Information Sciences 324 (2015) 286–309

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Learning from normalized local and global discriminative information for semi-supervised regression and dimensionality reduction

Mingbo Zhao [a,b], Tommy W.S. Chow [b], Zhou Wu [b,*], Zhao Zhang [a], Bing Li [b]

[a] School of Computer Science and Technology, Soochow University, P. R. China
[b] City University of Hong Kong, Department of Electronic Engineering, Kowloon, Hong Kong

ARTICLE INFO

Article history:
Received 30 October 2014
Revised 16 May 2015
Accepted 12 June 2015
Available online 25 June 2015

Keywords:
Dimensionality reduction
Semi-supervised learning
Local and Global Discriminative Information

ABSTRACT

Semi-supervised dimensionality reduction is one of the important topics in pattern recognition and machine learning. During the past decade, Laplacian Regularized Least Square (LapRLS) and Semi-supervised Discriminant Analysis (SDA) are the two widely-used semi-supervised dimensionality reduction methods. In this paper, we show that SDA and LapRLS can be unified into a constrained manifold regularized least square framework. The manifold term, however, cannot fully utilize the underlying discriminative information. We thus introduce a new and effective semi-supervised dimensionality reduction method, called Learning from Local and Global Information (LLGDI), to solve the problem. The proposed LLGDI method adopts a set of local classification functions to preserve both local geometrical and discriminative information of dataset. It also adopts a global classification function to preserve the global discriminative information, and an uncorrelated constraint to calculate the projection matrix for simultaneously solving regression and dimensionality reduction problem. As a result, the LLGDI method is able to preserve local discriminative, manifold information as well as the global discriminative information. Theoretical analysis and extensive simulations presented in the paper show the effectiveness of the LLGDI algorithm. The results also demonstrate LLGDI can achieve superior performance compared with other existing methods.

1. Introduction

Dealing with high-dimensional data has always been a major problem with the research of pattern recognition and machine learning. Typical applications include face recognition, document categorization, and image retrieval. Thus, finding a low-dimensional representation of high-dimensional space is of great practical importance. The goal of dimensionality reduction is to reduce the complexity of input space and embed high-dimensional space into a low-dimensional space while keeping most of the desired intrinsic information [16,18,28,36,38-43]. Among all the dimensionality reduction techniques, Principle Compent Analysis (PCA) [19] and Linear Discriminant Analysis (LDA) [1] are the two most popular methods. PCA pursues the direction of maximum variance for optimal reconstruction, while LDA, a supervised method, finds the optimal projection V maximizing the between-class scatter matrix S_b and minimizing the within-class scatter matrix S_w in a low-dimensional subspace. Due to the utilization of label information, LDA can achieve better classification results compared with PCA given sufficient labeled samples are provided [1,4].

* Corresponding author. Tel.: +852 34422874; fax: +852 34420437.
E-mail address: wuzhsky@gmail.com (Z. Wu).

http://dx.doi.org/10.1016/j.ins.2015.06.021
0020-0255/

In general, supervised methods can deliver better performance than unsupervised methods, but obtaining sufficient number of labeled data for training can be problematic because labeling large number of samples is costly and laborious. On the other hand, unlabeled samples are abundant and can easily be obtained in numerous real world cases. Compared to supervised learning approaches that only rely on labeled training data, the idea of semi-supervised learning is to incorporate labeled and unlabeled data together to improve learning performance [2,3,12–14,44,45]. In brief, semi-supervised learning can be perceived as a framework that can provide efficient alternative to labeling unlabeled data. Well-known semi-supervised learning methods include Gaussian Field and Harmonic Fuction (GFHF) [45], Learning from Local and Global Consistency (LLGC) [44] and Special Label Propagation (SLP) [12]. These methods work in a transductive way by propagating label information from labeled set into unlabeled set through label propagation. This approach is efficient but it cannot predict class labels of new-coming samples. This drawback usually results in the out-of-sample problem. In contrast, semi-supervised dimensionality reduction methods not only reduce the dimensionality, but also naturally solve the out-of-sample problem. Thus semi-supervised methods can usually deliver better results when dealing with real-world applications.

The two widely-used semi-supervised methods are Semi-supervised Discriminant Analysis (SDA) [3] and Laplacian Regularized Least Square (LapRLS) [9]. These two methods share the same concept of dimensionality reduction, i.e. they first construct a graph Laplacian matrix to approximate the manifold structure by using both labeled and unlabeled samples. They then perform dimensionality reduction by adding the graph Laplacian matrix as a regularized term to the original objective function of LDA and Regularized Least Square (RLS). As a result, the discriminative structure embedded in the labeled samples and the geometrical structure embedded in labeled and unlabeled data can be preserved. In fact Lap-RLS is essentially derived from the perspective of regression instead of classification. Lap-RLS can be perceived as a training method that is aimed at training a linear classification model by regressing labeled set on the class label, while SDA is a subspace learning method which is aimed for solving classification problems. Though they both are stemmed from different supervised methods, we in this paper show that both SDA and Lap-RLS can be unified under a regularized least square framework. As a result, both of them are able to solve regression as well as subspace learning problems.

The connection and theoretical similarities between SDA and LapRLS can be elaborated under the least square framework. It should be noted that the regression term in LapRLS and the least square framework is supervised, which mean these two methods utilize a labeled set to train a linear classification function. Since the number of labeled data is relatively small compared with unlabeled data, training a linear classification function under a small sample size can be ineffective [21]. Another issue of semi-supervised method is the utilization of data samples to construct a graph that is used for characterizing the local structure of data manifold. In SDA and Lap-RLS, local structure is preserved by using a manifold regularized term defined on the affinity matrix of Gaussian function. But these Laplacian matrixes cannot capture the discriminative information of classes. This is essential when handling classification problems. In addition, the Gaussian function based affinity matrix is found to be over sensitive to the Gaussian variance; only a slight variation on the variance may affect the results significantly. Thus, Gaussian function based affinity matrix is not a popular method for handling complicated image classification and visualization problems. Instead of using Gaussian function for graph construction, several methods including Locally Linear Reconstruction [22,23], Local Regression and Global Alignment [30,31] and Local Spline Regression [26,27] have then been proposed.

In this paper, we introduce a newly developed method, **L**earning from **L**ocal and **G**lobal **D**iscriminative **I**nformation (**LLGDI**), for solving the above semi-supervised dimensionality reduction problems. The proposed LLGDI aims to train a classification function by utilizing all availanle data points. Specifically, our proposed method first relaxes the original supervised regression term making it a loss term and a global regression regularized term. The loss term measures the inconsistency between the predicted and initial labels on a labeled set, while the global regression regularized term aims to train the classification function as well as to calculate the projection matrix for out-of-sample problem. In addition, in order to characterize both manifold and discriminative structure embedded in a dataset, LLGDI employs a set of local classification function for each data point to predict the label of its neighboring points. In this way, both local and global discriminative information of a dataset can be preserved by using the LLGDI method. Also, in order to handle the subspace learning problem, we have also introduced an uncorrelated constraint into the objective function of LLGDI. As a result, both regression and subspace learning problems can be solved at the same time.

The main contributions of this work are as follows. First, we address the SDA method into a least square framework and establish the connections between SDA and LapRLS. Second, in order to relaxe the limitations of the least square framework of SDA, we develop a new method, called LLGDI. The new method can preserve the local geometrical and discriminative information of a dataset by using a normalized local discriminative manifold regularization term. Third, we extend the LLGDI method to perform dimensionality reduction by including a relaxed uncorrelated constraint to the objective function. As a result, both regression and subspace learning problems can be solved simultaneously. Finally, the relationship between LLGDI and other state-of-the-art methods are analyzed. Theoretical analysis shows that many other semi-supervised methods are different the special cases of the LLGDI method.

This paper is organized as follows. In Section 2, the notations and a brief review of LDA, MR and SDA are detailed. In Section 3, the equivalence between SDA and Lap-RLS under a constrained regularized least square framework is derived. Section 4 presents the proposed LLGDI method for semi-supervised regression and dimensionality reduction through the introduction of a normalized local discriminative manifold regularization term. Discussion on the relationship between LLGDI and other state-of-the-art semi-supervised methods is also included. Section 5 demonstrates the extensive simulations and the final conclusions are drawn in Section 6.

Table 1
Abbreviations of algorithms used in our work.

Abbreviations	Initial Names
LLGDI	Learning From Local and Global Information
LDA [4]	Linear Discriminant Analysis
SDA [3]	Semi-supervised Discriminant Analysis
MR [9]	Manifold Regularization
LapRLS/L [9]	Linear Laplacian Regularized Least Square
LLGC [44]	Learning from Local and Global Consistency
GFHF [45]	Gausian Field and Harmonic Fuction
SLP [12]	Special Label Propagation
FME [13,14]	Flexiable Manifold Embedding

2. Notations and review of related work

In this section, we will first give some notations used in our work and briefly review several related works, which include Linear Discriminant Analysis (LDA), Manifold Regularization (MR) and Semi-supervised Discriminant Analysis (SDA). Let $X = \{X_l, X_u\} = \{x_1, x_2, \ldots, x_{l+u}\} \in R^{D\times(l+u)}$ be the data matrix where the first l and the remaining u columns are the labeled and unlabeled samples, respectively, $Y_l = \{y_1, y_2, \ldots, y_l\} \in R^{c\times l}$ be the binary label matrix with each column y_j representing the class assignment of x_j, i.e. $y_{ij} = 1$, as the class matrix, where $y_{ij} = 1$, if x_j belongs to the *ith* class; $y_{ij} = 0$, otherwise, D and c are the numbers of features and classes, respectively. We also let $\widetilde{G} = (\widetilde{V}, \widetilde{E})$ be an undirected weighted graph, where $\widetilde{V}$ is the vertex set of $\widetilde{G}$ representing the training samples, and $\widetilde{E}$ is the edge set of $\widetilde{G}$ associated with a weight matrix W containing the local information between two nearby samples. Then, the graph Laplacian matrix that is to approximate the geometrical structure of data manifold can be defined as $L = D - W$, where D is a diagonal matrix satisfying $D_{ii} = \sum_{j=1}^{l+u} w_{ij}$. In addition, some abbreviations of algorithms used in our paper are given in Table 1.

2.1. Linear Discriminant Analysis (LDA)

The objective of LDA is to find an optimal projection matrix $V^* \in R^{D\times d}$ maximizing between-class scatter matrix while minimizing within-class scatter matrix [4]. Let we denote $G = \{g_1, g_2, \ldots, g_l\} = (YY^T)^{-1/2}Y \in R^{c\times l}$ as the scaled class indicator matrix [15], where $g_{ij} = 1/\sqrt{l_i}$, if x_j belongs to the ith class; $g_{ij} = 0$, otherwise. Since YY^T is diagonal matrix, then $GG^T = (YY^T)^{-1/2}YY^T(YY^T)^{-1/2} = I$. Hence assuming the data matrix X_l is centered, the total-class, between-class and within-class scatter matrix S_t, S_b, S_w can be defined as

$$\begin{aligned} S_t &= \sum_{i=1}^{c}\sum_{x\in c_i}(x-\mu)(x-\mu)^T = X_l X_l^T \\ S_b &= \sum_{i=1}^{c} l_i(\mu_i-\mu)(\mu_i-\mu)^T = X_l G^T G X_l^T \\ S_w &= \sum_{i=1}^{c}\sum_{x\in c_i}(x-\mu_i)(x-\mu_i)^T = X_l X_l^T - X_l G^T G X_l^T, \end{aligned} \tag{1}$$

where l_i is the number of samples in the ith class, μ_i is the mean of samples in the ith class, and μ is the mean of all labeled samples. The optimal projection matrix V^*_{LDA} is then formed by eigenvectors corresponding to the d largest eigenvalues of $S_w^{-1}S_b$ or $S_t^{-1}S_b$.

2.2. Manifold Regularization (MR)

The MR method [3] extends many methods such as least square and SVM to semi-supervised learning methods by introducing a manifold regularized term to preserve the geometrical structure. Take the linear Laplacian regularized Least Squares method (referred as Lap-RLS/L) as an example. The goal of Lap-RLS/L is to fix a linear model $y_j = V^T x_j + b^T$ by regressing X on Y and simultaneously to preserve the manifold smoothness embedded in both labeled and unlabeled set, where $V \in R^{D\times d}$ is the projection matrix and $b \in R^{1\times c}$ is the bias term. The objective function of Lap-RLS/L can be given as

$$J(V,b) = \min \sum_{j=1}^{l} \left\| V^T x_j + b^T - y_j \right\|_F^2 + \alpha_t \|V\|_F^2 + \alpha_m Tr\left(V^T X L X^T V\right), \tag{2}$$

where $L = D - W$ is the graph Laplacian matrix associated with both labeled and unlabeled sets [6], W is the weight matrix defined as: $w_{ij} = \exp(-||x_i - x_j||^2/2\sigma^2)$, if x_i is within the k nearest neighbor of x_j or x_j is within the k nearest neighbor of x_i; $w_{ij} = 0$, otherwise, D is a diagonal matrix satisfying $D_{ii} = \sum_{j=1}^{l+u} w_{ij}$, α_m and α_t are the two parameters balance the tradeoff between manifold and Tikhonov regularized terms.

2.3. Semi-supervised Discriminant Analysis (SDA)

Motivated by Manifold Regularization (MR), SDA extends the conventional LDA to preserve geometric structure by adding a manifold regularized term to the objective function of LDA. The objective function of SDA can be given by

$$J(V) = \max Tr\left\{\left(V^T\left(S_t + \alpha_t I + \alpha_m XLX^T\right)V\right)^{-1}V^T S_b V\right\}. \tag{3}$$

Similar to LDA, the optimal solution of SDA can be obtained by the Generalized Eigenvalue Decomposition as

$$S_b V^*_{\mathrm{SDA}} = \left(S_t + \alpha_t I + \alpha_m XLX^T\right)V^*_{\mathrm{SDA}}\Lambda, \tag{4}$$

where $V^*_{\mathrm{SDA}} \in R^{D\times d}$ is the solution of SDA with each column being the eigenvector of $(S_t + \alpha_t I + \alpha_m XLX^T)^{-1}S_b$, and $\Lambda \in R^{d\times d}$ is the eigenvalue matrix. It can be observed that the solution of SDA or LDA (LDA is a special case of SDA when $\alpha_m = 0$) is not unique as $V^*\Xi$ is also its solution, where $\Xi \in R^{d\times d}$ is an arbitrary diagonal matrix. Hence, to make the solution unique, a typical uncorrelated constraint can be imposed to the objective function of SDA or LDA as

$$V^T\left(S_t + \alpha_t I + \alpha_m XLX^T\right)V = I. \tag{5}$$

In this paper, we mainly focus on the solution of SDA or LDA ($\alpha_m = 0$) with the above constraint, which can be reformulated as follows:

$$\begin{aligned} J(V) = \max\ & Tr\left\{\left(V^T\left(S_t + \alpha_t I + \alpha_m XLX^T\right)V\right)^{-1}V^T XG^T GX^T V\right\} \\ & s.t.\ V^T\left(S_t + \alpha_t I + \alpha_m XLX^T\right)V = I \end{aligned}. \tag{6}$$

This problem can be solved by a technique of simultaneous diagonalization of three scatter matrices [33].

3. On the equivalence between SDA and Lap-RLS/L under uncorrelated constraint

Previous work in [41] has established a relationship between SDA and Lap-RLS/L using a least square framework, but their equivalence is not clear. In this section, we will analyze the equivalent relationship between SDA and Lap-RLS/L under an uncorrelated constraint. Specifically, we will first introduce a class labeled induced semi-supervised discriminant analysis (C-SDA). By using C-SDA as a bridge, we then establish the equivalence between SDA and Lap-RLS/L.

3.1. Class label induced semi-supervised discriminant analysis

The objective function of C-SDA is first given as

$$\begin{aligned} J(V) = \max\ & Tr\left\{\left(V^T\left(S_t + \alpha_t I + \alpha_m XLX^T\right)V\right)^{-1}V^T XY^T YX^T V\right\} \\ & \text{s.t. } V^T\left(S_t + \alpha_t I + \alpha_m XLX^T\right)V = I \end{aligned}. \tag{7}$$

It can be observed from Eqs. (6) and (7) that the objective functions of SDA and C-SDA share the similar formulation and their solutions can be expressed as the eigenvectors to the top eigenvalues of the following matrix

$$\left(S_t + \alpha_t I + \alpha_m XLX^T\right)^{-1}XH^T HX^T, \tag{8}$$

where $H = G = Y(Y^TY)^{-1/2}$ for SDA and $H = Y$ for C-SDA. We then show how to calculate the eigenvectors of Eq. (8) [33]. By performing Singular Value Decomposition (SVD) to $S_t + \alpha_t I + \alpha_m XLX^T$, we have $S_t + \alpha_t I + \alpha_m XLX^T = \Theta\sum\Theta^T$, where $\Theta \in R^{D\times D}$ is an orthogonal matrix, $\sum \in R^{D\times D}$ is a diagonal matrix. Let $B = \sum^{-1/2}\Theta^T XH^T$ and perform the SVD of BB^T as:

$$BB^T = P\sum_b P^T, \tag{9}$$

where $P \in R^{D\times q}$ is an orthogonal matrix, $\sum_b \in R^{q\times q}$ is a diagonal matrix with rank q, then we have the following lemma:

Lemma 1. *The eigenvectors to the top $d(d \le q)$ eigenvalues of $(S_t + \alpha_t I + \alpha_m XLX^T)^{-1}XH^T HX^T$ are given by $V^* = \Theta\sum^{-1}P_d$, where P_d consists of the first d columns of P.*

Proof of Lemma 1. We decompose Eq. (8) as follows:

$$\begin{aligned} \left(S_t + \alpha_t I + \alpha_m XLX^T\right)^{-1}XH^T HX^T &= \Theta\sum{}^{-1/2}\sum{}^{-1/2}\Theta^T XH^T HX^T\Theta\sum{}^{-1/2}\sum{}^{1/2}\Theta^T \\ &= \Theta\sum{}^{-1/2}BB^T\sum{}^{1/2}\Theta^T \\ &= \Theta\sum{}^{-1/2}P\sum\nolimits_b P^T\sum{}^{1/2}\Theta^T. \end{aligned} \tag{10}$$

The second equation holds as $\Theta^T\Theta = \Theta\Theta^T = I$. Let $V_G = \Theta\sum^{-1}P$, it follows:

$$\left\{\left(S_t + \alpha_t I + \alpha_m XLX^T\right)^{-1} XH^T HX^T\right\} V_G = V_G \sum_b$$

$$V_G^T\left(S_t + \alpha_t I + \alpha_m XLX^T\right)V_G = I. \tag{11}$$

The first equation indicates that V_G is the eigenvector of Eq. (8), where the eigenvalue matrix is the diagonal matrix $\sum_b$, while the second equation indicates that V_G satisfies the uncorrelated constraint. Hence we prove Lemma 1.

3.2. Equivalence between SDA and C-SDA

Lemma 1 shows that both SDA and C-SDA have the same form of solution. We next show the solutions of SDA and C-SDA are equivalent. This equivalence is based on the following lemma:

Lemma 2. *Let $B_{\mathrm{SDA}} = \sum^{-1}\Theta^T XG^T$, $B_{\mathrm{C-SDA}} = \sum^{-1}\Theta^T XY^T$, then B_{SDA} and $B_{\mathrm{C-SDA}}$ have the same range space.*

The proof of Lemma 2 is straightforward. Since $G = Y(Y^TY)^{-1/2}$, we have $B_{\mathrm{C-SDA}} = B_{\mathrm{SDA}}(Y^TY)^{1/2}$ or $B_{\mathrm{SDA}} = B_{\mathrm{C-SDA}}(Y^TY)^{-1/2}$. Note $Y^TY \in R^{c\times c}$ is a diagonal matrix, i.e. $(Y^TY)_{jj} = l_j$, we then have the range space of B_{SDA} and $B_{\mathrm{C-SDA}}$ is the same.

With Lemma 2, we can directly establish the equivalence between SDA and C-SDA, which is based on the following theorem:

Theorem 1. *Let the SVD of $B_{\mathrm{SDA}}B_{\mathrm{SDA}}^T$ and $B_{\mathrm{C-SDA}}B_{\mathrm{C-SDA}}^T$ be*

$$B_{\mathrm{SDA}}B_{\mathrm{SDA}}^T = P_{\mathrm{SDA}}\sum_{\mathrm{SDA}} P_{\mathrm{SDA}}^T$$

$$B_{\mathrm{C-SDA}}B_{\mathrm{C-SDA}}^T = P_{\mathrm{C-SDA}}\sum_{\mathrm{C-SDA}} P_{\mathrm{C-SDA}}^T, \tag{12}$$

where P_{SDA}, $P_{\mathrm{C-SDA}} \in R^{D\times r}$, and $r = rank(P_{\mathrm{SDA}}) = rank(P_{\mathrm{C-SDA}})$. Then, there exists an orthogonal matrix $R \in R^{r\times r}$ satisfying $P_{\mathrm{SDA}} = P_{\mathrm{C-SDA}}R$. More importantly, if we let $V_{\mathrm{SDA}} = \Theta\sum^{-1}P_{\mathrm{SDA}}$ and $V_{\mathrm{C-SDA}} = \Theta\sum^{-1}P_{\mathrm{C-SDA}}$, we have $V_{\mathrm{SDA}} = V_{\mathrm{C-SDA}}R$.

Proof of Theorem 1. It is certainly that $P_{\mathrm{SDA}}P_{\mathrm{SDA}}^T$ and $P_{\mathrm{C-SDA}}P_{\mathrm{C-SDA}}^T$ are the orthogonal projections on the same range space of B_{SDA} or $B_{\mathrm{C-SDA}}$. Following Lemma 2, we have both $P_{\mathrm{SDA}}P_{\mathrm{SDA}}^T$ and $P_{\mathrm{C-SDA}}P_{\mathrm{C-SDA}}^T$ are orthogonal projections on the same subspace. Since the orthogonal projections on a subspace are unique, we have $P_{\mathrm{SDA}}P_{\mathrm{SDA}}^T = P_{\mathrm{C-SDA}}P_{\mathrm{C-SDA}}^T$, and it can also be noted that

$$P_{\mathrm{SDA}} = P_{\mathrm{C-SDA}}P_{\mathrm{C-SDA}}^T P_{\mathrm{SDA}} = P_{\mathrm{C-SDA}}R. \tag{13}$$

where $R = P_{\mathrm{C-SDA}}^T P_{\mathrm{SDA}} \in R^{r\times r}$ satisfying $RR^T = R^TR = I$. In addition, following Lemma 1, since SDA and C-SDA have the same form of solution, i.e $V^* = \Theta\sum^{-1}P_d$, we have $V_{\mathrm{SDA}} = V_{\mathrm{C-SDA}}R$. We thus prove Theorem 1.

Theorem 1 indicates that if we retain all the eigenvectors to the nonzero eigenvalues, i.e. $d = r$, then the difference between SDA and C-SDA lies in the orthogonal transformation R. Thus, SDA and C-SDA are equivalent because the orthogonal transformation can be neglected when we apply a distance-based classifier (such as k nearest neighbor classifier).

3.3. Equivalence between C-SDA and Lap-RLS/L with uncorrelated constraint

The equivalence between C-SDA and Lap-RLS/L under the uncorrelaed constraint is based on the following theorem [17,34,37]:

Theorem 2. *Let M be the auxiliary matrix defined as follows:*

$$M = YX^T\left(S_t + \lambda_m XLX^T + \lambda_t I\right)^{-1}XY^T = \Omega\Sigma_M\Omega^T. \tag{14}$$

where $\Omega\Sigma_M\Omega^T$ is the Singular Value Decomposition of M. We also denote V_E^* as:

$$V_E^* = \left(S_t + \lambda_m XLX^T + \lambda_t I\right)^{-1}XY^T\Omega\Sigma_M^{-1/2}. \tag{15}$$

Then, V_E^* is the optimal solution of C-SDA, $V_E^* = V_{\mathrm{C-SDA}}^*$.

To prove Theorem 2, we first give the following lemma:

Lemma 3. *Given two matrixes A and B, then AB and BA have the same non-zero eigenvalues. For each nonzero eigenvalue of AB, if the corresponding eigenvector of AB is v, then the corresponding eigenvector of BA is u = Bv.*

Proof of Theorem 2. Recall that the solution of C-SDA is formed by the eigenvectors of $(S_t + \lambda_m XLX^T + \lambda_t I)^{-1}XY^TYX^T$. Based on Lemma 3, it has the same nonzero eigenvalues to the auxiliary matrix M; According to Lemma 3 again, if Ω is the eigenvector to the nonzero eigenvalues of M, $(S_t + \lambda_m XLX^T + \lambda_t I)^{-1}XY^T\Omega\Xi$ is the eigenvector of the matrix $(S_t + \lambda_m XLX^T + \lambda_t I)^{-1}XY^TYX^T$, where Ξ is an arbitrary diagonal matrix as the eigenvectors of Eq. (8) are not unique. Here, if we let $\Xi = \Sigma_M^{-1/2}$ and $V_E^* = (S_t + \lambda_m XLX^T + \lambda_t I)^{-1}XY^T\Omega\Sigma_M^{-1/2}$, then it follows $V_E^{*T}(S_t + \alpha_t I + \alpha_m XLX^T)V_E^* = I$, which indicates V_E^* satisfies the uncorrelated constraint and is the solution of C-SDA in Eq. (7). We thus prove Theorem 2.

Table 2
Least square framework for solving SDA.

Input: Data matrix $X \in R^{D\times(l+u)}$, reduced matrix d and other related parameters.
Output: The projection matrix $V \in R^{D\times d}$.

Algorithm:
1. Solve the least square problem of Lap-RLS/L in Eq. (2):

$$J(V,b) = \min \sum_{j=1}^{l} \left\| V^T x_j + b^T - y_j \right\|_F^2 + \alpha_t \|V\|_F^2 + \alpha_m Tr\left(V^T X L X^T V\right),$$

and obtain the optimal solution $V^*_{\text{Lap}-\text{RLS/L}} = (S_t + \lambda_m XLX^T + \lambda_t I)^{-1} XY^T$.
2. Perform the SVD of the auxiliary matrix as in Eq. (14) $M = YX^T(S_t + \lambda_m XLX^T + \lambda_t I)^{-1}XY^T = \Omega\Sigma_M\Omega^T$.
3. Output $V^*_{\text{SDA}} = V^*_{\text{Lap}-\text{RLS/L}}\Omega\Sigma_M^{-1/2}$.

We next establish the equivalence between SDA and Lap-RLS/L by using C-SDA. Following Eq. (2), it can be easily noted that $(S_t + \lambda_m XLX^T + \lambda_t I)^{-1}XY^T$ is the optimal solution of Lap-RLS/L, then, we have $V^*_E = V^*_{\text{C}-\text{SDA}} = V^*_{\text{Lap}-\text{RLS/L}}\Omega\Sigma_M^{-1/2}$. This indicates V^*_E is also the optimal solution of Lap-RLS/L with the uncorrelated constraint, which can be given as follows:

$$\begin{aligned} J(V,b) = \min & \sum_{j=1}^{l} \left\| V^T x_j + b^T - y_j \right\|_F^2 + \alpha_t \|V\|_F^2 + \alpha_m Tr\left(V^T X L X^T V\right) \\ s.t.\ & V^T\left(S_t + \alpha_t I + \alpha_m XLX^T\right)V = I \end{aligned} . \tag{16}$$

In addition, since the optimal solutions of SDA and C-SDA are the same as analyzed in Section 3.2, i.e. $V^*_{\text{SDA}} = V^*_{\text{C}-\text{SDA}}R$, hence by using C-SDA as a bridge, we can establish the equivalence between SDA and Lap-RLS/L under the uncorrelated constraint, and the problem of SDA in Eq. (6) can be equivalently solved by Eq. (16). The basic steps to solve SDA based on a least square framework are shown in Table 2. For simplicity, we refer it as LS-SDA.

4. Learning from local and global discriminative information

The connection between SDA and LapRLS/L throws light on their relationship for semisupervised learning. Two issuess still need to be addressed: (1) the regression term in both LapRLS and LS-SDA is supervised and it only utilizes the labeled set to train the linear classification function. Since the number of labeled set is small compared with that of unlabeled data, this can be problematic that the linear classification function can be underfit because of small sample size [21]; (2) the proposed LS-SDA utilizes Gaussian function based graph Laplacian matrix to characterize the local structure of data manifold. But these Laplacian matrixes cannot capture the discriminative information of classes for data, which results in degrading classification performance. In addition, the Gaussian function based affinity matrix is over sensitive to the Gaussian variance. Therefore, Gaussian function based affinity matrix is less effective for image classification and visualization. In this session, we describe our newly developed LLGDI to solve the problem.

4.1. Local and global discriminative information embedding

The least square regression term in Lap-RLS/L and LS-SDA in Eq. (16) is

$$\min \sum_{j=1}^{l} \left\| V^T x_j + b^T - y_j \right\|_F^2 + \alpha_t \|V\|_F^2. \tag{17}$$

Clearly, Eq. (17) is a supervised formulation, as the label y_j of x_j ($j \le l$) has already been known. However, since l is usually very small, the classification function $z_j = V^T x_j + b$ may not be sufficiently trained due to the small sample size. To solve this problem, we introduce a set of estimated labels $Z = \{Z_l, Z_u\} = \{z_1, z_2, \ldots, z_{l+u}\} \in R^{c\times(l+u)}$ by replacing $V^T x_j + b$ with z_j. A regression term is then added to Eq. (17) as follows:

$$\min \sum_{i=1}^{l} \|z_i - y_i\|_F^2 + \alpha_r\left(\sum_{j=1}^{l+u} \left\| V^T x_j + b^T - z_j \right\|_F^2 + \eta\|V\|_F^2\right). \tag{18}$$

Following Eq. (18), the classification function $z_j = V^T x_j + b$ can well be trained by utilizing all the estimated labels as well as fixing to its initial labels. Here, since Z can be viewed as the global label matrix, by regressing X on Z, the projection matrix V and bias b actually capture the globally discriminative direction of each class. In other words, the global discriminative information can be characterized by the regression residual term of Eq. (18). In order to capture the locally discriminative information, we adopt a local regression model for each data sample x_j. Specifically, let $N_k(x_j)$ be the k neighborhood set of x_j including itself, we denote $X_j = \{x_{j_0}, x_{j_1}, \ldots, x_{j_{k-1}}\} \in R^{D\times k}$ as the local data matrix formed by all samples in $N_k(x_j)$, where $\{j_1, j_1, \ldots, j_k\}$ is the index set of $N_k(x_j)$ and $j_1 = j$, $x_{j_1} = x_j$. We also denote $Z_j = \{z_{j_1}, z_{j_2}, \ldots, z_{j_k}\} \in R^{c\times k}$ as local low-dimensional label matrix in $N_k(x_j)$. Then, the local regression function for all data samples can be given as follows:

$$\min_{Z_j, V_j, b_j} \sum_{j=1}^{l+u} \left(\sum_{i=1}^{k} \left\| V_j^T x_{j_i} + b_j^T - z_{j_i} \right\|_F^2 + \eta \left\| V_j \right\|_F^2\right). \tag{19}$$

However, minimizing the above total errors over all samples in each local patch tends to force each local error $\alpha_{j_i} = \|V_j^T x_{j_i} + b_j^T - z_{j_i}\|_F$ to become similar to each other. Given some dataset with multi-density distribution, where the data sampling is uniform, treating all the local errors as in Eq. (19) equally may strengthen the contributions of samples in dense distribution while weaken those in sparse distribution, causing bias exists. To solve this problem, we add a weight vector $\Gamma_j = \{\tau_{j_1}, \tau_{j_2}, \ldots, \tau_{j_k}\} \in R^{1\times k}$ for each local data patch X_j in order to penalize each regression error, which can be shown as follows:

$$\min_{Z_j, V_j, b_j} \sum_{j=1}^{l+u} \left(\sum_{i=1}^{k} \tau_{j_i} \left\| V_j^T x_{j_i} + b_j^T - z_{j_i} \right\|_F^2 + \eta \left\| V_j \right\|_F^2 \right) \tag{20}$$

We will show in Section 4.3 that by choosing a special Γ_j, the graph Laplacian matrix derived from Eq. (20) can also be a normalized graph Laplacian matrix, which is useful for handling the multi-density dataset. To calculate the local projection matrix V_j and bias b_j in Eq. (20), we set the derivatives to Eq. (20) with respect to V_j and b_j to zeros. Then, we have:

$$\begin{cases} b_j = e_k \Delta_j \left(Z_j^T - X_j^T V_j \right) / \left(e_k \Delta_j e_k^T \right) \\ V_j = \left(X_j H_j X_j^T + \lambda I \right)^{-1} X_i H_j Z_j^T \end{cases} \tag{21}$$

where $e_k \in R^{1\times k}$ is a unit vector with size k, $\Delta_j \in R^{k\times k}$ is a diagonal matrix with each element being $(\Delta_j)_{ii} = \tau_{j_i}$, $H_j \in R^{k\times k}$ is the local weighted center matrix defined as $H_j = \Delta_j - (\Delta_j e_k^T e_k \Delta_j)/(e_k \Delta_j e_k^T)$. By substituting V_j and b_j in Eq. (21) with Eq. (20), Eq. (20) will be reduced to:

$$\min_{Z_j} \sum_{j=1}^{l+u} Tr\left(Z_j L_j Z_j^T \right). \tag{22}$$

where $L_j = H_j - H_j X_j^T (X_j H_j X_j^T + \eta I)^{-1} X_j H_j$. Here, let $S_j \in R^{(l+u)\times k}$ be the selected matrix with each element satisfying $(S_j)_{pq} = 1$, if $p = i_q$; $(S_j)_{pq} = 0$, otherwise. Then, Z_j can be viewed as a selection from Z as $Z_j = ZS_j$, and Eq. (22) will be reduced to:

$$\min_{Z} \sum_{j=1}^{l+u} Tr\left(Z S_j L_j S_j^T Z^T \right) = \min_{Z} Tr\left(Z L_d Z^T \right). \tag{23}$$

where $L_d = \sum_{j=1}^{l+u} (S_j L_j S_j^T)$. In this paper, by integrating Eq. (23) into the objective function of Eq. (17), we formulate our proposed method as follows:

$$J(V, Z, b) = \min \sum_{i=1}^{l} \|z_i - y_i\|_F^2 + \alpha_m Tr\left(Z L_d Z^T \right) + \alpha_r \left(\left\| V^T X + b^T e - Z \right\|_F^2 + \eta \|V\|_F^2 \right). \tag{24}$$

The second term in Eq. (24) characterizes the local discriminative and manifold structure of dataset while the third term characterizes the global discriminative structure of dataset, α_m and α_r are two parameters balancing the tradeoff between three terms.

4.3. Solution of LLGDI for simutaneous megression and dimensionality reduction

In this subsection, we will show how to calculate the optimal solution of the proposed LLGDI in Eq. (24). In addition, we will also discuss how to realize subspace for the proposed LLGDI. By setting the derivatives of Eq. (23) with respect to V and b to zero, we have:

$$\begin{cases} b = \left(eZ^T - eX^T V \right) / ee^T \\ V = \left(X L_c X^T + \eta I \right)^{-1} X L_c Z^T \end{cases}. \tag{25}$$

where $e \in R^{1\times(l+u)}$ is a unit vector and $L_c = I - e^T e / ee^T$ is used for centering the samples by subtracting the mean of all samples. Similarly, with b and V in Eq. (25), the global regression term can be written as:

$$\left\| V^T X + b^T e - Z \right\|_F^2 + \eta \|V\|_F^2 = Tr\left(Z L_g Z^T \right), \tag{26}$$

where $L_g = L_c - L_c X^T (X L_c X^T + \eta I)^{-1} X L_c$. Then, Eq. (24) can be rewritten as:

$$J(Z) = \min_{Z} Tr\left((Z - Y) U (Z - Y)^T \right) + \alpha_m Tr\left(Z L_d Z^T \right) + \alpha_r Tr\left(Z L_g Z^T \right). \tag{27}$$

where $U \in R^{(l+u)\times(l+u)}$ is a diagonal matrix with the first l and remaining u diagonal elements as 1 and 0, respectively. However, it should be noted that the problems in Eqs. (24) and (27) are derived from regression problem instead of subspace learning problem (they can only reduce the dimensionality to c, where c is the number of classes). Hence in order to solve both regression as well as subspace learning problems, we can add an uncorrelated constraint to Eq. (24) which is similar to the LS-SDA as in Eq. (16). Specifically, recall the uncorrelated constraint in Eq. (16) as

$$V^T \left(S_t + \alpha_t I + \alpha_m X L X^T \right) V = I, \tag{28}$$

Table 3
Algorithm of LLGDI.

Input: Data matrix $X \in R^{D\times(l+u)}$, the initial label matrix $Y \in R^{c\times(l+u)}$, number of neighbors k, reduced matrix d and other related parameters.
Output: The projection matrix $V \in R^{D\times d}$.

Algorithm:

1. Form the normalized local regression regularized term $Tr(ZL_dZ^T) = \sum_{j=1}^{l+u} Tr(ZS_jL_jS_j^TZ^T)$ as in Eq. (23), where L_j is a graph Laplacian matrix satisfying $L_j = H_j - H_jX_j^T(X_jH_jX_j^T + \eta I)^{-1}X_jH_j$, S_j is the selected matrix with each element satisfying $(S_j)_{pq} = 1$, if $p = i_q$; $(S_j)_{pq} = 0$, otherwise.
2. Form the global regression regularized term $Tr(ZL_gZ^T) = \|V^TX + b^Te - Z\|_F^2 + \eta\|V\|_F^2$ as in Eq. (26), where L_g is a graph Laplacian matrix satisfying $L_g = L_c - L_cX^T(XL_cX^T + \eta I)^{-1}XL_c$.
3. Solving the regression problem as in Eq. (24):

$$J(Z) = \min_Z Tr\big((Z-Y)U(Z-Y)^T\big) + \alpha_m Tr\big(ZL_dZ^T\big) + \alpha_r Tr\big(ZL_gZ^T\big)$$

and obtain the optimal the estimated label matrix $Z_r = YU(U + \alpha_mL_d + \alpha_rL_g)^{-1}$.
4. Perform EVD of the auxiliary matrix: $YU(U + \alpha_mL_d + \alpha_rL_g)^{-1}UY^T = \widetilde{\Omega}\widetilde{\sum}\widetilde{\Omega}^T \in R^{c\times c}$ as in Eq. (31).
5. Obtain the optimal projection matrix V^* in Eq. (32) as $Z^* = \widetilde{\sum}^{-1/2}\widetilde{\Omega}^TYU(U + \alpha_mL_d + \alpha_rL_g)^{-1}$. Output V^*.

we observe the global scatter matrix $V^TS_tV = V^TXUX^TV$ may not well charactize the global structure of dataset by only utilizing labeled samples. Here, similar to Eq. (18), we reduce the effect of such global scatter matrix by replacing V^Tx_j with z_j. We also replace the graph Laplacian matrix L with the local and global discriminative Laplacian matrix $\alpha_mL_d + \alpha_rL_g$, then, the uncorrelated constraint in Eq. (28) can be relaxed to:

$$ZUZ^T + \alpha_mZL_dZ^T + \alpha_rZL_gZ^T = I. \tag{29}$$

We then integrate the above relaxed uncorrelated constraint to Eq. (24) and form the problem as follows:

$$\begin{aligned} J(Z) = \min_Z & Tr\big((Z-Y)U(Z-Y)^T\big) + \alpha_m Tr\big(ZL_dZ^T\big) + \alpha_r Tr\big(Z(L_c-N)Z^T\big) \\ s.t.\ & ZUZ^T + \alpha_mZL_dZ^T + \alpha_rZL_gZ^T = I. \end{aligned} \tag{30}$$

Here, with the constraint added into Eq. (29), Z actually represents the low-dimensional embedding to X and has no meaning of the estimated label matrix of X. To solve the problem in Eq. (30), we adopt a two-stage approach to calculate the optimal solution. Specifically, we first set the derivatives of $J(Z)$ with respect to Z to zero and obtain the regression solution as $Z_r = YU(U + \alpha_mL_d + \alpha_rL_g)^{-1}$. We then perform the Eigen-value Decomposition (EVD) of the following auxiliary matrix:

$$YU(U + \alpha_mL_d + \alpha_rL_g)^{-1}UY^T = \widetilde{\Omega}\widetilde{\sum}\widetilde{\Omega}^T \in R^{c\times c}. \tag{31}$$

and let the optimal solution Z^* as:

$$Z^* = \widetilde{\sum}^{-1/2}\widetilde{\Omega}^TYU(U + \alpha_mL_d + \alpha_rL_g)^{-1}. \tag{32}$$

It can be easily verified that Z^* satisfies $Z^*(U + \alpha_mL_d + \alpha_rL_g)Z^{*T} = I$, hence Z^* is the optimal solution of Eq. (30). Finally, the projection matrix V^* can be calculated by replacing Z^* into Eq. (25). Here, since we have added an relaxed uncorrelated constraint in Eq. (29) to the objective function of LLGDI as in Eq. (24), by solving this uncorrelated constrained regression problem, the obtained projection matrix V^* can reduce the dimensionality to less than c while keeping most desired discriminative information. As a result, both regression and subspace learning problems can be solved by Eq. (30). The basic steps of the proposed LLGDI method are shown in Table 3.

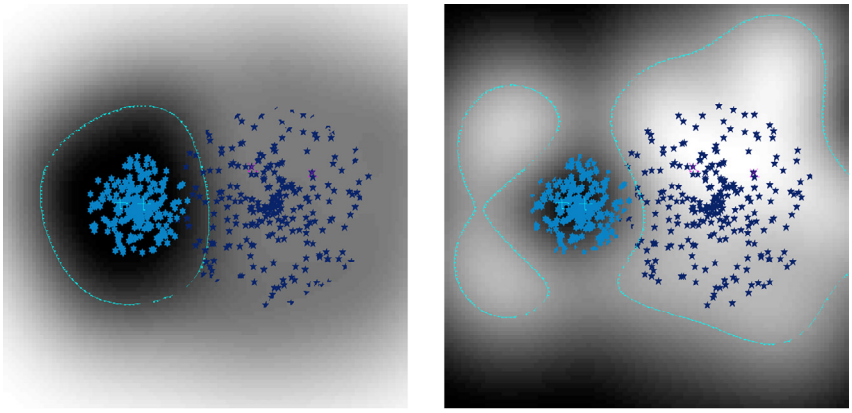
4.3. Normalized local discriminative graph Laplacian matrix

It can be easily proved that L_d is a graph Laplacian matrix (seen in the Appendix). But L_d may not be a normalized graph Laplacian matrix. As pointed in [12], the normalization can strengthen the local regressions in the low-density region and weaken those in the high density region. Since the data sampling is usually uniform in practice, normalization is useful for handling the case when the density of dataset varies dramatically. In this subsection, we show that by choosing a special weight vector Γ_j for each X_j, L_d can be a normalized graph Laplacian matrix.

Specifically, let we consider a data sample x_l and let K_l be the index set of those neighborhood set $N_k(x_j)$, which contains x_l as a neighbor of x_j, i.e. if $j \in K_l$, then $x_l \in N_k(x_j)$, where x_l can be denoted as x_{j_i} in the neighborhood set $N_k(x_j)$ and $i = i(l, j)$ is a local index depending on l and j. Obviously, if x_l is in the low-density area, it has sparse neighbors and K_l is relatively small. As a result, its connections to other samples will be weaker than that which has large K_l. Here, to strengthen the connections of samples in the low-density area, we need to normalize the weights corresponding to each K_l. Let τ_j^l be the weight of x_{j_i} and l be the global index of x_{j_i}, we then define $\tau_{j_i} = \tau_j^l$ as follows:

$$\tau_{j_i} = \tau_j^l \leftarrow \frac{1}{|K_l|}. \tag{33}$$

where $|K_l|$ is the total index number in K_l. Hence, based on this definition, we have the following theorem:



(a) (b)

Fig. 1. Gray image of reduced space learned by LLGDI without normalization and LLGDI with normalization: two-plate dataset (a) LLGDI without normalization and (b) LLGDI with normalization.

Theorem 4. *With the normalization for each w_{j_i} as in Eq. (33), L_d is both graph Laplaican matrix and normalized graph Laplacian matrix.*

Proof of Theorem 4. The proof that L_d is a graph Laplacian matrix can be seen in the Appendix. In order to prove L_d is a normalized graph Laplacian matrix, we need prove that L_d can be reformulated in a form of $L_d = I - W_d$ and the sum of each row or column of the affinity matrix W_d is equal to 1. Note that $L_d = \sum_{j=1}^{l+u}(S_j L_j S_j^T)$ and $L_j = H_j - H_j X_j^T (X_j H_j X_j^T + \eta I)^{-1} X_j H_j$, where $H_j = \Delta_j - (\Delta_j e_k^T e_k \Delta_j)/(e_k \Delta_j e_k^T)$, we first define the affinity matrix W_d as follows:

$$W_d = \sum_{j=1}^{l+u}\left(S_j W_j^d S_j^T\right), \tag{34}$$

where each W_j^d satisfies:

$$W_j^d = \left(\Delta_j e_k^T e_k \Delta_j\right)/\left(e_k \Delta_j e_k^T\right) - H_j X_j^T \left(X_j H_j X_j^T + \eta I\right)^{-1} X_j H_j. \tag{35}$$

Then, L_d can be reformulated as:

$$L_d = \sum_{j=1}^{l+u}\left(S_j \Delta_j S_j^T\right) - \sum_{j=1}^{l+u}\left(S_j W_j^d S_j^T\right). \tag{36}$$

Here for each $S_j \Delta_j S_j^T$, we have $S_j^T e^T = e_k^T \Rightarrow S_j \Delta_j S_j^T e^T = S_j \Gamma_j^T$, where $S_j \Gamma_j^T \in R^{(l+u)\times 1}$ is a column vector by putting each τ_j^l to its global index l corresponding to x_{j_i}. We then have:

$$\left\{\sum_{j=1}^{l+u}\left(S_j \Delta_j S_j^T\right)\right\} e^T = \sum_{j=1}^{l+u}\left(S_j \Gamma_j^T\right) = e^T, \tag{37}$$

The second equation holds as $\sum_{j\in K_l} \tau_j^l = \sum_{j\in K_l} 1/|K_l| = 1$ hence the sum of all $S_j \Gamma_j^T$ in each element is equal to 1. Then, following Eq. (37), it indicates $\sum_{j=1}^{l+u}(S_j \Delta_j S_j^T)$ is an identity matrix, i.e. $\sum_{j=1}^{l+u}(S_j \Delta_j S_j^T) = I$. Then based on the above analysis, we can reformulate L_d in a form of $L_d = I - W_d$. In addition, since L_d is a graph Laplaican matrix (proved in the Appendix), it satisfies $L_d e^T = 0$, then we have

$$\begin{aligned} L_d e^T = 0 &\Rightarrow \left\{\sum_{j=1}^{l+u}\left(S_j \Delta_j S_j^T\right) - \sum_{j=1}^{l+u}\left(S_j W_j^d S_j^T\right)\right\} e^T = 0 \\ &\Rightarrow \left\{\sum_{j=1}^{l+u}\left(S_j \Delta_j S_j^T\right)\right\} e^T = \left\{\sum_{j=1}^{l+u}\left(S_j W_j^d S_j^T\right)\right\} e^T \\ &\Rightarrow e^T = \left\{\sum_{j=1}^{l+u}\left(S_j W_j^d S_j^T\right)\right\} e^T \\ &\Rightarrow W_d e^T = e^T \text{ or } e W_d = e. \end{aligned} \tag{38}$$

which indicates that the sum of each column or row of W_d is equal to 1. We thus prove the theorem. Theorem 4 indicates that by choosing a special weight vector τ_{j_i} for each x_{j_i}, L_d can be both graph Laplacian matrix and normalized graph Laplacian matrix.

To show the merit of normalization, we generate a dataset with two classes. Each class follows a Guassian distribution but with different cores and density. In each class, two samples are selected as labeled set and the remainings are as unlabeled set. Our goal is to show LLGDI can handle multi-density dataset. Fig. 1 shows the gray images of decision surfaces and boundaries learned by LLGDI without normalization and LLGDI with normalization. The gray value of each pixel represents the difference of

distance from the pixel to its nearest labeled samples in different classes after dimensionality reduction. The decision boundaries are then formed by the pixels with the values equal to 0. In this example, we set the reduced dimensionality as 1. Here, LLGDI without normalization derives the local graph Laplacian matrix L_d directly from Eq. (19); while LLGDI with normalization derives the local graph Laplacian matrix L_d from Eq. (20), where $\tau_{j_i} = \tau_j^l \leftarrow 1/|K_l|$ is set as in Eq. (33). From Fig. 1, we can observe that LLGDI without normalization cannot find proper boundary. However, LLGDI with normalization can achieve better performance, as there are less missing-classified data points separated by the decision boundary, which becomes more distinctive and accurate. The improved result is believed to be due to the fact that normalization can strengthen the local regressions in the low-density region and weaken those in the high density region. This is proved to be advantageous to be used for multi-density dataset.

4.4. Discussion and relative work

In this subsection, we discuss the relationship of LLGDI with LS-SDA in Eq. (16) and other state-of-the-art methods including FME and LRGA.

4.4.1. Relationship to LS-SDA in **Eq. (16)**

We will first show LS-SDA in Eq. (16) is only a special case of LLGDI. Let $\alpha_t \to \infty, \alpha_r\eta = \alpha_t$ and $L_d \to L$, we have $||V^TX + b^Te - Z||_F^2 \to 0$ or $V^TX + b^Te = Z$. By replacing Z into Eq. (24), the objective function $J(V, Z, b)$ in Eq. (24) will be reduced to Lap-RLS/L as in Eq. (2). If we further fix the bias term as $b = -eUX^TV/(eUe^T)$, the constraint in Eq. (35) will be relaxed to:

$$\begin{aligned} & ZUZ^T + \alpha_m ZLZ^T + \alpha_r\Big(\big(V^TX + b^Te - Z\big)\big(V^TX + b^Te - Z\big)^T + \eta V^TV\Big) = I \\ & \to V^TX\big(I - Ue^Te/eUe^T\big)U\big(I - Ue^Te/eUe^T\big)X^TV + \alpha_m V^TXLX^TV + \alpha_t V^TV = I \\ & \to V^TX\big(U - Ue^TeU/eUe^T\big)X^TV + \alpha_m V^TXLX^TV + \alpha_t V^TV = I \\ & \to V^T\big(S_t + \alpha_m XLX^T + \alpha_t I\big)V = I. \end{aligned} \tag{39}$$

Following Eq. (39), it indicates if we let $\alpha_r \to \infty, \alpha_r\eta = \alpha_t$ and $b = -V^TXUe^T/(eUe^T)$, the problem in Eq. (24) can be reduced to that in Eq. (16). The problem of Eq. (16) and Lap-RLS/L are only special cases of LLGDI. But since LLGDI is aimed at characterizing local and global discriminative information embedded in a dataset, LLGDI is preferable to handle classification problem than the least square framework in Eq. (16).

4.4.2. Relationship to FME [13,14]

Nie et al. have proposed another unified framework, i.e. Flexiable Manifold Embedding (FME) [13,14], for semi-supervised dimensionality reduction, in which they verify that LLGC, GFHF and Lap-RLS/L are only special cases in the framework. The basic objective function of FME can be given as

$$J(V, Z, b) = \min \sum_{i=1}^{l} \|z_i - y_i\|_F^2 + \alpha_m Tr\big(ZLZ^T\big) + \alpha_r\Big(\big\|V^TX + b^Te - Z\big\|_F^2 + \eta\|V\|_F^2\Big), \tag{40}$$

where $L = D - W$ is the graph Laplacian matrix and W is Gaussian function based affinity matrix. It can be observed that Eq. (40) is almost the same as the objective function of LLGDI in Eq. (24), when we let $L_d \to L$. But FME is essentially derived for handling regression problem (it can only reduce the dimensionality to c), whereas it cannot solve subspace learning problem. For the case of LLGDI, by adding an uncorrelated constraint to Eq. (24), LLGDI can solve both regression and subspace learning problems. In addition, LLGDI has utilized a normalized local discriminative Laplacian matrix to preserve manifold and discriminative structure in a dataset. This is a better way than only relying on neighborhood graph.

4.4.3. Relationship to LRGA [30,31]

Recently, Yang et al. have proposed semi-supervised transductive learning method, namely, Local Regression and Global Alignment (LRGA) [30,31], for multimedia retrieval. They share the similar concept with the proposed method. The basic objective function of LRGA can be given as:

$$J(Z) = \min_{Z, V_j, b_j} \sum_{i=1}^{l} \|z_i - y_i\|_F^2 + \alpha_m \sum_{j=1}^{l+u}\Big(\sum_{i=1}^{k} \big\|V_j^Tx_{j_i} + b_j^T - z_{j_i}\big\|_F^2 + \eta\big\|V_j\big\|_F^2\Big). \tag{41}$$

It can be noted that LRGA is a special case of LLGDI when $\alpha_r = 0$. Therefore, LRGA is only a transductive learning method and cannot handle out-of-sample problem, while LLGDI is a transductive and inductive learning method. Another superiority of LLGDI over LRGA is that LLGDI normalized each local regression term. Thus as shown in the simulation results, LLGDI can handle multi-density dataset remarkably.

4.4.4. Relation to the eigen-image [25]

The paper in [25] presents a method to learn eigen-images from an image to be segmented. It shares the similar concept with the model in Eq. (19) in the case of graph construction. However, LLGDI is to perform graph construction based on the model of Eq. (20) which is not Eq. (19). The main difference between Eq. (20) and Eq. (19) is that Eq. (20) normalizes each local regression

Table 4
Computational complexities of different methods.

Algorithms	RLDA/SDA	LS-SDA	Lap-RLS/L	FME	LLGDI
Computation complexity	$O(D^2d)$	$O(D^3)$	$O(D^3)$	$\min(O((l+u)^3), O(D^3))$	$\min(O((l+u)^3), O(D^3))$

term, while Eq. (19) considers each local regression term equally without normalization. This can be useful as sampling is usually not uniform in practice, and over-emphasizing neighborhoods with high densities may occlude useful information in sparse regions. Hence the model in Eq. (20) is superior to the one of Eq. (19). In addition, it should also be noted that in the case of eigen-images in [40], Eq. (20) can be equal to Eq. (19). This is because for constructing eigen-images, each pixel only connects to its eight neighborhood pixels, making the total number of connections for each pixel equal in graph construction. In such cases, normalization can be neglect and Eq. (20) is equal to Eq. (19). In other words, the model of Eq. (19) can be considered a special case of Eq. (20) when the total number of connections for each data samples or pixels is the same. Thus, the model described in Eq. (20) can be seen as an extension to the one of Eq. (19).

4.5. Analysis of computational complexity

In this subsection, we will analyze the computational complexity of the proposed LLGDI and compare those with other state-of-the-art methods. Note that to calculate the projection matrix for the proposed LLGDI method, one needs to calculate the local regression regularized term L_d, the global regression residual term L_g and the low-dimensional representation Z as in Eq. (30). (1) For calculating the local regression regularized term $L_d = \sum_{j=1}^{l+u}(S_jL_jS_j^T)$, where $L_j = \eta G_j(G_j^TX_j^TX_jG_j + \eta I)^{-1}G_j^T$, one needs to calculate the inverse of $(G_j^TX_j^TX_jG_j + \eta I)^{-1} \in R^{k\times k}$ for each L_j, where the computational complexity for calculating each L_j is $O(k^3)$ (k is the number of neighborhoods) and the total complexity for calculating L_d is $O((l+u)k^3)$; (2) for calculating the global regression residual term $L_g = N - L_c = L_cX^TXL_c(L_cX^TXL_c + \eta I)^{-1} - L_c$, we need to calculate inverse of $L_cX^TXL_c + \eta I \in R^{(l+u)\times(l+u)}$ or $XL_cX^T + \eta I \in R^{D\times D}$, and the computational complexity of this step is $O(\min((l+u)^3, D^3))$; (3) then the computational complexity for calculating the low-dimensional representation Z in Eq. (30) is $O(\min((l+u)^2c, D^2c) + c^2d)$, where $O(\min((l+u)^2c, D^2c))$ is that for solving the least square regression problem of Eq. (27), c^2d is that for performing the SVD of the auxiliary matrix in Eq. (31), i.e. $YU(U + \alpha_mL_d + \alpha_rL_g)^{-1}UY^T = \widetilde{\Omega}\widetilde{\sum}\widetilde{\Omega}^T \in R^{c\times c}$; finally, the optimal projection matrix V^* can be obtained by $(XL_cX^T + \eta I)^{-1}XL_cZ^T = XL_c(L_cX^TXL_c + \eta I)^{-1}Z^T$, and the computation complexity can be neglected, as the inverse of $L_cX^TXL_c + \eta I$ or $XL_cX^T + \eta I$ has already been obtained. As a result, since $d \le c \ll \min(l+u, D)$ and $k \ll l+u$, the total computational complexity for the proposed LLGDI is $O((l+u)k^3) + \min(O(D^3), O((l+u)^3)) + \min(O((l+u)^2d), O(D^2d)) + O(c^2d) \approx \min(O(D^3), O((l+u)^3))$.

We next analyze the computational complexity of other supervised and semi-supervised methods. For RLDA and SDA, both of them need to perform the generalized eigenvalue decomposition (GEVD) in order to calculate the projection matrix, and the computational complexities of RLDA and SDA are $O(D^3)$ or $O(D^2d)$, if only d eigenvectors are involved; for Lap-RLS/L and LS-SDA, both of the algorithms need to calculate the inverse of regularized Laplacian matrix, which is $XL_tX^T + \alpha_tI + \alpha_mXLX^T \in R^{D\times D}$, hence the computational complexity of this step is $O(D^3)$. In addition, for LS-SDA, it further needs to perform the eigenvalue decomposition of $YX^T(S_t + \lambda_mXLX^T + \lambda_tI)^{-1}XY^T \in R^{c\times c}$, and the computational complexity is $O(c^2d)$. Since $d \le c \ll D$, we have $O(c^2d) \ll O(D^3)$. Hence the total computational complexities of Lap-RLS/L and LS-SDA are almost the same, which are $O(D^3)$; for FME, following the work in [9], it is with similar computational complexity of the proposed LLGDI, which is also $\min(O((l+u)^3), O(D^3))$. This is because both two methods need to calculate the regression residual term and low-dimensional representation z. The computational complexities of different algorithms can be seen in Table 4.

From Table 3, we have the following observations: (1) given the dataset is with low dimensionality, i.e. $D \ll l+u$, the computational complexities of different methods are almost the same, as in such case, the computational complexities of different methods are close related to the number of dimensionality; (2) given the dataset is with high dimensionality, i.e. $l+u \ll D$, FME and the proposed LLGDI are with the smallest computational complexities, as in such case, their computational complexities are close related to the number of dataset. Hence the proposed LLGDI method is more suitable for the dataset with high dimensionality.

Here, it should be noted that Table 3 only shows the computational complexities of different methods **in one simulation**. Though FME and the proposed LLGDI share the same computational complexity as shown in Table 3, we have to say that FME has one more parameter that needs to be adjusted, i.e. Gaussian covariance in Gaussian function based affinity matrix. However, the Gaussian function based affinity matrix is sensitive to Gaussian variance and even a small variation can cause the results dramatically. Hence in practice, some approaches for parameter selection, such as five-fold cross validation, are needed in order to choose the best value of Gaussian variance, which will greatly increase the computational efforts. But for the proposed LLGDI, there is no such parameter, as LLGDI is to construct the graph by aligning all local regressions instead of relying on Gaussian function. As a result, the proposed LLGDI will need much less computational complexity than FME, which is more efficient in real-world application such as image classification and visualization.

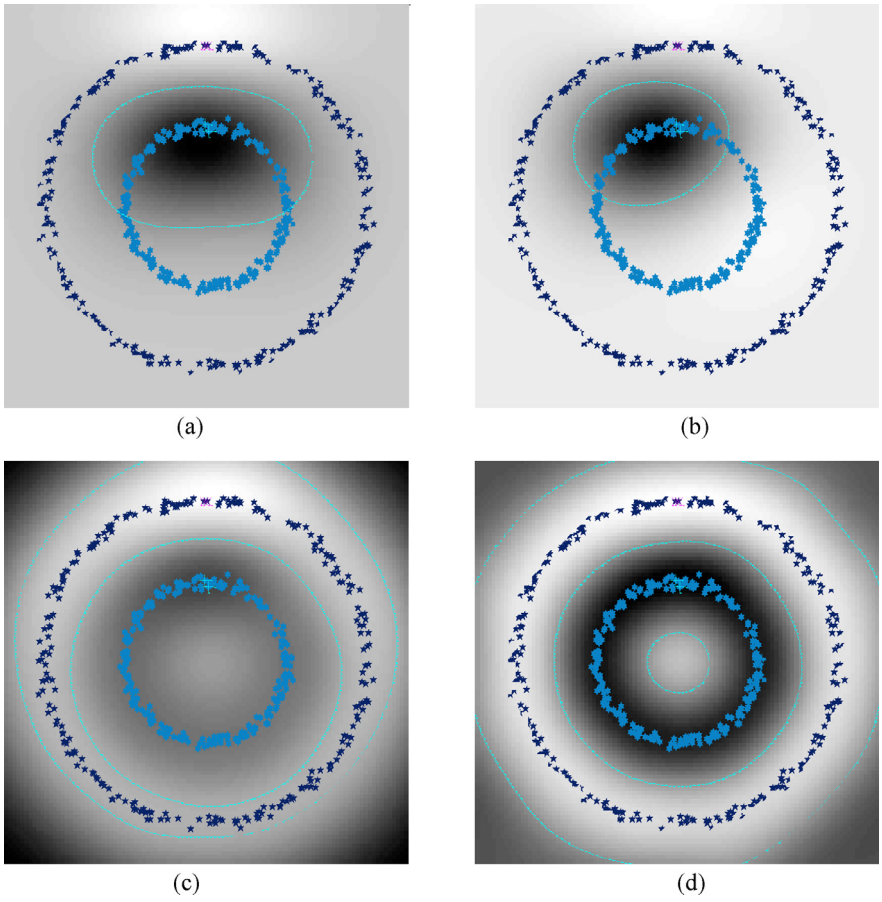


Fig. 2. Gray image of reduced space learned by KPCA, LDA,SDA and LLGDI: two-cycle dataset (a) KPCA, (b) KPCA+LDA, (c) KPCA+SDA, (d) KPCA+LLGDI.

5. Simulations

In this section, we evaluate our algorithms with three synthetic datasets and several real-world datasets. For the synthetic datasets, we evaluate the proposed method using two-cycle, two-Swiss-roll and two-plate datasets. For real-world datasets, we focus on solving the classification problems based on six real-world datasets which are all benchmark datasets. For classification problem, we use 8 real-world datasets to evaluate the performance of methods, which include UMNIST [5], Extended Yale-B [9], MIT-CBCL [24], COIL100 [11], ETH80 [10], USPS [8] datasets. Furthermore, we compare our algorithm with state-of-the-art supervised and semi-supervised algorithms. In the comparative study, we randomly split each dataset into training set and test set. We also randomly select samples from the training set to form labeled and unlabeled sets. All the training sets are preliminarily processed with a PCA operator to eliminate the null space before performing dimensionality reduction [19]. All algorithms used the training set in the output reduced space to train a nearest neighborhood classifier for evaluating the accuracy of test set.

5.1. Toy examples for synthetic datasets

In this toy example, we generate a dataset with two classes; each follows a cycle distribution with the same core but different radius. In each class, one sample is selected as labeled set and the remaining as unlabeled set. We then in Fig. 2 investigate the effectiveness of different methods based on the dataset. Since the distribution of two-cycle dataset is nonlinear, to handle this problem, we first perform KPCA to the two-cycle dataset; we then use the output in the full-rank KPCA to train the linear methods [29,35]. Other parameters are set the same as in two-plate dataset. Fig. 2 shows the gray images of decision surfaces and boundaries obtained by KPCA, LDA, SDA and LLGDI. From Fig. 2 we can observe that for the two-cycle dataset, KPCA fails to discover the decision boundary in two-cycle datasets. The main reason is that KPCA is an unsupervised method, which cannot grasp the discriminative structure embedded in the training set. In addition, though LDA is a supervised method, it cannot achieve better performance to KPCA given insufficient labeled samples, as it still cannot learn a better boundary that well separates the two classes. In contrast, by using the unlabeled samples to construct the manifold term for preserving the geometrical structure

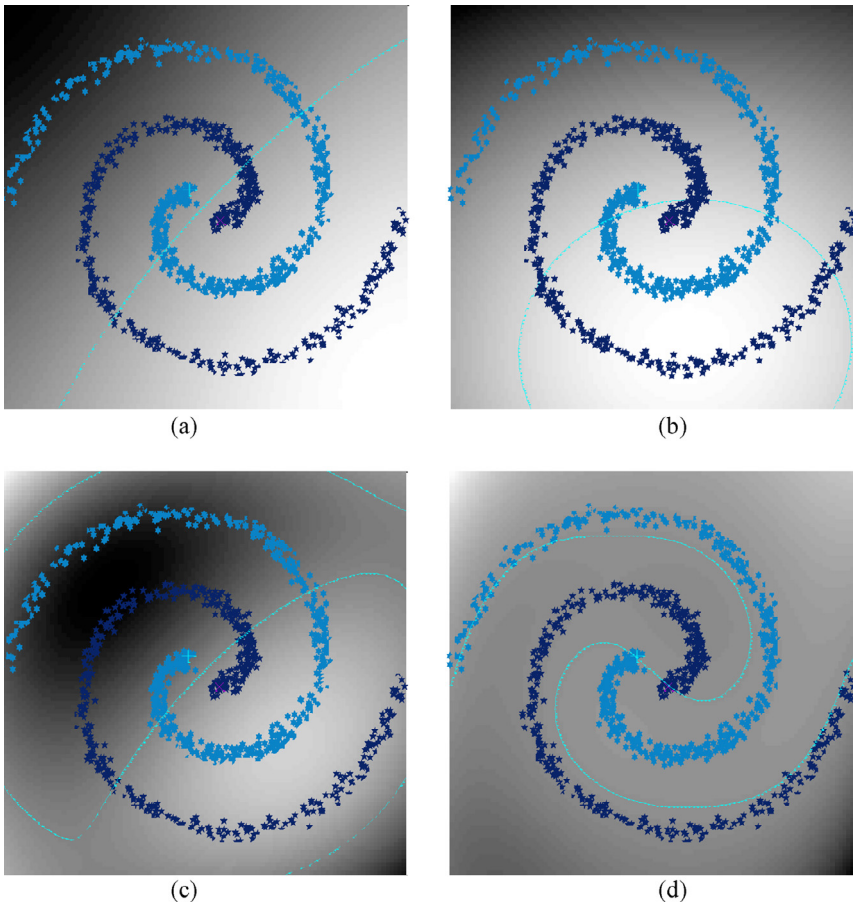


Fig. 3. Gray image of reduced space learned by KPCA, LDA,SDA and LLGDI: two-cycle dataset (a) KPCA, (b) KPCA+LDA, (c) KPCA+SDA, (d) KPCA+LLGDI.

embedded into the dataset, SDA can find the precise decision boundary. In practice, since labeling a large number of samples are time-consuming and impractical, semi-supervised DR methods have become more effective instead of only relying on supervised methods. In addition, the proposed LLGDI can achieve the best performance, as the decision boundary learned by LLGDI is more precise than those obtained by SDA. The improvement is reliable due to the fact that LLGDI preserves both local and global discriminative information embedded in dataset.

In order to further evaluate the effectiveness of the proposed methods, we generate a more challenging two-Swiss-roll dataset with two classes and each follows a Swiss-roll distribution with the same core but increased radius. In each class, only one sample is selected as labeled set and the remaining as unlabeled set. Note that the two-Swiss-roll dataset is also nonlinear. We use the same method to handle the nonlinear problem as in two-cycle dataset. Other parameters are set the same as in two-plate and two-cycle dataset. Fig. 3 shows the gray images of decision surfaces and boundaries obtained by KPCA, LDA, SDA and the proposed LLGDI, where the gray value of each pixel and boundary represents the same means as in two-moon and two-cycle dataset. In Fig. 3(a), we can see that similar to two-cycle dataset, KPCA still fails to discover the decision boundary in two-Swiss-roll dataset. In Fig. 3(b) and (c), we can also see that LDA and SDA have failed to discover the boundaries between two classes. This fact verifies that for two-Swiss-roll dataset, the limited labeled samples cannot provide sufficient discriminative information, hence causing the performance of classification unsatisfied. Moreover, in Fig. 3(c), we can see that even with sufficient unlabeled samples to preserve the geometrical structure, SDA does not perform well. This is mainly because compared with the two-moon dataset, the two-Swiss-roll dataset has more complex geometrical structure as the samples in two classes parallel revolve around the same core. Hence, there is not a clear boundary that can divide the two classes into two sides. In Fig. 3(d), we can see that though there are still some miss-classified samples, the proposed LLGDI can achieve the best performance in a way that the boundaries learned by LLGDI can well separate the samples in different classes by parallel revolving the same core. This enhancement is mainly due to LLGDI can preserve more discriminative information embedded in dataset.

5.2. Image classification

For classification problem, we use 8 real-world datasets to evaluate the performance of methods, which include UMNIST, Extended Yale-B, MIT-CBCL, COIL-100, ETH-80, USPS datasets. The UMIST dataset is a multi-view face dataset, consisting of 1012

Table 5
Dataset information and data partition for each dataset.

Dataset	Database type	#Samples	#Dim	#Class	#Training per class	#Test per class
UMNIST [5]	Face	1012	1024	20	20	Remains
Extended Yale-B [9]	Face	16,123	1024	38	50	Remains
MIT-CBCL [24]	Face	3240	1024	10	30	30
COIL100 [11]	Object	7200	1024	100	40	Remains
ETH80 [10]	Object	3280	1024	80	21	Remains
CASIA-HWDB [8]	Hand-written digit	2381	196	10	100	100

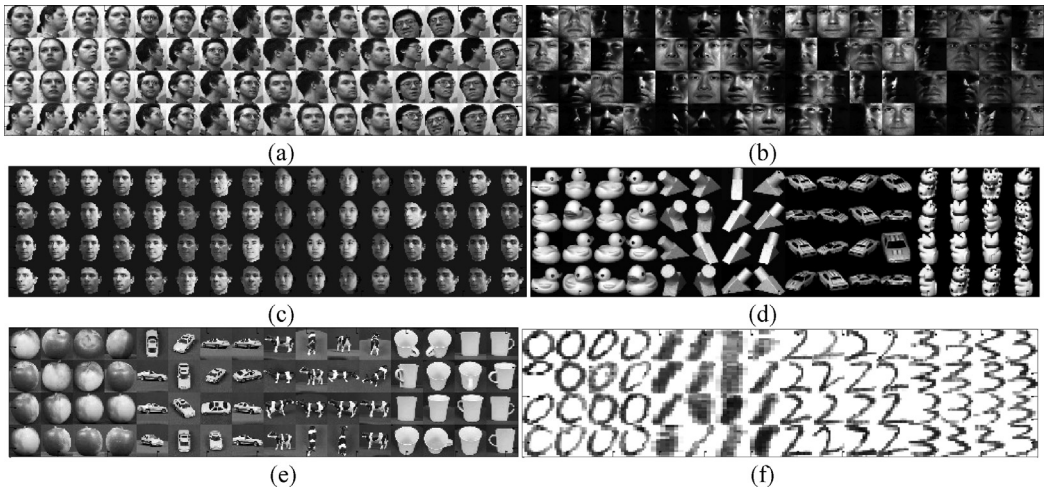


Fig. 4. Sample images of real-world datasets: (a) UMNIST dataset, (b) Extended Yale-B dataset, (c) MIT-CBCL dataset, (d) COIL100 dataset, (e) ETH80 dataset, (f) CASIA-HWDB dataset.

images of 20 people, each covering a wide range of poses from profile to frontal views. The size of each image is 112 × 92 with 256 gray levels per pixel. In our simulation, we down-sample the size of each image to 28 × 23 and no other preprocessing is performed. The Extended Yale-B dataset contains 16,123 images of 38 human subjects under nine poses and 64 illumination conditions. Similar to the MIT-CBCL dataset, the images are also cropped and resized to 32 × 32 pixels. This dataset now has around 64 near frontal images under different illuminations per individual. The MIT-CBCL dataset provides 3240 synthetic images rendered from 3D head models of 10 people. The head models are generated by fitting a morphable model to the high-resolution training images. The size of each image is originally 200 × 200 with 256 gray levels per pixel. In our simulation, we down-sample the size of each image to 32 × 32 and no other preprocessing is performed. The COIL100 dataset consists of images of 100 objects viewed from varying angles at the interval of five degrees, resulting in 72 images per object. The size of each cropped image is 128 × 128 with 24 bit color levels per pixel. In our simulation, we down-sample the size of image to 32 × 32 and transfer each image to 256 gray levels. ETH80 dataset contains 80 objects with each object represented by 41 views of images. The original size of each image is 128 × 128 with 24bit color levels per pixel. Similar to COIL100 dataset, we down-sample the size of image to 32 × 32 and transfer it to 256 gray levels. The CASIA-HWDB dataset is a handwritten image dataset which include both isolated characters and handwritten texts. In our work, we choose a subset from it which includes 10-digit images from 0 to 9. Then, the subset has a training set of 2381 samples with an image size of 14 × 14 in 256 gray levels. The detailed information of dataset and some sampled images of real-world datasets can be shown in Table 5 and Fig. 4. For each dataset, we randomly select 10, 50, 30, 40, 21 and 100 samples from each class as training samples for UMNIST, Extended Yale-B, MIT-CBCL, COIL-100, ETH-80, USPS datasets. The test set is then formed by the selected or all remaining samples. The data partitioning for each dataset is also given in Table 5.

Next, we compare our method with other supervised and semi-supervised dimension reduction methods. These methods include RLDA [1], SDA [3], Lap-RLS/L [2], least square solution for solving SDA in Eq. (16) (Table 1, we refer it as LS-SDA), FME [13,14] and the proposed LLGDI. Note that PCA is an unsupervised method while RLDA is a supervised method, and the remaining methods LLGDI are all semi-supervised methods. The simulation settings are as follows: For SDA, Lap-RLS/L, two parameters, i.e. α_t and α_m, need to be determined for balancing the trade-off between the manifold and Tikhonov terms. We use five-fold cross validation to determine the best values and the candidate set is $\{10^{-9}, 10^{-6}, 10^{-3}, 10^{0}, 10^{3}, 10^{6}, 10^{9}\}$. For RLDA, only the Tikhonov term parameter α_t involves, we use the above candidate set to determine the best value. For FME and the proposed LLGDI, an addition regularized parameter α_r is involved for balancing the trade-off between the regression residual term and other terms. We also use the same candidate set to determine the best value. The training set in all datasets are preliminarily processed with PCA operator to eliminate the null space before performing dimension reduction. For supervised methods such

Table 6
Average classification accuracy over 20 random splits on unlabeled set and test set of different datasets (means ± standard derivations).

Dataset	Method	4 labeled samples per class		7 labeled samples per class		10 labeled samples per class	
		Unlabeled Mean ± SD	Test Mean ± SD	Unlabeled Mean ± SD	Test Mean ± SD	Unlabeled Mean ± SD	Test Mean ± SD
UMNIST	Baseline	81.1 ± 0.9	80.2 ± 1.0	88.6 ± 0.7	88.3 ± 0.7	93.1 ± 0.6	93.0 ± 0.7
	RLDA	85.2 ± 0.6	85.0 ± 0.7	90.7 ± 0.5	90.4 ± 0.6	95.3 ± 0.4	94.4 ± 0.5
	SDA	86.4 ± 0.7	86.3 ± 0.7	92.1 ± 0.6	91.7 ± 0.7	96.2 ± 0.4	95.4 ± 0.5
	LS-SDA	86.4 ± 0.7	86.3 ± 0.7	92.1 ± 0.6	91.7 ± 0.7	96.2 ± 0.4	95.4 ± 0.5
	Lap-RLS/L	86.6 ± 0.7	86.0 ± 0.8	91.9 ± 0.3	91.9 ± 0.4	95.7 ± 0.5	95.3 ± 0.6
	FME	88.2 ± 0.6	87.7 ± 0.6	93.1 ± 0.3	92.9 ± 0.4	96.7 ± 0.5	96.1 ± 0.5
	LLGDI	89.0 ± 0.5	88.7 ± 0.5	94.1 ± 0.3	93.7 ± 0.4	97.6 ± 0.5	97.0 ± 0.5
Extended Yale-B	Baseline	50.5 ± 1.9	50.1 ± 1.9	63.9 ± 1.5	63.6 ± 1.7	69.8 ± 1.5	69.4 ± 1.6
	RLDA	74.8 ± 2.0	74.5 ± 2.1	81.6 ± 1.6	81.1 ± 1.6	89.6 ± 1.5	89.3 ± 1.7
	SDA	86.4 ± 1.9	86.0 ± 2.0	89.8 ± 1.7	89.5 ± 1.8	92.8 ± 1.5	92.6 ± 1.5
	LS-SDA	86.4 ± 1.9	86.0 ± 82.0	89.8 ± 1.7	89.5 ± 1.8	92.8 ± 1.5	92.6 ± 1.5
	Lap-RLS/L	86.7 ± 2.0	86.2 ± 2.2	90.1 ± 1.8	89.7 ± 1.9	93.1 ± 1.6	92.8 ± 1.7
	FME	88.2 ± 1.8	88.0 ± 2.0	91.3 ± 1.7	91.1 ± 1.9	93.9 ± 1.5	93.6 ± 1.6
	LLGDI	89.4 ± 1.9	89.0 ± 2.0	92.2 ± 1.7	92.0 ± 1.7	94.7 ± 1.5	94.4 ± 1.9
MIT-CBCL	Baseline	69.5 ± 3.5	68.9 ± 3.9	80.5 ± 6.9	79.8 ± 7.0	90.3 ± 4.0	90.1 ± 3.8
	RLDA	72.6 ± 2.9	73.8 ± 4.6	82.5 ± 7.6	82.2 ± 7.2	92.1 ± 3.2	92.9 ± 3.0
	SDA	75.7 ± 2.9	76.9 ± 4.7	84.7 ± 6.7	84.2 ± 6.5	95.0 ± 3.3	94.9 ± 3.1
	LS-SDA	75.7 ± 2.9	76.9 ± 4.7	84.7 ± 6.7	84.2 ± 6.5	95.0 ± 3.3	94.9 ± 3.1
	Lap-RLS/L	75.2 ± 4.0	76.1 ± 4.1	84.7 ± 6.5	84.1 ± 6.5	93.7 ± 4.3	93.7 ± 3.6
	FME	78.4 ± 2.5	78.1 ± 3.2	85.3 ± 5.4	85.1 ± 5.0	95.1 ± 3.7	94.9 ± 3.1
	LLGDI	80.9 ± 2.2	80.7 ± 3.1	87.1 ± 5.0	86.4 ± 5.2	96.4 ± 3.5	96.2 ± 3.1
COIL100	Baseline	69.7 ± 3.3	70.1 ± 2.8	77.2 ± 1.9	77.6 ± 1.5	79.8 ± 1.7	80.4 ± 1.8
	RLDA	72.5 ± 3.1	73.1 ± 2.7	79.5 ± 1.9	79.3 ± 2.3	82.8 ± 1.7	83.7 ± 1.7
	SDA	78.3 ± 2.9	75.4 ± 2.4	81.9 ± 2.0	79.9 ± 2.0	84.0 ± 1.2	83.8 ± 1.7
	LS-SDA	78.3 ± 2.9	75.4 ± 2.4	81.9 ± 2.0	79.9 ± 2.0	84.0 ± 1.2	83.8 ± 1.7
	Lap-RLS/L	77.4 ± 2.9	75.8 ± 2.6	81.2 ± 1.8	80.0 ± 1.7	83.4 ± 1.4	83.0 ± 1.7
	FME	79.6 ± 2.7	79.3 ± 2.3	83.3 ± 1.8	82.7 ± 1.9	85.7 ± 1.4	85.9 ± 1.5
	LLGDI	81.8 ± 2.4	81.5 ± 2.4	84.8 ± 1.4	83.3 ± 1.9	86.5 ± 1.5	86.5 ± 1.7
ETH80	Baseline	55.8 ± 3.5	55.2 ± 4.9	65.7 ± 4.4	65.2 ± 4.9	70.9 ± 3.0	71.0 ± 4.6
	RLDA	60.4 ± 3.1	60.6 ± 4.3	69.6 ± 5.0	69.8 ± 4.7	74.6 ± 2.9	75.8 ± 4.6
	SDA	64.1 ± 3.0	63.7 ± 4.3	71.6 ± 5.0	70.7 ± 4.8	75.7 ± 2.9	76.9 ± 4.7
	LS-SDA	64.1 ± 3.0	63.7 ± 4.3	71.6 ± 5.0	70.7 ± 4.8	75.7 ± 2.9	76.9 ± 4.7
	Lap-RLS/L	64.9 ± 3.4	65.2 ± 4.0	71.9 ± 5.0	70.6 ± 4.8	75.2 ± 4.0	76.1 ± 4.1
	FME	71.2 ± 4.6	71.0 ± 4.2	74.0 ± 4.2	72.8 ± 3.6	77.6 ± 3.5	77.6 ± 3.5
	LLGDI	73.6 ± 4.4	73.2 ± 3.8	75.8 ± 3.9	74.5 ± 3.1	79.4 ± 3.0	79.2 ± 3.3
CASIA-HWDB	Baseline	56.5 ± 4.8	56.2 ± 3.6	69.1 ± 3.0	68.8 ± 3.3	76.0 ± 3.5	75.6 ± 2.8
	RLDA	59.6 ± 4.2	59.9 ± 3.1	73.1 ± 2.8	73.5 ± 3.1	79.6 ± 3.3	79.1 ± 2.5
	SDA	62.6 ± 4.2	62.9 ± 3.2	75.1 ± 2.8	75.5 ± 3.1	80.7 ± 3.3	80.1 ± 2.5
	LS-SDA	62.6 ± 4.2	62.9 ± 3.2	75.1 ± 2.8	75.5 ± 3.1	80.7 ± 3.3	80.1 ± 2.5
	Lap-RLS/L	63.3 ± 4.4	63.5 ± 3.1	73.0 ± 2.7	72.8 ± 3.7	77.4 ± 3.2	77.2 ± 2.3
	FME	66.9 ± 3.6	66.9 ± 3.6	77.2 ± 2.5	77.3 ± 2.8	81.5 ± 3.3	80.1 ± 2.4
	LLGDI	68.7 ± 3.5	68.1 ± 3.6	78.5 ± 2.6	78.2 ± 2.9	82.8 ± 3.2	82.5 ± 2.2

as RLDA, we use only labeled set to train the learner. For semi-supervised dimension reduction methods, we use all the training set with both labeled and unlabeled sets to train the learner. Since most of the methods, such as RLDA, SDA, Lap-RLS/L and FME and the proposed LLGDI have a limited rank of $c-1$, we simply reduce the dimensionality of all methods to $c-1$. All methods used labeled set in the output reduced subspace to train a nearest neighborhood classifier in order to evaluate the classification accuracy of test set. We also compare the performance of nearest neighborhood classifier with other state-of-the-art methods as a baseline.

The average accuracies over 20 random splits with the above parameters for each dataset are shown in Table 6. From the simulation results, we can obtain the following observation: (1) given sufficient labeled samples, all the supervised and semi-supervised dimension reduction methods outperform nearest neighborhood classifier due to the utilization of label information and feature extraction; (2) the semi-supervised dimension reduction methods are better than the corresponding supervised methods. For example, SDA outperforms RLDA by about 5–6% in COIL100 dataset with 2 labeled samples per class. For other datasets, it can outperform by 2–3%. This indicates that by incorporating the unlabeled set into the training procedure, the classification performance can be markedly improved, as the manifold structure embedded in the dataset is preserved; (3) we also observe that both SDA and the least square solution in Table 1 can achieve the same classification results due to the reason as analyzed in Section 3; (4) the proposed LLGDI can deliver better accuracies than those delivered by other semi-supervised dimension reduction methods such as SDA and Lap-RLS/L by about 3–4% in most datasets. The improvement can even achieve

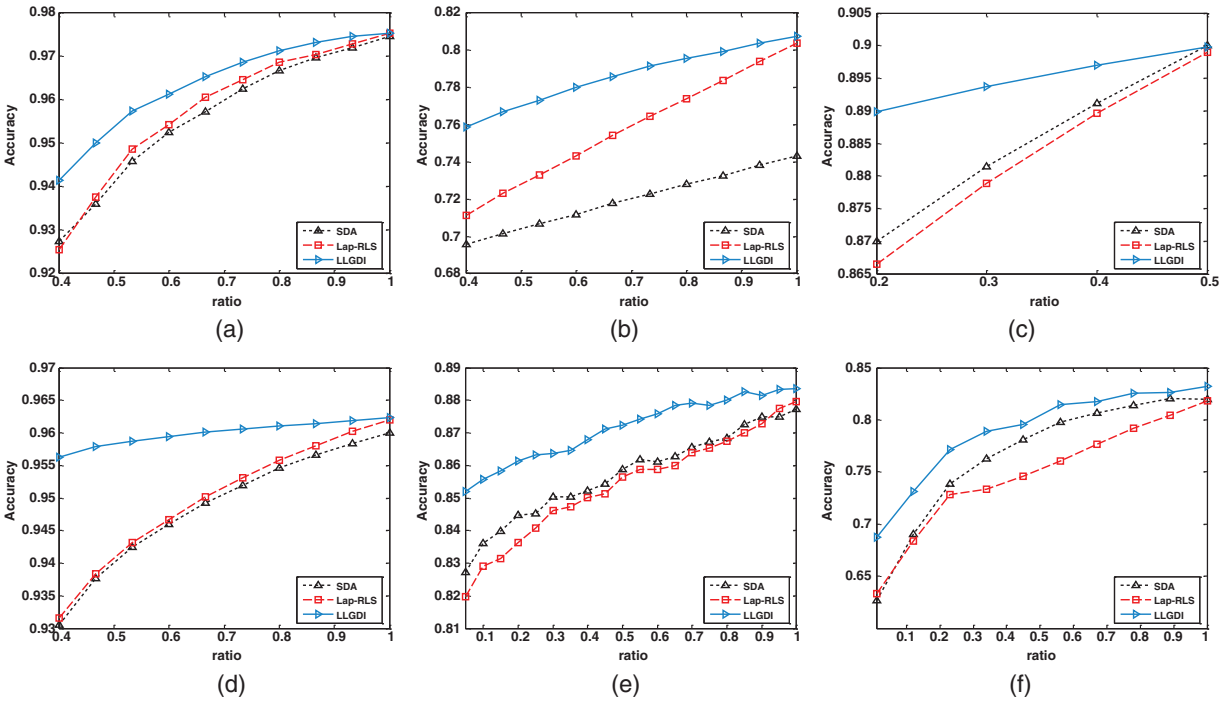


Fig. 5. Accuracy vs. ratio between the numbers of labeled set and that of training set: (a) UMNIST dataset, (b) Extended Yale-B dataset, (c) MIT dataset, (d) COIL100 dataset, (e) ETH80 dataset, (f) CASIA-HWDB dataset.

almost 8% in ETH80 dataset with 2 labeled samples per class. The improvement is believed to be true that LLGDI aims to characterize both local and global discriminative information embedded in dataset, which is better to handle classification problem; (5) we observe that LLGDI outperforms FME by about 2% in most cases. The main reason is that LLGDI has utilized a normalized local discriminative Laplacian matrix to preserve both manifold and discriminative structure in dataset, which is better than only relying on neighborhood graph; (6) we also evaluate LLGDI and compare it with SDA and Lap-RLS/L by fixing the number of training set and increasing the number of labeled set. The simulation results can be seen in Fig. 5. Following Fig. 5, we can observe that with the increase of labeled samples, the accuracies of three methods are all improved. However, LLGDI is more robust to the increase of labeled samples, specifically in COIL100 and USPS datasets. Another observation is that LLGDI can achieve better performances than SDA and Lap-RLS/L given few labeled samples. The reason for it is LLGDI incorporates local discriminative information into learning hence is more suitable for handling classification problem.

5.3. Image visualization

In order to further show the superiority of the proposed LLGDI, we demonstrate the visualization of the proposed method and compare it with other state-of-the-art methods such as PCA, LPP, LDA, SDA in Eq. (16), SDA2 in [41] and the proposed LLGDI. In this study, we randomly choose the first five classes of COIL100, UMINST, MIT and USPS dataset for simulation. In this dataset, we randomly choose the training set and testing set as in Table 3. In each training set, we randomly select 2 samples per class as labeled set for COIL100, UMNIST and MIT datasets while select 5 samples per class for CASIA-HWDB dataset. The remaining samples per class are selected as unlabeled set. Hence the number of labeled samples is quite few compared with that unlabeled samples. Our goal is to visualize the test set in the 2d subspace by projecting the test set on the 2d projection matrix learned by different methods. We do not compare the proposed LLGDI with Lap-RLS/L and FME for the sub-manifold visualization, as both of the latter methods are actually derived from regression problem and can only reduce the dimensionality reduction to the number of class c (c is usually larger than 2). The simulation results are shown in Figs. 6–9. From the simulation results, we can observe: (1) for unsupervised methods such as PCA and LPP, the sub-manifold structure can be well preserved. LPP delivers better performance than PCA due to the characteristics of LPP that the local information embedded in dataset is preserved; (2) all the supervised and semi-supervised methods outperform unsupervised methods such as PCA and LPP due to the utilization of label information; (3) the proposed LLGDI can achieve the much better performance than other methods (especially in COIL100 and MIT datasets), in a way that the sub-manifold of each object is closely conglomerated, while those belonging to different objects are clearly separated. The main reason is that LLGDI has incorporated local discriminative information into learning hence is more suitable for handling classification problem.

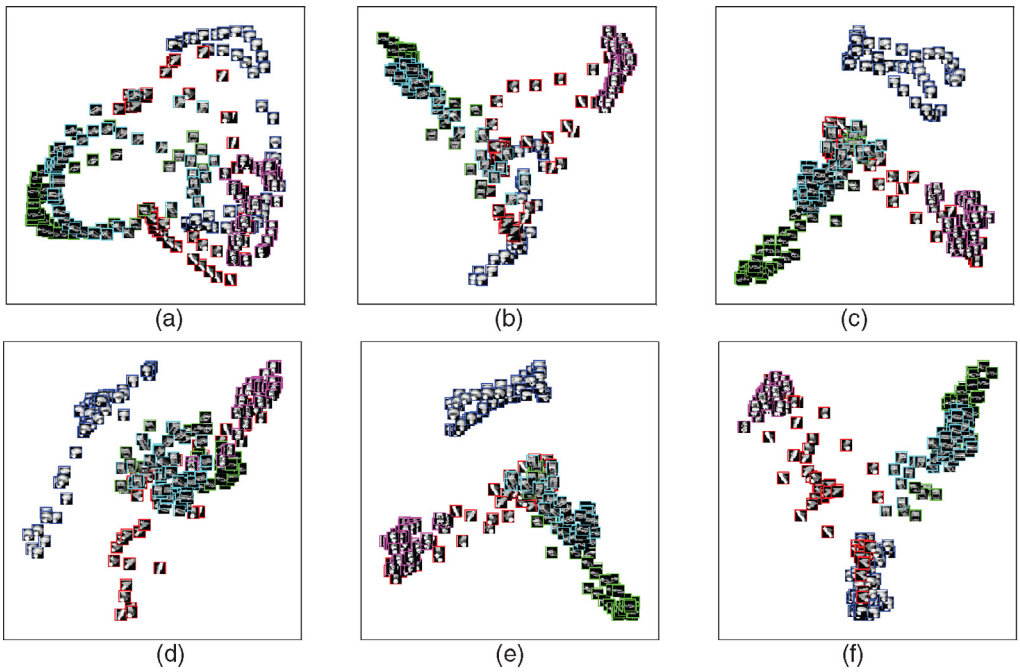


Fig. 6. Visualization performance of different methods: COIL100 dataset (first 5 classes, each color represents an object) (a) PCA, (b) LPP, (c) LDA, (d) SDA, (e) SDA2, (f) LLGDI.

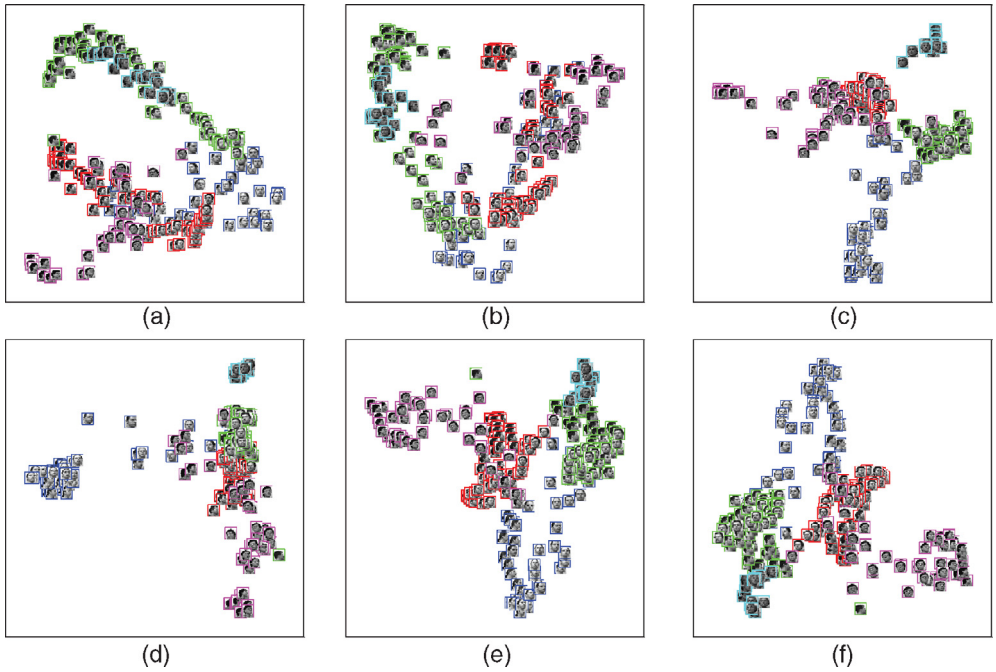


Fig. 7. Visualization performance of different methods: UMNIST dataset (first 5 classes, each color represents an object) (a) PCA, (b) LPP, (c) LDA, (d) SDA, (e) SDA2, (f) LLGDI.

5.4. Parameter analysis

In this subsection, we will give the parameter analysis for the LLGDI method. Note that following Eq. (30), the LLGDI method includes three main parameters, i.e. the local discriminative regularized parameter α_m, the global discriminative regularized parameter α_r and the number of neighborhoods k. We only need to provide the detailed analysis for the three parameters. Here, we first fix the number of neighborhoods k to a certain value, i.e. $k = 10$, and then evaluate the classification performance by

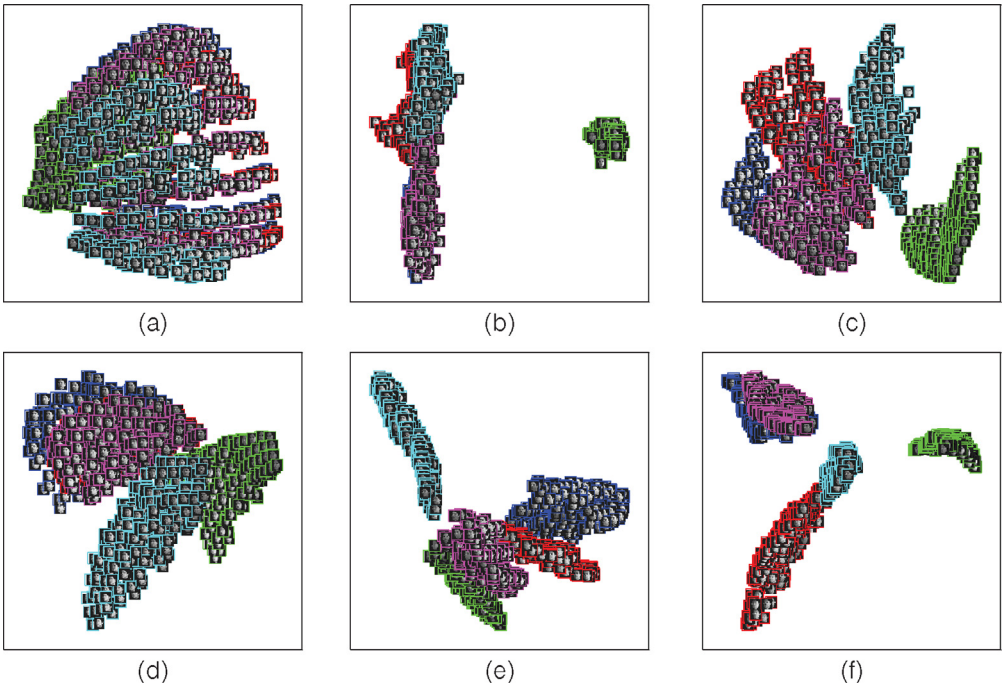


Fig. 8. Visualization performance of different methods: MIT dataset (first 5 classes, each color represents an object) (a) PCA, (b) LPP, (c) LDA, (d) SDA, (e) SDA2, (f) LLGDI.

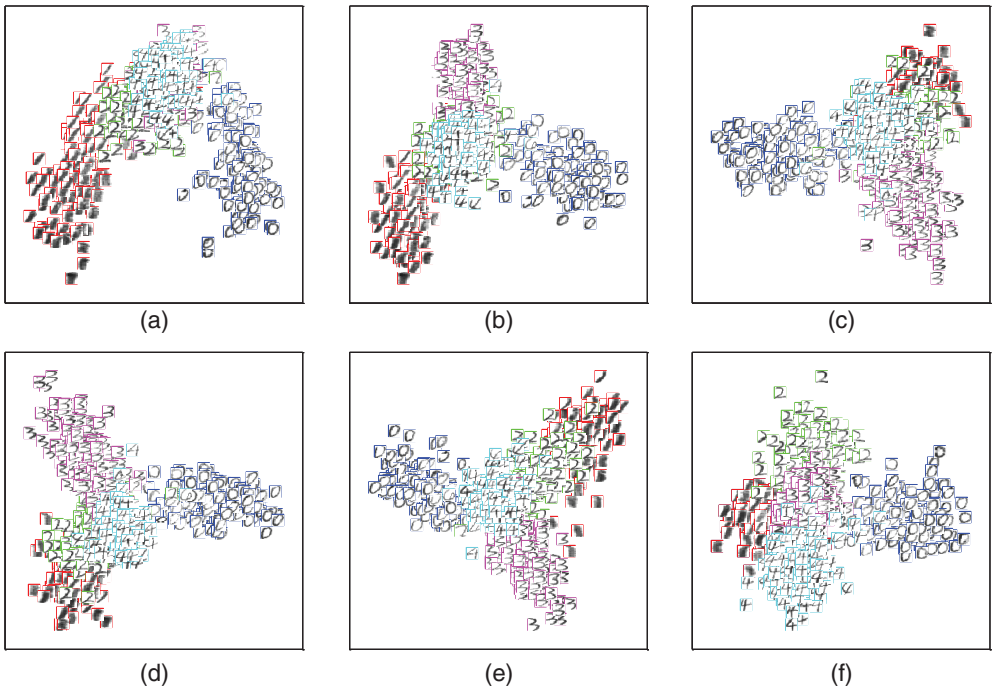
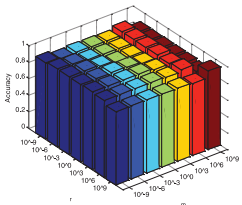


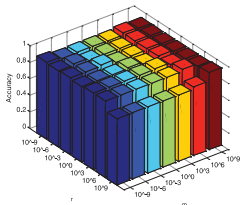
Fig. 9. Visualization performance of different methods: CASIA-HWDB dataset (first 5 classes, each color represents an object) (a) PCA, (b) LPP, (c) LDA, (d) SDA, (e) SDA2, (f) LLGDI.

choosing different values of α_m and α_r, where the candidate set for α_m and α_r is from $\{10^{-9}, 10^{-6}, 10^{-3}, 10^{0}, 10^{3}, 10^{6}, 10^{9}\}$. Next, after α_m and α_r have been tuned to their best value, we study the variation of performance by tuning the number of k from 4 to 40.

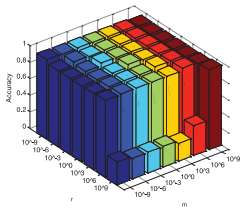
The average classification accuracies over random splits of six different datasets with varied α_m and α_r ($k = 10$) are shown in Fig. 10, and the detailed parameter analysis are given in Section 5.4. But in summary, we can draw the conclusion that for



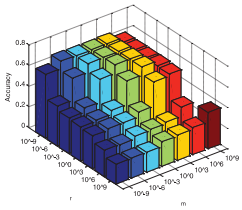
(a) 4 labels per class
UMNIST dataset



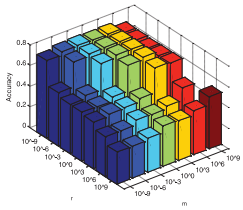
(b) 7 labels per class
UMNIST dataset



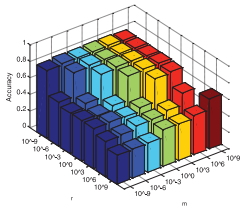
(c) 10 labels per class
UMNIST dataset



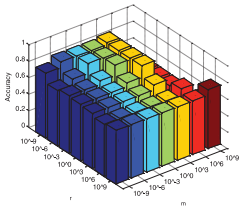
(a) 4 labels per class
Extended Yale-Bdataset



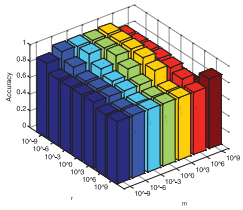
(b) 7 labels per class
Extended Yale-B dataset



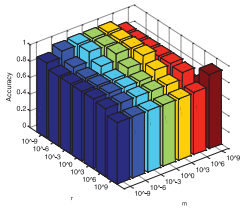
(c) 10 labels per class
Extended Yale-B dataset



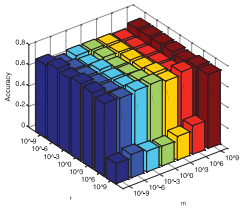
(a) 2 labels per class
MIT-CBCL dataset



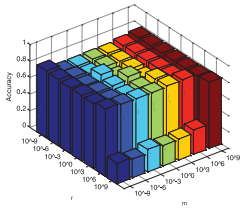
(b) 3 labels per class
MIT-CBCL dataset



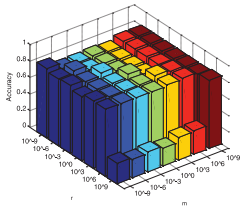
(c) 4 labels per class
MIT-CBCL dataset



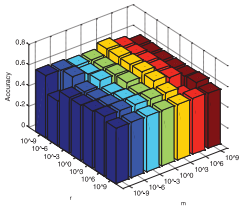
(a) 2 labels per class
COIL100 dataset



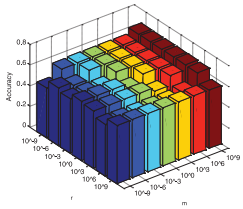
(b) 3 labels per class
COIL100 dataset



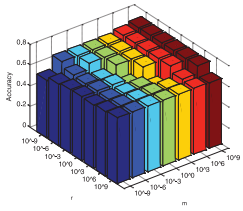
(c) 4 labels per class
COIL100 dataset



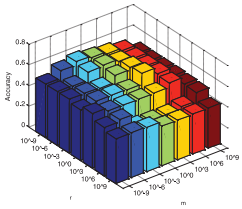
(a) 2 labels per class
ETH80 dataset



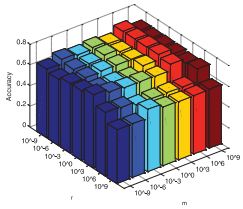
(b) 3 labels per class
ETH80 dataset



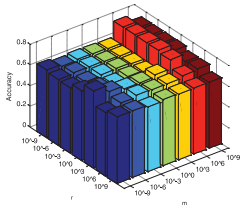
(c) 4 labels per class
ETH80 dataset



(a) 2 labels per class
CASIA-HWDB dataset



(b) 5 labels per class
CASIA-HWDB dataset



(c) 8 labels per class
CASIA-HWDB dataset

Fig. 10. The average accuracies over random splits of UMNIST, Extended Yale-B, MIT-CBCL, COIL100, ETH80 and Digit datasets with varied $\alpha_m \in \{10^{-9}, 10^{-6}, 10^{-3}, 10^{0}, 10^{3}, 10^{6}, 10^{9}\}$ and $\alpha_r \in \{10^{-9}, 10^{-6}, 10^{-3}, 10^{0}, 10^{3}, 10^{6}, 10^{9}\}$.

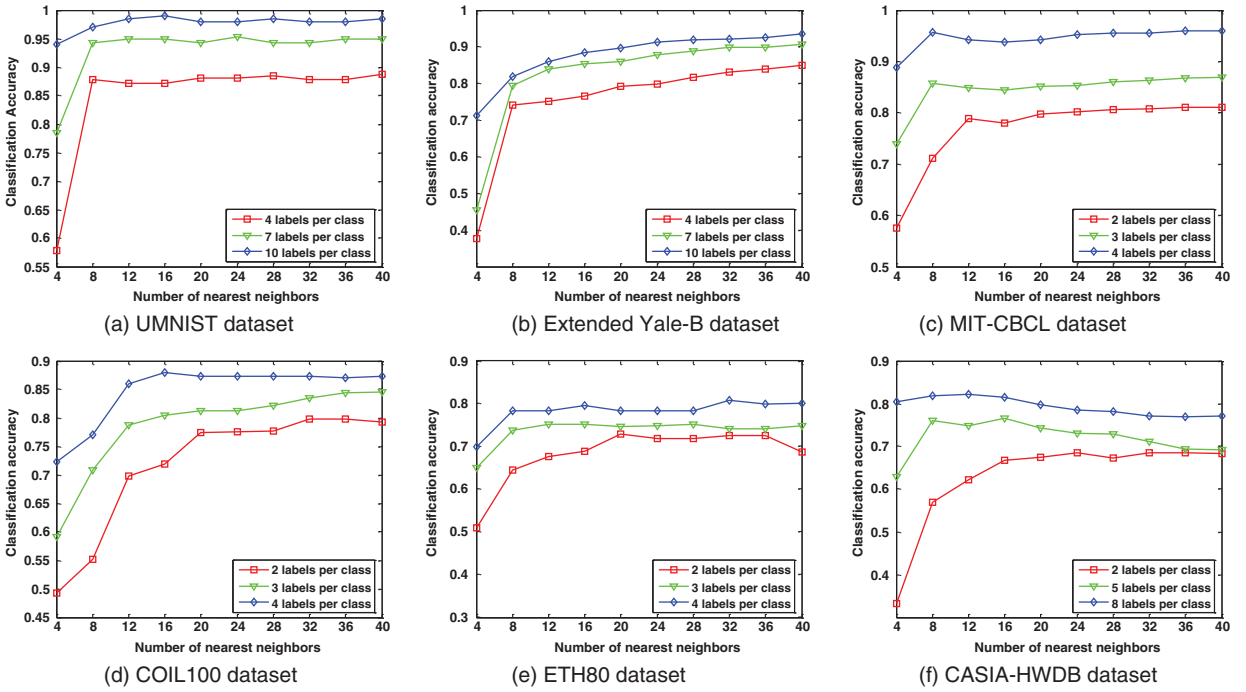


Fig. 11. The average accuracies over random splits of UMNIST, Extended Yale-B, MIT-CBCL, COIL100, ETH80 and Digit datasets with varied number of nearest neighbors k from 4 to 40.

most datasets, LLGDI method works better when α_m is large and α_r is small. For example, for UMNIST, COIL100 and ETH80 datasets, LLGDI can achieve much better classification accuracies when $\alpha_m \geq 1$ and $\alpha_r \leq 10^{-3}$; for Extended Yale-B, MIT-CBCL and CASIA-HWDB datasets, such good results can be obtained when $1 \leq \alpha_m \leq 10^6$ and $\alpha_r \leq 10^{-3}$. The above observations indicate that for LLGDI the local regression regularized term plays a more important role in calculating the projection matrix than the global regression regularized term. This is reasonable because the local regression regularized term characterizes discriminative information and local geometrical structure of a dataset. Thus LLGDI can preserve more discriminative information instead of only relying on the global regression regularized term. In practice, since the proposed LLGDI is relatively robust to the parameter α_r. When α_r is small ($\alpha_r \leq 10^{-3}$), we can simply set $\alpha_r = 10^{-9}$ and adjust α_m from $1 \leq \alpha_m \leq 10^6$ for conducting the simulations.

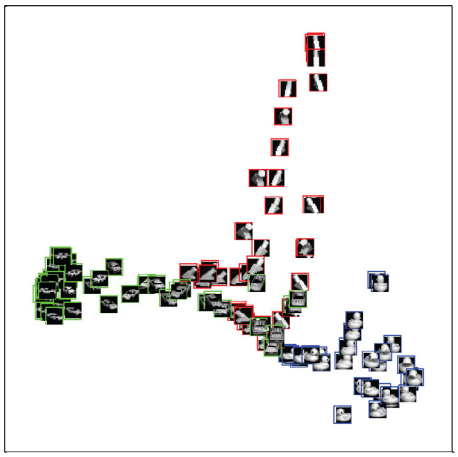
After α_m and α_r are best tuned according to the above process, we then study the variation of classification performance by adjusting the number of k from 4 to 40. Fig. 11 shows the average classification accuracies over a number of random splits of six different datasets with varied k. From Fig. 11, we can see that for most datasets, LLGDI is robust to the parameter k if k is not too small. For example, in UMNIST, MIT-CBCL and COIL100 datasets, the classification accuracies will almost have no change when $k \geq 16$. Such trends can also be observed in other datasets. In practice, since most of the datasets have wide intervals of k, we can easily set k for simulations given k is not too small. Clearly, a method that has wide intervals of parameters means it is more robust to the parameters, which is more suitable and impractical for real-world image classification.

5.5. Analysis of normalized weight vector

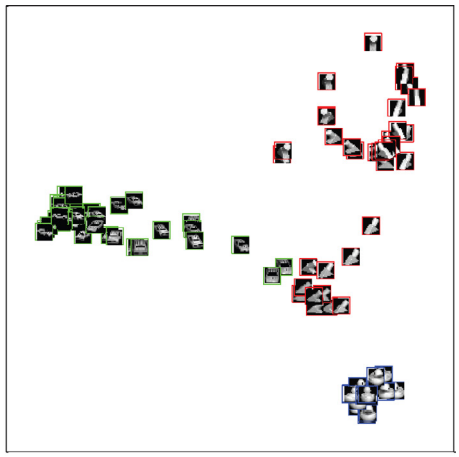
In this subsection, we will evaluate the effectiveness of the normalized weight vector. Note that the special weight vector is to normalize the graph Laplacian matrix of local regression term. The normalization can strengthen the local regressions in the low-density region but weaken those in the high density region, which is to handle the dataset with multi-density distribution or imbalanced dataset. This can be useful as sampling is usually not uniform in practice, and over-emphasizing the neighborhoods with high densities may occlude the information in sparse regions. Our objective is to show the effectiveness of the special weight vector for handling multi-density distributed dataset from a view of visualization instead of classification.

In this study, we first choose three classes from COIL100, MIT-CBCL and CASIA-HWDB datasets to form training set. We also let the number of sampled data in each class different from each other, which is used to make the dataset imbalanced or multi-density distributed. In each training set, we randomly select 2 samples per class as labeled set for COIL100 and MIT datasets while select 5 samples per class for CASIA-HWDB dataset. The remaining samples per class are selected as unlabeled set. We are then visualize the training set in the 2d subspace by projecting the training set on the 2d projection matrix learned by LLGDI without normalization and LLGDI with normalization.

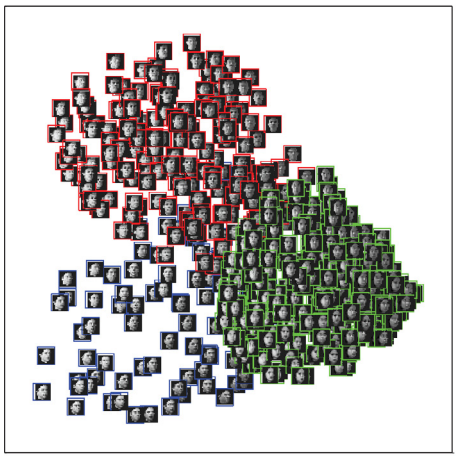
Fig. 12 shows simulation results. From Fig. 2, we can observe that the LLGDI with normalization can achieve much better performance than LLGDI without normalization, in a way that the sub-manifold of each object is closely conglomerated, while those belonging to different objects are clearly separated. For example in MIT-CBCL dataset, the green object is with high-density



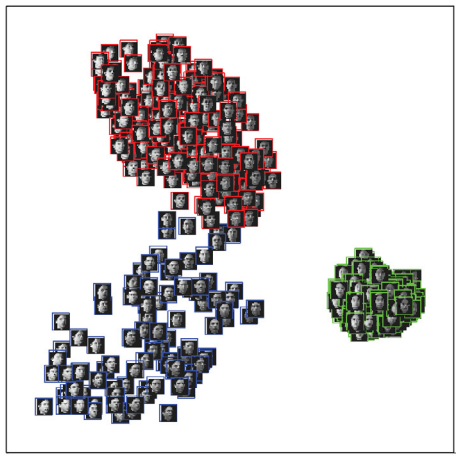
(a) LLGDI without normalization
COIL100 dataset



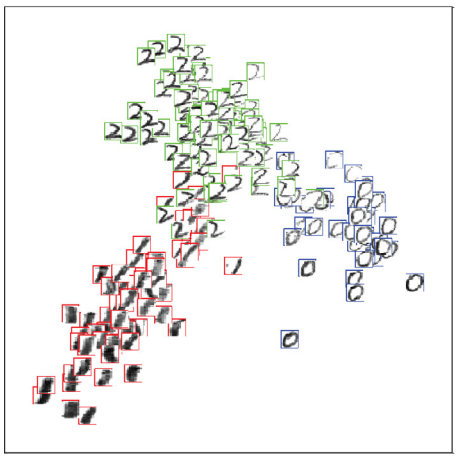
(b) LLGDI with normalization
COIL100 dataset



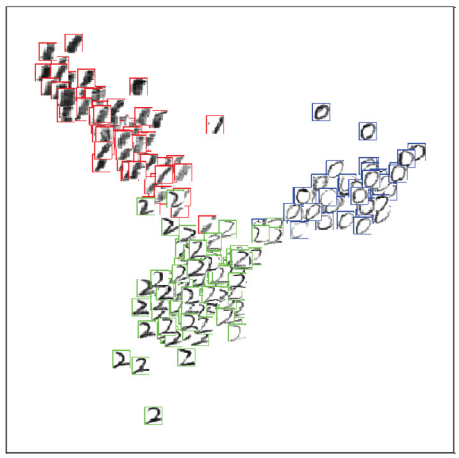
(a) LLGDI without normalization
MIT-CBCLdataset



(b) LLGDI with normalization
MIT-CBCLdataset



(a) LLGDI without normalization
CASIA-HWDB dataset



(b) LLGDI with normalization
CASIA-HWDB dataset

Fig. 12. Visualization performance compared between LLGDI without normalization and LLGDI with normalization of COIL100, MIT-CBCL and CASIA-HWDB datasets (first 3 classes with different class numbers, each color represents an object). (For interpretation of the references to color in the text, the reader is referred to the web version of this article.)

distribution. Without normalization, the red and blue objects, which are with low-density distribution, will get close to the green object. However, after normalization by utilizing the special weight vector, the local regressions in the low-density region (red and blue objects) will be strengthened while those in the high-density region (green object) will be weakened. As a result, the red and blue objects will be closely conglomerated to their own classes, making the three classes well separated from each other. The similar observations can also be obtained from other datasets, which verify the effectiveness of the special weight vector for handling multi-density dataset.

6. Conclusion

In this paper, we propose an effective LLGDI method for semi-supervised regression and dimensionality reduction. LLGDI is aimed at characterizing local and global discriminative manifold structure in a given dataset. This paper theoretically shows SDA can be addressed as a least square framework. An interesting equivalent relationship between SDA and Lap-RLS/L is derived under the uncorrelated constraint. As a result, the least square solution can be used for regression as well as subspace learning problem.

We propose the LLGDI method to preserve local discriminative, manifold information as well as the global discriminative information. Our study shows LLGDI can achieve better performances in almost all the studied cases. By adding an uncorrelated constraint to the objective function, LLGDI is extended to a dimensionality reduction method. As a result, LLGDI is able to solve regression and dimensionality reduction problem simultaneously. It is useful to show the connections between LLGDI and other methods. Theoretical results indicate other semi-supervised methods such as Lap-RLS/L, FME can be the special cases of LLGDI.

Despite being able to deliver promising results for image classification and visualization, LLGDI can be further improved. First, LLGDI method uses two parameters to balance the tradeoff between normalized local discriminative regularized term and global discriminative regularized term. In this study, they are determined using five-fold cross validation. Introducing a way to adaptively adjust the parameters would provide computational advantageous and performance improvement [20]. Second, we have added a special weight vector to each local regression error for normalizing graph Laplacian matrix. This approach is effective for handling multi-density dataset, but deriving a new weight vector to eliminate the effects of outliers before normalization would be useful especially when dataset has outlier problem. Finally, instead of solving image classification and visualization problem, how to utilize LLGDI for some real-world applications, such as content based image retrieval (CBIR) or even multimedia retrieval [30,31], are other challenges and of great importance. Our proposed method can be extended or improved to be applicable for these tasks.

Acknowledgment

This work was partly supported by the National Natural Science Foundation of China (Grant No. 61300209), partly supported by major program of National Natural Science Foundation of China (Grant No. 61033013) and also partly supported by the National Natural Science Foundation of China (Grant No. 61402310).

Appendix

In order to prove that L_d is graph Laplacian matrix, we need to prove L_d is positive semi-definite matrix and the sum of each row or column of L_d is equal to zero. We first have the following lemmas:

Lemma 3. *For each local patch X_j, L_j can be reformulated as follows:*

$$L_j = \eta G_j \left(G_j^T X_j^T X_j G_j + \eta I\right)^{-1} G_j^T, \tag{41}$$

where $G_j = (I - \Delta_j e_k^T e_k / (e_k \Delta_j e_k^T)) \Delta_j^{-1/2} \in R^{k \times k}$.

Proof of Lemma 3. First, it can be easily noted that $G_j G_j^T = H_j$, which is verified as follows:

$$\begin{aligned} G_j G_j^T &= \left(I - \Delta_j e_k^T e_k / \left(e_k \Delta_j e_k^T\right)\right) \Delta_j \left(I - \Delta_j e_k^T e_k / \left(e_k \Delta_j e_k^T\right)\right)^T \\ &= \left(\Delta_j - \Delta_j e_k^T e_k \Delta_j / \left(e_k \Delta_j e_k^T\right)\right) \left(I - \Delta_j e_k^T e_k / \left(e_k \Delta_j e_k^T\right)\right)^T \\ &= \Delta_j - 2\Delta_j e_k^T e_k \Delta_j / \left(e_k \Delta_j e_k^T\right) + \Delta_j e_k^T \left(e_k \Delta_j e_k^T\right) e_k \Delta_j / \left(e_k \Delta_j e_k^T\right)^2 \\ &= \Delta_j - \Delta_j e_k^T e_k \Delta_j / \left(e_k \Delta_j e_k^T\right) = H_j \end{aligned} \tag{42}$$

Then, we have:

$$\begin{aligned} L_j &= H_j - H_j X_j^T (X_j H_j X_j^T + \eta I)^{-1} X_j H_j \\ &= G_j G_j^T - G_j G_j^T X_j^T (X_j G_j G_j^T X_j^T + \eta I)^{-1} X_j G_j G_j^T \\ &= G_j G_j^T - G_j G_j^T X_j^T X_j G_j \left(G_j^T X_j^T X_j G_j + \eta I\right)^{-1} G_j^T \end{aligned}$$

$$= G_j G_j^T - G_j\left(G_j^T X_j^T X_j G_j + \eta I - \eta I\right)\left(G_j^T X_j^T X_j G_j + \eta I\right)^{-1} G_j^T$$
$$= \eta G_j\left(G_j^T X_j^T X_j G_j + \eta I\right)^{-1} G_j^T. \tag{43}$$

The second equation holds as $A(A^TA + \lambda I)^{-1} = (AA^T + \lambda I)^{-1}A$, for any matrix A. We thus prove Lemma 3.

Lemma 4. *Given a positive semi-definite matrix C, DCD^T is a positive semi-definite matrix for any matrix D.*

Lemma 5. *Given a set of positive semi-definite matrixes $\{C_1, C_2 \ldots, C_n\}$, then $\sum_{j=1}^{n} C_j$ is a positive semi-definite matrix.*

We neglect the proofs of Lemmas 4 and 5 as they can be seen in [7,32]. Then with Lemmas 3–5, we can easily prove Theorem 4 as follows:

Proof of Theorem 4. Note that following Lemma 3, we reformulate each L_j as $L_j = \eta G_j(G_j^T X_j^T X_j G_j + \eta I)^{-1} G_j^T$. It can be noted $(G_j^T X_j^T X_j G_j + \eta I)^{-1}$ is a positive semi-definite matrix, then, following Lemmas 4 and 5, we have each $\eta S_j G_j (G_j^T X_j^T X_j G_j + \eta I)^{-1} G_j^T S_j^T$ is positive semi-definite matrix and L_d, i.e.

$$L_d = \sum\nolimits_{j=1}^{l+u}\left(S_j L_j S_j^T\right) = \sum\nolimits_{j=1}^{l+u}\left(\eta S_j G_j\left(G_j^T X_j^T X_j G_j + \eta I\right)^{-1} G_j^T S_j^T\right). \tag{44}$$

is also positive semi-definite matrix.
In addition, for each $\eta S_j G_j (G_j^T X_j^T X_j G_j + \eta I)^{-1} G_j^T S_j^T$, we have $S_j^T e^T = e_k^T$ and

$$G_j^T e_k^T = \left(e_k^T - \left(e_k^T \Delta_j e_k\right) e_k^T / \left(e_k \Delta_j e_k^T\right)\right)\Delta_j^{-1/2} = \left(e_k^T - e_k^T\right)\Delta_j^{-1/2} = 0$$
$$\Rightarrow \eta S_j G_j\left(G_j^T X_j^T X_j G_j + \eta I\right)^{-1} G_j^T S_j^T e^T = 0$$
$$\Rightarrow L_d e^T = \sum\nolimits_{j=1}^{l+u}\left(\eta S_j G_j\left(G_j^T X_j^T X_j G_j + \eta I\right)^{-1} G_j^T S_j^T\right) e^T = 0. \tag{45}$$

which indicates that the sum of each row or column of L_d is equal to zero. We thus prove L_d is graph Laplacian matrix.

References

[1] P.N. Belhumeur, J.P. Hespanha, D.J. Kriegman, Eigenfaces vs. Fisherfaces: recognition using class specific linear projection, IEEE Trans. Pattern Anal. Mach. Intell. 19 (7) (1997) 711–720.
[2] M. Belkin, P. Niyogi, V. Sindhwani, Manifold regularization: a geometric framework for learning from labeled and unlabeled examples, J. Mach. Learn. Res. 7 (2006) 2399–2434.
[3] D. Cai, X. He, J. Han, Semi-supervised discriminant analysis, in: IEEE International Conference on Computer Vision (ICCV), 2007, pp. 1–7.
[4] K. Fukunaga, Introduction to Statistical Pattern Classification, Academic Press, 1990.
[5] D.B. Graham, N.M. Allinson, Characterizing virtual eigensignatures for general purpose face recognition in face recognition: from theory to application, NATO ASI Ser. F, Comput. Syst. Sci. 163 (1998) 446–456.
[6] X. He, S. Yan, Y. Hu, P. Niyogi, H. Zhang, Face recognition using Laplacianfaces, IEEE Trans. Pattern Anal. Mach. Intell. 27 (3) (2005) 328–340.
[7] R.A. Horn, C.R. Johnson, Matrix Analysis, Cambridge University Press, Cambridge, 1990.
[8] J. Hull, A database for handwritten text recognition research, IEEE Trans. Pattern Anal. Mach. Intell. 16 (5) (1994) 550–554.
[9] K.C. Lee, J. Ho, D. Kriegman, Acquiring linear subspaces for face recognition under variable lighting, IEEE Trans. Pattern Anal. Mach. Intell. 27 (5) (2005) 947–963.
[10] B. Leibe, B. Schiele, Analyzing appearance and contour based methods for object categorization, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 2, 2003, pp. 409–415.
[11] S.A. Nene, S.K. Nayar, H. Murase, Columbia Object Image Library (COIL-100), Columbia University, 1996 Technical report CUCS-005-96.
[12] F. Nie, S. Xiang, Y. Liu, C. Zhang, A general graph-based semi-supervised learning with novel class discovery, Neural Comput. Appl. 19 (4) (2010) 549–555.
[13] F. Nie, D. Xu, I.W.H. Tsang, C. Zhang, Flexible manifold embedding: a framework for semi-supervised and unsupervised dimensionality reduction, IEEE Trans. Image Process. 19 (7) (2010) 1921–1932.
[14] F. Nie, D. Xu, I.W.H. Tsang, C. Zhang, A flexible and effective linearization method for subspace learning, Graph Embedding for Pattern Analysis, 2013, pp. 177-203 (Book Chapter).
[15] F. Nie, Z. Zeng, I.W. Tsang, D. Xu, C. Zhang, Spectral embedded clustering: a framework for in-sample and out-of-sample spectral clustering, IEEE Trans. Neural Networks Learn. Syst. 22 (11) (2011) 1796–1808.
[16] S. Roweis, L. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (2000) 2323–2326.
[17] L. Sun, B. Ceran, J. Ye, A scalable two-stage approach for a class of dimensionality reduction techniques, in: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 2010, pp. 313–322.
[18] J.B. Tenenbaum, V. de Silva, J.C. Langford, A global geometric framework for nonlinear dimensionality reduction, Science 290 (2000) 2319–2323.
[19] M. Turk, A. Pentland, Face recognition using Eigenfaces, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 1991, pp. 586–591.
[20] D. Wang, F. Nie, H. Huang, Large-scale adaptive semi-supervised learning via unified inductive and transductive model, in: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 2014, pp. 482–491.
[21] F. Wang, A general learning framework using local and global regularization, Pattern Recognit. 43 (9) (2010) 3120–3129.
[22] F. Wang, C. Zhang, Label propagation through linear neighborhoods, IEEE Trans. Knowl. Data Eng. 20 (1) (2008) 55–67.
[23] J. Wang, F. Wang, C. Zhang, H.C. Shen, L. Quan, Linear neighborhood propagation and its applications, IEEE Trans. Pattern Anal. Mach. Intell. 31 (9) (2009) 1600–1615.
[24] B. Weyrauch, J. Huang, B. Heisele, V. Blanz, Component-based face recognition with 3D morphable models, in: First IEEE Workshop on Face Processing in Video, IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004, p. 85.
[25] S. Xiang, C. Pan, F. Nie, C. Zhang, TurboPixel segmentation using eigen-images, IEEE Trans. Image Process. 19 (1) (2010) 3024–3034.
[26] S. Xiang, F. Nie, C. Zhang, Semi-supervised classification via local spline regression, IEEE Trans. Pattern Anal. Mach. Intell. 32 (11) (2010) 2039–2053.
[27] S. Xiang, F. Nie, C. Zhang, Nonlinear dimensionality reduction with local spline embedding, IEEE Trans. Knowl. Data Eng. 21 (9) (2009) 1285–1298.

[28] S. Yan, D. Xu, B. Zhang, H. Zhang, Q. Yang, S. Lin, Graph embedding and extensions: a general framework for dimensionality reduction, IEEE Trans. Pattern Anal. Mach. Intell. 29 (1) (2007) 40–51.
[29] J. Yang, A.F. Frangi, J. Yang, D. Zhang, Z. Jin, KPCA plus LDA: a complete kernel Fisher discriminant framework for feature extraction and recognition, IEEE Trans. Pattern Anal. Mach. Intell. 27 (2) (2005) 230–244.
[30] Y. Yang, D. Xu, F. Nie, J. Luo, Y. Zhuang, Ranking with local regression and global alignment for cross media retrieval, in: ACM International Conference on Multimedia (MM), 2009, pp. 175–184.
[31] Y. Yang, F. Nie, D. Xu, J. Luo, Y. Zhuang, Y. Pan, A multimedia retrieval framework based on semi-supervised ranking and relevance feedback, IEEE Trans. Pattern Anal. Mach. Intell. 34 (5) (2012) 723–742.
[32] Y. Yang, D. Xu, F. Nie, S. Yan, Y. Zhuang, Image clustering using local discriminant models and global integration, IEEE Trans. Image Process. 19 (10) (2010) 2761–2773.
[33] J. Ye, Characterization of a family of algorithms for generalized discriminant analysis on undersampled problems, J. Mach. Learn. Res. 6 (2005) 483–502.
[34] J. Ye, Least square linear discriminant analysis, in: IEEE International Conference on Machine Learing (ICML), 2007, pp. 1087–1093.
[35] C. Zhang, F. Nie, S. Xiang, A general kernelization framework for learning algorithms based on kernel PCA, Neurocomputing 73 (4–6) (2010) 959–967.
[36] T. Zhang, D. Tao, X. Li, J. Yang, Patch alignment for dimensionality reduction, IEEE Trans. Knowl. Data Eng. 21 (9) (2009) 1299–1313.
[37] Z. Zhang, G. Dai, C. Xu, M.I. Jordan, Regularized discriminant analysis, ridge regression and beyond, J. Mach. Learn. Res. 11 (2010) 2199–2228.
[38] Z. Zhang, T.W.S. Chow, M. Zhao, Trace ratio optimization-based semi-supervised nonlinear dimensionality reduction for marginal manifold visualization, IEEE Trans. Knowl. Data Eng. 25 (5) (2013) 1148–1161.
[39] Z. Zhang, M. Zhao, T.W.S. Chow, Marginal semi-supervised sub-manifold projections with informative constraints for dimensionality reduction and recognition, Neural Networks 36 (2012) 97–111.
[40] M. Zhao, Z. Zhang, T.W.S. Chow, Trace ratio criterion based generalized discriminative learning for semi-supervised dimensionality reduction, Pattern Recognit. 45 (4) (2012) 1482–1499.
[41] M. Zhao, Z. Zhang, H. Zhang, Learning from local and global discriminative information for semi-supervised dimensionality reduction, in: The International Joint Conference on Neural Networks (IJCNN), 2013, pp. 1–8.
[42] M. Zhao, Z. Zhang, T.W.S. Chow, B. Li, Soft label based linear discriminant analysis for image recognition and retrieval, Comput. Vision Image Understanding 121 (2014) 86–99.
[43] M. Zhao, Z. Zhang, T.W.S. Chow, B. Li, A general soft label based linear discriminant analysis for semi-supervised dimension reduction, Neural Networks 55 (2014) 83–97.
[44] D. Zhou, O. Bousquet, T.N. Lal, J. Weston, B. Scholkopf, Learning with local and global consistency, in: Advances in Neural Information Processing Systems (NIPS), 2004, pp. 321–328.
[45] X. Zhu, Z. Ghahramani, J.D. Lafferty, Semi-supervised learning using Gaussian fields and harmonic functions, in: IEEE International Conference on Machine Learing (ICML), 2003, pp. 912–919.