Signed Laplacian Embedding for Supervised Dimension Reduction

Chen Gong[†,⋆] and **Dacheng Tao**[⋆] and **Jie Yang**[†] and **Keren Fu**[†]
[†]Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University
[⋆]Centre for Quantum Computation and Intelligent Systems, University of Technology Sydney
{goodgongchen, jieyang, fkrsuper}@sjtu.edu.cn
dacheng.tao@uts.edu.au

Abstract

Manifold learning is a powerful tool for solving nonlinear dimension reduction problems. By assuming that the high-dimensional data usually lie on a low-dimensional manifold, many algorithms have been proposed. However, most algorithms simply adopt the traditional graph Laplacian to encode the data locality, so the discriminative ability is limited and the embedding results are not always suitable for the subsequent classification. Instead, this paper deploys the signed graph Laplacian and proposes Signed Laplacian Embedding (SLE) for supervised dimension reduction. By exploring the label information, SLE comprehensively transfers the discrimination carried by the original data to the embedded low-dimensional space. Without perturbing the discrimination structure, SLE also retains the locality. Theoretically, we prove the immersion property by computing the rank of projection, and relate SLE to existing algorithms in the frame of patch alignment. Thorough empirical studies on synthetic and real datasets demonstrate the effectiveness of SLE.

Introduction

High-dimensional data often make data analysis intractable because of the unbearable computational burden. One common way to overcome this difficulty is to preprocess the data by dimension reduction, which seeks a low-dimensional representation of the original dataset while duly preserving its structural information, e.g. discrimination and locality. So far massive dimension reduction algorithms have been developed and demonstrated to be effective in various utilizations.

Traditional dimension reduction methods include Principal Component Analysis (PCA) (Hotelling 1933) and Linear Discriminative Analysis (LDA) (Fisher 1936). Both PCA and LDA assume that examples are drawn from Gaussian distributions.

To discover the locality encoded in non-Gaussian distributed data, manifold learning has been intensively investigated, which assumes that the high-dimensional data are often embedded in a low-dimensional intrinsic manifold $\mathcal{M}$. Typical algorithms include ISOMAP (Tenenbaum, Silva, and Langford 2000), Locally Linear Embedding (LLE) (Roweis and Saul 2000), Laplacian Eigenmaps (LE) (Belkin and Niyogi 2001), Locality Preserving Projections (LPP) (He and Niyogi 2004), Local Tangent Space Alignment (LTSA) (Zhang and Zha 2005), Structure Preserving Embedding (SPE) (Shaw and Jebara 2009), Maximum Variance Correction (MVC) (Chen, Chen, and Weinberger 2013), and Multi-scale Manifold Learning (Wang and Mahadevan 2013).

However, the manifold learning methods do not take the supervised information into account and only retains the locality in the low-dimensional space, so they may not improve the performance for classification tasks. Although Manifold Elastic Net (Zhou, Tao, and Wu 2011), Marginal Fisher Analysis (MFA) (Yan et al. 2007) and Discriminative Locality Alignment (DLA) (Zhang et al. 2009) encode the label information for graph construction, they simply preserve the local proximity relationship between similar or dis-similar pairs by implementations on a k nearest neighborhood (kNN) graph.

To address the defects of existing methodologies, we propose a novel approach called Signed Laplacian Embedding (SLE). Specifically, the signed graph Laplacian (Kunegis et al. 2010) is incorporated to equip SLE with better discriminative ability. Compared with the traditional graph Laplacian that simply encodes the pairwise similarity between data points, the signed graph Laplacian also exploits the label information, which guarantees the examples of different classes in the original space still separable in the embedded subspace. As a result, SLE is a global method since it treats all the examples belonging to one class together, and projects them as a whole when implements reduction. In this way, the discrimination in the original data can be well retained in the embedded low-dimensional space. Besides, we show that SLE has a strong connection with other typical manifold learning algorithms from the perspective of the patch alignment framework.

Model Description

Given a set of training examples $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n) \in \mathbb{R}^{m\times n}$ in which each column $\{\mathbf{x}_i\}_{i=1}^n \in \mathbb{R}^m$ represents m-dimensional data, then the target of dimension reduction is to find a projection matrix $\mathbf{A} \in \mathbb{R}^{m\times d}$ that properly maps these n examples to a smooth Riemannian manifold $\mathcal{M}$ em-

bedded in a low-dimensional ambient space $\mathbb{R}^d$ with $d < m$. Moreover, a dimension reduction method should have the generalizability that accurately predicts the low-dimensional representation of unseen test data $\mathbf{x}_t \in \mathbb{R}^m$ as $\mathbf{y}_t = \mathbf{A}^T\mathbf{x}_t$.

Signed Graph Laplacian

Popular manifold learning algorithms, such as LPP (He and Niyogi 2004) and LE (Belkin and Niyogi 2001), usually use a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ to represent the training set $\mathbf{X}$, in which $\mathcal{V}$ is the vertex set composed of the examples $\{\mathbf{x}_i\}_{i=1}^n$, and $\mathcal{E}$ is the edge set recording the relationship between them. The (i,j)-th element of the $\mathcal{G}$'s adjacency matrix $\mathbf{W}$ is computed as $W_{ij} = \exp(\|\mathbf{x}_i - \mathbf{x}_j\|^2/\sigma^2)$ (σ is the kernel width) if $\mathbf{x}_j$ is one of the k nearest neighbors of $\mathbf{x}_i$, and $W_{ij} = 0$ otherwise. Furthermore, by defining a diagonal matrix $\mathbf{D}$ as $D_{ii} = \sum_j W_{ij}$ for $i = 1, 2, \cdots, n$, the traditional graph Laplacian is formulated as $\mathbf{L} = \mathbf{D} - \mathbf{W}$.

Note that $\mathbf{L}$ above is only suitable for the $\mathbf{W}$ with elements in the range $[0, 1]$. In contrast, Kunegis et al. (2010) studies the signed graph $\tilde{\mathcal{G}}$, and defines a novel adjacency matrix $\tilde{\mathbf{W}}$ which contains both positive and negative elements. A diagonal matrix $\tilde{\mathbf{D}}$ is then represented as $\tilde{D}_{ii} = \sum_j \left|\tilde{W}_{ij}\right|$ ($i = 1, 2, \cdots, n$), and the *signed graph Laplacian* is given by $\tilde{\mathbf{L}} = \tilde{\mathbf{D}} - \tilde{\mathbf{W}}$. Similar to $\mathbf{L}$, it is easy to verify that $\tilde{\mathbf{L}}$ is also positive semi-definite.

The Algorithm

Based on the signed graph Laplacian introduced above, SLE is proposed and is divided into two major steps:

1. **Signed graph construction**: The vertices of $\tilde{\mathcal{G}}$ correspond to the training examples $\{\mathbf{x}_i\}_{i=1}^n$. $\mathbf{x}_i$ and $\mathbf{x}_j$ are connected by a positive edge if they belong to the same class, while they are linked by a negative edge if they come from different classes. Therefore, the elements in the adjacency matrix $\tilde{\mathbf{W}}$ are

$$\tilde{W}_{ij} = \begin{cases} 1, \mathbf{x}_i \; and \; \mathbf{x}_j \; are \; in \; the \; same \; class \\ -1, \mathbf{x}_i \; and \; \mathbf{x}_j \; are \; in \; the \; different \; classes \end{cases}. \tag{1}$$

2. **Finding the projection matrix**: The following generalized eigenvalue problem (2) is to be solved:

$$\mathbf{X}\tilde{\mathbf{L}}\mathbf{X}^T\mathbf{a} = \lambda\mathbf{X}\mathbf{X}^T\mathbf{a}. \tag{2}$$

Suppose $\lambda_1, \lambda_2, \cdots, \lambda_d$ are the eigenvalues obtained and arranged in ascending order, and the corresponding eigenvectors are $\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_d$, respectively, then the projection matrix is $\mathbf{A}_{m\times d} = (\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_d)$. Let $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_n) \in \mathbb{R}^{d\times n}$ denote the embedding results of the training data, then the projection on $\mathbf{X}$ can be easily expressed as $\mathbf{Y} = \mathbf{A}^T\mathbf{X}$.

Model Justification

As introduced above, the kNN graph is usually adopted by the existing manifold-based methods, and requires that the adjacent examples in $\mathbb{R}^m$ are not mapped far apart in $\mathbb{R}^d$. In other words, their target is to minimize $\sum_{ij} (\mathbf{y}_i - \mathbf{y}_j)^2 W_{ij}$. However, this measure simply focuses on preserving the local structure among examples, and W_{ij} ($i, j = 1, 2, \cdots, n$) are required to be nonnegative.

In contrast, we claim that: 1) The examples belonging to the same class, not simply originally nearby, are better to be projected together for the subsequent classification; 2) The elements in the adjacency matrix are allowed to take negative values to enable the incorporation of both similarity and dissimilarity information, so that better discriminative performance can be achieved. Therefore, taking the two-class situation as an example, we aim to minimize the following objective function:

$$Q = \frac{1}{2}\sum_{i=1}^n \sum_{j=1}^n (\mathbf{y}_i - s_{ij}\mathbf{y}_j)^2 \left|\tilde{W}_{ij}\right|, \tag{3}$$

in which $s_{ij} = 1$ if $\tilde{W}_{ij} \geq 0$ and $s_{ij} = -1$ if $\tilde{W}_{ij} < 0$. Suppose $\mathbf{a}$ is a projection vector, namely $\mathbf{y}_i^T = \mathbf{a}^T\mathbf{x}_i$, and considering that $s_{ij}^2 = 1$ and $\left|\tilde{W}_{ij}\right| s_{ij} = \tilde{W}_{ij}$, (3) is reformulated as

$$\begin{aligned} Q &= \frac{1}{2}\sum_{i=1}^n \sum_{j=1}^n (\mathbf{a}^T\mathbf{x}_i - \mathbf{a}^T\mathbf{x}_j s_{ij})^2 \left|\tilde{W}_{ij}\right| \\ &= \sum_{i=1}^n \mathbf{a}^T\mathbf{x}_i\tilde{D}_{ii}\mathbf{x}_i^T\mathbf{a} - \sum_{i=1}^n\sum_{j=1}^n \mathbf{a}^T\mathbf{x}_i \left|\tilde{W}_{ij}\right| s_{ij}\mathbf{x}_j^T\mathbf{a}. \\ &= \mathbf{a}^T\mathbf{X}(\tilde{\mathbf{D}} - \tilde{\mathbf{W}})\mathbf{X}^T\mathbf{a} \\ &= \mathbf{a}^T\mathbf{X}\tilde{\mathbf{L}}\mathbf{X}^T\mathbf{a} \end{aligned} \tag{4}$$

Moreover, to uniquely determine the minimizer $\mathbf{a}$, an additional constraint is imposed as

$$\mathbf{a}^T\mathbf{X}\mathbf{X}^T\mathbf{a} = 1. \tag{5}$$

Consequently, our task is to find the solution to the following optimization problem:

$$\begin{aligned} &\min_{\mathbf{a}} \; \mathbf{a}^T\mathbf{X}\tilde{\mathbf{L}}\mathbf{X}^T\mathbf{a} \\ &s.t. \; \mathbf{a}^T\mathbf{X}\mathbf{X}^T\mathbf{a} = 1 \end{aligned}. \tag{6}$$

By applying the Lagrangian multiplier method (Boyd and Vandenberghe 2004), (6) can be transformed into a generalized eigenvalue problem which shares the same formation as (2).

Extension to Multi-class Situations

The current model (6) is only suitable for two-class situations because its objective function (3) simply "pushes" the $\mathbf{x}_j$'s low-dimensional embedding $\mathbf{y}_j$ to $-\mathbf{y}_i$ if $\mathbf{x}_j$ belongs to a different class from $\mathbf{x}_i$, which will definitely cause ambiguity in multi-class situations.

We use the "one-versus-the-rest" strategy to adapt SLE to multiple classes. Suppose there are C classes in total, a projection matrix $\mathbf{A}_i$ is explicitly trained by (6) for each class c_i ($i = 1, 2, \cdots, C$), so that all the examples belonging to c_i (denoted as $\mathbf{X}^{(i)}$) are mapped to $\mathbf{Y}^{(i)} = \mathbf{A}_i^T\mathbf{X}^{(i)}$, while the examples of other classes (denoted as $\bar{\mathbf{X}}^{(i)}$) are embedded together as $\bar{\mathbf{Y}}^{(i)} = \mathbf{A}_i^T\bar{\mathbf{X}}^{(i)}$. To address the example imbalance problem during this process, we re-weight the $\tilde{W}_{ij}$ in (1) to $1/n_{c_i}$ (n_{c_i} is the number of examples in c_i) if $\mathbf{x}_i$ and $\mathbf{x}_j$ share the same label. Then, for test data $\mathbf{x}_t$ we implement dimension reduction using $\mathbf{A}_1, \mathbf{A}_2, \cdots, \mathbf{A}_C$ respec-

tively. The corresponding results are $\mathbf{y}_t^{(1)}, \mathbf{y}_t^{(2)}, \cdots, \mathbf{y}_t^{(C)}$, and the *membership degree* that $\mathbf{x}_t$ belongs to the i-th class is then calculated by

$$\theta(i) = \frac{\min_j d(\mathbf{y}_t^{(i)}, \bar{\mathbf{Y}}_{\cdot j}^{(i)})}{\min_j d(\mathbf{y}_t^{(i)}, \mathbf{Y}_{\cdot j}^{(i)})}. \tag{7}$$

In (7), $\mathbf{Y}_{\cdot j}^{(i)}$ denotes the j-th column of the matrix $\mathbf{Y}^{(i)}$, and $d(\cdot,\cdot)$ represents the Euclidean distance between two vectors. Consequently, $\mathbf{x}_t$ is classified into the class $c_t = \arg\max_{1\le i\le C}\theta(i)$, and its corresponding projection matrix is $\mathbf{A}_{c_t}$.

Theoretical Analyses

This section studies some theoretical properties related to SLE. We reveal that SLE has a solid statistical background, and prove that SLE is essentially an immersion between two smooth manifolds.

Statistical Viewpoint of SLE

SLE can also be understood from the statistical viewpoint. Suppose all the examples obey an underlying distribution P, and $\mathbf{x}_1, \mathbf{x}_2$ are two data points randomly drawn from P, then a projection $\mathbf{x} \to \mathbf{a}^T\mathbf{x}$ is defined that perfectly preserves the main structure of the whole dataset in the $\mathcal{L}^2$ space. If $\mathcal{S}$ and $\mathcal{D}$ are used to denote the positive and negative edge sets, the following mathematical expectation regarding $\mathbf{a}$ should be minimized:

$$E\left(\left|\mathbf{a}^T\mathbf{x}_1 - s_{12}\mathbf{a}^T\mathbf{x}_2\right|^2 | (\mathbf{x}_1,\mathbf{x}_2)\in\mathcal{S} \text{ or } (\mathbf{x}_1,\mathbf{x}_2)\in\mathcal{D}\right). \tag{8}$$

By denoting $\mathbf{z}_{12} = \mathbf{x}_1 - s_{12}\mathbf{x}_2$, (8) can be reformulated as

$$\begin{aligned} &E\left(\left|\mathbf{a}^T\mathbf{x}_1 - s_{12}\mathbf{a}^T\mathbf{x}_2\right|^2 | (\mathbf{x}_1,\mathbf{x}_2)\in\mathcal{S} \text{ or } (\mathbf{x}_1,\mathbf{x}_2)\in\mathcal{D}\right) \\ &= E\left(\left|\mathbf{a}^T\mathbf{z}_{12}\right|^2 | (\mathbf{x}_1,\mathbf{x}_2)\in\mathcal{S} \text{ or } (\mathbf{x}_1,\mathbf{x}_2)\in\mathcal{D}\right) \\ &= \mathbf{a}^T E\left(\mathbf{z}_{12}\mathbf{z}_{12}^T | (\mathbf{x}_1,\mathbf{x}_2)\in\mathcal{S} \text{ or } (\mathbf{x}_1,\mathbf{x}_2)\in\mathcal{D}\right)\mathbf{a}^T \end{aligned} \tag{9}$$

According to the Law of Large Numbers, the expectation in (9) can be estimated from the massive data as follows:

$$\begin{aligned} &E\left(\mathbf{z}_{12}\mathbf{z}_{12}^T | (\mathbf{x}_1,\mathbf{x}_2)\in\mathcal{S} \text{ or } (\mathbf{x}_1,\mathbf{x}_2)\in\mathcal{D}\right) \\ &\approx \frac{1}{|\mathcal{S}|+|\mathcal{D}|}\sum\nolimits_{(\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{S} \text{ or } (\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{D}} \mathbf{z}_{ij}\mathbf{z}_{ij}^T \\ &= \frac{1}{|\mathcal{S}|+|\mathcal{D}|}\sum\nolimits_{i=1}^n\sum\nolimits_{j=1}^n (\mathbf{x}_i - s_{ij}\mathbf{x}_j)(\mathbf{x}_i - s_{ij}\mathbf{x}_j)^T\left|\tilde{W}_{ij}\right| \\ &= \frac{4}{|\mathcal{S}|+|\mathcal{D}|}\left(\sum\nolimits_{i=1}^n \mathbf{x}_i\tilde{D}_{ii}\mathbf{x}_i^T - \sum\nolimits_{i=1}^n\sum\nolimits_{j=1}^n \mathbf{x}_i\mathbf{x}_j^T s_{ij}\left|\tilde{W}_{ij}\right|\right)' \\ &= \frac{4}{|\mathcal{S}|+|\mathcal{D}|}(\mathbf{X}\tilde{\mathbf{D}}\mathbf{X}^T - \mathbf{X}\tilde{\mathbf{W}}\mathbf{X}^T) \\ &= \frac{4}{|\mathcal{S}|+|\mathcal{D}|}\mathbf{X}\tilde{\mathbf{L}}\mathbf{X}^T \end{aligned} \tag{10}$$

where $|\cdot|$ denotes the size of a set, and $\tilde{\mathbf{L}}$ has the same formation as defined above. By plugging (10) into (9), and adding the constraint (5), we finally obtain the equivalent optimization problem as (6).

Rank of Projection

Suppose $\mathcal{M}$ and $\mathcal{N}$ are two manifolds, and $f : \mathcal{M} \to \mathcal{N}$ is a smooth mapping between them, then for each point $p \in \mathcal{M}$, the Jacobian df_p of f defines a linear mapping from the tangent plane of $\mathcal{M}$ at p (denoted as $T_p(\mathcal{M})$), to the tangent plane of $\mathcal{N}$ at $f(p)$ (denoted as $T_{f(p)}(\mathcal{N})$). The definition for the rank of f is provided in Definition 1, based on which we investigate the rank of SLE in this section.

Definition 1 (Rank, (Joncas and Meila 2013)) *A smooth mapping $f_i(x_1,\cdots,x_m) : \mathcal{M} \to \mathcal{N}$ with $i = 1, 2, \cdots, d$ has $rank(f) = r$ if the Jacobian $df_p : T_p\mathcal{M} \to T_{f(p)}\mathcal{N}$ of the map has rank r for all points $p = (x_1^p, \cdots, x_m^p) \in \mathcal{M}$, which is defined by*

$$\mathbf{J} = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_m} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_d}{\partial x_1} & \cdots & \frac{\partial f_d}{\partial x_m} \end{pmatrix}. \tag{11}$$

According to Definition 1, the Jacobian for SLE is exactly $\mathbf{A}^T$, so the next step is to find the rank of $\mathbf{A}$. Since $\mathbf{A}$ is composed of the generalized eigenvectors of (2), we first give an important theorem related to the generalized eigenvalue problem.

Theorem 2 (Q-orthogonal, (Golub and Loan 2012)) *For a generalized eigenvalue problem $\mathbf{P}\mathbf{v} = \lambda\mathbf{Q}\mathbf{v}$ with λ, $\mathbf{v}$ denoting the eigenvalue and eigenvector, respectively, if $\mathbf{P} = \mathbf{P}^T \in \mathbb{R}^{m\times m}$ and $\mathbf{Q} = \mathbf{Q}^T \in \mathbb{R}^{m\times m}$ with $\mathbf{Q}$ positive definite, then all eigenvalues λ_i ($i = 1, 2, \cdots m$) are real, and all eigenvectors are $\mathbf{Q}$-orthogonal, i.e., $\mathbf{v}_i^T\mathbf{Q}\mathbf{v}_j = 0$ for $i \neq j$ and $\mathbf{v}_i^T\mathbf{Q}\mathbf{v}_j \neq 0$ for $i = j$.*

Based on Theorem 2, we have the following theorem:

Theorem 3 *All the eigenvectors of* (2) *constitute a linearly independent set, so SLE is a full-rank projection.*

Proof: In SLE, $\tilde{\mathbf{L}}$ is symmetric and $\tilde{\mathbf{D}}$ is symmetric and positive definite, so it is easy to verify that the matrices $\mathbf{X}\tilde{\mathbf{L}}\mathbf{X}^T$ and $\mathbf{X}\tilde{\mathbf{D}}\mathbf{X}^T$ are equivalent to $\mathbf{P}$ and $\mathbf{Q}$ in Theorem 2. Therefore, the eigenvectors of (2) satisfy $\mathbf{a}_i^T\mathbf{X}\tilde{\mathbf{D}}\mathbf{X}^T\mathbf{a}_j = 0$ for $i \neq j$. Suppose the eigenvectors are linearly dependent, then without loss of generality, we may assume that $\mathbf{a}_1$ can be explicitly represented by other eigenvectors, namely $\mathbf{a}_1 = \alpha_2\mathbf{a}_2 + \cdots + \alpha_m\mathbf{a}_m$ ($\alpha_2, \cdots, \alpha_m$ cannot be all zeros). According to Theorem 2, $\mathbf{a}_1$ is $\mathbf{X}\tilde{\mathbf{D}}\mathbf{X}^T$-orthogonal to $\{\mathbf{a}_i\}_{i=2}^m$, so we have $(\alpha_2\mathbf{a}_2 + \cdots + \alpha_i\mathbf{a}_i + \cdots + \alpha_m\mathbf{a}_m)^T\mathbf{X}\tilde{\mathbf{D}}\mathbf{X}^T\mathbf{a}_i = \alpha_i\mathbf{a}_i^T\mathbf{X}\tilde{\mathbf{D}}\mathbf{X}^T\mathbf{a}_i = 0$. However, Theorem 2 reveals that $\mathbf{a}_i^T\mathbf{X}\tilde{\mathbf{D}}\mathbf{X}^T\mathbf{a}_i \neq 0$, so all the coefficients $\{\alpha_i\}_{i=2}^m$ should be 0, which means that $\mathbf{a}_1$ cannot be linearly represented by other eigenvectors. Therefore, all the eigenvectors of (2) are linearly independent, and the projection of SLE has $rank(f_{SLE}) = m$.

Theorem 3 reveals that the rank of f_{SLE} always equals the dimension of original space, so f_{SLE} is a valid embedding from $\mathcal{M}$ to $\mathcal{N}$, and the mapping f_{SLE} is called an *immersion* (Joncas and Meila 2013).

Relationship with Other Methods

Zhang et al. (2009) proposed a patch alignment framework which summarizes various dimension reduction algorithms into one unified formulation, such as PCA (Hotelling 1933), LLE (Roweis and Saul 2000), LPP (He and Niyogi 2004), and DLA (Zhang et al. 2009). Here we demonstrate that the proposed SLE can also be unified into this framework, based on which we further analyze the relationship between SLE and other approaches.

Theorem 4 *Let $\tilde{\mathbf{W}}_{\cdot i}$ be the i-th column of the generalized adjacency matrix $\tilde{\mathbf{W}}$, and $\mathbf{\Lambda}_{|\tilde{\mathbf{W}}_{\cdot i}|}$ is a diagonal matrix defined by $\mathbf{\Lambda}_{|\tilde{\mathbf{W}}_{\cdot i}|} = diag\Big(\Big|(\tilde{\mathbf{W}}_{\cdot i})_1\Big|, \Big|(\tilde{\mathbf{W}}_{\cdot i})_2\Big|, \cdots, \Big|(\tilde{\mathbf{W}}_{\cdot i})_n\Big|\Big) \in \mathbb{R}^{n\times n}$, then SLE can be regarded as a special case of patch alignment framework when the part optimization matrix is*

$$\tilde{\mathbf{L}}_i = \begin{pmatrix} \tilde{D}_{ii} & -\tilde{\mathbf{W}}_{\cdot i}^T \\ -\tilde{\mathbf{W}}_{\cdot i} & \mathbf{\Lambda}_{|\tilde{\mathbf{W}}_{\cdot i}|} \end{pmatrix}.$$

Proof: The patch alignment framework contains two steps: part optimization and whole alignment. In the part optimization step, we rewrite (3) as

$$\min_{\mathbf{y}_i} \frac{1}{2}\sum\nolimits_{i=1}^{n}\sum\nolimits_{j=1}^{n}(\mathbf{y}_i - s_{ij}\mathbf{y}_{ij})^2\Big|(\tilde{\mathbf{W}}_{\cdot i})_j\Big|, \quad (12)$$

where $\mathbf{y}_{ij}$ $(j = 1, 2, \cdots, n)$ are the n connected points[1] of the central example $\mathbf{y}_i$. Therefore, (12) can be rearranged as

$$\min_{\mathbf{y}_i}\frac{1}{2}\sum_{i=1}^{n} tr\left[\begin{pmatrix}(\mathbf{y}_i - s_{i1}\mathbf{y}_{s1})^T\\ \vdots \\ (\mathbf{y}_i - s_{in}\mathbf{y}_{sn})^T\end{pmatrix}\begin{pmatrix}\mathbf{y}_i - s_{i1}\mathbf{y}_{s1} & \cdots & \mathbf{y}_i - s_{in}\mathbf{y}_{sn}\end{pmatrix} diag\Big(\Big|(\tilde{\mathbf{W}}_{\cdot i})_1\Big|, \cdots, \Big|(\tilde{\mathbf{W}}_{\cdot i})_n\Big|\Big)\right]. \quad (13)$$

By denoting $\mathbf{Y}_i = (\mathbf{y}_i, \mathbf{y}_{i1}, \cdots \mathbf{y}_{in}) \in \mathbb{R}^{d\times(n+1)}$, $\mathbf{e}_n = [1, \cdots, 1] \in \mathbb{R}^n$, $\mathbf{S}_n = diag(s_{i1}, s_{i2}, \ldots, s_{in}) \in \mathbb{R}^{n\times n}$, and $\mathbf{I}_n$ as an $n \times n$ identity matrix, (13) can be further formulated as

$$\begin{aligned}&\min_{\mathbf{Y}_i} \sum\nolimits_{i=1}^{n} tr\left[\mathbf{Y}_i\begin{pmatrix}\mathbf{e}_l^T \\ -\mathbf{S}_n\mathbf{I}_n\end{pmatrix}\mathbf{\Lambda}_{|\tilde{\mathbf{W}}_{\cdot i}|}\begin{pmatrix}\mathbf{e}_n & -\mathbf{S}_n\mathbf{I}_n\end{pmatrix}\mathbf{Y}_i^T\right],\\ &\Leftrightarrow \min_{\mathbf{Y}_i} \sum\nolimits_{i=1}^{n} tr(\mathbf{Y}_i\tilde{\mathbf{L}}_i\mathbf{Y}_i^T)\end{aligned} \quad (14)$$

in which the part optimization matrix

$$\tilde{\mathbf{L}}_i = \begin{pmatrix}\mathbf{e}_n^T \\ -\mathbf{S}_n\mathbf{I}_n\end{pmatrix}\mathbf{\Lambda}_{|\tilde{\mathbf{W}}|}\begin{pmatrix}\mathbf{e}_n & -\mathbf{S}_n\mathbf{I}_n\end{pmatrix} = \begin{pmatrix}\tilde{D}_{ii} & -\tilde{\mathbf{W}}_{\cdot i}^T \\ -\tilde{\mathbf{W}}_{\cdot i} & \mathbf{\Lambda}_{|\tilde{\mathbf{W}}_{\cdot i}|}\end{pmatrix}. \quad (15)$$

(14) is used to compute the part optimizations of all the examples, which reveals that the *patch* of $\mathbf{x}_i$ can be properly established by $\mathbf{x}_i$ and the remaining data. By using the iterative method as Zhang et al. (2009) suggests, the alignment matrix $\tilde{\mathbf{L}}$ can finally be obtained in the whole alignment step, which is the same as the signed graph Laplacian adopted by SLE. This completes the proof.

Based on Theorem 4, the characteristics of SLE and other dimension reduction methods are summarized in Table 1, from which we observe that SLE mainly differs from other methods in the construction of patches and the representation of part optimization. In detail, the patches of PCA and SLE are globally constructed by using all the examples in the dataset, while LPP, LLE and DLA establish each patch by an example and its nearest neighbours. Besides, Table 1 reveals that LPP, DLA and SLE preserve the proximity relationship in a patch through the adjacency matrices, which are different from LLE that preserves reconstruction coefficients obtained in the original high-dimensional space. The superiority of SLE brought about by these differences will be empirically demonstrated in the experiments.

Experimental Results

This section evaluates the performance of SLE on synthetic and practical datasets. Two unsupervised methods, PCA (Hotelling 1933) and LPP (He and Niyogi 2004), and two supervised methods, MFA (Yan et al. 2007) and DLA (Zhang et al. 2009), serve as the baselines for comparisons. Of these, PCA is a global method, while the others simply use the kNN graph to preserve the local structure of the data. In this section, each of the adopted real-world datasets has 10 different partitions, and the original dataset in each partition was randomly split into a training set and a test set. All the algorithms were independently tested on these 10 partitions and the final results were averaged over the outputs of the 10 tests. Note that in these 10 tests, the ratio of the sizes of training and test sets for all the algorithms was identical.

Toy Data

The *SwissRoll* synthetic dataset was used to visualize the performances of SLE and other baselines. For *SwissRoll*, the data cloud in the 3D space is shaped like a roll. The examples with z-coordinates less than -5 are regarded as positive, while the examples with z-coordinates above -5 constitute the negative class. Of particular note is the fact that the points belonging to one class are distributed discontinuously along the high-dimensional manifold, as shown in Fig. 1 (a).

UCI Data

Six UCI datasets (Frank and Asuncion 2010), *BreastCancer*, *Musk*, *Waveform*, *Seeds*, *SPECT* and *Wine*, were adopted to test the dimension reduction performances of the algorithms compared.

All the feature vectors in these six databases were normalized to $[0, 1]$, and the RBF kernel width σ in LPP, MFA and DLA was chosen from the set $\{0.01, 0.1, 0.5, 1, 10\}$. The numbers of neighbors for graph construction, such as k in LPP, k_1 and k_2 in MFA and DLA, were chosen from $\{5, 10, 15, 20, 25, 30\}$. The optimal results were then reported over the different selections of these parameters. Similar to (Yan et al. 2007) and (Zhang et al. 2009), we adopted PCA to preprocess the features before implementing the

[1]For convenience, we simply regard $\mathbf{y}_{ii} = \mathbf{y}_i$ as one of $\mathbf{y}_i$'s connective points. This will not influence the final optimization result according to (12).

Table 1: Summary of various dimension reduction algorithms. ($\mathbf{c}_i$ is the reconstruction coefficient vector in LLE.)

Algorithms	Patch	Representation of part optimization	Objective function
PCA	Given example and the rest	$\frac{1}{n^2}\begin{pmatrix}(n-1)^2 & -(n-1)\mathbf{e}_{n-1}^T \\ -(n-1)\mathbf{e}_{n-1} & \mathbf{e}_{n-1}\mathbf{e}_{n-1}^T\end{pmatrix}$	Orthogonal linear
LPP	Given example and its neighbors	$\begin{pmatrix}\sum_{j=1}^{k}(\mathbf{W}_{\cdot i})_j & -\mathbf{W}_{\cdot i}^T \\ -\mathbf{W}_{\cdot i} & diag(\mathbf{W}_{\cdot i})\end{pmatrix}$	Linear
LLE	Given example and its neighbors	$\begin{pmatrix}1 & -\mathbf{c}_i^T \\ -\mathbf{c}_i & \mathbf{c}_i\mathbf{c}_i^T\end{pmatrix}$	Non-linear
DLA	Given example with its k_1 nearest neighbors in the same class, and k_2 nearest neighbors in the different class	$\begin{pmatrix}\sum_{j=1}^{k_1+k_2}(\mathbf{W}_{\cdot i})_j & -\mathbf{W}_{\cdot i}^T \\ -\mathbf{W}_{\cdot i} & diag(\mathbf{W}_{\cdot i})\end{pmatrix}$	Linear
SLE	Given example and the rest	$\begin{pmatrix}\tilde{D}_{ii} & -\tilde{\mathbf{W}}_{\cdot i}^T \\ -\tilde{\mathbf{W}}_{\cdot i} & \mathbf{\Lambda}_{\lvert\tilde{\mathbf{W}}_{\cdot i}\rvert}\end{pmatrix}$	Linear

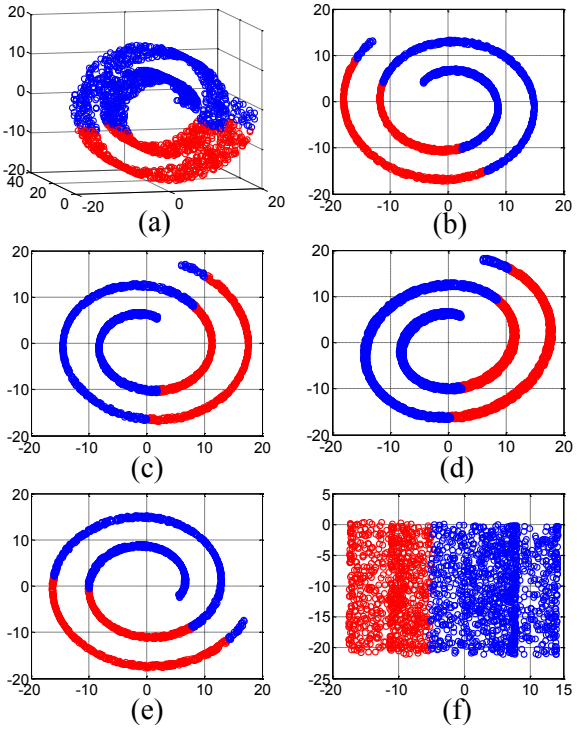

Figure 1: The performance of algorithms on the *SwissRoll* dataset. (a) is the original data distributed in 3D space, and (b)-(f) are the 2D embedding results of PCA, LPP, MFA, DLA and SLE, respectively. The red circles denote positive examples and the blue circles represent negative examples.

compared dimension reduction algorithms, with the exception of PCA itself, to eliminate the useless information and prevent the singularity problem. For multi-class situations, (7) was employed to decide the labels of the test examples, and the Nearest Neighborhood Classifier (NNC) was directly adopted to handle the two-class classifications.

For every algorithm compared, a projection matrix was trained on the training set and was then applied to reduce the dimensions of a set of test examples. Finally, these test data were classified based on their dimension reduction results. The classification accuracy on the test sets, in particular, was observed to evaluate the dimension reduction performance of the algorithms. The experimental results are presented in Fig. 2, in which the y-axes denote the accuracy and the x-axes stand for the reduced dimensions d. We see that the proposed SLE is superior to other baselines in most cases, and the performance of SLE is comparable with PCA and LPP on the *BreastCancer* dataset. Furthermore, Fig. 2 suggests that the output of SLE is not very sensitive to the choice of d in all datasets, so this parameter can be tuned easily for practical use.

Face Data

It has been extensively demonstrated that there is an underlying manifold embedded in face images, though they may look different in appearance, illumination and angle of observation (Roweis and Saul 2000; Tenenbaum, Silva, and Langford 2000). This section uses two popular face databases, *Yale* (Georghiades, Belhumeur, and Kriegman 2001) and *Labeled Face in the Wild* (*LFW*) (Gary et al. 2007), to compare the performances of SLE and other baselines.

Yale contains 165 gray images of 15 individuals. Each individual has 11 face images covering a variety of facial expressions and configurations including: center-light, wearing glasses, happy, left-light, wearing no glasses, normal, right-light, sad, sleepy, surprised, and wink. Some representative examples are provided in Fig. 3 (a) . Every image has a resolution of 64×64 and is therefore represented by a 4096-dimensional pixel-wise long vector. In each implementation, we randomly chose 6 images of every individual as the training set, and the remaining 5 images were used for testing, so the sizes of the training set and test set were $6 \times 15 = 90$ and $5 \times 15 = 75$, respectively. We ran all the algorithms 10 times based on the different partitions of the training and test sets, and the recognition accuracies w.r.t. the increase of d were reported in Fig. 3 (b).

We built a 5NN graph for LPP with kernel width $\sigma = 10$.

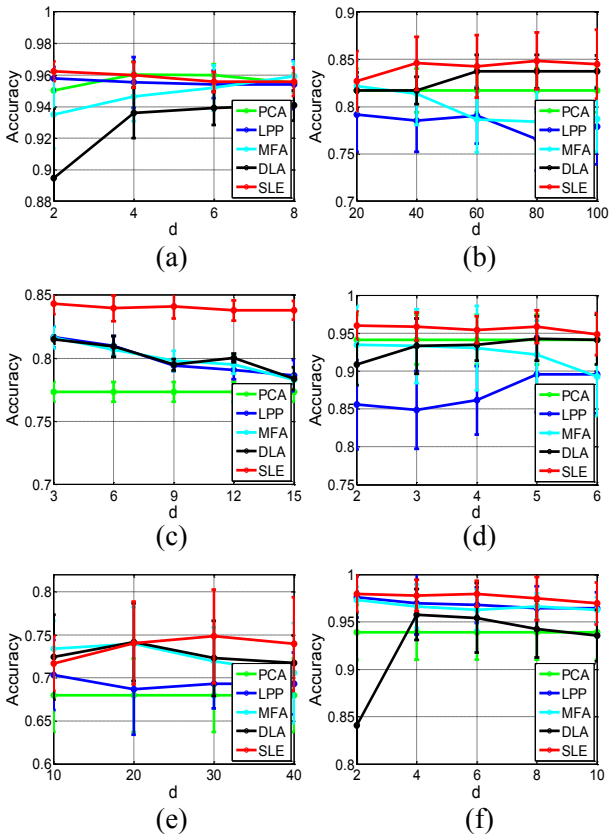

Figure 2: Experimental results on six UCI datasets. (a) denotes *BreastCancer*, (b) represents *Musk*, (c) shows *Waveform*, (d) denotes *Seeds*, (e) illustrates *SPECT*, and (f) is *Wine*.

In MFA and DLA, we set $k_1 = 5$ and $k_2 = 10$ to obtain the best performances. We see that PCA, DLA and MFA achieve similar performances on the *Yale* data with approximately 60% accuracy. Comparatively, the recognition rate of SLE reaches almost 80%, so the effectiveness of the proposed algorithm is demonstrated.

LFW contains a large number of face images gathered from the web. Because these face images are directly collected from natural scenes, the unconstrained facial expressions, unsuitable observation angles, undesirable illumination conditions and complicated background settings create difficulties for accurate face recognition. We used a subset of *LFW* by choosing face images of Toledo, Sharon, Schwarzenegger, Powell, Rumsfeld, Bush, Arroyo, Agassi, Beckham, and Hewitt, giving 392 examples in total belonging to 10 people in the subset. We adopted the 73-dimensional features built by Kumar et al. (2009) to characterize every face image, and reduced their dimensions via PCA, LPP, MFA, DLA, and SLE, respectively. The projection matrices were learnt from the training set with 200 images and were employed to compute the embeddings of the 192 test examples.

To obtain optimal performance, k and σ were adjusted to 10 and 1 for LPP, and we set $k_1 = k_2 = 10$ for MFA and DLA. Fig. 4 (b) shows the results, which reveal that SLE

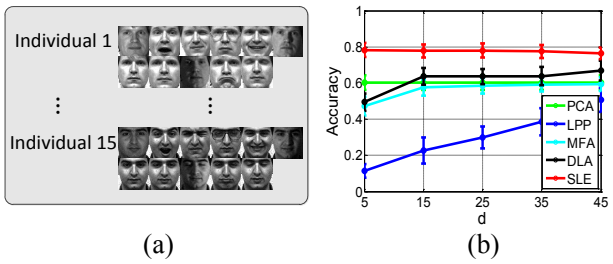

Figure 3: Comparison of algorithms on *Yale* data. (a) shows representative examples. (b) plots the accuracies of algorithms under different choices of reduced dimensions d.

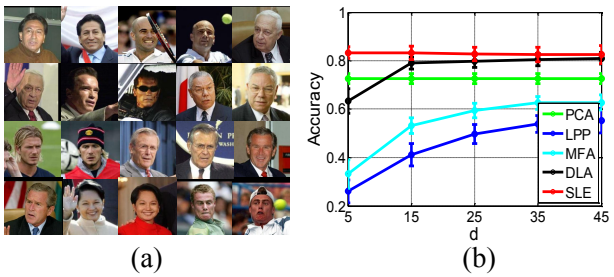

Figure 4: Experimental results of algorithms on *LFW* dataset. (a) shows some typical examples. (b) reports the obtained accuracies w.r.t. different d.

obtains the highest accuracy generally. DLA achieves comparable performance with SLE when d is larger than 25, but only achieves around 60% accuracy when $d = 5$, which is worse than SLE with an approximate 80% recognition rate. Though recognizing faces in *LFW* is a very challenging task, SLE obtains very promising results on this dataset.

Conclusion

This paper has proposed a novel supervised manifold learning algorithm called Signed Laplacian Embedding (SLE). By constructing the binary signed graph and employing the signed graph Laplacian, SLE utilizes the label information and preserves the global data locality of examples simultaneously. Numerous experiments have demonstrated that this improves embedding performance. SLE has been proven from both spectral graph and statistical theories, and is related to other algorithms from the viewpoint of the patch alignment framework. The core purpose of SLE is to solve a generalized eigenvalue problem, which can be efficiently computed by using existing numerical methods such as the QZ algorithm or Krylov subspace algorithm. Furthermore, we note that d is the only parameter to be tuned, so SLE can be easily implemented.

Acknowledgments

This research is supported by NSFC, China (No: 6127325861375048), Ph.D. Programs Foundation of Ministry of Education of China (No.20120073110018), and Australian Research Council Discovery Project (No: DP-140102164).

References

Belkin, M., and Niyogi, P. 2001. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Advances in Neural Information Processing System*, volume 14, 585–591.

Boyd, S., and Vandenberghe, L. 2004. *Convex optimization*. Cambridge university press.

Chen, W.; Chen, Y.; and Weinberger, K. 2013. Maximum variance correction with application to A* search. In *Proceedings of the 30th International Conference on Machine Learning*, 302–310.

Fisher, R. 1936. The use of multiple measurements in taxonomic problems. *Annals of eugenics* 7(2):179–188.

Frank, A., and Asuncion, A. 2010. UCI machine learning repository.

Gary, B.; Manu, R.; Tamara, B.; and Erik, L. 2007. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst.

Georghiades, A.; Belhumeur, P.; and Kriegman, D. 2001. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Trans. Pattern Anal. Mach. Intelligence* 23(6):643–660.

Golub, G., and Loan, C. 2012. *Matrix computations*, volume 3. JHU Press.

He, X., and Niyogi, P. 2004. Locality preserving projections. In *Advances in Neural Information Processing System*, volume 16, 153–160.

Hotelling, H. 1933. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology* 24(6):417–441.

Joncas, D., and Meila, M. 2013. Non-linear dimensionality reduction: Riemannian metric estimation and the problem of geometric discovery. *arXiv:1305.7255*.

Kumar, N.; Berg, A.; Belhumeur, P.; and Nayar, S. 2009. Attribute and simile classifiers for face verification. In *IEEE 12th International Conference on Computer Vision and Pattern Recognition*, 365–372. IEEE.

Kunegis, J.; Schmidt, S.; Lommatzsch, A.; Lerner, J.; Luca, E. D.; and Albayrak, S. 2010. Spectral analysis of signed graphs for clustering, prediction and visualization. In *SDM*, volume 10, 559–570.

Roweis, S., and Saul, L. 2000. Nonlinear dimensionality reduction by locally linear embedding. *Science* 290(5500):2323–2326.

Shaw, B., and Jebara, T. 2009. Structure preserving embedding. In *Proceedings of the 26th Annual International Conference on Machine Learning*, 937–944. ACM.

Tenenbaum, J.; Silva, V.; and Langford, J. 2000. A global geometric framework for nonlinear dimensionality reduction. *Science* 290(5500):2319–2323.

Wang, C., and Mahadevan, S. 2013. Multiscale manifold learning. In *The 27th AAAI Conference on Artificial Intelligence*.

Yan, S.; Xu, D.; Zhang, B.; Zhang, H.; Yang, Q.; and Lin, S. 2007. Graph embedding and extensions: a general framework for dimensionality reduction. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 29(1):40–51.

Zhang, Z., and Zha, H. 2005. Principal manifolds and nonlinear dimension reduction via local tangent space alignment. *SIAM J. Scientific Computing* 26(1):313–338.

Zhang, T.; Tao, D.; Li, X.; and Yang, J. 2009. Patch alignment for dimensionality reduction. *Knowledge and Data Engineering, IEEE Transactions on* 21(9):1299–1313.

Zhou, T.; Tao, D.; and Wu, X. 2011. Manifold elastic net: a unified framework for sparse dimension reduction. *Data Mining and Knowledge Discovery* 22(3):340–371.