LETTER **Communicated by Xiaojun Chang**

Unsupervised 2D Dimensionality Reduction with Adaptive Structure Learning

Xiaowei Zhao
xiaoweizhao4@gmail.com
School of Information Science and Technology, Northwest University, Xian 71027, China

Feiping Nie
feipingnie@gmail.com
Center for Optical Imagery Analysis and Learning, Northwestern Polytechnical University, Xian 710072, China

Sen Wang
sen.wang@griffith.edu.au
School of Information and Communication Technology, Griffith University, Southport, 4222, Australia

Jun Guo*
guojun@nwu.edu.cn
Pengfei Xu
pfxu@nwu.edu.cn
Xiaojiang Chen
xjchen@nwu.edu.cn
School of Information Science and Technology, Northwest University, Xian 71027, China

In recent years, unsupervised two-dimensional (2D) dimensionality reduction methods for unlabeled large-scale data have made progress. However, performance of these degrades when the learning of similarity matrix is at the beginning of the dimensionality reduction process. A similarity matrix is used to reveal the underlying geometry structure of data in unsupervised dimensionality reduction methods. Because of noise data, it is difficult to learn the optimal similarity matrix. In this letter, we propose a new dimensionality reduction model for 2D image matrices: unsupervised 2D dimensionality reduction with adaptive structure learning (DRASL). Instead of using a predetermined similarity matrix to characterize the underlying geometry structure of the original 2D image space, our proposed approach involves the learning of a similarity matrix in the

*Corresponding author.

Neural Computation **29, 1352–1374** (2017)
doi:10.1162/NECO_a_00950

procedure of dimensionality reduction. To realize a desirable neighbors assignment after dimensionality reduction, we add a constraint to our model such that there are exact c connected components in the final subspace. To accomplish these goals, we propose a unified objective function to integrate dimensionality reduction, the learning of the similarity matrix, and the adaptive learning of neighbors assignment into it. An iterative optimization algorithm is proposed to solve the objective function. We compare the proposed method with several 2D unsupervised dimensionality methods. K-means is used to evaluate the clustering performance. We conduct extensive experiments on Coil20, AT&T, FERET, USPS, and Yale data sets to verify the effectiveness of our proposed method.

1 Introduction

High-dimensional feature data frequently appear in many fields of scientific research such as image and video category recognition, gene expression; and time series prediction (Lakshmanan, Sadtler, Tyler-Kabara, Batista, & Yu, 2015). However, directly handling these high-dimensional data inevitably suffers from the curse of dimensionality and massive storage cost (Kadir, Goodman, & Harris, 2014; Yamada, Jitkrittum, Sigal, Xing, & Sugiyama, 2014). In the past, many dimensionality reduction techniques that aim to learn an intrinsic low-dimensional compact representation have received considerable attention. Among these diverse approaches, unsupervised dimensionality reduction is appealing because it drops some irrelevant features while keeping the interpretation of dimension-reduced processing. Considering that large-scale data are usually collected without labels in practice, in this letter, we focus on the challenging problem of unsupervised dimensionality reduction.

Without the discriminative information from labels, the key step of unsupervised dimensionality reduction lies in preserving the intrinsic structure of original input data after dimensionality reduction (Dy & Brodley, 2004). Typically the local geometric structure is characterized by a pairwise similarity matrix based on a kernel-defined function. Once the similarity matrix is predetermined in the original high-dimensional space, it is fixed in the procedure of dimensionality reduction, such as locality preserving projection (LPP) (Niyogi, 2004) and locally linear embedding (LLE) (Roweis & Saul, 2000). However, this strategy might degrade performance because the kernel-defined weights used for calculating a similarity matrix are sensitive to the hyperparameter (such as the width in gaussian function) and lack meaningful interpretation. Importantly, the pairwise similarity matrix is learned at the beginning of the procedure of dimensionality reduction. Thus, it might not be the optimal one and fails to capture the intrinsic structure accurately (Gao et al., 2016). Recently, some efforts have been

devoted to exploiting an optimal underlying structure in the procedure of dimensionality reduction. For example, Nie, Wang, and Huang (2014) proposed learning the data similarity matrix by assigning adaptive and optimal neighbors for each data point. Du and Shen (2015) looked at both the global and local structures of data and performed adaptive structure learning by iteratively improving the probabilistic neighborhood relationship Kodirov, Xiang, Fu, and Gong (2016) integrated the graph's learning into an ℓ_1-norm graph regularized optimization problem for robust subspace clustering.

It is noteworthy that the state-of-the-art dimensionality reduction models for image and video representation commonly transform the two-dimensional image matrices into one-dimensional image vectors. This matrix-to-vector transformation not only leads to an extremely high-dimensional image vector space but also ruins the specific structure of the original 2D images (Yang, Zhang, Frangi, & Yu Yang, 2004; Zhang & Zhou, 2005; Bennamoun, Guo, & Sohel, 2015; Koch & Naito, 2007; Chang et al., 2015). Instead of using the matrix-to-vector transformation, in this letter, we propose a straightforward 2D unsupervised dimensionality reduction model with adaptive structure learning. This method not only mitigates the negative impact of the predetermined similarity matrix but also involves the optimal structure learning of 2D images into the procedure of dimensionality reduction. The main contributions of this letter are summarized as follows:

- A novel 2D unsupervised dimensionality reduction model is proposed by performing dimensionality reduction and optimal underlying geometry structure learning simultaneously.
- To achieve a desirable neighbors assignment, we impose a structure regularization on the graph of 2D data such that the number of connected components in the optimal graph equals the number of clusters.
- We exploit an efficient algorithm to solve the proposed challenging problem and conduct extensive experiments on benchmark data sets to illustrate the superiority of the proposed algorithm.

Notations and definitions: We use some special symbols in the formulas. For any matrix $P = [P_{ij}] \in \mathbb{R}^{m\times n}$, let P_j is the jth column of matrix P. $tr(A)$ refers to the trace of matrix A. $r(A)$ denotes the rank of matrix A. $\mathbf{1} \in \mathbb{R}^n$ is a column vector with all the elements are 1.

2 Related Work

Considerable effort has been devoted to improving the performance of image representation and recognition through dimensionality reduction techniques (Turk & Pentland, 1991; Roweis & Saul, 2000; Yang et al., 2004;

Zhang & Zhou, 2005; Hu, Feng, & Zhou, 2007; Kambhatla & Leen, 1997). Among these approaches, principal component analysis (PCA), widely used for image recognition, transforms the 2D image matrix into a 1D vector. When the dimensionality of the instance is high, it is hard to evaluate the total covariance matrix accurately due to its large size (Luo et al., 2016; Gao, Song, Liu et al., 2015). To solve this problem, Yang et al. (2004) extended conventional PCA to its 2D version, (2D)PCA and computed the image covariance matrix based on the original image matrices. This strategy not only enhances the evaluation of covariance matrix but also saves considerable time in determining the corresponding eigenvectors. However, this algorithm works in the row direction of images and requires more coefficients for image representation (Zhang & Zhou, 2005). Because 1D dimensionality reduction is inflexible, it is hard to achieve a smaller and more representative projection space. Zhang and Zhou (2005) developed a two-dimensional PCA model, (2D)2PCA, by examining the row and column directions simultaneously. Note that all the PCA-based approaches to dimensionality reduction depend on the total variance, which might fail to obtain a desirable representation when the distance between two clusters is shorter than that of intracluster (Welling, 2005; Dong, Huang, & Wen, 2010; Hosoya & Hyvärinen, 2016).

Unsupervised dimensionality reduction aims to find the most discriminative features that preserve the underlying geometry structure of original feature space as much as possible (Du & Shen, 2015; Bennamoun et al., 2015; Dy & Brodley, 2004). In the absence of label information, the local manifold structure is typically captured through a graph of data points with pairwise similarity (weighted) matrix (He, Ji, Zhang, & Bao, 2011; Hou, Nie, Li, Yi, & Wu, 2014; Dash & Liu, 2000; Niyogi, 2004; Cai, Zhang, & He, 2010; He, Yan, Hu, Niyogi, & Zhang, 2005). The Laplacian of this graph indeed reveals the adjacency relations of the data points (Belkin & Niyogi, 2001, 2003; Niyogi, 2004). It facilitates mapping the original data in high-dimensional space into a more representative subspace. However, it is a necessary step for previous feature selection and extraction models to transform the 2D image matrices into 1D vectors. Few studies are conducted to select or extract the most informative features from 2D image matrices in a straightforwardway. To the best of our knowledge, Hu et al. (2007) extended the conventional LPP model to its 2D version by predetermining the adjacency relation in the original 2D image space. As we mentioned, this method might fail to characterize the underlying local structure accurately due to the noise in the original space (Nie et al., 2014; Dy & Brodley, 2004). For the annotation of image and video, Gao, Song, Nie et al. (2015) proposed optimal graph learning (OGL), a method that constructs a similarity graph on multiple features and partial tags (Wang, Zhang, Song, Sebe, & Shen, 2016). OGL is a semisupervised method that learns the optimal graph based on a theory that the shorter the distance is between two samples, the greater the possibility is for them to be neighbors. However, if we have less labeled

images or videos, the performance of annotation may be degraded. Song et al. (2016) proposed joint graph learning and video segmentation (JGLVS), an unsupervised framework that learns the optimal similarity graph and video segmentation simultaneously. For JGLVS, the similarity matrix of superpixels is learned by using the location information of superpixels and spatial information, and all the superpixels have K connected components. Beyond the previous approaches, we aim in this letter to involve the optimal intrinsic structure learning into the procedure of 2D dimensionality reduction by assigning the optimal neighbors to each data point adaptively.

3 The Proposed Model

In this letter, we suppose that a 2D image data set consists of N samples $X_1, X_2, \ldots, X_N$. Each sample is presented by a matrix of size $m \times n$. Let P_{ij} $(i, j = 1, 2, \ldots, N)$ be the probability of data points X_i and X_j to be a neighbor. $U \in \mathbb{R}^{m\times u}$ and $V \in \mathbb{R}^{n\times v}$ are the transformation matrices that map the original data points in high-dimensional space to a lower-dimensional space. For better representation, we collect all the probabilities of data points to be neighbors into matrix $P = [P_{ij}] \in \mathbb{R}^{N\times N}$. To involve the optimal underlying structure learning into the procedure of dimensionality reduction (Kokiopoulou & Saad, 2007; Zhao, Wang, Liu, & Ye, 2013), we propose to estimate the transformation matrices U, V and the probability matrix P simultaneously by solving the following optimization problem,

$$\begin{aligned} \min_{P,U,V} \quad & \sum_{i,j=1}^{N} \|U^T X_i V - U^T X_j V\|_F^2 P_{ij}, \\ s.t. \quad & P_i^T \mathbf{1} = 1, 0 \le P_i \le 1, U^T U = I, V^T V = I \end{aligned} \tag{3.1}$$

where the objective function 3.1 encourages the low-dimensional representation of samples that obey the structure of original feature space. Unlike the previous methods that predetermine the pairwise weighted matrix P to characterize the underlying structure (Niyogi, 2004), our proposed model updates the weighted matrix iteratively to achieve an optimal one in the procedure of dimensionality reduction.

Note that optimization equation 3.1 employs a probability matrix to reveal the local geometry structure of input data. However, this strategy ignores the group structure and supposes that each data point in the adjacent graph with matrix P has only one neighbor. For a task of c clustering, it is reasonable to encourage the graph to have exact c connected components (Nie et al., 2014). Assume that each data point is given a function value $\mathbf{f}_i \in \mathbb{R}^{1\times c}$ for $i = 1, 2, \cdots, N$. According to the theorem in Fan (1950), we have

$$\sum_{i,j=1}^{N} \|\mathbf{f}_i - \mathbf{f}_j\|_2^2 P_{ij} = 2Tr(F^T L F), \tag{3.2}$$

where matrix $F = [\mathbf{f}_1^T, \mathbf{f}_2^T, \cdots, \mathbf{f}_N^T]^\top \in \mathbb{R}^{N\times c}$. $L \in \mathbb{R}^{N\times N}$ is the Laplacian matrix, which is defined as $L = D - \frac{P^T+P}{2}$, where D is a diagonal matrix with its ii-th element $D_{ii} = \sum_j (P_{ij} + P_{ji})/2$ for $j = 1, 2, \cdots, N$ (Chang, Nie, Ma, Yang, & Zhou, 2014). Based on these definitions, if matrix P is nonnegative, there is an important theorem (Chung, 1997; Mohar, 1991):

Theorem 1. *After the eigenvalue decomposition of the matrix L, the multiplicity c of the eigenvalue 0 is equal to the number of connected components in the graph associated with P.*

According to theorem 1, if the rank of matrix L is $N - c$, that is, $r(L) = N - c$, data points could be assigned to c clusters. To assign the projected data points in the reduced subspace to c clusters, we involve an additional rank constraint in the construction of graph such that the optimization problem, equation 3.1, turns to

$$\begin{aligned} &\min_{P,U,V} \quad \sum_{i,j=1}^{N} \|U^T X_i V - U^T X_j V\|_F^2 P_{ij}, \\ &s.t. \quad P_i^T \mathbf{1} = 1, 0 \le P_i \le 1, U^T U = I, V^T V = I, r(L) = N - c. \end{aligned} \tag{3.3}$$

Suppose σ_i is the ith smallest eigenvalue of L; the constraint $r(L) = N - c$ of optimization equation 3.3 is equal to the solution of the following problem:

$$\sum_{i=1}^{c} \sigma_i = \min_{F\in\mathbb{R}^{N\times c}, F^T F = I} tr\left(F^T L F\right). \tag{3.4}$$

Combining the optimization equations 3.3 to 3.4, the idea of adaptive graph learning with consideration of group structure is formulated as the following problem:

$$\begin{aligned} &\min_{P,U,V,F} \quad \sum_{i,j=1}^{N} \|U^T X_i V - U^T X_j V\|_F^2 P_{ij} + \lambda_\infty \|\mathbf{f}_i - \mathbf{f}_j\|_2^2 P_{ij}, \\ &s.t. \quad P_i^T \mathbf{1} = 1, 0 \le P_i \le 1, U^T U = I, V^T V = I, \end{aligned} \tag{3.5}$$

where λ_∞ is a large enough value, which can keep the c smallest eigenvalue of L equal to zero.

Compared with equation 3.1, equation 3.5 can ensure that the projected data points can clustered be better. In order to avoid a trivial solution, we impose a penalty $\gamma \sum_{i,j=1}^{N} P_{ij}^2$ into the objective function, equation 3.5, such that our model becomes

$$\begin{aligned} &\min_{P,U,V,F} \quad \sum_{i,j=1}^{N} \|U^T X_i V - U^T X_j V\|_F^2 P_{ij} + \gamma P_{ij}^2 + \lambda_\infty \|\mathbf{f}_i - \mathbf{f}_j\|_2^2 P_{ij}, \\ &s.t. \quad U^T U = I, V^T V = I, P_i^T \mathbf{1} = 1, 0 \le P_i \le 1, F^T F = I \end{aligned} \tag{3.6}$$

where γ is a regularization parameter (Chang, Yang, Long, Zhang, & Hauptmann, 2016).

3.1 Optimization Algorithm for Problem 3.6. It is obvious that the proposed model is jointly convex. In this section, we use an alternative optimization algorithm to solve the proposed model. When U, V, and P are fixed, the optimal F can be obtained through solving the following optimization problem:

$$\min_{F \in \mathbb{R}^{N \times c}, F^T F = I} tr\left(F^T L F\right). \tag{3.7}$$

It is evident that the variable F can be solved by generalizing eigenvalue decomposition.

When P and F are fixed, the optimization problem 3.6 associated with variables U and V becomes

$$\min_{U^T U = I, V^T V = I} \sum_{i,j=1}^{N} ||U^T X_i V - U^T X_j V||_F^2 P_{ij}. \tag{3.8}$$

For convenience, we denote the objective function of equation 3.8 as

$$\mathbf{G}(\mathbf{U}, \mathbf{V}) = \sum_{i,j=1}^{N} ||U^T X_i V - U^T X_j V||_F^2 P_{ij}. \tag{3.9}$$

Considering the definition $\|A\|_F^2 = tr(A^T A)$ for any matrix A, we deduce the following equations, respectively:

$$W^v = \sum_{i,j=1}^{N} P_{ij} (X_i - X_j) V V^T (X_i - X_j)^T, \tag{3.10}$$

$$W^u = \sum_{i,j=1}^{N} P_{ij}(X_i - X_j)^T UU^T(X_i - X_j). \tag{3.11}$$

As a result, the objective function of optimization equation 3.8 can be rewritten as

$$\mathbf{G}(\mathbf{U}, \mathbf{V}) = tr\left(U^T W^v U\right) = tr\left(V^T W^u V\right). \tag{3.12}$$

When the variable V is fixed, optimization equation 3.8 is equivalent to the following form:

$$\min_{U^T U = I} tr\left(U^T W^v U\right). \tag{3.13}$$

The optimal solution U to optimization problem 3.13 is the orthogonal generalized eigenvectors of W^v corresponding to the u smallest generalized eigenvalues. Similarly, if variable U is fixed, optimization problem 3.8 with respect to variable V is equivalent to the following form:

$$\min_{V^T V = I} tr\left(V^T W^u V\right). \tag{3.14}$$

The optimal solution V to problem 3.14 is achieved by the v eigenvectors corresponding to the v smallest eigenvalues of W^u.

When U, V, and F are fixed, the optimal solution P to optimization problem, equation 3.6, can be obtained by solving the following problem:

$$\begin{aligned} &\min_{P} \quad \sum_{i,j=1}^{N} ||U^T X_i V - U^T X_j V||_F^2 P_{ij} + \gamma P_{ij}^2 + \lambda_\infty ||\mathbf{f}_i - \mathbf{f}_j||_2^2 P_{ij}. \\ &s.t. \quad P_i^T \mathbf{1} = 1, 0 \le P_i \le 1 \end{aligned} \tag{3.15}$$

In order to simplify the solution process, denote $d_{ij}^1 = \|U^T X_i V - U^T X_j V\|_F^2$, $d_{ij}^2 = \|f_i - f_j\|_2^2$ and $d_{ij} = d_{ij}^1 + d_{ij}^2$. Then the optimization problem 3.5 can be rewritten as

$$\min_{P_i^T \mathbf{1} = 1, 0 \le P_i \le 1} \left\| P_i + \frac{1}{2\gamma} d_i \right\|_2^2. \tag{3.16}$$

This problem turns to a conventional Euclidean projection problem in the simplex space, which can be solved efficiently by using the algorithm proposed in Nie et al. (2014).

Algorithm 1: The Optimal Algorithm of DRASL.

Data: Data points X_1, X_2, ... X_N, the parameters $k, c, r, u, v, \lambda_\infty$.

Result: Projection matrices $U \in \mathbb{R}^{m\times u}$ and $V \in \mathbb{R}^{n\times v}$.

Initialize each column of P by solving the optimization problem

$$\min_{P_i^T \mathbf{1}=1, 0\le P_i\le 1} \sum_{j=1}^{N} ||X_i - X_j||_F^2 P_{ij} + \gamma P_{ij}^2;$$

The initial matrices of V and U are set as an arbitrary column orthogonal matrix;

Set t=0;

repeat

  Update $L^t = D^t - \frac{P^{tT}+P^t}{2}$, where $D^t \in \mathbb{R}^{N\times N}$ is diagonal matrix with the i-th diagonal element as $\sum_j (P_{ij}^t + P_{ji}^t)/2$;

  Update F^t, whose columns are the c eigenvectors of L^t corresponding to its c smallest eigenvalues;

  Update U^t, whose columns are the u eigenvectors of W_t^v corresponding to the u smallest eigenvalues in equation 3.13;

  Update V^t, whose columns are the v eigenvectors of W_t^u corresponding to the v smallest eigenvalues in equation 3.14;

  Update the ith column of P^t by solving the equation 3.15, where $d_i \in \mathbb{R}^{N\times 1}$ is a vector with the jth element is $d_{ij} = d_{ij}^1 + d_{ij}^2$;

  $t = t + 1$;

until *Convergence*;

Return the projection matrices U and V.

In summary, the alternative algorithm for optimization problem, equation 3.6, is shown in algorithm 1.

3.2 The Solution of Parameter γ and Matrix P. In objective function 3.6, γ is an important parameter that is connected with the learning of underlying geometry structure. However, the value of γ, which is among zero to infinite, is difficult to determine. Here, we provide an efficient method to obtain its value. According to Nie et al. (2014), the Lagrangian function of problem 3.16 can be formulated as

$$\frac{1}{2}\left\|P_i + \frac{d_i}{2\gamma_i}\right\|_2^2 - \alpha(P_i^T\mathbf{1} - 1) - \beta_i^T P_i. \tag{3.17}$$

According to KKT condition (Boyd, Vandenberghe, & Faybusovich, 2013), P_{ij} can be solved by

$$P_{ij} = \left(-\frac{d_{ij}}{2\gamma_i} + \alpha\right)_+. \tag{3.18}$$

The shorter the distance between two samples is, the greater the possibility is for them to be neighbors. In order to detemine the top-k neighbors of each sample, we sort each row of D to Q in ascending order, where D is a matrix with iith element d_{ij}. After that, we have

$$\begin{cases} -\dfrac{Q_{it}}{2\gamma_i} + \alpha > 0 \quad t = 1, \ldots, k \\ -\dfrac{Q_{it}}{2\gamma_i} + \alpha \le 0 \quad t = k+1, \ldots, N, \end{cases}. \tag{3.19}$$

By imposing a constraint $\sum_{j=1}^{N} P_{ij} = 1$ on equation 3.18, the following criterion is adopted to determine α:

$$\alpha = \frac{1}{k} + \frac{1}{2k\gamma_i}\sum_{t=1}^{k} Q_{it}. \tag{3.20}$$

For γ_i, the optimal solution satisfies

$$\frac{k}{2}Q_{ik} - \frac{1}{2}\sum_{t=1}^{k} Q_{it} < \gamma_i \le \frac{k}{2}Q_{i,k+1} - \frac{1}{2}\sum_{t=1}^{k} Q_{it}. \tag{3.21}$$

Then, if the above inequality is satisfied, each sample will have only k neighbors. Without loss of generality, we set

$$\gamma_i = \frac{k}{2} Q_{i,k+1} - \frac{1}{2} \sum_{t=1}^{k} Q_{it}. \tag{3.22}$$

Finally, γ is easily solved by

$$\gamma = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{k}{2} Q_{i,k+1} - \frac{1}{2} \sum_{t=1}^{k} Q_{it} \right). \tag{3.23}$$

3.3 Convergence Analysis. By the following theorem, the convergence of our proposed objective function can be proved and the global optimal solution of projection matrices U and V can be obtained.

Theorem 2. *The value of our objective function decreases constantly until convergence in the process of iteration in algorithm 1.*

Lemma 1. *If three matrices of the four matrices P, F, U, and V can be fixed, then another matrix can be obtained.*

Proof. First, by fixing P, F, and U, the objective function in equation 3.6 will become the form of equation 3.13, a convex optimization problem. Thus we can get the global solution of V by taking the derivative of V in equation 3.14 and setting, it to zero. Second, in the same way, if P, F, and V are fixed, we can obtain the global solution of U by solving the convex optimization problem shown in equation 3.13. Third, when P, U, and V are fixed, it is easy to obtain the global solution of F. Finally, if U, V, and F are fixed, it can be seen that equation 3.15 is a convex function, and the global solution of P can be solved easily.

Based on lemma 1, we verify theorem 2 as follows. According to algorithm 1, after the tth iteration, we obtain $U = U^t$, $V = V^t$, $P = P^t$, and $F = F^t$. In the next iteration, $U = U^{t+1}$, $V = V^{t+1}$, $P = P^{t+1}$, and $F = F^{t+1}$.

If P^t, V^t, F^t is fixed, we have

$$\begin{aligned} &\sum_{i,j=1}^{N} \|{U^{t+1}}^T X_i V^t - {U^{t+1}}^T X_j V^t\|_F^2 P_{ij}^t + \gamma {P_{ij}^t}^2 + \lambda_\infty \|\mathbf{f}_i^t - \mathbf{f}_j^t\|_2^2 P_{ij}^t \\ &\leq \sum_{i,j=1}^{N} \|{U^t}^T X_i V^t - {U^t}^T X_j V^t\|_F^2 P_{ij}^t + \gamma {P_{ij}^t}^2 + \lambda_\infty \|\mathbf{f}_i^t - \mathbf{f}_j^t\|_2^2 P_{ij}^t. \end{aligned} \tag{3.24}$$

Similarly, if P^t, U^t, F^t is fixed, we have

$$\sum_{i,j=1}^{N} \|U^{t^T} X_i V^{t+1} - U^{t^T} X_j V^{t+1}\|_F^2 P_{ij}^t + \gamma {P_{ij}^t}^2 + \lambda_\infty \|\mathbf{f}_i^t - \mathbf{f}_j^t\|_2^2 P_{ij}^t$$

$$\leq \sum_{i,j=1}^{N} \|U^{t^T} X_i V^t - U^{t^T} X_j V^t\|_F^2 P_{ij}^t + \gamma {P_{ij}^t}^2 + \lambda_\infty \|\mathbf{f}_i^t - \mathbf{f}_j^t\|_2^2 P_{ij}^t. \tag{3.25}$$

If U^t, V^t, F^t is fixed,

$$\sum_{i,j=1}^{N} \|U^{t^T} X_i V^t - U^{t^T} X_j V^t\|_F^2 P_{ij}^{t+1} + \gamma {P_{ij}^{t+1}}^2 + \lambda_\infty \|\mathbf{f}_i^t - \mathbf{f}_j^t\|_2^2 P_{ij}^{t+1}$$

$$\leq \sum_{i,j=1}^{N} \|U^{t^T} X_i V^t - U^{t^T} X_j V^t\|_F^2 P_{ij}^t + \gamma {P_{ij}^t}^2 + \lambda_\infty \|\mathbf{f}_i^t - \mathbf{f}_j^t\|_2^2 P_{ij}^t. \tag{3.26}$$

If P^t, V^t, U^t is fixed,

$$\sum_{i,j=1}^{N} \|U^{t^T} X_i V^t - U^{t^T} X_j V^t\|_F^2 P_{ij}^t + \gamma {P_{ij}^t}^2 + \lambda_\infty \|\mathbf{f}_i^{t+1} - \mathbf{f}_j^{t+1}\|_2^2 P_{ij}^t$$

$$\leq \sum_{i,j=1}^{N} \|U^{t^T} X_i V^t - U^{t^T} X_j V^t\|_F^2 P_{ij}^t + \gamma {P_{ij}^t}^2 + \lambda_\infty \|\mathbf{f}_i^t - \mathbf{f}_j^t\|_2^2 P_{ij}^t. \tag{3.27}$$

Consider

$$\sum_{i,j=1}^{N} \|U^{t+1^T} X_i V^{t+1} - U^{t+1^T} X_j V^{t+1}\|_F^2 P_{ij}^{t+1}$$

$$\leq \sum_{i,j=1}^{N} \|U^{t^T} X_i V^t - U^{t^T} X_j V^t\|_F^2 P_{ij}^t, \tag{3.28}$$

and

$$\|\mathbf{f}_i^{t+1} - \mathbf{f}_j^{t+1}\|_2^2 P_{ij}^{t+1} \leq \|\mathbf{f}_i^t - \mathbf{f}_j^t\|_2^2 P_{ij}^t. \tag{3.29}$$

Combining the formulas between equations 3.24 and 3.29, we can sum up the following formula:

$$\sum_{i,j=1}^{N} \|U^{t+1^T} X_i V^{t+1} - U^{t+1^T} X_j V^{t+1}\|_F^2 P_{ij}^{t+1} + \gamma P_{ij}^{t+1^2} + \lambda_\infty \|\mathbf{f}_i^{t+1} - \mathbf{f}_j^{t+1}\|_2^2 P_{ij}^{t+1}$$

$$\leq \sum_{i,j=1}^{N} \|U^{t^T} X_i V^t - U^{t^T} X_j V^t\|_F^2 P_{ij}^t + \gamma P_{ij}^{t^2} + \lambda_\infty \|\mathbf{f}_i^t - \mathbf{f}_j^t\|_2^2 P_{ij}^t. \tag{3.30}$$

□

Theorem 2 is proved; thus, the value of objective function 3.6 decreases constantly until convergence by using algorithm 1.

4 Experimental Analysis

In this section, we evaluate the performance of our proposed unsupervised 2D dimensionality reduction with adaptive structure learning algorithm (DRASL) on five popular data sets: Coil20 (Nene, Nayar, & Murase, 1996), AT&T (Samaria & Harter, 1994), Yale (Belhumeur, Hespanha, & Kriegman, 1997), USPS, and FERET (Phillips, Wechsler, Huang, & Rauss, 1998).

4.1 Experiment Data Sets. The Coil20 (Nene et al., 1996) data set consists of 1440 images—72 different images per object for 20 objects. These images were taken at five different degrees. To simplify the computation of experiments, we crop each image to 32×32 pixels and use the pixel values as features. The AT&T (Samaria & Harter, 1994) data set consists of 400 images from 40 different subjects, and each subject has 10 distinct images. Some images were taken at different times and in different lighting, and have various facial expressions (e.g., smiling or nonsmiling, open or closed eyes). Each image has is 112×92, and we use the pixel values as features. The Yale (Belhumeur et al., 1997) data set consists of 165 gray-scale images of 15 individuals. Each individual has 11 images with distinct facial expression (e.g., sad, happy, supervised), different lighting (e.g., center-light, left-light, right-light), and other configurations. The FERET (Phillips et al., 1998) data set consists of 1400 images of 200 subjects. Each subject has 7 images. To facilitate computation, we select 490 images of the data set, each image is downsampled to the size of 80×80, and the pixel values are used as features. The subset of the USPS data set contains 1854 gray-scale handwritten digit images, and each image is cropped to 16×16 to evaluate the performance of handwritten digit recognition. The detailed information about these data sets is summarized in Table 1.

Table 1: Description of Data Sets.

Data Set	Sample	Feature	Class	Dimensions of Feature Vector
USPS	1854	256	10	$\{2, 4, \ldots, 14\}$
Coil20	1440	1024	20	$\{4, 8, \ldots, 28\}$
FERET	490	6400	70	$\{10, 20, \ldots, 70\}$
AT&T	400	10,304	40	$\{10, 20, \ldots, 80\}$
Yale	165	77,760	15	$\{20, 50, \ldots, 200\}$

4.2 Evaluation Matrices. In our experiments, accuracy (ACC) and normalized mutual information (NMI) are used to evaluate the clustering performance (Cai, He, & Han, 2005):

- ACC: For ith image, we denote g_i as the obtained clustering label and h_i as the truth label, respectively. The calculation formula of ACC is

$$ACC = \frac{\sum_{i=1}^{N} \delta(h_i, map(g_i))}{N}, \tag{4.1}$$

  where N is the number of images, $map(g_i)$ is a permutation function that maps the obtained clustering label to the truth label, and δ is a function that accomplishes the matching of x and y if $x = y$, $\delta(x, y) = 1$ and equals 0 otherwise.
- NMI: Normalized mutual information is another standard that can be used to assess the performance of clustering. For any two arbitrary variables C and D, $NMI(C, D)$ can be used as

$$NMI(C, D) = \frac{I(C, D)}{\sqrt{H(C)H(D)}}, \tag{4.2}$$

  where $I(C, D)$ is a function that computes the mutual information between C and D, $H(C)$, and $H(D)$ represent the entropies of C and D, respectively. We denote t_l as the number of samples in cluster ζ and $\widetilde{t}_h$ as the number of samples of hth truth class. NMI can be computed by the following formula:

$$NMI = \frac{\sum_{l=1}^{c} \sum_{h=1}^{c} t_{l,h} \log\left(\frac{N \times t_{l,h}}{t_l \widetilde{t}_h}\right)}{\sqrt{\left(\sum_{l=1}^{c} t_l \log \frac{t_l}{N}\right)\left(\sum_{h=1}^{c} \widetilde{t}_h \log \frac{\widetilde{t}_h}{N}\right)}}, \tag{4.3}$$

  where $t_{l,h}$ is the intersectional samples number of cluster ζ and the hth ground truth class.

4.3 Compared Algorithms. To evaluate the effectiveness of our proposed DRASL, we compare it with the following 2D unsupervised dimensionality reduction algorithms.

4.3.1 (2D)PCA (Yang et al., 2004). This extracts the eigenvectors of image matrix in the row direction of 2D image matrices. The solution of the projection matrix relies on a covariance matrix *RO*, which can be obtained by the following formula,

$$RO = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^T(X_i - \bar{X}), \tag{4.4}$$

where $\bar{X}$ is the average image of all images; projection matrix A of (2D)PCA is composed of r eigenvectors corresponding to the r largest eigenvalues of *RO*.

4.3.2 (2D)2PCA (Zhang and Zhou, 2005). This considers both row and column directions to extract the eigenvectors of a 2D image matrix. In (2D)2PCA, the alternative (2D)PCA is used to extract the features of column direction. Assume $X_i^{(j)}$ and $\bar{X}^{(j)}$ are the jth column vectors of X_i and $\bar{X}$, respectively. Then the covariance matrix *CO* of alternative (2D)PCA can be defined as

$$CO = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{n}(X_i^{(j)} - \bar{X}^{(j)})(X_i^{(j)} - \bar{X}^{(j)})^T. \tag{4.5}$$

The projection matrix B of alternative (2D)PCA consists of the s eigenvectors, which correspond to the s largest eigenvalues of *CO*. Considering equations 4.34 and 4.35, the projection matrix C of (2D)2PCA is

$$C = B^T XA. \tag{4.6}$$

4.3.3 (2D)LPP (Hu et al., 2007). This is based directly on 2D image matrices to conduct dimensionality reduction rather than 1D vectors as conventional LPP does. The basic idea of the (2D)LPP algorithm is to find a subspace that can not only reduce the dimensionality, but also preserve the local manifold structure of data. The important part of (2D)LPP is to construct the nearest neighbor graph, which reveals the adjacent relationships of samples. The (2D)LPP procedure that extracts features is summarized as follows:

1. Construct adjacency graph *G*.

2. Assign weights.
3. Construct. the covariance matrix. The eigenvectors of (2D)LPP are computed by

$$H^T(L \otimes I_m)Hw = \lambda H^T(D \otimes I_m)Hw, \tag{4.7}$$

where $\otimes$ means the Kronecker product of two matrices. D is a diagonal matrix, and the iith element is $D_{ii} = \sum_{j=1}^{n} S_{ij}$ (S_{ij} is the weight between the ith and the jth data points). H comprises all the image matrices and $H = [X_1^T, X_2^T, X_3^T, \ldots, X_N^T]^T$. Then the projection matrix J of (2D)LPP consists of k eigenvectors that correspond to the k smallest eigenvalues.

4.3.4 I(2D)PCA (Woraratpanya et al., 2015). This is extension of (2D)PCA that excludes the influence of the illumination effect based on (2D)PCA. The difference between (2D)PCA and I(2D)PCA is that the construction of a covariance matrix is based on only one zero mean image by I(2D)PCA when (2D)PCA relies on N zero mean images to construct it.

4.4 Experiment Settings. In our experiment, we first obtain the projection matrices by various dimensionality reduction methods, and then evaluate the clustering performance by K-means in terms of accuracy (ACC) and normalized mutual information (NMI). Since the final values of the clustering are determined by the initial values, for different dimensionality reduction methods, we repeat the K-means 30 times with random initial values and take the average values as the final results.

There are some suggestions for parameter settings. For (2D)LPP, the neighbor structure graph is constructed by using the self-tune gaussian method (Zelnik-Manor & Perona, 2005). For DRASL, γ is determined by equation 3.23 with $k = 5$, and c is set to be the number of classes. Since λ_∞ is a very large number, it is unsuitable to use it as a fixed value in the algorithm. In our experiments, we dynamically change the value of λ_∞ according to the value of c: λ_∞ is initialized with γ; if the number of clusters is larger than c, λ_∞ will be divided by two; λ_∞ doubles its value otherwise.

4.5 Experimental Results. For evaluating the performance of our algorithm in clustering, we conduct experiments on five data sets. To be fair to all the comparison algorithms, the dimensions of feature vectors in (2D)2PCA and DRASL are set to $(c-1) \times (c-1)$ and $(c-1)$ in (2D)PCA, (2D)LPP, and I(2D)PCA, where c is the number of classes. The comparison results are summarized in Tables 2 and 3, in which each element is the sum of the average value and the standard deviation. The baseline is all features.

From Tables 2 and 3 we can observe that the DRASL algorithm can obtain better clustering results than the baseline in terms of ACC and NMI. The reason is that after the dimensionality reduction, the noisy and irrelevant

Table 2: Clustering Results (ACC%±Standard Deviation) of Different Dimensionality Reduction Algorithms on Five Data Sets.

Data Set	Yale	FERET	AT&T	Coil20	USPS
Baseline	51.49 ± 5.61	33.10 ± 1.71	58.21 ± 3.92	59.07 ± 5.14	62.07 ± 3.86
(2D)PCA	52.28 ± 6.49	33.58 ± 1.26	59.35 ± 4.01	59.65 ± 3.76	63.00 ± 3.08
(2D)2PCA	53.88 ± 6.03	33.63 ± 1.32	59.67 ± 4.07	60.38 ± 5.79	62.81 ± 3.71
I(2D)PCA	54.47 ± 5.07	33.30 ± 1.83	59.04 ± 3.94	59.81 ± 5.42	62.62 ± 4.20
(2D)LPP	53.29 ± 4.76	33.56 ± 1.3	61.92 ± 3.56	60.85 ± 5.8	63.06 ± 5.11
DRASL	**56.03 ± 3.86**	**35.9 ± 1.98**	**63.50 ± 3.32**	**62.39 ± 5.75**	**64.87 ± 4.95**

Note: Numbers in bold are the result of our proposed method.

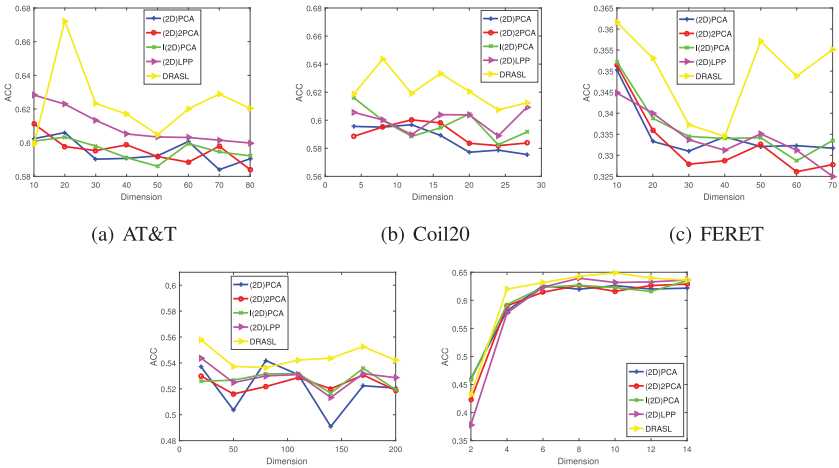
Table 3: Clustering Results (NMI% ± Standard Deviation) of Different Dimensionality Reduction Algorithms on Five Data Sets.

Data Set	Yale	FERET	AT&T	Coil20	USPS
Baseline	60.68 ± 3.71	63.08 ± 1.04	79.01 ± 1.64	73.04 ± 2.33	61.03 ± 1.94
(2D)PCA	62.23 ± 4.36	63.17 ± 1.18	79.43 ± 1.92	74.09 ± 2.07	61.20 ± 1.55
(2D)2PCA	61.33 ± 3.81	63.46 ± 0.86	78.87 ± 2.03	72.78 ± 2.39	61.67 ± 1.58
I(2D)PCA	62.31 ± 2.73	63.34 ± 1.41	79.16 ± 2.15	74.11 ± 2.42	61.30 ± 2.06
(2D)LPP	61.40 ± 3.50	63.60 ± 1.21	81.05 ± 1.87	74.10 ± 2.92	61.08 ± 2.37
DRASL	**65.29 ± 2.91**	**65.47 ± 0.97**	**81.98 ± 1.96**	**76.02 ± 2.63**	**63.83 ± 2.47**

Note: Numbers in bold are the result of our proposed method.

features can mostly be discarded. Therefore, the dimensionality reduction technique is necessary and effective for the clustering task. Not only can the informative features be preserved, but the performance of clustering is also improved. Second, as shown in Tables 2 and 3, the performance of DRASL outperforms the other four comparison algorithms. This is attributed to the process of constructing the neighbor structure graph and the learning of neighbors' assignment. Since the construction of a neighbor graph is a dynamic process, the proposed algorithm can effectively preserve the adjacency relations of data in the subspace by adaptive learning, and then it guarantees that the final results have only c connected clusters. Meanwhile, the better adjacency graph is helpful for us to reduce the dimensionality of images. However, (2D)PCA-based approaches neglect the importance of the underlying geometry structure and pay attention only to the total scatter of data. For (2D)LPP, its neighbor structure graph cannot reveal the actual relationships of data in the noise. Thus, the underlying geometry structure is crucial in the process of 2D unsupervised dimensionality reduction and our algorithm can obtain better results than others.

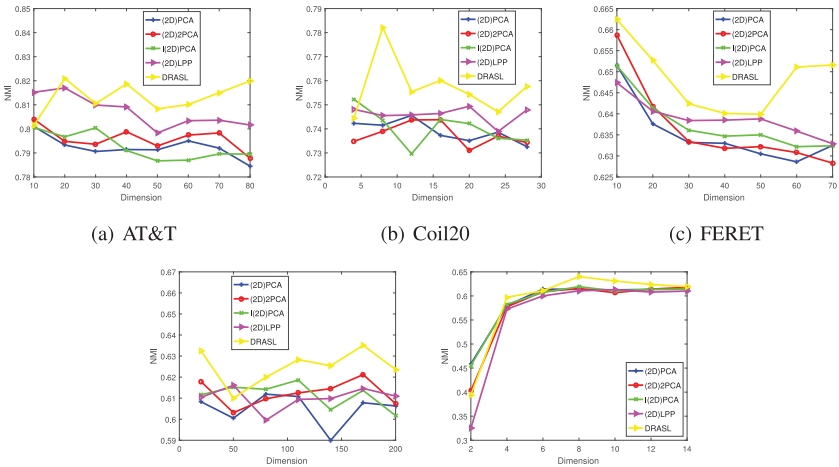
The second experiments aim to demonstrate the clustering performance of diverse dimensionality reduction algorithms with different numbers of



(a) AT&T (b) Coil20 (c) FERET

(d) Yale (e) USPS

Figure 1: Clustering ACC of five algorithms on five data sets.



(a) AT&T (b) Coil20 (c) FERET

(d) Yale (e) USPS

Figure 2: Clustering NMI of five algorithms on five data sets.

dimensions of feature vector. The experiments are conducted in five data sets, and the ACC and NMI results are in Figures 1 and 2. For the dimensionality reduction algorithms, we regularly select the dimensions of feature vectors for testing; the specific information is shown in Table 1.

Table 4: Clustering Results (ACC% ± Standard Deviation and NMI% ± Standard Deviation) of PCA and LPP on Three Data Sets.

Method	Data Set	FERET	Yale	AT&T
PCA	ACC	33.13 ± 1.12	54.79 ± 2.46	61.05 ± 3.03
	NMI	64.01 ± 1.10	61.28 ± 1.17	80.52 ± 1.73
LPP	ACC	32.53 ± 1.45	52.76 ± 4.85	62.21 ± 4.00
	NMI	63.42 ± 1.16	61.88 ± 1.99	80.95 ± 2.10

We find that our algorithm can get better results than (2D)PCA, (2D)2PCA, I(2D)PCA, and (2D)LPP in a number of dimensions of feature vectors. The results prove that our algorithm outperforms others.

In the experiments, we also compare our approach with the 1D image dimensional reduction algorithm PCA and LPP. To be fair, we transform the 2D image matrices to 1D vectors, and the dimensions of feature vector are set to $(c-1)$. For LPP, the neighbor structure graph is constructed by using the self-tuned gaussian method (Zelnik-Manor & Perona, 2005). The results are summarized in Table 4. For example, for the Yale data set, Table 3 shows the highest clustering NMI results, 65.29% of DRASL, and Table 4 shows the highest clustering NMI results, 61.88%. The results indicate that DRASL can get better performance than conventional PCA and LPP approaches.

Furthermore, as mentioned in algorithm 1, DRASL can achieve convergence by continuous iteration. In practice, our algorithm usually converges within 15 iterations, and the time complexity of our algorithm is $O(m^2u + n^2v + N^2c)$. However, a major drawback cannot be ignored. In the iteration procedure of algorithm 1, the deduction of matrices U, V, and F is derived from eigendecomposition and shows that the solution of projection matrices consumes much more time than others. In the experiment, all the algorithms are implemented with Matlab programming and run on CPU $i5-6200U$ with 2.40 GHz and 8 GB RAM. The average running time of our algorithm is about 14 minutes. Thus, we still have much work to do to improve our algorithm.

5 Conclusion

To address the challenging task of large-scale high-dimensional data dimensionality reduction, we propose a novel dimensionality reduction algorithm, DRASL. To construct an optimal similarity matrix, we involve the learning of a similarity matrix into the procedure of dimensionality reduction. To obtain a desirable neighbors' assignment after dimensionality reduction, we introduce the adaptive structure learning to the proposed model. An efficient iterative optimization algorithm has also been proposed to solve our objective function. We conduct rich experiments on five data

sets, and the experimental results indicate that DRASL can obtain better clustering results when the similarity matrix is optimal. But every coin has two sides; our algorithm also has some drawbacks, such as that DRASL will need more run time and some parameters need to be tuned in the experimental process. These deficiencies should be considered carefully in future work.

Acknowledgments

This work was jointly supported by the National Natural Science Foundations of China under grant 61502387, the Natural Science Foundations of Shaanxi under grant 2016JQ6029, and Educational Commission of Shaanxi Province, China under Grant 11JK1062.

References

Belhumeur, P., Hespanha, J. P., & Kriegman, D. J. (1997). Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection. *IEEE Trans. Pattern Anal. Mach. Intelligence*, *9*(7), 711–720.

Belkin, M., & Niyogi, P. (2001). Laplacian eigenmaps and spectral techniques for embedding and clustering. In T. G. Dietlerich, S. Becker, & Z. Ghahramani (Eds.), *Advances in neural information processing systems, 14* (pp. 585–591). Cambridge, MA: MIT Press.

Belkin, M., & Niyogi, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, *15*(6), 1373–1396.

Bennamoun, M., Guo, Y., & Sohel, F. (2015). Feature selection for 2d and 3d face recognition. In *Wiley encyclopedia of electrical and electronics engineering* (pp. 1–54). Hoboken, NJ: Wiley.

Boyd, S., Vandenberghe, L., & Faybusovich (2013). Convex optimization. *IEEE Transactions on Automatic Control*, *51*(11), 1859–1859.

Cai, D., He, X., & Han, J. (2005). Document clustering using locality preserving indexing. *IEEE Transactions on Knowledge and Data Engineering*, *17*(12), 1624–1637.

Cai, D., Zhang, C. Y., & He, X. F. (2010). Unsupervised feature selection for multi-cluster data. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 333–342). New York: ACM.

Chang, X., Nie, F., Ma, Z., Yang, Y., & Zhou, X. (2014). *A convex formulation for spectral shrunk clustering*. arXiv:1411.6308

Chang, X., Nie, F., Wang, S., Yang, Y., Zhou, X., & Zhang, C. (2015). Compound rank-k projections for bilinear analysis. *IEEE Transactions on Neural Networks and Learning Systems*, *27*(7), 1.

Chang, X., Yang, Y., Long, G., Zhang, C., & Hauptmann, A. G. (2016). *Dynamic concept composition for zero-example event detection*. arXiv:1601.03679

Chung, F. R. (1997). *Spectral graph theory*. Providence, RI: American Mathematical Society.

Dash, M., & Liu, H. (2000). Feature selection for clustering. *Encyclopedia of Database Systems*, *21*(3), 110–121.

Dong, X., Huang, H., & Wen, H. (2010). A comparative study of several face recognition algorithms based on PCA. In *Proceedings of the Third International Symposium on Computer Science and Computational Technology* (p. 443). N.p.

Du, L., & Shen, Y.-D. (2015). Unsupervised feature selection with adaptive structure learning. In *Proceedings of the ACM-SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 209–218). New York: ACM.

Dy, J. G., & Brodley, C. E. (2004). Feature selection for unsupervised learning. *Journal of Machine Learning Research*, *5*(4), 845–889.

Fan, K. (1950). On a theorem of Weyl concerning eigenvalues of linear transformations I. *PNAS*, *35*(11), 652.

Gao, L., Song, J., Liu, X., Shao, J., Liu, J., & Shao, J. (2015). Learning in high-dimensional multimedia data: The state of the art. *Multimedia Systems*, *21*, 1–11.

Gao, L., Song, J., Nie, F., Yan, Y., Sebe, N., & Heng, T. S. (2015). Optimal graph learning with partial tags and multiple features for image and video annotation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* Piscataway, NJ: IEEE.

Gao, L., Song, J., Nie, F., Zou, F., Sebe, N., & Shen, H. T. (2016). Graph-without-cut: An ideal graph learning for image segmentation. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence.* Cambridge, MA: AAAI Press.

He, X., Ji, M., Zhang, C., & Bao, H. (2011). A variance minimization criterion to feature selection using Laplacian regularization. *IEEE Trans. Pattern Anal. Mach. Intelligence*, *33*(10), 2013–2025.

He, X., Yan, S., Hu, Y., Niyogi, P., & Zhang, H.-J. (2005). Face recognition using Laplacianfaces. *IEEE Trans. Pattern Anal. Mach. Intelligence*, *27*(3), 328–340.

Hosoya, H., & Hyvärinen, A. (2016). Learning visual spatial pooling by strong PCA dimension reduction. *Neural Computation*, *28*(7), 1–16.

Hou, C., Nie, F., Li, X., Yi, D., & Wu, Y. (2014). Joint embedding learning and sparse regression: A framework for unsupervised feature selection. *IEEE Trans. Cybern.*, *44*(6), 793–804.

Hu, D., Feng, G., & Zhou, Z. (2007). Two-dimensional locality preserving projections (2DLPP) with its application to palmprint recognition. *Pattern Recognition*, *40*(1), 339–342.

Kadir, S. N., Goodman, D. F., & Harris, K. D. (2014). High-dimensional cluster analysis with the masked EM algorithm. *Neural Computation*, *26*(11), 2379–2394.

Kambhatla, N., & Leen, T. K. (1997). Dimension reduction by local PCA. *Neural Computation*, *9*, 1493–1516.

Koch, I., & Naito, K. (2007). Dimension selection for feature selection and dimension reduction with principal and independent component analysis. *Neural Computation*, *19*(2), 513–545.

Kodirov, E., Xiang, T., Fu, Z., & Gong, S. (2016). Learning robust graph regularisation for subspace clustering. In *Proceedings of the British Machine Conference.* N.p.

Kokiopoulou, E., & Saad, Y. (2007). Orthogonal neighborhood preserving projections: A projection-based dimensionality reduction technique. *IEEE Trans. Pattern Anal. Mach. Intelligence*, *29*(12), 2143–2156.

Lakshmanan, K. C., Sadtler, P. T., Tyler-Kabara, E. C., Batista, A. P., & Yu, B. M. (2015). Extracting low-dimensional latent structure from time series in the presence of delays. *Neural Computation*, *27*(9), 1–32.

Luo, M., Nie, F., Chang, X., Yang, Y., Hauptmann, A., & Zheng, Q. (2016). Avoiding optimal mean robust PCA/2DPCA with non-greedy l1-norm maximization. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence*. N.p.

Mohar, B. (1991). The laplacian spectrum of graphs. In Y. Alavi, G. Chartrand, O. Oellermann, & A. J. Schenk, *Graph theory, Combinatorics, and Applications* (pp. 871–898). New York: Wiley.

Nene, S. A., Nayar, S. K., & Murase, H. (1996). *Columbia object image library (Coil-20)* (Technical Report CUCS-005-96). New York: Columbia University.

Nie, F., Wang, X., & Huang, H. (2014). Clustering and projected clustering with adaptive neighbors. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (pp. 977–986). New York: ACM.

Niyogi, X. (2004). Locality preserving projections. In S. Thrun, L. K. Saul, & B. Schölkopf (Eds.), *Advances in neural information processing systems*, *16*. Cambridge, MA: MIT Press.

Phillips, P. J., Wechsler, H., Huang, J., & Rauss, P. J. (1998). The FERET database and evaluation procedure for face-recognition algorithms. *Image and Vision Computing*, *16*, 295–306.

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

Samaria, F. S., & Harter, A. C. (1994). Parameterisation of a stochastic model for human face identification. In *Proceedings of the Second IEEE Workshop on Applications of Computer Vision* (pp. 138–142). Washington, DC: IEEE Computer Society.

Song, J., Gao, L., Puscas, M. M., Nie, F., Shen, F., & Sebe, N. (2016). Joint graph learning and video segmentation via multiple cues and topology calibration. In *Proceedings of the 2016 ACM on Multimedia Conference* (pp. 831–840). New York: ACM.

Turk, M. A., & Pentland, A. P. (1991). Face recognition using eigenfaces. In *Proceedings of the Conference on Computer Vision and Pattern Recogintion* (pp. 586–591). Washington, DC: IEEE Computer Society.

Wang, J., Zhang, T., Song, J., Sebe, N., & Shen, H. (2016). *A survey on learning to hash.* arXiv:1606.00185.

Welling, M. (2005). *Fisher linear discriminant analysis*. Toronto: Department of Computer Science, University of Toronto.

Woraratpanya, K., Sornnoi, M., Leelaburanapong, S., Titijaroonroj, T., Varakulsiripunt, R., Kuroki, Y., & Kato, Y. (2015). An improved 2DPCA for face recognition under illumination effects. In *Proceedings of the International Conference on Information Technology and Electrical Engineering* (pp. 448–452). Piscatway, NJ: IEEE.

Yamada, M., Jitkrittum, W., Sigal, L., Xing, E. P., & Sugiyama, M. (2014). High-dimensional feature selection by feature-wise kernelized lasso. *Neural Computation*, *26*(1), 185–207.

Yang, J., Zhang, D., Frangi, A. F., & Yu Yang, J. (2004). Two-dimensional PCA: A new approach to appearance-based face representation and recognition. *IEEE Trans. Pattern Anal. Mach. Intelligence*, *26*(1), 131–137.

Zelnik-Manor, L., & Perona, P. (2005). Self-tuning spectral clustering. In L. K. Saul, Y. Weiss, & L. Bottou (Eds.), *Advances in neural information processing systems* (pp. 1601–1608). Cambridge, MA: MIT Press.

Zhang, D., and Zhou, Z.-H. (2005). $(2d)^2$PCA: Two-directional two-dimensional PCA for efficient face representation and recognition. *Neurocomputing*, *69*(1–3), 224–231.

Zhao, Z., Wang, L., Liu, H., & Ye, J. (2013). On similarity preserving feature selection. *IEEE Transactions on Knowledge and Data Engineering*, *25*(3), 619–632.

Received November 15, 2016; accepted December 19, 2016.