Spectral Feature Scaling Method for Supervised Dimensionality Reduction

Momo Matsuda[1], **Keiichi Morikuni**[1], **Tetsuya Sakurai**[1,2]

[1] University of Tsukuba

[2] JST/CREST

matsuda@mma.cs.tsukuba.ac.jp, morikuni@cs.tsukuba.ac.jp, sakurai@cs.tsukuba.ac.jp

Abstract

Spectral dimensionality reduction methods enable linear separations of complex data with high-dimensional features in a reduced space. However, these methods do not always give the desired results due to irregularities or uncertainties of the data. Thus, we consider aggressively modifying the scales of the features to obtain the desired classification. Using prior knowledge on the labels of partial samples to specify the Fiedler vector, we formulate an eigenvalue problem of a linear matrix pencil whose eigenvector has the feature scaling factors. The resulting factors can modify the features of entire samples to form clusters in the reduced space, according to the known labels. In this study, we propose new dimensionality reduction methods supervised using the feature scaling associated with the spectral clustering. Numerical experiments show that the proposed methods outperform well-established supervised methods for toy problems with more samples than features, and are more robust regarding clustering than existing methods. Also, the proposed methods outperform existing methods regarding classification for real-world problems with more features than samples of gene expression profiles of cancer diseases. Furthermore, the feature scaling tends to improve the clustering and classification accuracies of existing unsupervised methods, as the proportion of training data increases.

1 Introduction

Consider clustering a set of data samples with high-dimensional features into mutually exclusive subsets, called clusters, and classifying these clusters when given prior knowledge on the class labels of partial samples. These kinds of problems arise in pathological diagnoses using gene expression data [Tarca *et al.*, 2006], the analysis of chemical sensor data [Jurs *et al.*, 2000], the community detection in social networks [Tichy *et al.*, 1979], the analyses of neural spike sorting [Bar-Hillel *et al.*, 2006], and so on [Sogh, 2006]. Spectral clustering is an unsupervised technique that projects the data samples to the eigenspace of the Laplacian matrix, in which the data samples are linearly separated if they are well clustered. This technique is more effective in complicated clustering than existing methods. For example, the k-means algorithm induces partitions using hyperplanes, but may fail to give satisfactory results due to irregularities and uncertainties of the features.

Here, we consider aggressively modifying the scales of the features to improve the separation in the reduced space, and obtain the desired classification. We derive the factors for scaling the features (scaling factors) using a learning machine based on spectral dimensionality reduction; namely, the inputs are unscaled and labeled samples, and the outputs are the scaling factors. To this end, by exploiting the prior knowledge on the labels of partial samples, we specify the Fiedler vector and reformulate the Laplacian eigenproblem as an eigenproblem of a linear matrix pencil whose eigenvector has the scaling factors. The obtained factors can modify the features of the entire samples to form clusters in the reduced dimensionality space according to the known labels. Thus, we use the prior knowledge to implicitly specify the phenomenon of interest, and supervise dimensionality reduction methods for the scaling factors. These approaches yield the desired clusters, incorporated with unsupervised spectral clustering, for test data scaled by the obtained factors for training data. Numerical experiments on artificial data and real-world data from gene expression profiles show that the feature scaling improves the accuracy of the spectral clustering. In addition, the spectral dimensionality reduction methods supervised using the feature scaling outperform existing methods in some cases, and are more robust than existing methods.

We review related work on supervised spectral dimensionality reduction methods. Learning a similarity matrix from training data is effective in spectral clustering [Bach and Jordan, 2003; Kamvar *et al.*, 2003]. The proposed classification method can be considered a supervised method, incorporated with the kernel version of the locality preserving projections (LPP) [He and Niyogi, 2003]. LPP is comparable with the linear discriminant analyisis (LDA) [Belhumeur *et al.*, 1997], the local Fisher discriminant analysis (LFDA) [Sugiyama, 2007], and the locality adaptive discriminant analysis (LADA) [Li *et al.*, 2017]. Kernel versions of LDA, LPP and LFDA aim at nonlinear dimensionality reduction. These methods have semi-supervised variants [Song *et al.*, 2008; Kulis *et al.*, 2005; Cui and Fan, 2012].

1.1 Spectral Clustering

The underlying idea of the proposed methods follows from spectral clustering, which is induced by graph partitioning, where the graph is weighted and undirected. The weight $w_{i,j}$ of the edge between nodes i and j represents the similarity between samples i and j for $i, j = 1, 2, \ldots, n$, where n is the number of samples. By dividing the graph into mutually exclusive subgraphs, the corresponding samples form clusters. This discrete problem can be continuously relaxed to a matrix eigenvalue problem. Let

$$W = \{w_{i,j}\} \in \mathbb{R}^{n \times n},$$

$$D = \mathrm{diag}\,(d_1, d_2, \ldots, d_n) \in \mathbb{R}^{n \times n} \text{ with } d_i = \sum_{i=1}^{n} w_{i,j},$$

$e \in \mathbb{R}^n$ be a vector with all elements 1, and t be an indicator vector with the ith entry $t_i = 1$ if sample i is in a cluster, and $t_i = -1$ if sample i is in another cluster. Then, the constrained minimization of the Ncut function [Shi and Malik, 2000]

$$\min_{v} \frac{v^{\mathsf{T}}(D - W)v}{v^{\mathsf{T}} D v}, \quad \text{subject to} \quad e^{\mathsf{T}} D v = 0 \qquad (1)$$

over $v_i \in \{1, -b\}$ with $b = \sum_{t_i > 0} d_i / \sum_{t_i < 0} d_i$, is relaxed to finding the Fiedler vector $v \in \mathbb{R}^n \setminus \{0\}$ [Fiedler, 1973] associated with the smallest eigenvalue of the constrained generalized eigenvalue problem

$$Lv = \lambda D v, \quad \text{subject to} \quad e^{\mathsf{T}} D v = 0, \qquad (2)$$

where $L = D - W$ is a Laplacian matrix and $\lambda \in \mathbb{R}$. The latter problem (2) is tractable, while the former (1) is NP-hard. Moreover, a few eigenvectors of (2) form clusters that are separable using hyperplanes if the samples form well-separated clusters. A sophisticated algorithm for (2) is nearly linear time [Spielman and Teng, 2014].

2 Proposed Method

Irregular scales and uncertainty of features prevent well-established dimensionality reduction methods from clustering data samples into the desired disjoint subsets. In spectral clustering, the Ncut function is popular by virtue of the non-linear separability, but is not versatile. To cope with these issues, we propose a remedy to aggressively modify the scales of the features in the eigenspace where a linear separation works, based on prior knowledge of the partial n samples $X = [x_1, x_2, \ldots, x_n]^{\mathsf{T}}, x_i \in \mathbb{R}^m$ with m features. If the classes of partial samples are known, we can estimate the entries of the Fiedler vector $v \in \mathbb{R}^n \setminus \{0\}$ of the Laplacian eigenvalue problem

$$L_s v = \lambda_s D_s v, \quad e^{\mathsf{T}} D_s v = 0, \quad \lambda_s \in \mathbb{R}, \qquad (3)$$

where $L_s = D_s - W_s$, and $W_s = \{w_{i,j}^{(s)}\} \in \mathbb{R}^{n \times n}$ depends on the feature scaling factors $s \in \mathbb{R}^m$. We obtain the scaling factors $s \in \mathbb{R}^m$ in (3) by solving an eigenproblem of a linear matrix pencil. Then, applying the obtained scaling factors to the features of the entire data $Y = [X^{\mathsf{T}}, R^{\mathsf{T}}]^{\mathsf{T}} \in \mathbb{R}^{N \times m}$

with N samples, we can extend the prior knowledge on the partial samples to overall samples, where $R \in \mathbb{R}^{(N-n) \times m}$ denotes the remaining data. In statistical terms, the scaling changes the mean $\bar{y}_j = \frac{1}{n} \sum_{i=1}^{n} y_{i,j}$ and the variance $\sigma_j^2 = \frac{1}{n} \sum_{i=1}^{n} (y_{i,j} - \bar{y}_j)^2$ of the jth feature to $s_j^{1/2} \bar{y}_j$ and $|s_j| \sigma_j^2$, respectively, for $j = 1, 2, \ldots, m$. Note that we allow the scaling factors to have negative values. See [Schleif and Tino, 2015]. Finding appropriate scaling factors is also considered metric learning [Yang, 2006]. Recall that the center of a cluster associated with a Bregman divergence is equal to the centroid of the samples of the cluster [Banerjee et al., 2005].

Now, we reformulate (3) as another eigenproblem to extract the scaling factors $s_i \in \mathbb{R}, i = 1, 2, \ldots, m$ as an eigenvector. Denote the scaling matrix by $S^{\frac{1}{2}} = \mathrm{diag}\,(s_1, s_2, \ldots, s_m)^{\frac{1}{2}} \in \mathbb{R}^{m \times m}$ and denote the (i, j) entry of the similarity matrix W_s for the scaled data $X S^{1/2}$ by

$$w_{i,j}^{(s)} = \begin{cases} 1 - \frac{(x_i - x_j)^{\mathsf{T}} S (x_i - x_j)}{2\sigma^2} = 1 - \frac{s^{\mathsf{T}} x_{i,j}}{2\sigma^2}, \\ \quad \simeq \exp\left(-\frac{(x_i - x_j)^{\mathsf{T}} S (x_i - x_j)}{2\sigma^2}\right), \quad i \neq j, \\ 0, \quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad i = j, \end{cases} \quad (4)$$

where $s = [s_1, s_2, \ldots, s_m]^{\mathsf{T}} \in \mathbb{R}^m$ and the kth entry of $x_{i,j} \in \mathbb{R}^m$ is $(x_{i,k} - x_{j,k})^2$. Here, we used the first-order approximation of the exponential function $\exp(-x) \approx 1 - x$ for $0 < x < 1$. Then, the ith row of W_s is

$$w_i^{(s)\mathsf{T}} = [1, \ldots, 1, 0, 1, \ldots, 1] - \frac{s^{\mathsf{T}}}{2\sigma^2} [x_{i,1}, x_{i,2}, \ldots, x_{i,n}]$$
$$= \tilde{e}_i^{\mathsf{T}} - s^{\mathsf{T}} X_i,$$

where $\tilde{e}_i$ is the n-dimensional vector with the ith entry equal to zero and the remaining entries ones, and $X_i = \frac{1}{2\sigma^2} [x_{i,1}, x_{i,2}, \ldots, x_{i,n}] \in \mathbb{R}^{m \times n}$. Hence, we have

$$W_s v = [\tilde{e}_1^{\mathsf{T}}, \tilde{e}_2^{\mathsf{T}}, \ldots, \tilde{e}_n^{\mathsf{T}}]^{\mathsf{T}} v - [X_1 v, X_2 v, \ldots, X_n v]^{\mathsf{T}} s.$$

Let $\hat{x}_i = \frac{1}{2\sigma^2}(x_{i,1} + x_{i,2} + \cdots + x_{i,n})$. Then, the ith diagonal entry $d_i^{(s)}$ of D_s is

$$d_i^{(s)} = \sum_{j=1}^{n} w_{i,j}^{(s)} = (n-1) - s^{\mathsf{T}} \hat{x}_i.$$

Hence, denoting the Fiedler vector by $v = [v_1, v_2, \ldots, v_n]^{\mathsf{T}}$, we have

$$D_s v = (n-1)v - [v_1 \hat{x}_1, v_2 \hat{x}_2, \ldots, v_n \hat{x}_n]^{\mathsf{T}} s.$$

Thus, (3) is written as

$$L_s v = \lambda_s D_s v \quad \Leftrightarrow \quad W_s v = (1 - \lambda_s) D_s v$$

$$\Leftrightarrow \quad [A \;\; \alpha] \begin{bmatrix} s \\ -1 \end{bmatrix} = \mu [B \;\; \beta] \begin{bmatrix} s \\ -1 \end{bmatrix},$$

where $\mu = 1 - \lambda_s$,

$$A = \begin{bmatrix} (X_1 v)^{\mathsf{T}} \\ (X_2 v)^{\mathsf{T}} \\ \vdots \\ (X_n v)^{\mathsf{T}} \end{bmatrix}, \quad B = \begin{bmatrix} v_1 \hat{x}_1^{\mathsf{T}} \\ v_2 \hat{x}_2^{\mathsf{T}} \\ \vdots \\ v_n \hat{x}_n^{\mathsf{T}} \end{bmatrix} \in \mathbb{R}^{n \times m}, \qquad (5)$$

$$\boldsymbol{\alpha} = [\tilde{\boldsymbol{e}}_1, \tilde{\boldsymbol{e}}_2, \ldots, \tilde{\boldsymbol{e}}_n]^{\mathsf{T}} \boldsymbol{v} \in \mathbb{R}^n, \quad \boldsymbol{\beta} = (n-1)\boldsymbol{v} \in \mathbb{R}^n. \quad (6)$$

Furthermore, the feature scaling factors $\boldsymbol{s}$ have the constraint

$$\boldsymbol{e}^{\mathsf{T}} D_s \boldsymbol{v} = \sum_{i=1}^{n} \left((n-1)v_i - v_i \boldsymbol{s}^{\mathsf{T}} \hat{\boldsymbol{x}}_i \right) = 0$$

$$\Leftrightarrow \left(\sum_{i=1}^{n} v_i \hat{\boldsymbol{x}}_i^{\mathsf{T}} \right) \boldsymbol{s} - (n-1) \sum_{i=1}^{n} v_i = 0.$$

Therefore, we can obtain the scaling factors $\boldsymbol{s}$ by solving the generalized eigenproblem of a linear matrix pencil

$$\begin{bmatrix} A & \boldsymbol{\alpha} \\ \boldsymbol{\gamma}^{\mathsf{T}} & \rho \end{bmatrix} \begin{bmatrix} \boldsymbol{s} \\ -1 \end{bmatrix} = \mu \begin{bmatrix} B & \boldsymbol{\beta} \\ \boldsymbol{0}^{\mathsf{T}} & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{s} \\ -1 \end{bmatrix}, \ \boldsymbol{s} \in \mathbb{R}^m, \mu \in \mathbb{R}, \quad (7)$$

where

$$\boldsymbol{\gamma} = \sum_{i=1}^{n} v_i \hat{\boldsymbol{x}}_i \in \mathbb{R}^m, \quad \rho = (n-1) \sum_{i=1}^{n} v_i \in \mathbb{R}. \quad (8)$$

It follows from the equality $(A - B)^{\mathsf{T}} \boldsymbol{e} = \boldsymbol{0}$ that there exists a real number μ that satisfies (7). To solve (7), we can use the contour integral method [Morikuni, 2016] for $n < m$; otherwise, we can also use the minimal perturbation approach [Ito and Murota, 2016]. When $\boldsymbol{s}$ is obtained as a complex vector, its real and imaginary parts are also eigenvectors of (7). We took the real part in the experiments given in section 3. For $\mu = 1 - \lambda$, we adopt the eigenvector of (7) associated with the eigenvalue μ closest to one as the scaling factors. Then, we apply the squared feature scaling factors $\boldsymbol{s}^{1/2}$ to the entire data Y, such that $Z = Y S^{1/2}$. Finally, we perform the spectral clustering or classification on the scaled data $Z \in \mathbb{R}^{N \times m}$ by solving the Laplacian eigenvalue problem

$$L' \boldsymbol{u} = \lambda' D' \boldsymbol{u}, \quad \boldsymbol{e}^{\mathsf{T}} D' \boldsymbol{u} = 0, \quad \boldsymbol{u} \in \mathbb{R}^N \setminus \{\boldsymbol{0}\}, \quad \lambda' \in \mathbb{R}$$

for a few eigenvectors corresponding to the smallest eigenvalues, where $L' = D' - W' \in \mathbb{R}^{N \times N}, W' = \{w'_{i,j}\} \in \mathbb{R}^{N \times N}$,

$$w'_{i,j} = \exp \left(-\frac{\|\boldsymbol{z}_i - \boldsymbol{z}_j\|_2^2}{2\sigma^2} \right), \quad i, j = 1, 2 \ldots, N, \quad (9)$$

$\boldsymbol{z}_i \in \mathbb{R}^m$ is the ith row of Z, $d'_i = \sum_{j=1}^{n} w'_{i,j}$, and $D' = \text{diag}(d'_1, d'_2, \ldots, d'_N)$. We summarize the procedures of the proposed methods in Algorithm 1.

3 Numerical Experiments

Numerical experiments on test problems were performed to compare the proposed methods with existing methods in terms of accuracy. For clustering problems, we compared the proposed method (spectral clustering method supervised using the feature scaling, SC-S) with the unsupervised spectral clustering method (SC) [Shi and Malik, 2000], LPP, its kernel version (KLPP) [He and Niyogi, 2003], and two supervised methods, LFDA and its kernel version (KLFDA) [Sugiyama, 2007]. For classification problems, we compared the proposed method (spectral classification method supervised using the feature scaling, SC-S) with LPP, KLPP, LFDA, and KLFDA. All programs were coded and run in Matlab 2016b.

Algorithm 1 Spectral clustering/classification methods supervised using the feature scaling

Input: Training data $X \in \mathbb{R}^{n \times m}$, entire data $Y \in \mathbb{R}^{N \times m}$, estimated Fiedler vector $\boldsymbol{v}$, dimension of the reduced space ℓ.
1: Compute $A, B, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \rho$ in (5), (6), (8).
2: Solve (7) for an eigenvector $\boldsymbol{s} = [s_1, s_2, \ldots, s_m]^{\mathsf{T}}$.
3: Set the scaling matrix $S^{1/2} = \text{diag}(s_1, s_2, \ldots, s_m)^{1/2}$.
4: Compute the scaled data matrix $Z = Y S^{1/2}$.
5: Compute the similarity matrix $W' \in \mathbb{R}^{N \times N}$ (9) and D', using the k-nearest neighbors, and set $L' = D' - W'$.
6: Solve $L' \boldsymbol{u} = \lambda' D' \boldsymbol{u}$ for the eigenvectors $\boldsymbol{u}_1, \boldsymbol{u}_2 \ldots \boldsymbol{u}_\ell$ corresponding to the ℓ nonzero smallest eigenvalues.
7: Cluster/classify the reduced data samples $[\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_\ell] \in \mathbb{R}^{N \times \ell}$ rowwise by using the k-means algorithm/the one-nearest neighborhood method, respectively.

We employed the LPP code in the Matlab Toolbox for Dimensionality Reduction[1] and the LFDA and KLFDA codes in the Sugiyama-Sato-Honda Lab site[2]. In the compared methods except for LFDA, we chose the optimal value of the parameter σ among $1, 10^{\pm 1}$, and $10^{\pm 2}$ to achieve the best accuracy. We formed the similarity matrix in (9) by using the seven-nearest neighbor and taking the symmetric part. We set the dimensions of the reduced space to one for clustering problems, and one, two and three for classification problems. It is known that using one eigenvector is sufficient for binary clustering if the clusters are well-separated [Chung, 1996]. The reduced data samples to a low-dimension were clustered using the k-means algorithm, and classified using the one-nearest neighborhood method.

The performance measures we used were the normalized mutual information (NMI) [Strehl and Ghosh, 2002] and the Rand index (RI) [Rand, 1971] for evaluating the clustering methods, and RI for evaluating the classification methods. Let n_i be the number of samples in cluster/class i, $\hat{n}_i$ be the number of samples clustered/classified by a method to cluster/class i, and $n_{i,j}$ be the number of samples in cluster/class i clustered/classified by a method to cluster/class j. Then, the NMI measure for binary clustering is defined by

$$\text{NMI} = \frac{\sum_{i=1}^{2} \sum_{j=1}^{2} n_{i,j} \log \left(n \cdot n_{i,j} / \left(n_i \cdot n'_j \right) \right)}{\sqrt{\left(\sum_{i=1}^{2} n_i \log \left(n_i/n \right) \right) \left(\sum_{j=1}^{2} n'_j \log \left(n'_j/n \right) \right)}}.$$

Next, let TP (true positive) denote the number of samples which are correctly identified as class 1, FP (false positive) denote the number of samples which are incorrectly identified as class 1, TN (true negative) denote the number of samples which are correctly identified as class 2, FN (false negative) denote the number of samples which are incorrectly identified as class 2. Then, the RI measure is defined by

$$\text{RI} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}.$$

---

[1]https://lvdmaaten.github.io/drtoolbox/
[2]http://www.ms.k.u-tokyo.ac.jp/software.html

As the obtained clustering and classification become more accurate, the values of NMI and RI become large.

3.1 Artificial Data

In this subsection, we use toy problems of data sets consisting of 800 samples with 10 features. Figure 1 shows three of the 10 features, where the symbols $\circ$ and $+$ denote the data samples in different classes, and the remaining seven features were given uniformly distributed random numbers in the interval $[0, 1]$. The original data were normalized to have a mean of zero and variance one. We set the entries of the Fiedler vector v to $v_i = 1$ if sample i is in a cluster, and $v_i = -0.2$ if sample i is in another cluster (see (1)). We repeated clustering the same reduced data samples 20 times using the k-means algorithm and the one-nearest neighborhood method, starting with different random initial guesses. Moreover, we tested on different proportions of training data from 5% to 50%, and repeated each test 10 times with different random choices of training data. Thus, we report the average performance by taking the arithmetic mean of RI and NMI.

Figures 2 and 3 show the means and standard deviations of RI and NMI, respectively, for the clustering problems. SC-S performed the best in terms of accuracy among the compared methods. The value of RI for SC-S, LFDA, and KLFDA tended to increase with the proportion of training data used. On the other hand, the accuracies of clustering using SC, LPP, and KLPP were similar for different proportions of training data.

Figure 4 shows the means and standard deviations of RI for the classification problems, where the reduced dimensions are $\ell = 1, 2$, and, 3. SC-S performed the best in terms of accuracy for proportions of training data of 5–50% with few exceptions among the compared methods. As the proportion of training data increased, the accuracy of the compared methods tended to increase.

Figure 5 shows the values of the scaling factors for each feature when using 50% of the entire data for training. Since the fourth to tenth features given by random numbers were not involved in the classification, the values of the scaling factors of these seven features were small. Thus, the values of the scaling factors indicated which features were effective for the desired classification.

Figure 6 shows the data samples reduced to a one-dimensional space for each method when using 50% of all data for training. SC-S reduced all data to one dimension, and made it linearly separable to the true clusters. KLFDA failed to reduce the data samples to be linearly separable to the true clusters.

Table 1 gives the means and standard deviations of RI through the leave-one-out cross-validation for the classification problems for reduced dimensions $\ell = 1, 2$, and 3. Table 1 shows that SC-S was immune from an overfitting in the learning process.
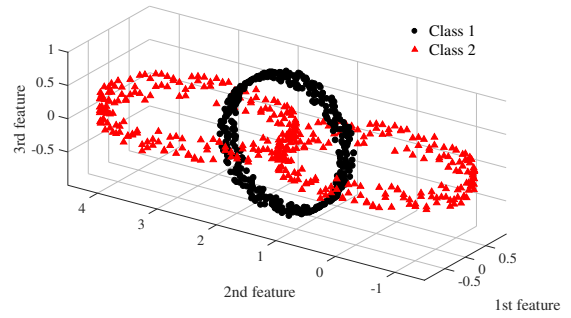

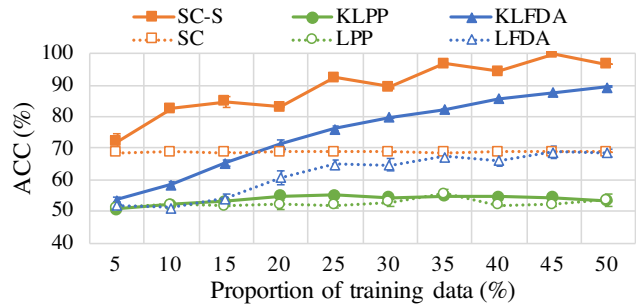Figure 1: Visualization in three dimensions of the artificial data

Figure 2: Proportion of training data vs. RI for clustering artificial data
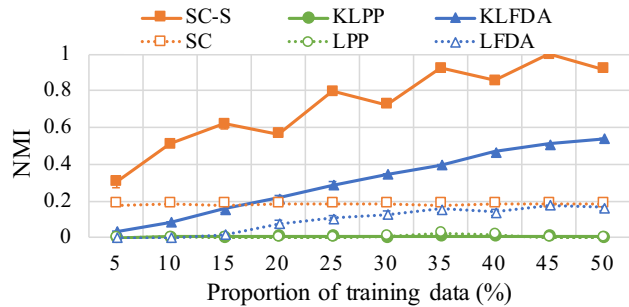

Figure 3: Proportion of training data vs. NMI for clustering artificial data

ℓ	1	2	3
SC-S	*100.0± 0.0	*100.0± 0.0	*100.0± 0.0
LPP	77.0±40.1	89.6±30.5	99.9± 3.5
KLPP	52.4±49.9	52.4±50.0	52.4±49.9
LFDA	78.4±41.2	83.9±36.8	81.4±38.9
KLFDA	86.8±33.9	86.9±33.8	87.0±33.6

Table 1: RI (mean% ± std) for classification of artificial data sets

3.2 GEO Data Sets

In this subsection, we use four data sets in the Gene Expression Omunibus[3] from real-world problems with more features

---

[3]https://www.ncbi.nlm.nih.gov/

than samples. Table 2 gives the specifications of the data sets, where IPF denotes Idiopathic Pulmonary Fibrosis. Each data set has two binary classes: diseased patients and healthy peoples. We chose 50% of the entire data for training and repeated each test 10 times with different choices of the training data. The entries of the Fiedler vector v were set to $v_i = 1$ if

sample i is in a cluster, and $v_i = -1$ if sample i is in another cluster (see (1)).

Tables 3 and 4 give the means and standard deviations of RI and NMI, respectively, for each data set for the clustering problems. The symbol $*$ denotes the largest values of RI and NMI for each data set. SC-S performed the best in terms of accuracy for two data sets, and second best for two data sets among the compared methods. LFDA performed the best in terms of accuracy for one data set, second best for one data set but worst for one data set among the compared methods. These results showed that SC-S was more robust than other methods.

Table 5 gives the means and standard deviations of RI for the classification problems for the reduced dimensions $\ell = 1, 2,$ and 3. SC-S was best in average in terms of accuracy for

(a) One dimension

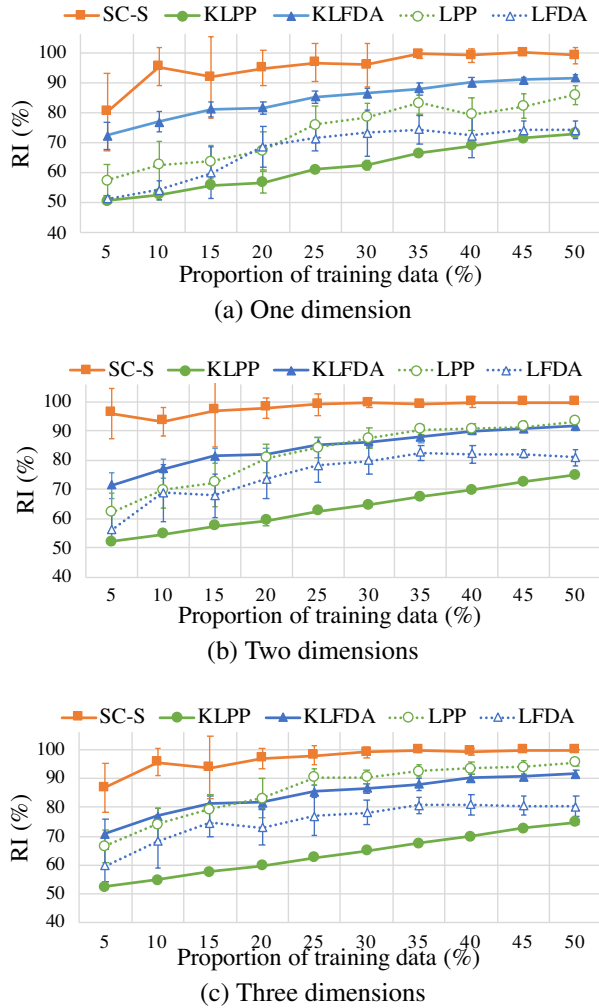

(b) Two dimensions

(c) Three dimensions

Figure 4: Proportion of training data vs. RI for classification of artificial data for reduced dimensions $\ell = 1, 2,$ and 3

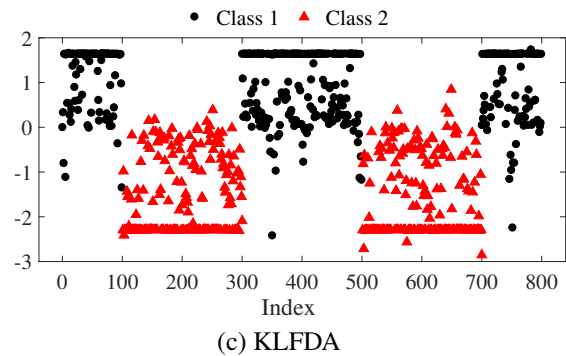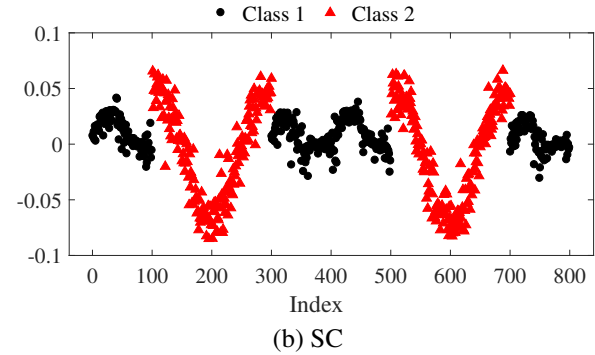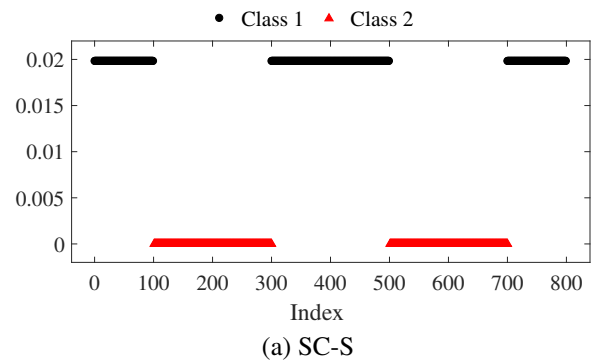Figure 5: Values of the scaling factors

(a) SC-S

(b) SC

(c) KLFDA

Figure 6: Values of the reduced one-dimensional data samples

Data set	# of training samples	# of entire samples	# of features
Ovarian	11	22	54675
Pancreatic	12	24	54675
IPF	12	23	54613
Colorectral	11	22	54675

Table 2: Specifications of data sets

Data set	Ovarian	Pancreatic
SC	52.6 ± 0.05	54.2 ± 0.00
SC-S	$*60.2 \pm 1.11$	56.0 ± 1.22
LPP	57.0 ± 1.47	$*59.5 \pm 1.20$
KLPP	54.6 ± 0.00	54.2 ± 0.00
LFDA	57.9 ± 1.77	55.1 ± 1.57
KLFDA	56.4 ± 0.05	51.0 ± 0.08

Data set	IPF	Colorectal
SC	69.6 ± 0.00	50.0 ± 0.00
SC-S	76.1 ± 0.00	$*74.1 \pm 1.10$
LPP	76.7 ± 2.12	57.8 ± 1.33
KLPP	71.3 ± 0.00	53.2 ± 0.00
LFDA	70.0 ± 4.34	54.3 ± 2.01
KLFDA	$*79.9 \pm 0.12$	63.4 ± 0.25

Table 3: RI (mean% $\pm$ std) for clustering GEO data sets

three data sets, except for Ovarian with $\ell = 1$. LFDA was best in terms of accuracy for Ovarian with $\ell = 1$, and for IPF with $\ell = 3$. KLFDA was best in terms of accuracy for IPF. SC-S became more accurate, as the number of eigenvectors ℓ used increased for Ovarian.

It is known that the cross-validation is unreliable in small sample classification [Isaksson *et al.*, 2008]. Indeed, we observed significantly large values of the standard deviation for the compared methods for the data sets.

4 Conclusions

We considered the dimensionality reduction of high-dimensional data. To deal with irregularity or uncertainty in features, we proposed supervised dimensionality reduction methods that exploit knowledge on the labels of partial samples. With the supervision, we modify the features' variances and means. Moreover, the feature scaling can reduce those features that prevent us from obtaining the desired clusters. To obtain the factors used to scale the features, we formulated an eigenproblem of a linear matrix pencil whose eigenvector has the feature scaling factors, and described the procedures for the proposed methods. Numerical experiments showed that the feature scaling is effective when combined with spectral clustering and classification methods. For toy problems with more samples than features, the feature scaling improved the accuracy of the unsupervised spectral clustering, and the proposed methods outperformed several existing methods. For real-world problems with more features than samples from gene expression profiles, the proposed method was more robust in terms of clustering than well-established methods, and outperformed existing methods in some cases.

Data set	Ovarian	Pancreatic
SC	0.004 ± 0.000	0.043 ± 0.000
SC-S	$*0.082 \pm 0.019$	0.042 ± 0.010
LPP	0.044 ± 0.024	$*0.053 \pm 0.022$
KLPP	0.006 ± 0.000	0.039 ± 0.000
LFDA	0.057 ± 0.021	0.040 ± 0.021
KLFDA	0.067 ± 0.001	0.005 ± 0.000

Data set	IPF	Colorectal
SC	0.024 ± 0.000	0.041 ± 0.000
SC-S	0.093 ± 0.000	$*0.323 \pm 0.016$
LPP	0.040 ± 0.014	0.048 ± 0.024
KLPP	0.040 ± 0.000	0.052 ± 0.000
LFDA	0.077 ± 0.041	0.030 ± 0.013
KLFDA	$*0.148 \pm 0.005$	0.109 ± 0.003

Table 4: NMI (mean $\pm$ std) for clustering GEO data sets

(a) Ovarian

ℓ	1	2	3
SC-S	75.0 ± 6.82	$*90.5 \pm 6.88$	$*90.0 \pm 6.03$
LPP	76.4 ± 4.90	75.5 ± 7.93	75.9 ± 8.87
KLPP	55.5 ± 4.45	56.4 ± 6.17	57.7 ± 5.40
LFDA	$*77.3 \pm 7.87$	70.9 ± 4.64	69.6 ± 5.40
KLFDA	55.9 ± 6.12	55.0 ± 3.78	56.4 ± 4.17

(b) Pancreatic

ℓ	1	2	3
SC-S	$*77.5 \pm 8.58$	$*77.9 \pm 7.69$	$*73.3 \pm 7.50$
LPP	74.6 ± 8.00	73.8 ± 5.61	72.1 ± 7.23
KLPP	55.4 ± 4.19	56.7 ± 3.33	54.6 ± 3.46
LFDA	50.0 ± 0.00	50.0 ± 0.00	50.0 ± 0.00
KLFDA	54.2 ± 2.64	54.2 ± 2.64	54.2 ± 2.64

(c) IPF

ℓ	1	2	3
SC-S	$*90.4 \pm 7.48$	$*96.5 \pm 3.79$	$*97.4 \pm 4.43$
LPP	74.4 ± 9.01	76.5 ± 8.06	76.5 ± 8.30
KLPP	68.7 ± 3.25	65.2 ± 4.76	67.8 ± 9.76
LFDA	75.2 ± 8.26	84.8 ± 6.52	87.0 ± 0.00
KLFDA	87.0 ± 0.00	87.0 ± 0.00	87.0 ± 0.00

(d) Colorectal

ℓ	1	2	3
SC-S	$*82.3 \pm 8.73$	$*85.5 \pm 6.98$	$*82.7 \pm 6.98$
LPP	72.3 ± 9.63	75.0 ± 6.18	80.0 ± 4.64
KLPP	57.3 ± 7.39	60.9 ± 6.80	64.6 ± 7.27
LFDA	70.5 ± 3.05	77.3 ± 4.07	79.1 ± 5.82
KLFDA	77.3 ± 0.00	77.3 ± 0.00	77.3 ± 0.00

Table 5: RI (mean% $\pm$ std) for classification of GEO data sets

References

[Bach and Jordan, 2003] Francis R. Bach and Michael I. Jordan. Learning spectral clustering. In *Advances in Neural Information Processing Systems 16*, pages 305–312, Vancouver, Canada, June 2003.

[Banerjee et al., 2005] Arindam Banerjee, Srujana Merugu, Inderjit S. Dhillon, and Joydeep Ghosh. Clustering with bregman divergences. *Journal of Machine Learning Research*, 6:1705–1749, October 2005.

[Bar-Hillel et al., 2006] Aharon Bar-Hillel, Adam Spiro, and Eran Stark. Spike sorting: Bayesian clustering of non-stationary data. *Journal of Neuroscience Methods*, 157(2):303–316, October 2006.

[Belhumeur et al., 1997] Peter N. Belhumeur, João P. Hespanha, and David J. Kriegman. Eigenfaces vs. fisherfaces: recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):711–720, July 1997.

[Chung, 1996] Fan R. K. Chung. *Spectral Graph Theory*. American Mathematical Society, Providence, USA, 1996.

[Cui and Fan, 2012] Yan Cui and Liya Fan. A novel supervised dimensionality reduction algorithm: Graph-based Fisher analysis. *Pattern Recognition*, 45(4):1471–1481, April 2012.

[Fiedler, 1973] Miroslav Fiedler. Algebraic connectivity of graphs. *Czechoslovak Mathematical Journal*, 23(98):298–305, 1973.

[He and Niyogi, 2003] Xiaofei He and Partha Niyogi. Locality preserving projections. In *Advances in Neural Information Processing Systems 16*, pages 153–160, Vancouver, Canada, June 2003.

[Isaksson et al., 2008] Anders Isaksson, Mikael Wallman, Hanna Göransson, and Mats G. Gustafsson. Cross-validation and bootstrapping are unreliable in small sample classification. *Pattern Recognition Letters*, 29(14):1960–1965, July 2008.

[Ito and Murota, 2016] Shinji Ito and Kazuo Murota. An algorithm for the generalized eigenvalue problem for nonsquare matrix pencils by minimal perturbation approach. *SIAM Journal on Matrix Analysis and Applications*, 37(1):409–419, June 2016.

[Jurs et al., 2000] Peter C. Jurs, Gregory A. Bakke, and Heidi E. McClellands. Computational methods for the analysis of chemical sensor array data from volatile analytes. *Chemical Reviews*, 100(7):2649–2678, June 2000.

[Kamvar et al., 2003] Sepandar D. Kamvar, Dan Klein, and Christopher D. Manning. Spectral learning. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence*, pages 561–566, Acapulco, Mexico, August 2003.

[Kulis et al., 2005] Brian Kulis, Sugato Basu, Inderjit Dhillon, and Raymond Mooney. Semi-supervised graph clustering: a kernel approach. In *Proceedings of the 22th International Conference on Machine Learning*, pages 457–464, Bonn, Germany, July 2005.

[Li et al., 2017] Xuelong Li, Mulin Chen, Feiping Nie, and Qi Wang. Locality adaptive discriminant analysis. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 2201–2207, Melbourne, Australia, August 2017.

[Morikuni, 2016] Keiichi Morikuni. Contour integral-type method for eigenproblems of rectangular matrix pencils. In *Annual Meeting 2016 of the Japan Society for Industrial and Applied Mathematics*, pages 352–353, Kitakyushu, Japan, September 2016.

[Rand, 1971] William M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850, December 1971.

[Schleif and Tino, 2015] Frank-Michael Schleif and Peter Tino. Indefinite proximity learning: A review. *Journal of Neural Computation*, 27(10):2039–2096, October 2015.

[Shi and Malik, 2000] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, August 2000.

[Sogh, 2006] Tarek S. Sogh. Wired and wireless intrusion detection system: Classifications, good characteristics and state-of-the-art. *Computer Standards and Interfaces*, 28(6):670–694, September 2006.

[Song et al., 2008] Yangqiu Song, Feiping Nie, Changshui Zhang, and Shiming Xiang. A unified framework for semi-supervised dimensionality reduction. *Pattern Recognition*, 41(9):2789–2799, September 2008.

[Spielman and Teng, 2014] Daniel A. Spielman and Shang-Hua Teng. Nearly linear time algorithms for preconditioning and solving symmetric, diagonally dominant linear systems. *SIAM Journal on Matrix Analysis and Applications*, 35(3):853–885, July 2014.

[Strehl and Ghosh, 2002] Alexander Strehl and Joydeep Ghosh. Cluster ensembles — a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3:583–617, December 2002.

[Sugiyama, 2007] Masashi Sugiyama. Dimensionality reduction of multimodal labeled data by local Fisher discriminant analysis. *Journal of Machine Learning Research*, 8:1027–1061, May 2007.

[Tarca et al., 2006] Adi L. Tarca, Robert Remero, and Sorin Draghici. Analysis of microarray experiments of gene expression profiling. *American Journal of Obstetrics and Gynecology,*, 195(2):373–388, August 2006.

[Tichy et al., 1979] Noel M. Tichy, Michael L. Tushman, and Charles Fombrun. Social network analysis for organizations. *Academy of Management Review*, 4(4):507–519, October 1979.

[Yang, 2006] Liu Yang. Distance metric learning: A comprehensive survey. *Michigan State University*, pages 1–51, May 2006.