provide more supervision information when labeled number is limited [18], [22]. Therefore, it is a great advantage to utilize PC for discriminant semi-supervised manifold learning, especially when the labeled number is few. Two representative works of PC derived algorithms are *Semi-Supervised Dimensionality Reduction* (SSDR) [15] and *Semi-Supervised Metric Learning* (SSML) [19] which incorporate PC with abundant unlabeled data. SSDR and SSML can keep the intrinsic structures of unlabeled samples and PC defined on the labeled data, and considerable improvements in embeddings are exhibited. But note that SSDR is a global algorithm and the terms involving PC in SSML are also global. This paper applies the neighborhood graph induced PC to guide the manifold learning and develops new algorithms. In this paper, there are four major contributions. First, we propose two effective and novel LAE and LLE criteria-based semi-supervised manifold learning techniques called S^2LAE and S^2LLE under a trace ratio criterion [28], [30]. S^2LAE and S^2LLE are naturally different from virtually all previous semi-supervised manifold learning methodologies. By utilizing the graph-induced *ML* and *CL* constraints, the types of neighboring pairs are categorized to intra- and interclass. By defining reasonable criteria, large margins between intra- and interclass clusters are organized and enhanced compactness of intracluster neighbors can be obtained at the same time. Second, compared with utilizing the class labels, the pairwise *ML* and *CL* constraints can provide us with more supervised information. More importantly, the PC sets are flexible in providing more degree of freedom for generalization. In other words, we can employ either partial or all constraints for optimization. Note that when the number of labeled samples or available constraints is small, unlabeled samples can help boosting the performance. Third, practical approaches are developed to extend S^2LAE and S^2LLE to kernelized scenarios. Fourth, TR optimization [29], [33] is used to solve our developed problems delivering more specific solution according to the stronger orthogonal constraints.

The outline of the paper is given as follows: In Section 2, we briefly reviews LAE and LLE. In Section 3, we formulate the proposed algorithms and their extensions. In Section 4, we present the solution schemes of our problems. Subsequently, in Section 5, we describe the simulation settings and evaluate our algorithms using benchmark MIT CBCL, CMU PIE, MNIST, and USPS databases. Finally, the concluding remarks are drawn in Section 6.

2 Preliminaries

Given a data graph $G = (V, E)$, where V is the set of vertices $\{x_i\}_{i=1}^N$ in n-dimensional space $\mathbb{R}^n$ and E is the set of edges. Then, LAE and LLE can be defined as follows:

2.1 Laplacian Eigenamps Revisited

LAE [10] starts by constructing a neighborhood graph by k nearest neighbor search (NNS), that is x_i and x_j are connected by an edge if they are neighbors. Note that the heat kernel method, the local scaling heuristic method [23] and the simple-minded method can be used to define the weights. For the simple-minded method, $W_{i,j} = W_{j,i} = 1$ if x_i and x_j are neighbors. The criterion for computing the optimal embeddings is to solve the following problem:

$$\underset{Y}{Min} \frac{1}{2}\sum_{i,j=1}^{N} \left\|y_i - y_j\right\|^2 W_{i,j} = \underset{YDY^{\mathrm{T}}=I}{Min} Tr\big(Y(D-W)Y^{\mathrm{T}}\big), \quad (1)$$

where I is an identity matrix, y_i is the dimension-reduced representation of x_i and weight $W_{i,j}$ incurs a heavy penalty if neighboring pairs x_i and x_j are mapped far apart. The notation $\|\cdot\|$ is the l^2-norm, notation $^{\mathrm{T}}$ denotes the transpose of a vector or a matrix and $Tr(\cdot)$ is trace operator. Thus, minimizing (1) can ensure that if x_i and x_j are close, then y_i and y_j are also close. Let $L = D - W$ be the Laplacian matrix of W over the data in $X = [x_1|x_2|\ldots x_N]$ and D be an N-dimensional diagonal matrix with ith (or jth, since W is symmetric) element being $D_{ii} = \sum_j W_{i,j}$, then the solution $Y \in^{d\times N}$ $(d \le n)$ can be achieved as the eigenvectors corresponding to with the d smallest eigenvalues of the generalized eigen-problem:

$$(D-W)\mu_j = \lambda_j D\mu_j; \;\; Y = [u_2, \ldots, u_{d+1}]^{\mathrm{T}} \in \mathbb{R}^{d\times N}. \quad (2)$$

2.2 Locally Linear Embedding Revisited

LLE [1] works in a similar manner to LAE. The first step is to determine the k neighbors of each point x_i by NNS, and then compute the weights that can linearly reconstruct x_i in the best possible way from its neighbors by

$$\varepsilon(\Delta) = \sum_{i=1}^{N} \left\| x_i - \sum_{x_j \in N_+^{(x_i)}} \Delta_{i,j} x_j \right\|^2, \;\; s.t.\ \forall_i \sum_j \Delta_{i,j} = 1, \quad (3)$$

where $N_+^{(x_i)}$ denotes the k nearest neighbor set of each vertex x_i. The weights $\Delta_{i,j}$ summarize the effect of the jth point on constructing the ith point [1], [34], satisfying

$$\Delta_{i,j} = \frac{\sum_{r=1}^{N} \chi_{jr}^{(i)}}{\sum_{u=1}^{N}\sum_{t=1}^{N} \chi_{ut}^{(i)}}, \quad (4)$$

where $\chi^{(i)} = (\aleph^{(i)})^{-1}$ and $\aleph^{(i)}$ satisfy

$$\aleph_{jr}^{(i)} = (x_i - x_j)^{\mathrm{T}}(x_i - x_r), \quad (5)$$

where x_j and x_r are neighbors of x_i. Thus, the optimal low-dimensional embedding can be achieved by

$$\underset{Y}{Min} \sum_{i=1}^{N} \left\| y_i - \sum_{x_j \in N_+^{(x_i)}} \Delta_{i,j} y_j \right\|^2. \quad (6)$$

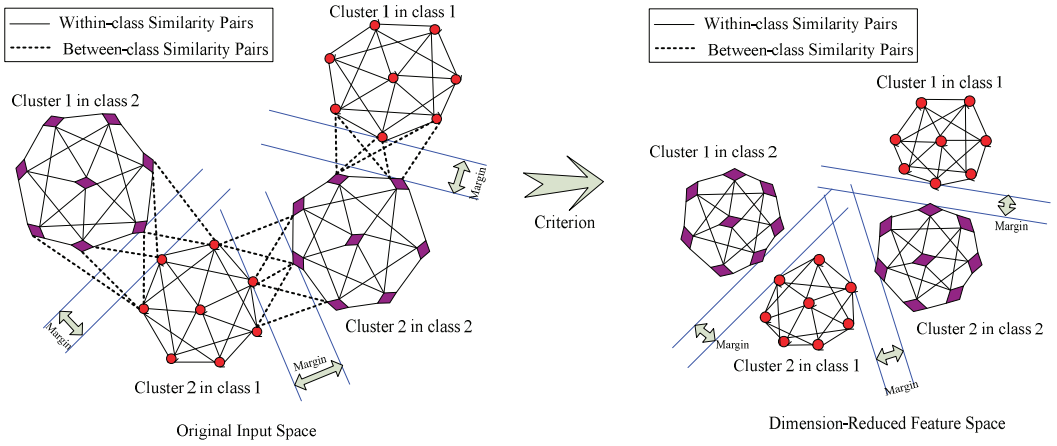


Fig. 1. Geometrical interpretations of neighborhood preserving LAE and LLE criteria.

Since $\forall_i \sum_j \Delta_{i,j} = 1$, the LLE objective function can be reformulated as the following problem [24]:

$$\begin{aligned} &\arg\min_{YY^{\mathrm{T}}=I} Tr(Y(I-\Delta)^{\mathrm{T}}(I-\Delta)Y^{\mathrm{T}}) \\ &\quad = \arg\min_{YY^{\mathrm{T}}=I} Tr(YLY^{\mathrm{T}}), \end{aligned} \tag{7}$$

where $L = D - W$, with $W_{i,j} = (\Delta + \Delta^{\mathrm{T}} - \Delta^{\mathrm{T}}\Delta)_{i,j}$ if $i \neq j$, else it equals to zero. D is a diagonal matrix with $D_{ii} = \sum_j W_{i,j}$. Note that the matrix W in (7) is different from the matrix W in LAE, but we still use the same notation hereinafter for facilitating the descriptions. To solve the above problem, the covariance matrix can be constrained to be identity, i.e., $(1/N)\sum_{i=1}^{N} y_i y_i^{\mathrm{T}} = I$, otherwise $Y = 0$ is optimal. From (7), Y spanned by the basis vectors can be obtained as eigenvectors $\{u_j^{\mathrm{T}}\}_{j=2}^{d+1}$ associated with d smallest eigenvalues of the eigen-problem: $L\mu_j = \lambda_j \mu_j$.

Note that LAE is similar to LLE from the following aspects. Both the LLE matrix $\widehat{Q} = (I-\Delta)^{\mathrm{T}}(I-\Delta)$ and LAE matrix $D - W$ are symmetric positive semidefinite and are related to the local preservation of data. As indicated in [20], problem (2) can be written as $Min_{Tr(\widehat{YY^{\mathrm{T}}})=c} Tr(\widehat{Y}\widehat{L}\widehat{Y}^{\mathrm{T}})$, where the normalized Laplacian

$$\widehat{L} = I - \widehat{W} = D^{-1/2}LD^{-1/2}, \widehat{W} = D^{-1/2}WD^{-1/2}$$

and $\widehat{Y} = YD^{1/2}$, then the only difference is that $\widehat{L}$ is the normalized graph Laplacian and Δ is an affinity matrix [20]. Note that when matrix $\Delta = W = (1/n)ee^{\mathrm{T}}$, we can have $(I-\Delta)^{\mathrm{T}}(I-\Delta) = I - \widehat{W}$, which means that LAE is equivalent to LLE in this case [20]. The detailed analyses between LAE and LLE can be referred to [20].

3 Semi-Supervised Multimodal Nonlinear Dimensionality Reduction

3.1 Motivation and Objective

LAE and LLE are two most representative local structure preserving multivariate visualization approaches. Next, we describe the principles of LAE and LLE criteria from the marginal perspective. Recall the LAE criterion in (1). If we substitute the squared euclidean distance $d^2(y_i, y_j)$ into (1), the problem (1) can be rewritten as

$$\varepsilon_{LAE}(Y) = \underset{Y}{Min} \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} d^2(y_i, y_j)W_{i,j}, \tag{8}$$

where $d^2(y_i, y_j) = \|y_i - y_j\|^2$ denotes the squared euclidean distance between the low-dimensional embeddings y_i and y_j. Similarly, we can obtain the LLE embeddings by solving the above problem with $W_{i,j} = (\Delta + \Delta^{\mathrm{T}} - \Delta^{\mathrm{T}}\Delta)_{i,j}$ if $i \neq j$, else it equals to zero. We consider a binary-class case shown in Fig. 1, in which each class has two isolated clusters, i.e., *multimodal*. The solid lines and dotted lines connect within-class and between-class neighbors in the undirected neighborhood graph. Note that we only show partial edges connecting neighboring pairs. In LAE and LLE, weights $W_{i,j}$ are used to reflect the proximity relations in the neighborhood graph. Obviously, when the weight value $W_{i,j}$ increases, distance $d^2(y_i, y_j)$ must decrease to minimize the summation. As LAE and LLE keep the local distances of all neighboring pairs, a nonzero weight $W_{i,j}$ will be set in all the solid lines and dotted lines as shown in Fig. 1. When weight $W_{i,j}$ becomes heavier, distance $d^2(y_i, y_j)$ between the similarity pairs has to be minimized for balancing the value of the objective. In other words, all margins between the similarity pairs in the reduced output spaces of LAE and LLE will be smaller than that in the original space. Thus, by optimizing the criteria of LAE and LLE, the embeddings of the clusters can be geometrically displayed in Fig. 1. For DR, obtaining high separation of interclass neighbors in addition to preserving the local information is important. But it is worth noting that the criteria of the regular LAE and LLE are unable to achieve this objective.

The problem of previous manifold learning methods, including LAE and LLE, stem from its inability to take interclass similarity data separation into account, as they only focus on preserving the geometrical structures of all similarity points. So, it is natural to define more efficient criteria to improve the tightness of intracluster similarity pairs and push intercluster neighbors far apart to achieve larger margins for feature extraction. To tackle the shortcomings of LAE and LLE, this work considers new criteria and proposes a solution to this problem with geometrical interpretations based on the marginal perspective.

We describe our criterion from a pairwise constrained perspective. The neighborhood graph-induced pairwise *ML* and *CL* constraints are employed for identifying the types of the neighboring pairs. The definitions of *ML* and *CL*

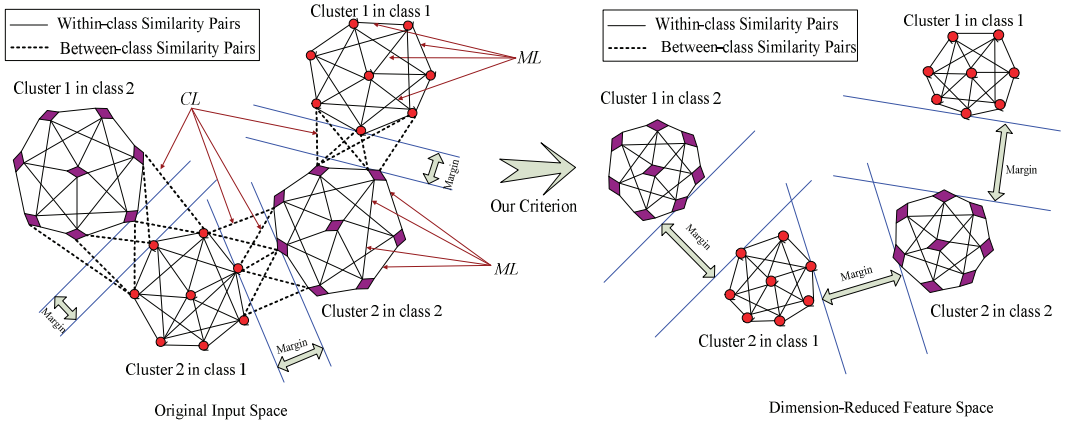


Fig. 2. Geometrical interpretation of our proposed learning criterion.

constraints will be detailed in the next section. *ML* and *CL* reflect the supervised information of samples, which is more practical way than trying to obtain the class labels [15], [18]. Intuitively, the proximity relations of neighbors constrained by *ML* should be enhanced, while the proximity relations between similarity pairs in *CL* should be weakened as much as possible, because we aim at separating them. Note that to balance the objective functional value, $W_{i,j}$ and $d^2(y_i, y_j)$ in LAE and LLE have the opposite meanings. Motivated by these analyses, for LAE, when weight value $W_{i,j}$ increases, we can reduce the distance $d^2(y_i, y_j)$ in order to minimize the summation $\sum_{i=1}^{N}\sum_{j=1}^{N} d^2(y_i, y_j)W_{i,j}$ if x_i and x_j are constrained by *ML*. On the contrary, if vertices x_i and x_j are constrained by *CL*, a heavy penalty $W_{i,j}$ will be imposed to maximize the summation $\sum_{i=1}^{N}\sum_{j=1}^{N} d^2(y_i, y_j)W_{i,j}$, implying that the distance $d^2(y_i, y_j)$ will be significantly expanded in the feature space. Note that similar discussions exist for LLE. In Fig. 2, we show some typical examples of *ML* and *CL* constraints. Based on the above criteria, margins between interclass clusters can be significantly broadened and the margins between points of the intraclass clusters are significantly shrunk in the reduced space as shown in Fig. 2. Most importantly, the intrinsic multimodal structures can be efficiently kept, because margins between intraclass clusters are enlarged. In this study, we focus on addressing this issue to achieving the marginal discriminant manifold learning as described in later sections.

3.2 Graph-Induced Pairwise ML and CL Constraints

Based on the definition of local neighborhood, we compute the constraint sets in a graph-induced approach. First, data graph $G = (V, E)$ with N vertices and $N(N-1)/2$ edges is obtained. Denote by $N_+^{(x_i)}$ the k nearest neighbor set of x_i. A weight is put on the edge $e(x_i, x_j) \in E$ between x_i and x_j if $x_j \in V$ and $x_i \in V$. Then, define

$$\begin{cases} e(x_i, x_j) = 1, \ if \ x_j \in N_+^{(x_i)} \ or \ x_i \in N_+^{(x_j)}, La(x_j) = La(x_i) \\ e(x_i, x_j) = -1, \ if \ x_j \in N_+^{(x_i)} \ or \ x_i \in N_+^{(x_j)}, La(x_j) \neq La(x_i) \\ e(x_i, x_j) = 0, \ if \ x_j \notin N_+^{(x_i)} \ and \ x_i \notin N_+^{(x_j)}, \end{cases} \tag{9}$$

where $La(x_i)$ denotes the class label of point x_i. In this way, a neighborhood graph $\widetilde{G}_N = (\widetilde{V}_N, \widetilde{E}_N)$ with nonzero weights is formed, satisfying $\widetilde{V}_N = V$. The purpose of constructing graph $\widetilde{G}_N$ is to represent the similarities between each vertex pair, where similarity is measured by $e_N(x_i, x_j) \in \widetilde{E}_N$. If the index and corresponding points of graph $\widetilde{G}_N$ are recorded, intra- and interclass neighborhood graphs can be obtained. We refer the neighborhood graphs constrained by *ML* and *CL* to as *ML*-graph and *CL*-graph, respectively. Then, the pairwise *ML* and *CL* constraint sets are defined as

$$ML = \left\{ \begin{array}{l} (x_i, x_j) | e_N(x_i, x_j) = 1, \ v(x_i) \in \widetilde{V}_N, \\ v(x_j) \in \widetilde{V}_N, La(x_j) = La(x_i) \end{array} \right\}, \tag{10}$$

$$CL = \left\{ \begin{array}{l} (x_i, x_j) | e_N(x_i, x_j) = -1, v(x_i) \in \widetilde{V}_N, \\ v(x_j) \ \in \widetilde{V}_N, La(x_j) \neq La(x_i) \end{array} \right\}, \tag{11}$$

where $v(x_i)$ is vertex x_i in graph $\widetilde{G}_N$. Note that $\widetilde{G}_N$ and G have the same number of vertices and nonzero edges. Note that all edge lengths in *ML*-graph should be "minimized," while all edge lengths in *CL*-graph should be "maximized." Based on the *ML* and *CL* constraints, more separate embeddings of interclass neighbors can be delivered. More importantly, natural clusters within each class or object, i.e., intrinsic multimodality, can be preserved.

3.3 Semi-Supervised LAE (S^{2}LAE)

A reasonable criterion for our proposed trace ratio criterion-based semi-supervised LAE (S^{2}LAE) algorithm is to maximize the following objective function:

$$\underset{Y}{Max} \frac{\frac{\ell}{2}\sum_{i,j} \|y_i - y_j\|^2 \Xi_{i,j} + \frac{(1-\ell)}{2}\sum_{(x_i,x_j)\in CL} \|y_i - y_j\|^2 \widetilde{W}_{i,j}^{CL}}{\frac{1}{2}\sum_{(x_i,x_j)\in ML} \|y_i - y_j\|^2 \widetilde{W}_{i,j}^{ML}}, \tag{12}$$

where Ξ is an $N \times N$ matrix with entries $\Xi_{i,j} = 1/N$ and $\ell_-(\in [0, 1])$ is a tradeoff parameter for balancing the two terms in the numerator of (12). Note that adjacency matrices $\widetilde{W}^{ML}$ and $\widetilde{W}^{CL}$ are defined for keeping the local relationships between data points constrained by *ML* and *CL*, respectively. Matrices $\widetilde{W}^{ML}$ and $\widetilde{W}^{CL}$ will be used in later sections without introductions. Note that the number of *CL*

constraints is usually smaller than that of the *ML* constraints, term $(1/2)\sum_{i,j}\|y_i - y_j\|^2\Xi_{i,j}$ over all data points is added to make the above problem more stable. Similar to LAE, optimizing (12) is equivalent to ensuring that if pairs x_i and x_j are originally close and are constrained by *ML*, then y_i and y_j will be close as well; else if x_i and x_j are close but are constrained by *CL*, then y_i and y_j should be separated. The first term in the numerator of (12) plays a significant role in preserving the global covariance structures of both labeled and unlabeled samples. The "unlabeled" means the data have not been accompanied with class labels and are not involved in the pairwise constraints. The motivation for exploiting unlabeled samples is to use them to boost the performance when the supervised information is fewer. When all data points are used to construct $\widetilde{W}^{ML}$, $\widetilde{W}^{CL}$ and constrained data matrices, that is $\widetilde{W}^{ML}$, $\widetilde{W}^{CL}$, *ML*, and *CL* constrained data matrices have the same size N, then the problem (12) can be formulated as the following TR problem [29], [30]:

$$\begin{aligned}\underset{Y}{Max}\frac{\frac{1}{2}\sum_{i,j}\|y_i - y_j\|^2(\ell\Xi_{i,j} + (1-\ell)\widetilde{W}^{CL}_{i,j})}{\frac{1}{2}\sum_{i,j}\|y_i - y_j\|^2\widetilde{W}^{ML}_{i,j}} \\ = \underset{YY^{\mathrm{T}}=I}{Max}\frac{Tr(Y\widetilde{L}^{CL}Y^{\mathrm{T}})}{Tr(Y\widetilde{L}^{ML}Y^{\mathrm{T}})},\end{aligned} \tag{13}$$

where $\widetilde{\Theta}^{CL}_{i,j} = (\ell\Xi_{i,j} + (1-\ell)\widetilde{W}^{CL}_{i,j})$, matrices $\widetilde{L}^{CL} = \widetilde{D}^{CL} - \widetilde{\Theta}^{CL}$ and $\widetilde{L}^{ML} = \widetilde{D}^{ML} - \widetilde{W}^{ML}$. Note that we only weight the similarity pairs constrained by *ML* and *CL* when defining the matrices $\widetilde{W}^{ML}$ and $\widetilde{W}^{CL}$. After conducting k NNS over the total data matrix, we can then apply similar methods to construct the weights. For the simple-minded method, $\widetilde{W}^{ML}_{i,j} = \widetilde{W}^{ML}_{j,i} = 1$ and $\widetilde{W}^{CL}_{i,j} = \widetilde{W}^{CL}_{j,i} = 1$ if vertices i and j are constrained, and else 0. $\widetilde{D}^{ML}$ and $\widetilde{D}^{CL}$ are diagonal matrices with $\widetilde{D}^{ML}_{ii} = \sum_j \widetilde{W}^{ML}_{i,j}$ and $\widetilde{D}^{CL}_{ii} = \sum_j \widetilde{\Theta}^{CL}_{i,j}$. Then, the solution Y can be obtained by solving the above TR problem.

3.4 Semi-Supervised LLE (S^2LLE)

A constrained extension of LLE is also presented. We refer this invariant to as trace ratio criterion-based semi-supervised LLE. S^2LLE shares similar implementation procedures to LLE. If all data points are used to construct the adjacency matrices and constrained data matrices, the reconstruction errors in S^2LLE can be measured by using the following criterion:

$$\begin{aligned}&\varepsilon(\widetilde{\Delta}^{ML}, \widetilde{W}^{CL}) \\ &= \frac{\frac{\ell}{2}\sum_{i,j}\|x_i - x_j\|^2\Xi_{i,j} + \frac{(1-\ell)}{2}\sum_{(x_i,x_j)\in CL}\|x_i - x_j\|^2\widetilde{W}^{CL}_{i,j}}{\sum_{(x_i,x_j)\in ML}\|x_i - \sum_{x_j\in N_+^{(x_i)}}\widetilde{\Delta}^{ML}_{i,j}x_j\|^2},\end{aligned} \tag{14}$$

with respect to $\sum_j \widetilde{\Delta}^{ML}_{i,j} = 1$, where $\widetilde{W}^{CL}$ denotes the *CL* constrained weight matrix applied in S^2LAE, Ξ is similarly defined as in S^2LAE and $(1/2)\sum_{i,j}\|x_i - x_j\|^2\Xi_{i,j} = S^{(t)}$ is the total scatter matrix. Then, the weight matrix $\widetilde{\Delta}^{ML}$ with entries $\widetilde{\Delta}^{ML}_{i,j}$ summarizes the contribution of the jth sample to the ith reconstruction, which satisfies

$$\widetilde{\Delta}^{ML}_{i,j} = \sum_{r=1}^{N}\widehat{\mathcal{X}}^{(i)}_{jr} \Big/ \sum_{u=1}^{N}\sum_{t=1}^{N}\widehat{\mathcal{X}}^{(i)}_{ut}, \tag{15}$$

where $\widehat{\mathcal{X}}^{(i)} = (\widehat{\aleph}^{(i)})^{-1}$ and local covariance matrix $\widehat{\aleph}^{(i)}$ satisfies the following formulation:

$$\widehat{\aleph}^{(i)}_{jr} = (x_i - x_j)^{\mathrm{T}}(x_i - x_r), \text{ where } (x_i, x_j), (x_i, x_r) \in ML, \tag{16}$$

where x_j and x_r are neighbors of x_i. Then, each observation x_i is mapped to a low-dimensional vector y_i. The low-dimensional coordinates can be determined by optimizing the following optimization problem for S^2LLE:

$$\varepsilon(Y) = \underset{Y}{Max}\frac{\frac{1}{2}\sum_{i,j}\|y_i - y_j\|^2(\ell\Xi_{i,j} + (1-\ell)\widetilde{W}^{CL}_{i,j})}{\frac{1}{2}\sum_{i,j}\|y_i - y_j\|^2\widetilde{W}^{ML}_{i,j}}, \tag{17}$$

where $\widetilde{W}^{ML}_{i,j} = (\widetilde{\Delta}^{ML} + \widetilde{\Delta}^{ML\mathrm{T}} - \widetilde{\Delta}^{ML\mathrm{T}}\widetilde{\Delta}^{ML})_{i,j}$ when $i \neq j$, and else 0. That is, optimizing the above criterion can ensure that if sample pair x_i and x_j are neighbors in the original space and are constrained by *ML*, then y_i and y_j should be close in the reduced space as well; otherwise, if pairs x_i and x_j are neighbors constrained by *CL*, then y_i and y_j should be separated. Similarly, we can obtain

$$\underset{YY^{\mathrm{T}}=I}{Max}\frac{Tr(Y(\widetilde{D}^{CL} - \widetilde{\Theta}^{CL})Y^{\mathrm{T}})}{Tr(Y(\widetilde{D}^{ML} - \widetilde{W}^{ML})Y^{\mathrm{T}})} = \underset{YY^{\mathrm{T}}=I}{Max}\frac{Tr(Y\widetilde{L}^{CL}Y^{\mathrm{T}})}{Tr(Y\widetilde{L}^{ML}Y^{\mathrm{T}})}, \tag{18}$$

where $\widetilde{\Theta}^{CL} = (\ell\Xi + (1-\ell)\widetilde{W}^{CL})$. Matrices $\widetilde{D}^{ML}$ and $\widetilde{D}^{CL}$ have entries $\widetilde{D}^{ML}_{ii} = \sum_j \widetilde{W}^{ML}_{i,j}$ and $\widetilde{D}^{CL}_{ii} = \sum_j \widetilde{\Theta}^{CL}_{i,j}$. So, the embedding Y can be achieved from solving (18).

Note that if there exist vectors $P = [p_1, \ldots, p_d] \in \mathbb{R}^{n\times d}$ such that $Y = P^{\mathrm{T}}X$, our S^2LAE and S^2LLE methods can be effectively linearized to embed new points. The out-of-sample extrapolation method with a global regression regularization [6] can also be used to linearize our methods.

3.5 Kernelized Extensions of S^2LAE and S^2LLE

We in this section show the method of kernelizing S^2LAE and S^2LLE. Recently, kernelized LAE [25] was proposed for DR of nonvectorial data, in which W was replaced by a kernel matrix K with $K_{i,j} = \langle\phi(x_i), \phi(x_j)\rangle = \phi(x_i)^{\mathrm{T}}\phi(x_j)$ [12] to measure the similarities between points. For kernelized cases, the Gaussian RBF kernel is often used. As indicated in [25], weights $W_{i,j}$ are restricted to be euclidean distances between vectorized representations of data, while K contains the proximity relations among all input data and the relations are encoded in a reduced space in the sense of euclidean distance by a minimization procedure. With this method, we can similarly kernelize S^2LAE by replacing $\widetilde{W}^{CL}, \widetilde{W}^{ML}$ by $\widetilde{K}^{CL}, \widetilde{K}^{ML}$, where $\widetilde{K}^{CL}$ and $\widetilde{K}^{ML}$ are kernel matrices over points constrained by *CL* and *ML*, respectively. Because *CL* and *ML* reflect the local information of data, we can preprocess $\widetilde{K}^{CL}$ and $\widetilde{K}^{ML}$ to keep the local information as [25], i.e., $\widetilde{K}^{ML}_{i,j} = \widetilde{K}^{ML}_{j,i} = 0, \widetilde{K}^{CL}_{i,j} = \widetilde{K}^{CL}_{j,i} = 0$ when $x_j \notin N_+^{(x_i)}$ or $x_i \notin N_+^{(x_j)}$. With all points to construct $\widetilde{K}^{ML}, \widetilde{K}^{CL}$ and

constrained data matrices, kernelized S^2LAE (KS^2LAE) can be similarly solved as S^2LAE.

Next, we show the kernelized S^2LLE using the approach of kernelizing LLE [34]. Based on the kernel matrix K, which implicitly defines a mapping ϕ from original space to a kernel space, then the reconstruction error associated with $\phi(x_i)$ in kernlized S^2LLE can be defined by the same problem in (14) with respect to $\sum_j \widetilde{\Delta}_{i,j}^{ML} = 1$. Note that $\widetilde{W}^{CL}$ here is similarly replaced by $\widetilde{K}^{(CL)}$. In kernel space, the local Gram matrix $\widehat{\aleph}^{(i)}$ with elements is defined as

$$\begin{aligned}\widehat{\aleph}_{jr}^{(i)} &= (\phi(x_i) - \phi(x_j))^{\mathrm{T}}(\phi(x_i) - \phi(x_r)) \\ &= K_{ii} - K_{ir} - K_{ji} + K_{jr}, \\ &\quad \text{where } (\phi(x_i), \phi(x_j)),\ (\phi(x_i), \phi(x_r)) \in CL,\end{aligned} \tag{19}$$

where $\phi(x_j)$ and $\phi(x_r)$ are neighbors of $\phi(x_i)$ in kernel space. Similarly, to compute the embedding for KS^2LLE, we implicitly regard if $\phi(x_i)$ and $\phi(x_j)$ are neighbors constrained by *ML*, their embeddings will be close in the dimension-reduced space; otherwise, their embeddings should be separated. The solution of KS^2LLE can be obtained by solving the same problem in (18) with respect to the matrices $\widetilde{W}_{i,j}^{CL}$ and $\widetilde{\Delta}_{i,j}^{ML}$ defined above.

3.6 Efficient Solution for Trace Ratio Problem

We in this section show how to solve the problems of S^2LAE, S^2LLE and their kernelized extensions. We take the TR problem of S^2LAE in (13) for example, which is a typical nonconvex optimization problem and no close-form solution for this TR problem exists. To solve problem (13), it is usually transformed into the following simple but inexact *ratio trace* expression [21]:

$$Y^* = \arg\max_{Y}\ Tr[(Y\widetilde{L}^{ML}Y^{\mathrm{T}})^{-1}(Y\widetilde{L}^{CL}Y^{\mathrm{T}})], \tag{20}$$

which can be solved by applying the generalized eigen-decomposition (GED). But note that the obtained solution is not necessarily able to optimally solve the original TR problem and is not orthogonal [29]. As stated in [29], in linearized cases, when evaluating the similarities between data points based on euclidean distance, the nonorthogonal projections may put different weights on different projection directions thus changing the similarities, while for orthogonal projections, such similarities can be preserved. Hence, TR optimization is empirically better than the GED. Guo et al. [28] show that the global optimum of TR problem can be equivalently solved by using a *trace difference* (TD) problem. To find the best TR value λ^* and Y^*, it is equivalent to solve a TD problem, that is to find the zero point of $F(\lambda) = \arg\max_{YY^{\mathrm{T}}=I} Tr(Y(\widetilde{L}^{CL} - \lambda\widetilde{L}^{ML})Y^{\mathrm{T}}) = 0$. Thus, the optimal matrix Y^* is given by

$$Y^* = \arg\max_{YY^{\mathrm{T}}=I} Tr[Y(\widetilde{L}^{CL} - \lambda^*\widetilde{L}^{ML})Y^{\mathrm{T}}]. \tag{21}$$

Note that in TR optimization, the orthogonal constraint is always assumed. Another iterative method called ITR [29] is recently proposed to solve the TR problem. ITR tackles the TR problem by directly optimizing the objective $Tr(Y_v\widetilde{L}^{CL}Y_v^{\mathrm{T}})/Tr(Y_v\widetilde{L}^{ML}Y_v^{\mathrm{T}})$ when the row vectors of Y_v are orthogonal together. Given λ^v at step v, Y_v can be obtained from the following problem:

$$Y^v = \arg\max_{YY^{\mathrm{T}}=I} Tr(Y(\widetilde{L}^{CL} - \lambda^v\widetilde{L}^{ML})Y^{\mathrm{T}}), \tag{22}$$

and renew λ^{v+1} as the trace ratio value given by Y_v: $\lambda^{v+1} = Tr(Y_v\widetilde{L}^{CL}Y_v^{\mathrm{T}})/Tr(Y_v\widetilde{L}^{ML}Y_v^{\mathrm{T}})$ until convergence. Theoretical analyses show that ITR can converge to the global optimum [29]. As shown in [33], the eigen-decomposition step of ITR is very time consuming. Recently, a fast trace solver of the TR problem was proposed in [33]. The algorithm in [33] adopts an effective method to accelerate the convergence speed, so the time complexity is greatly reduced and the algorithm is proved to be faster than ITR [33]. In the study, the algorithm in [33] is used to solve the TR problems of our methods. In summary, the computational procedures can be performed as follows:

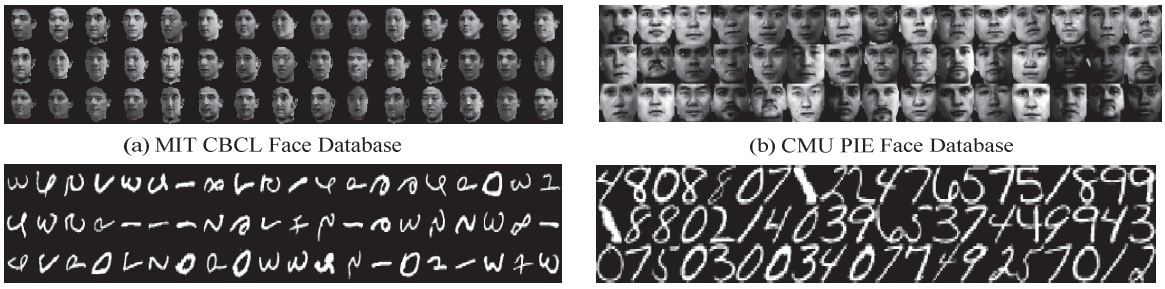
1. Initialize Y_0 as an arbitrary rowly orthogonal matrix such that $Y_0Y_0^{\mathrm{T}} = I$ and $v = 1$;
2. Repeat Steps 3 to 5 until convergence of the algorithm;
3. Compute $\lambda^v = Tr(Y_{v-1}\widetilde{L}^{CL}\ Y_{v-1}^{\mathrm{T}})/Tr(Y_{v-1}\widetilde{L}^{ML}\ Y_{v-1}^{\mathrm{T}})$;
4. Compute the d eigenvectors $\{\pi_\delta^v\}_{\delta=1}^d$ of $\widetilde{L}^{CL} - \lambda_v\widetilde{L}^{ML}$. Set $\eta = \lambda_v$ and repeat the following operations until there is no change to Y_v:
   a. Sort $(\pi_\delta^v)^{\mathrm{T}}(\widetilde{L}^{CL} - \lambda_v\widetilde{L}^{ML})\pi_\delta^v,\ \delta = 1, 2, \ldots, d$ in descending order and select the transpose of first d' eigenvectors to construct the matrix Y_v;
   b. Compute $\eta = Tr(Y_v\widetilde{L}^{CL}Y_v^{\mathrm{T}})/Tr(Y_v\widetilde{L}^{ML}Y_v^{\mathrm{T}})$;
5. Update $v = v + 1$;
6. Output $\lambda^* = \eta$ and the optimal matrix $Y^* = Y_v$.

Note that kernelized methods heavily rely on the kernels and parameters. In this work, we mainly evaluate the S^2LAE and S^2LLE algorithms for visualization.

4 Simulation Results and Analysis

Extensive settings are prepared to verify the efficiency of S^2LAE and S^2LLE. The performance of S^2LAE, S^2LLE is compared with PCA, LDA, LAE, LLE, SLLE [27], Hessian LLE (HLLE) [9], ISOMAP, S-ISOMAP, LTSA [11], SS-LLE, SS-LTSA, MMP, SSML, and SSDR. For LAE and LPP, the heat kernel, i.e., $\exp(-\|x_i - x_j\|^2/t)$, with $t = 5$ is used to define the adjacency matrix. The simple-minded method is used to define the adjacency matrix for MMP and our methods. For S-ISOMAP, parameters α and β applied to define the inter- and intraclass dissimilarity are set to 0.5 and the average euclidean distance between all pairs of samples, respectively. For our methods, ℓ_- is set to 0.5. A regularization term μI with $\mu = 0.001$ is applied in the GED type methods. For visualization, we mainly evaluate the embeddings in terms of interclass separability and intraclass compactness. We also numerically evaluate the embeddings using clustering similarity evaluation metric. All simulations were performed on a PC with Intel Core i5 CPU 650 at 3.20 GHz 3.19 GHz 4 G.

In our study, four real databases, including MIT CBCL face database (Available from http://cbcl.mit.edu/software-datasets/FaceData2.html), MNIST database (Available from http://yann.lecun.com/exdb/mnist/), CMU PIE database



(a) MIT CBCL Face Database (b) CMU PIE Face Database

(c) MNIST Handwritten Digits Database (d) USPS Handwritten Digits Database

Fig. 3. Typical sample images from the tested real-world databases.

[14], and USPS database [13] are presented. Some typical sample images of the databases are shown in Fig. 3.

4.1 Similarity Evaluation Metric

The clustering evaluation metric [32] is used for evaluating the performance of our methods. The clustering performance is evaluated by comparing the obtained cluster label of each digit or face data with that provided by the data corpus. Given a sample x_i, let r_i and f_i be the obtained cluster label and the class label provided by the data corpus. The clustering accuracy is defined as

$$AC = \frac{\sum_{i=1}^{N} \delta(f_i, Map(r_i))}{N}, \tag{23}$$

where N is the total amount of data, $\delta(p, q)$ is the delta function which equals one if $p = q$ and equals zero otherwise, and $Map(r_i)$ is the permutation mapping function, mapping each cluster label r_i to the equivalent class label from data corpus. The best mapping can be found by the Kuhn-Munkres algorithm [31]. Let C be the set of clusters obtained from the ground truth and C' from our method. Their mutual information metric $MI(C, C')$ is defined by

$$MI(C, C') = \sum_{c_i \in C, c'_j \in C'} Pr(c_i, c'_j).log_2 \frac{Pr(c_i, c'_j)}{Pr(c_i).Pr(c'_j)}, \tag{24}$$

where $Pr(c_i)$ and $Pr(c'_j)$ are, respectively, the probabilities that a point randomly selected from the data corpus belongs to the clusters c_i and c'_j, and $Pr(c_i, c'_j)$ is the joint probability that the arbitrarily selected point belongs to the clusters c_i and c'_j. $MI(C, C')$ takes values between zero and $\max(H(C), H(C'))$ as inputs, where $H(C)$ and $H(C')$ are the entropies of C and C'. In order to simplify the comparisons between different pairs of cluster sets, the normalized mutual information (Mutual_I) metric is employed, which is described as

$$\overline{MI}(C, C') = \frac{MI(C, C')}{\max(H(C), H(C'))}. \tag{25}$$

So, it is straightforward to check $\overline{MI}(C, C')$ ranges from 0 to 1, i.e., $\overline{MI}$ equals to one if the two sets of clusters are identical, and zero if they are completely independent.

4.2 Face Manifold Visualization Analysis

4.2.1 MIT CBCL Face Recognition Database

The MIT CBCL face recognition database consists of face images of 10 persons. In this study, the synthetic face set (324 images per person) from the database is tested. The images are captured under different illuminations, poses (up to about 30 degrees of rotation in depth and backgrounds. Each image is denoted by a 1,024-dimensional vector in the image space. In our simulations, 150 faces per person are randomly selected for manifold visualization. We test each method on the whole sampled set, i.e., 10 persons (1,500 images totally). Number k in NNS is set to 95 for S^{2}LAE and twice times for S^{2}LLE. For the other methods, k is fixed to 35. For SSML, SSDR and our methods, 50 percent constrains, which are randomly selected from the constraint sets, are applied in all simulations if without special remarks. In all our simulations, 40 images per person or digit of each data set are randomly selected as labeled in MMP and are used to compute the prior information for SS-LLE and SS-LTSA. The rest images are treated as unlabeled. The result of each method is illustrated in Fig. 4. We apply the clustering evaluation induced by the k-means clustering algorithm to compare the performance of each method. The clustering evaluation process is described as follows: First, after using DR to embed the face data into a low-dimensional face subspace, k-means clustering algorithm is applied. The cluster number, k, in k-means algorithm is set to the number of persons. For each setting, k-means algorithm is applied 100 times with different initializations, and the averaged clustering accuracies and Mutual_I over first 30 best results are recorded.

Observing from the visualizations, it is clear that SSDR, our S^{2}LAE and S^{2}LLE methods implicitly emphasize the natural clusters of faces and deliver the separated clusters between dissimilar faces. It can be noticed that LDA and MMP also work well in achieving enhanced interface separation and intraface compactness, but they project most persons to a compact area. This may result in relatively high clustering evaluation errors, while our S^{2}LAE and S^{2}LLE produce large margins between different faces and achieve enhanced compactness of intraperson faces. LAE, LLE, SLLE, S-ISOMAP, SS-LLE, LTSA, SS-LTSA, and SSML embed some faces appropriately, but most faces are mixed in their embeddings. PCA, ISOMAP, and HLLE perform relatively poorly on this data set, because they are incapable of identifying and separating different faces.

We report the clustering results in Table 1, in which *NumF* is face class number. From Table 1, we obtain similar observations to the visual results in Fig. 4. S^{2}LAE works

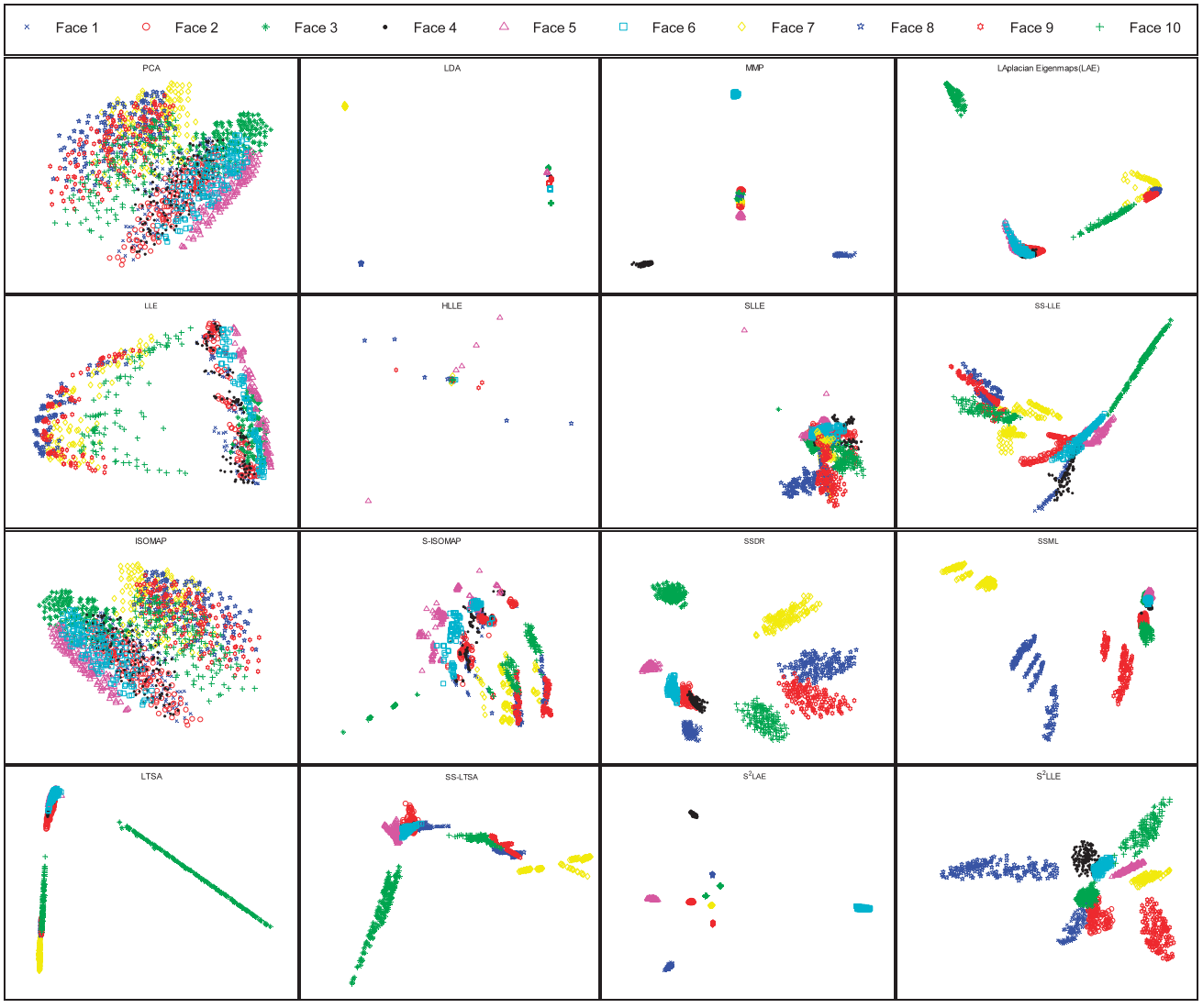


Fig. 4. The 2D manifold embedding obtained by each method on the MIT face database (10 persons).

TABLE 1
Performance Comparisons on the Real MIT CBCL, CMU PIE, MNIST, and USPS Databases

Result / Method	Simulation Setting							
	MIT CBCL(D=1024,d=2, N=1500, NumF=10)		*CMU (D=1024, d=2, N=1428, NumF=11)*		*UMIST (Dim=784, d=2, N=1500, NumD=10)*		*USPS (Dim=256, d=2, N=1500, NumD=10)*	
	Accuracy	Mutual_I	Accuracy	Mutual_I	Accuracy	Mutual_I	Accuracy	Mutual_I
PCA	0.4723	0.4973	0.4551	0.4260	0.5297	0.5398	0.5227	0.5351
LDA	0.8567	0.8598	0.8776	0.9153	0.5343	0.5559	0.8234	0.8541
MMP	0.7706	0.8350	0.8121	0.8962	0.6915	0.7102	0.8179	0.8350
LAE	0.5846	0.6604	0.4669	0.4840	0.6729	0.7259	0.6954	0.7282
LLE	0.5689	0.5790	0.4058	0.3824	0.6280	0.6472	0.7062	0.7651
HLLE	0.2600	0.2357	0.4394	0.4390	0.6089	0.6225	0.5449	0.5180
SLLE	0.6717	0.6698	0.6618	0.7165	0.5848	0.6330	0.4388	0.4164
SS-LLE	0.7767	0.8229	0.5768	0.5615	0.6774	0.7203	0.6898	0.7345
ISOMAP	0.4755	0.4951	0.4256	0.4357	0.5153	0.5276	0.5207	0.5357
S-ISOMAP	0.6878	0.6632	0.4617	0.4491	0.5125	0.5146	0.4208	0.4580
LTSA	0.4454	0.5778	0.4151	0.4170	0.6881	0.7103	0.6353	0.7017
SS-LTSA	0.7015	0.6890	0.4789	0.5021	0.6768	0.7116	0.7115	0.6892
SSML	0.6980	0.7151	0.6581	0.7468	0.5369	0.5612	0.4858	0.5126
SSDR	0.8126	0.8583	0.8268	0.8864	0.6785	0.7226	0.8129	0.8575
S^2LAE	0.8681	0.9229	0.9162	0.9249	0.8580	0.8742	0.8560	0.9138
S^2LLE	0.8041	0.8687	0.8671	0.9012	0.8072	0.8432	0.8110	0.8739

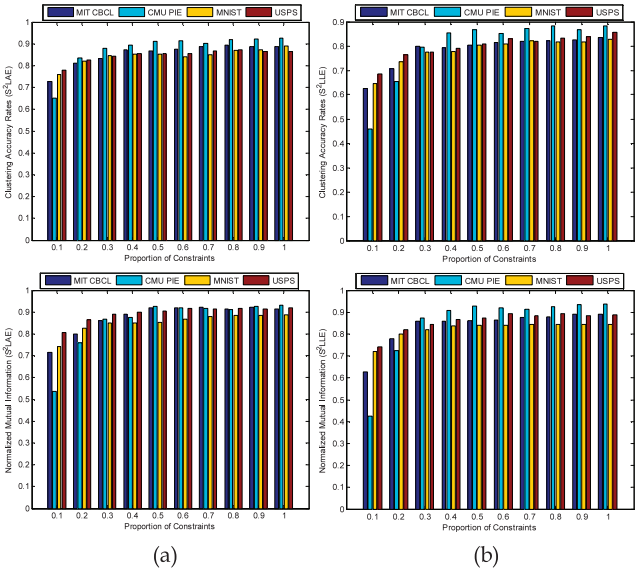


Fig. 5. Clustering results versus proportion of constraints on the four real databases. (a) S^2LAE and (b) S^2LLE.

remarkably well via delivering the highest clustering accuracy rates and mutual information. Also, SSDR and S^2LLE obtain comparable results to LDA. The results of SLLE and S-ISOMAP are comparative, and the performance of SS-LLE, SSML, and SS-LTSA is also comparable. HLLE delivers the lowest clustering accuracy compared with other methods. To show the effect of different proportions of constraints on the clustering performance, we show the clustering results of our methods based on different constraints in Fig. 5. It must be noted that the same simulation setting are used in these cases. For fixed proportion of constraints, the results are averaged by 30 random selections of constraints. It can be found that the overall clustering accuracies and mutual information increase with the increasing proportion of constraints. More importantly, S^2LAE and S^2LLE exhibit satisfactory results by applying relatively small proportion of constraints.

To demonstrate the locality preserving power of S^2LAE and S^2LLE, we choose the images of the first two persons (324 images per person) from the original face set as illustration. The results are compared with LAE and LLE. For LAE and S^2LAE, the k number in NNS is set to 85, and twice times for S^2LLE. The k number is set to 30 for LLE. The 2D embeddings are shown in Fig. 6. We can see that LAE, S^2LAE, and S^2LLE can deliver more separated manifolds and can organize the natural clusters of the faces. We can also conclude that S^2LAE and S^2LLE are capable of preserving the intrinsic local information of faces. From Fig. 6, it is clear that the poses and lighting conditions of the faces change continuously and smoothly from frontal to side, from dark to light.

4.2.2 CMU PIE Face Database

The CMU PIE database contains 68 individuals with 41,368 face images as a whole. The face images were captured under varying poses, illuminations and expressions. The sampled set, which is publicly available at http://www.zjucadcg.cn/dengcai/Data/FaceData.html/Pose27(lights-change) is tested. In this set, the pose and expressions are fixed, and there are 21 images per individual of 68 persons (totally 1,428 images) under different lighting conditions.

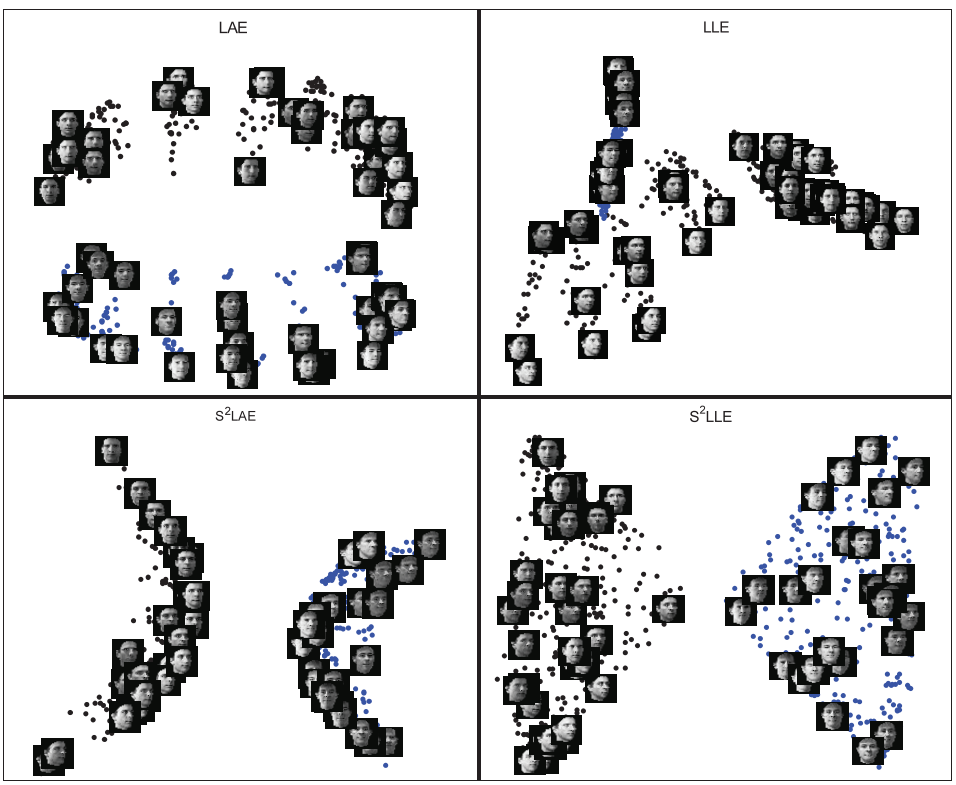


Fig. 6. Two-dimensional embedding of MIT CBCL face recognition database (two persons).

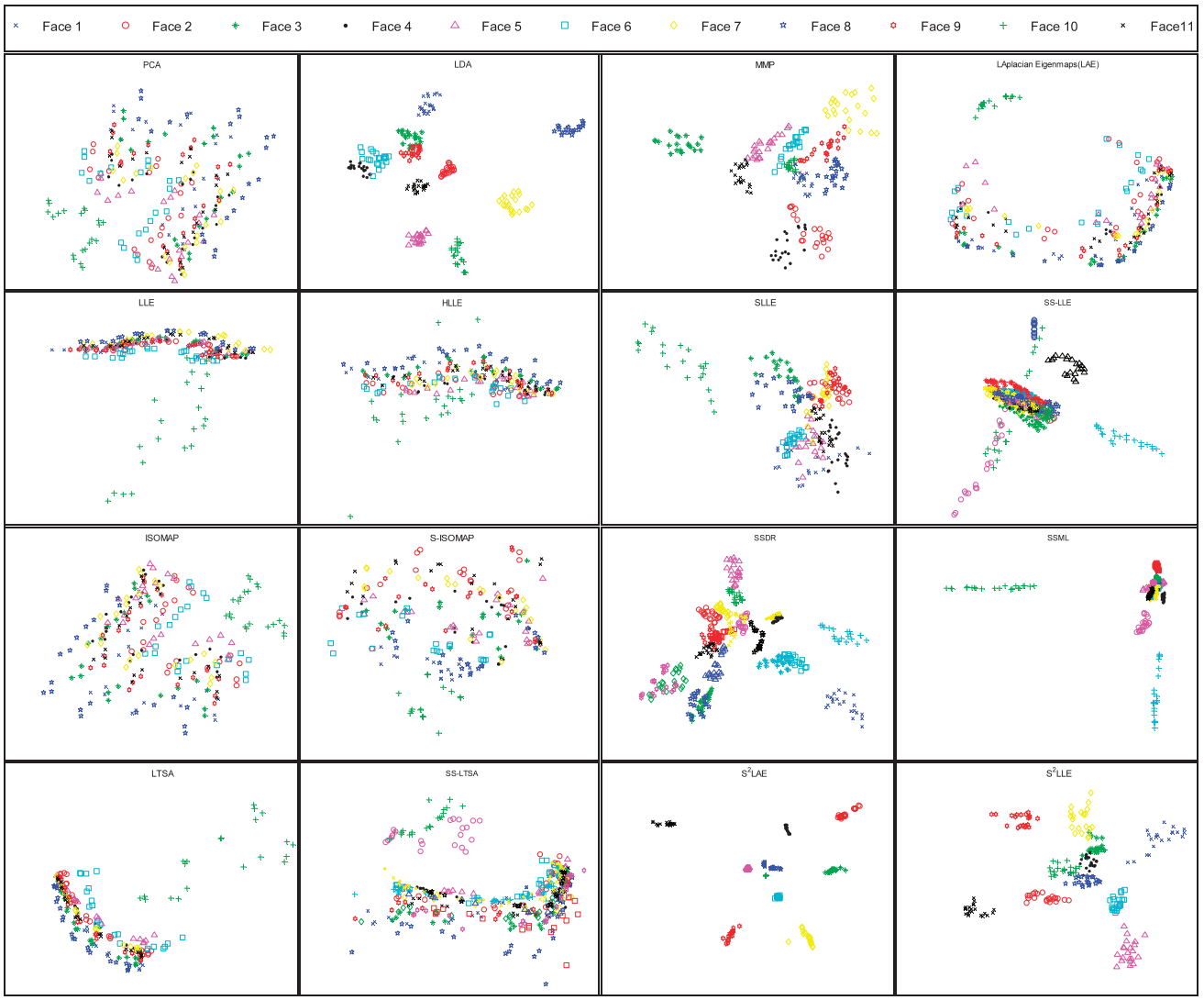


Fig. 7. The 2D manifold embedding obtained by each method on the CMU PIE database (11 persons).

Each face is denoted by a 1,024-dimensional vector in the image space. We choose 11 persons for this study. The number k in NNS is set to 45 for S^2LAE and S^2LLE. For other methods, k is set to 15. The sampled set is preprocessed by PCA to reduce the dimensionality of the data set to 200. The results are illustrated in Fig. 7. We obtain the following observations: 1) LDA, MMP, SSDR, S^2LAE, and S^2LLE can exhibit clear separate face manifolds of the 11 persons by comparing with other methods. Note that LDA and S^2LAE deliver the largest margins between different faces by organizing enhanced intraface compactness. 2) SLLE and SSML perform better than the remaining methods. 3) Unsatisfactory results are produced by PCA, LAE, LLE, ISOMAP, HLLE, LTSA, S-ISOMAP, SS-LLE, and SS-LTSA, because they are unable to identify the interperson faces. The clustering results are summarized in Table 1. We compute the averaged clustering accuracy rates and Mutual_I over the first 30 best records of 100 times initializations. From Table 1, we can observe that 1) Our S^2LAE, S^2LLE, and SSDR achieve the competitive or even better results to the supervised LDA and MMP. 2) The results obtained by PCA, LAE, LLE, ISOMAP, HLLE, LTSA, S-ISOMAP, SS-LLE, and SS-LTSA are close. SLLE and SSML deliver the higher accuracy rates and Mutual_I compared to these methods. The clustering results against the proportion of constraints are illustrated in Fig. 5, which shows the increasing proportion of constraints can enhance the clustering performance.

4.3 Handwritten Digital Manifold Visualization

4.3.1 MNIST Handwritten Digit Database

The MNIST database has 70,000 handwritten digit images. Each image has 28×28 pixels, so each image is denoted by a 784-dimensional vector. This study tests each method through visualizing the digits and comparing their clustering results. We randomly choose 150 images from digits "0-9" for simulations. The k value in NNS is set to 145 for our methods and is set to 35 for the other NNS type methods. We show the 2D digital embeddings in Fig. 8. The clustering results are described in Table 1, where *NumD* is the digital class number. The k-means clustering algorithm is again used for clustering evaluation. The averaged accuracy and Mutual_I

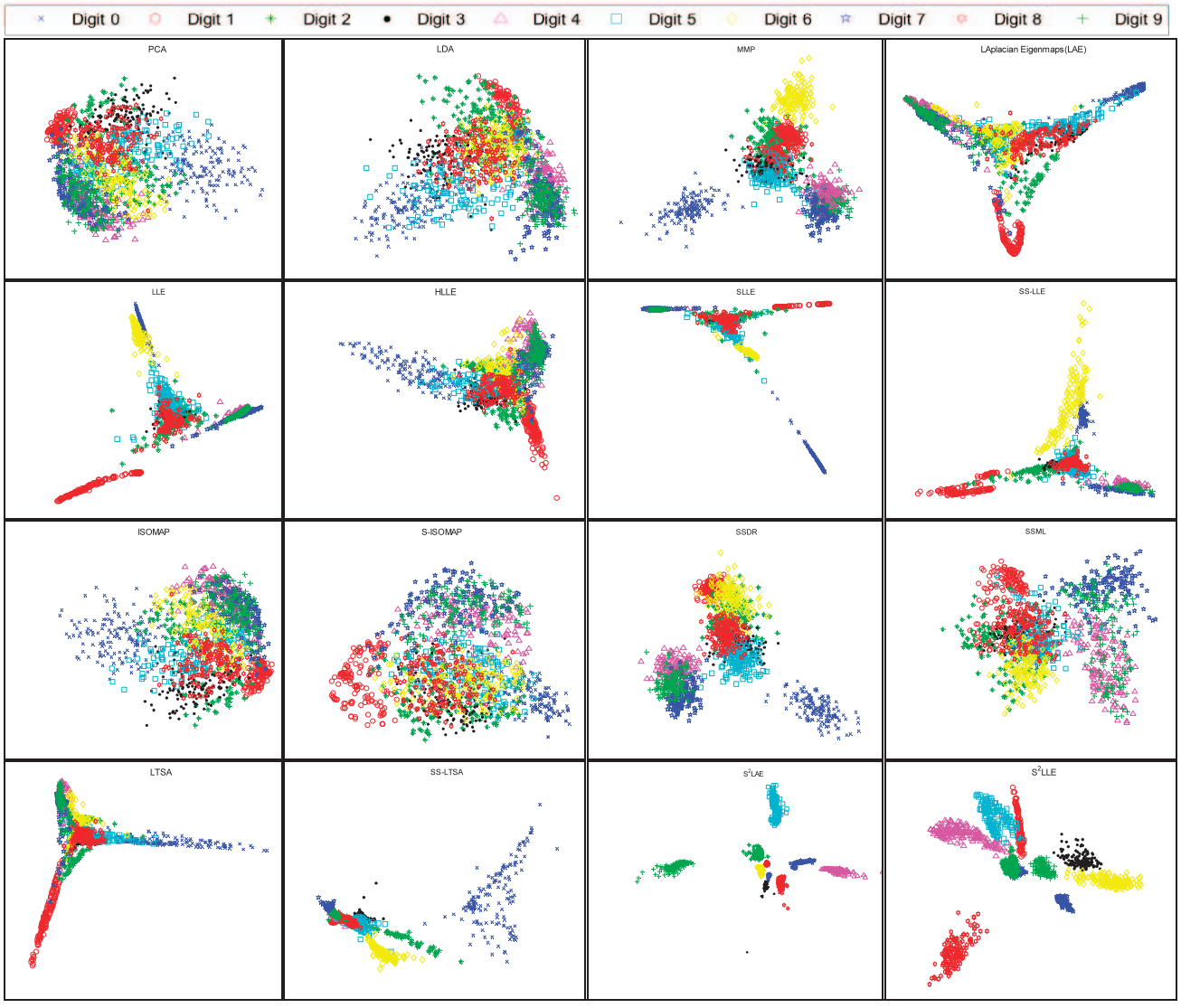


Fig. 8. The 2D manifold embedding obtained by each method on the MNIST database (digits "0-9").

of each method over the first 30 records are reported. We obtain the following observations:

1. Unsatisfactory visualization results appear for PCA, LDA, ISOMAP, HLLE, S-ISOMAP and SSML. They are unable to embed the intrinsic related low-dimensional digital manifolds respectably, because digital data of similar digits tend to have similar embeddings in the reduced output space. As a result, these digits are likely to be projected in their close vicinity, resulting in increasing errors in clustering evaluations.
2. LAE exhibits similar embedding to LLE, SLLE, LTSA, and SS-LLE. Note that they all fail to implicitly emphasize the natural clusters of digits and are unable to deliver separated clusters between dissimilar digit images.
3. Comparing with other techniques, the visual measurements reveal the strong performance of S^2LAE and S^2LLE. Most importantly, the results show that the intrinsic multimodal structures are effectively preserved. We also find that MMP, SS-LTSA, and SSDR can separate some digits from other clusters, but most of the digits are still mixed with each other.
4. The clustering results in Table 1 are consistent with the visual results of Fig. 8. Results show that, among all tested methods, S^2LAE and S^2LLE deliver the highest clustering accuracies and Mutual_I. The results of LAE, LTSA, LLE, SLLE, SS-LLE, SS-LTSA, and MMP are comparable and are higher than those of the remaining methods. The clustering results against proportions of constraints are illustrated in Fig. 5, which show similar result when the proportion of constraints increases.

4.3.2 USPS Handwritten Digit Database

The USPS database consists of 9,298 handwritten digits ("0-9"). In this study, the sample set available at http://cs. nyu.edu/~roweis/data.html is tested. The images are 16×16 pixels in 8-bit grayscale images of "0" through "9," and each digit has 1,100 images. So, each image is represented by a 256-dimensional vector in the original space. Some examples are shown in Fig. 3. As similar

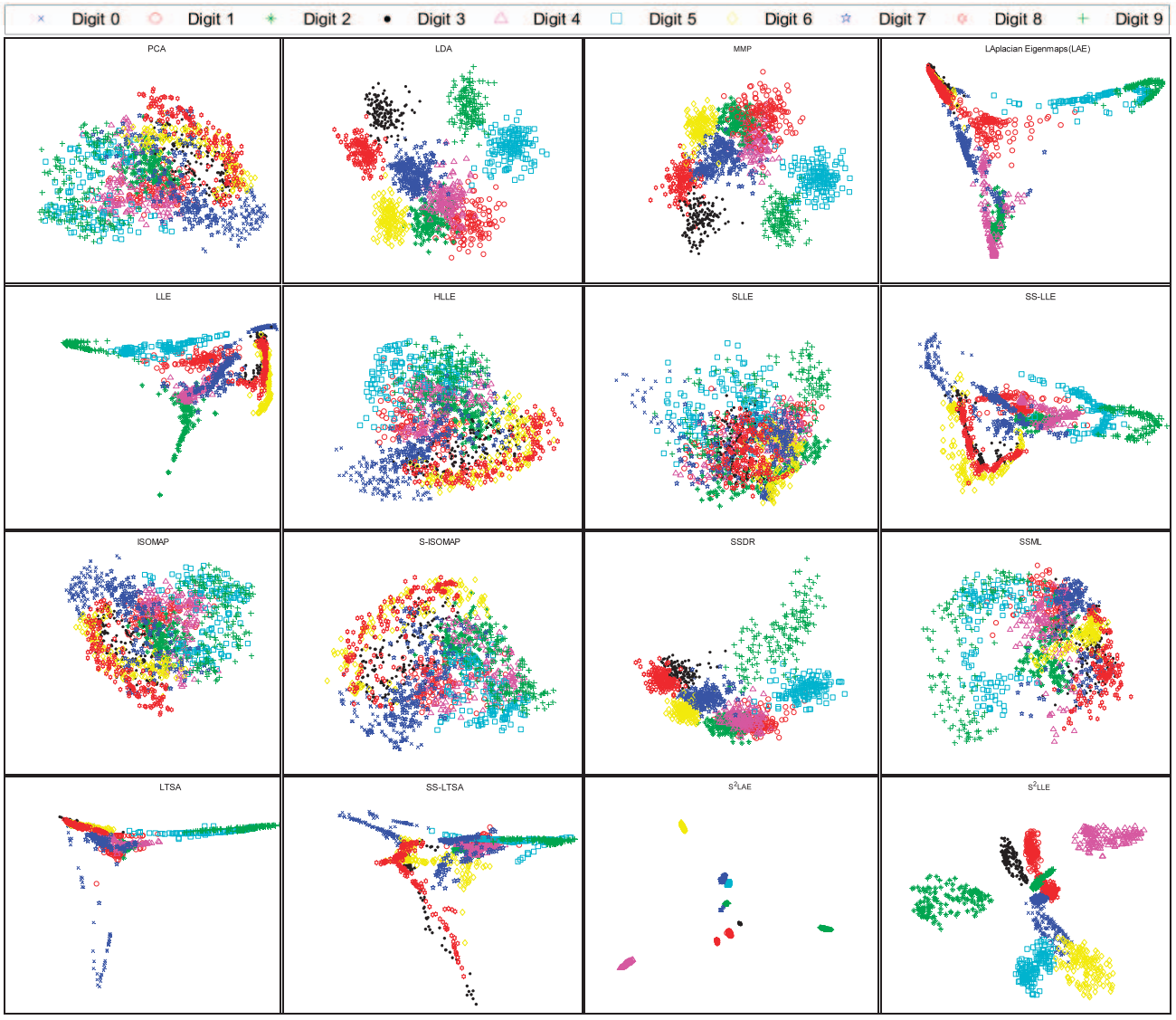


Fig. 9. The 2D manifold embedding obtained by each method on the USPS database (digits "0-9").

digits have similar embeddings, the data of similar digit exhibit similar manifolds. In this work, we randomly choose 150 images per digit for our simulations. The k value in NNS is set to 145, and 25 for our methods and other NNS type methods, respectively. The clustering evaluation system setting is the same as described in the above. We apply each method to visualize the digit data and compare their embeddings. Fig. 9 depicts the 2D embeddings of digits ("0-9"). We observe that

1. PCA, LAE, LLE, HLLE, ISOMAP, SLLE, S-ISOMAP, SS-LLE, SS-LTSA, and SSML cannot produce high interdigit separation and enhanced intradigit compactness, although they all seem to be able to preserve certain intrinsic structure characteristics.
2. LDA, MMP, and SSDR separate most of the digits, but they are incapable of improving the tightness of the same digits at the same time.
3. LAE, SS-LLE, LTSA, and SS-LTSA are capable of capturing the intrinsic manifold structures of the digits to some extent, but they cannot exhibit the separated embeddings of the digits and tend to mix some images of different digits into a single cluster.
4. Compared with the other methods, our S^2LAE and S^2LLE perform superiorly in organizing the enhanced intradigit compactness and interdigit separation without losing the multimodal structures.

The averaged clustering results are summarized in Table 1, which shows that S^2LAE delivers the best records compared with other methods. S^2LLE achieves comparable results to LDA, MMP, and SSDR. The performance of LTSA is slightly inferior to those delivered by LAE, LLE, SS-LLE, and SS-LTSA. PCA, ISOMAP, HLLE, SLLE and S-ISOMAP deliver the worst clustering results by comparing with the other methods. We evaluate the performance of our methods with varied proportions of constraints in Fig. 5. We find that our S^2LAE and S^2LLE methods can obtain satisfactory results with small proportion of constraints applied. Also, we find the increasing proportion of constraints can enhance the overall performance.

5 Concluding Remarks

Effective semisupervised extensions of LAE and LLE, namely S^2LAE and S^2LLE, are presented for multimodal nonlinear dimensionality reduction and marginal visualization. Different from virtually all existing discriminate manifold learning approaches, S^2LAE and S^2LLE use the neighborhood graph-induced pariwise constraints, which are derived from the labels of points, to guide the manifold learning. In extracting the representative features, S^2LAE and S^2LLE aim to preserve the discriminant manifold structures embedded in the pairwise constraints as well as global covariance structures of all training points. Also, S^2LAE and S^2LLE aim at preserving the local information of intraclass similarity pairs, and separating the embeddings of interclass neighbors. To solve the problems of S^2LAE and S^2LLE efficiently, the orthogonal trace ratio optimization is applied, leading to a specific solution with orthogonal projection axes. Based four real data sets, the manifold visualizations indicate that our S^2LAE and S^2LLE methods can provide more separations for the embeddings of multiple objects. Facial and handwritten digital data embeddings show S^2LAE and S^2LLE can reveal the local manifold and multimodal characteristics of faces and digits effectively. Because of the stronger constraints brought by the pairwise constraints, margins of different faces or digits are significantly enlarged in the projected spaces of S^2LAE and S^2LLE. These margins are larger than those produced by many previous supervised and semi-supervised algorithms. The clustering evaluations also examined the efficiency of our techniques. The numerical results show that our algorithms deliver better results than many state-of-the-art dimensionality reduction and data visualization techniques. We also observe that our S^2LAE and S^2LLE deliver satisfactory results by using relatively smaller proportions of constraints in most cases. For all k nearest neighbor search type algorithms, including our S^2LAE and S^2LLE, exploring determining an optimal k value for locality or neighborhood preservation is still an open problem. Another important future work is to theoretically extend our algorithms to linearized scenarios for handing pattern classification problems.

Acknowledgments

This work was partially supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (CityU8/CRF/09).

References

[1] S. Roweis and L. Saul, "Nonlinear Dimensionality Reduction by Locally Linear Embedding," *Science,* vol. 290, pp. 2323-2326, 2000.
[2] G.E. Hinton and R.R. Salakhutdinov, "Reducing the Dimensionality of Data with Neural Networks," *Science,* vol. 313, no. 5786, pp. 504-507, 2006.
[3] A.P. Dempster, "An Overview of Multivariate Data Analysis," *J. Multivariate Analysis,* vol. 1, no. 3, pp. 316-346, 1971.
[4] A.M. Martinez and A.C. Kak, "PCA versus LDA," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 23, no. 2, pp. 228-233, Feb. 2001.
[5] J.P. Maaten, E.O. Postma, and H.J. Herik, "Dimensionality Reduction: A Comparative Review," Technical Report TiCC-TR 2009-005, Tilburg Univ., 2009.
[6] Y. Yang, F.P. Nie, S.M. Xiang, Y.T. Zhuang, and W.H. Wang, "Local and Global Regressive Mapping for Manifold Learning with Out-of-Sample Extrapolation," *Proc. Am. Assoc. for Artificial Intelligence Conf.,* pp. 649-654, 2010.
[7] D. Zhou, O. Bousquet, T. Lal, J. Weston, and B. Schölkopf, "Learning with Local and Global Consistency," *Proc. Advances in Neural Information Processing Systems,* pp. 321-328, 2004.
[8] Y. Liu and J. Rong, "Distance Metric Learning: A Comprehensive Survey," technical report, Michigan State Univ., 2006.
[9] D. Donoho and C. Grimes, "Hessian Eigenmaps: Locally Linear Embedding Techniques for High-Dimensional Data," *Proc. Nat'l Academy Sciences USA,* vol. 100, pp. 5591-5596, 2003.
[10] M. Belkin and P. Niyogi, "Laplacian Eigenmaps for Dimensionality Reduction and Data Representation," *Neural Computation,* vol. 15, no. 6, pp. 1373-1396, 2003.
[11] Z.Y. Zhang and H.Y. Zha, "Principal Manifolds and Nonlinear Dimension Reduction by Local Tangent Space Alignment," *SIAM J. Scientific Computing,* vol. 26, no. 1, pp. 313-338, 2005.
[12] B. Schölkopf and A. Smola, *Learning with Kernels.* MIT Press, 2002.
[13] J.J. Hull, "A Database for Handwritten Text Recognition Research," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 16, no. 5, pp. 550-554, May 1994.
[14] T. Sim, S. Baker, and M. Bsat, "The CMU Pose, Illumination, and Expression Database," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 25, no. 12, pp. 1615-1618, Dec. 2003.
[15] D.Q. Zhang, Z.H. Zhou, and S.C. Chen, "Semi-Supervised Dimensionality Reduction," *Proc. SIAM Int'l Conf. Data Mining,* pp. 629-634, 2007.
[16] X. Yang, H.Y. Fu, H.Y. Zha, and L. Jesse, "Semi-Supervised Nonlinear Dimensionality Reduction," *Proc. Int'l Conf. Machine Learning,* pp. 1065-1072, 2006.
[17] Z.Y. Zhang, H.Y. Zha, and M. Zhang, "Spectral Methods for Semi-Supervised Manifold Learning," *Proc. Int'l Conf. Computer Vision and Pattern Recognition,* pp. 1-6, 2008.
[18] A. Bar-Hillel, T. Hertz, N. Shental, and D. Weinshall, "Learning a Mahalanobis Metric from Equivalence Constraints," *J. Machine Learning Research,* vol. 6, pp. 937-965, 2005.
[19] M.S. Baghshah and S.B. Shouraki, "Semi-Supervised Metric Learning Using Pairwise Constraints," *Proc. Int'l Joint Conf. Artificial Intelligence,* pp. 1217-1222, 2009.
[20] E. Kokiopoulou1, J. Chen, and Y. Saad, "Trace Optimization and Eigenproblems in Dimension Reduction Methods," *Numerical Linear Algebra with Applications,* vol. 18, pp. 565-602, 2011.
[21] K. Fukunaga, *Introduction to Statistical Pattern Recognition,* second ed. Academic Press, 1991.
[22] D.Q. Zhang, S.C. Chen, and Z.H. Zhou, "Constraint Score: A New Filter Method for Feature Selection with Pairwise Constraints," *Pattern Recognition,* vol. 41, no. 5, pp. 1440-1451, 2008.
[23] L. Zelnik-Manor and P. Perona, "Self-Tuning Spectral Clustering," *Proc. Advances in Neural Information Processing Systems,* pp. 1601-1608, 2005.
[24] S.C. Yan, D. Xu, B.Y. Zhang, H.J. Zhang, Q. Yang, and S. Lin, "Graph Embedding and Extensions: A General Framework for Dimensionality Reduction," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 29, no. 1, pp. 40-51, Jan. 2007.
[25] Y. Guo, J. Gao, and P.W. Kwan, "Kernel Laplacian Eigenmaps for Visualization of Non-Vectorial Data," *Proc. Australian Joint Conf. Artificial Intelligence: Advances in Artificial Intelligence,* pp. 1179-1183, 2006.
[26] X. Geng, D.C. Zhan, and Z.H. Zhou, "Supervised Nonlinear Dimensionality Reduction for Visualization and Classification," *IEEE Trans. Systems, Man, and Cybernetics-Part B: Cybernetics,* vol. 35, no. 6, pp. 1098-1107, Dec. 2005.
[27] D. Ridder, O. Kouropteva, O. Okun, M. Pietikäinen, and R.P.W. Duin, "Supervised Locally Linear Embedding," *Proc. Int'l Conf. Artificial Neural Networks,* pp. 333-341, 2003.
[28] Y. Guo, S. Li, J. Yang, T. Shu, and L. Wu, "A Generalized Foley-Sammon Transform Based on Generalized Fisher Discriminant Criterion and Its Application to Face Recognition," *Pattern Recognition Letter,* vol. 24, nos. 1-3, pp. 147-158, 2003.

[29] H. Wang, S. Yan, D. Xu, X. Tang, and T. Huang, "Trace Ratio vs. Ratio Trace for Dimensionality Reduction," *Proc. Int'l Conf. Computer Vision and Pattern Recognition*, pp. 1-8, 2007.
[30] Y. Jia, F. Nie, and C. Zhang, "Trace Ratio Problem Revisited," *IEEE Trans. Neural Network*, vol. 20, no. 4, pp. 729-735, Apr. 2009.
[31] L. Lovász and M.D. Plummer, *Matching Theory*. North Holland, 1986.
[32] D. Cai, X. He, and J.W. Han, "Document Clustering Using Locality Preserving Indexing," *IEEE Trans. Knowledge and Data Eng.*, vol. 17, no. 12, pp. 1624-1637, Dec. 2005.
[33] Y. Yang, F. Nie, D. Xu, J. Luo, Y. Zhuang, and Y. Pan, "A Multimedia Retrieval Framework Based on Semi-Supervised Ranking and Relevance Feedback," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 34, no. 4, pp. 723-742, Apr. 2012.
[34] O. Kayo, "Locally Linear Embedding Algorithm. Extensions and Applications," PhD thesis, Univ. of Oulu, Finland, 2006.
[35] X.F. He, D. Cai, and J. Han, "Learning a Maximum Margin Subspace for Image Retrieval," *IEEE Trans. Knowledge and Data Eng.*, vol. 20, no. 2, pp. 189-201, Feb. 2008.
[36] J.B. Tenenbaum, V. Silva, and J.C. Langford, "A Global Geometric Framework for Nonlinear Dimensionality Reduction," *Science*, vol. 290, no. 5500, pp. 2319-2323, 2000.

Zhao Zhang (S'11–) received the BEng (First Hons.) and master's degrees from the Department of Computer Science and Technology, Nanjing Forestry University, Nanjing, PR China, in 2008 and 2010, respectively. He is currently working toward the PhD degree in the Department of Electronic Engineering, City University of Hong Kong, Kowloon, Hong Kong SAR. He was a visiting research engineer at the Learning & Vision Research Group, Department of Electrical and Computer Engineering, National University of Singapore, under advisor Prof. Shuicheng Yan, from February to May 2012. He then visited the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences, under advisor Prof. Cheng-Lin Liu, from September to December 2012. His current interests are pattern recognition, machine learning, computer vision, and data mining. He is a student member of the IEEE.

Tommy W.S. Chow (M'93-SM'03–) received the BSc (First Hons.) and PhD degrees from the University of Sunderland, Sunderland, United Kingdom. He is currently a full professor in the Electronic Engineering Department, City University of Hong Kong, Hong Kong SAR. Dr. Chow has received an IEE U.K. undergraduate scholarship in 1983, and was awarded a first class Honours degree in 1984 at the University of Sunderland, United Kingdom. He has been an active committee member of HKIE (Hong Kong Institution of Engineers) Control Automation and Instrumentation (CAI) division since 1992, and the division chairman (1997-1998) for HKIE CAI division. He received the best paper award from the 28th Annual Conference of the IEEE Industrial Electronics Society (IECON 2002). He has authored or coauthored more than 160 technical papers in international journals, 5 book chapters, and more than 60 technical papers in conference proceedings. He is now serving as associate editor of the *International Journal of Information Technology* and *Pattern Analysis and Applications*. He is a senior member of the IEEE.

Mingbo Zhao received the BSc and master's degrees from the Department of Electronic Engineering, Shanxi University, Shanxi, PR China, in 2005 and 2008, respectively. He is currently working toward the PhD degree in the Department of Electronic Engineering, City University of Hong Kong. His current interests include machine learning, data mining, and pattern recognition and its applications. He is a student member of the IEEE.