Globally and Locally Consistent Unsupervised Projection

Hua Wang[†], Feiping Nie[‡], Heng Huang[‡*]
[†]Department of Electrical Engineering and Computer Science
Colorado School of Mines, Golden, Colorado 80401, USA
[‡]Department of Computer Science and Engineering
University of Texas at Arlington, Arlington, Texas 76019, USA
huawangcs@gmail.com, feipingnie@gmail.com, heng@uta.edu

Abstract

In this paper, we propose an unsupervised projection method for feature extraction to preserve both global and local consistencies of the input data in the projected space. Traditional unsupervised feature extraction methods, such as principal component analysis (PCA) and locality preserving projections (LPP), can only explore either the global or local geometric structures of the input data, but not the both at the same time. In our new method, we introduce a new measurement using the neighborhood data variances to assess the data locality, by which we propose to learn an optimal projection by rewarding both the global and local structures of the input data. The formulated optimization problem is challenging to solve, because it ends up a trace ratio minimization problem. In this paper, as an important theoretical contribution, we propose a simple yet efficient optimization algorithm to solve the trace ratio problem with theoretically proved convergence. Extensive experiments have been performed on six benchmark data sets, where the promising results validate the proposed method.

Dimensionality reduction is an important technique in statistical learning and pattern recognition, which has been widely applied to solve a variety of machine learning and computer vision problems, such as face recognition (Turk and Pentland 1991), image annotation (Wang, Huang, and Ding 2010b), to name a few. Dimensionality reduction algorithms usually seek to represent the input data in their lower-dimensional "intrinsic" subspace/sub-manifold, in which irrelevant features are pruned and inherent data structures are more lucid.

In the early ages, under the assumption that the input data objects are homogeneous but not relational, dimensionality reduction algorithms were often devised to be linear. For example, Principal Component Analysis (PCA) (Jolliffe 2002) attempts to maximize the covariance among the input data points, and Linear Discriminant Analysis (LDA) (Fukunaga 1990) aims at maximizing the class separability. In recent years, manifold learning motivates many nonlinear dimensionality reduction algorithms using pairwise similarities between data objects, either computed from data attributes or directly obtained from experimental observations, which are nonlinear. Successful attempts include ISOMAP (Tenenbaum, Silva, and Langford 2000), Locally Linear Embedding (LLE) (Roweis and Saul 2000), Laplacian Eigenmap (Belkin and Niyogi 2002) and Locality Preserving Projection (LPP) (He and Niyogi 2004), *etc*. These algorithms generally assume that the observed data are sampled from an underlying sub-manifold which are embedded in a high-dimensional observation space. Due to this reasonable assumption, manifold learning based nonlinear projection methods have demonstrated its usefulness in a number of real-world applications.

However, a critical problem in most existing manifold learning techniques often hinders their applications in many real machine learning tasks (Yang et al. 2007). That is, these methods explore the data locality but assume that there exists only one single manifold, which, however, is not true in reality. For example (Yang et al. 2007), although the face images of one individual person could exist on one single manifold, the face images of different persons typically lie on different manifolds. To recognize faces, it would be necessary to distinguish between images from different manifolds. To achieve the optimal recognition result, the recovered embeddings corresponding to different face manifolds should be as separate as possible in the final embedding space, which calls for a new projection method that is able to take into account the global structures of the input data as well. Yang *et al.* (Yang et al. 2007) made a first attempt to propose the unsupervised discriminant projection (UDP) method to incorporate both local and global geometrical information of the input data, which, however, as analyzed in (Deng et al. 2008), was not successful because they failed to recognize that the UDP method is a simplified version of LPP.

To tackle this important unsupervised dimensionality reduction problem, in this paper we propose a novel Globally and Locally consistent Unsupervised Projection (GLUP) method for feature extraction by rewarding both the global and local consistencies of the input data in the projected space, such that both the global and local geometrical structures of the original data can be utilized. Our new method is interesting from a number of perspectives as follows.

- We propose a new unsupervised learning objective for di-

[*]To whom all correspondence should be addressed. This work was partially supported by US NSF IIS-1117965, IIS-1302675, IIS-1344152.

mensionality reduction that takes into account both the global and local consistencies of the input data in the projected space.

- Instead of using the graph Laplacian to capture the data locality as in many existing methods (Belkin and Niyogi 2002; He and Niyogi 2004; Yang et al. 2007), we propose a new measurement for data locality by assessing the overall local data variance. Because this new measurement only uses the neighborhood structures of the input data, it involves less parameter, which makes it more stable and suitable for practical use.
- Despite its clear intuition, our objective is challenging to solve, because it ends up a trace ratio minimization problem. As an important theoretical contribution of this work, we propose a simple yet effective optimization framework to solve a general simultaneous minimization and maximization problem, by which we derive an efficient iterative solution algorithm to our objective with rigorously proved convergence.
- We performed extensive experiments to evaluate a variety of aspects of the proposed projection method under the unsupervised settings, in which our new method outperforms other state-of-the-art unsupervised dimensionality reductions methods. The promising experimental results in our empirical studies are consistent with our theoretical analysis and validates the proposed methods.

Learning Globally and Locally Consistent Unsupervised Projection

In this section, we will systematically develop the objective of the proposed GLUP method, where our goal is to learn a projection that is both globally and locally consistent with the input data under the unsupervised setting.

Throughout this paper, we will write matrices as bold uppercase letters and vectors as bold lowercase letters. Given a matrix $\mathbf{M} = [m_{ij}]$, its i-th column is denoted as and $\mathbf{m}_i$. The Frobenius norm of the matrix $\mathbf{M}$ is denoted as $\|\mathbf{M}\|_{\mathrm{F}}$, and the trace of $\mathbf{M}$ is defined as $\mathbf{tr}(\mathbf{M}) = \sum_i m_{ii}$.

Learning Global and Local Consistencies for Unsupervised Data

In this subsection, we will gradually develop the proposed GLUP objective for projection on unsupervised data.

Learning global consistency via PCA. In the setting of unsupervised learning, we are given a set of n data points $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \Re^{d \times n}$, without knowing their associated cluster memberships. Under the setting of unsupervised learning, clustering partitions the input data into a number of c groups (clusters), such that the data points in the same group are similar while those in different groups are dissimilar. Traditional clustering methods, such as K-means clustering, usually work well when the dimensionality of the input data is not very high. However, when the dimensionality is growing, these clustering methods, as well as other unsupervised or supervised learning methods, will fail due to the "curse of dimensionality" and intractable computational cost. As a result, learning a subspace with lower dimensionality while maintaining the original geometrical structures of the input data is desired for practical applications. To achieve this goal, PCA is the right tool that aims at preserving as much information as possible by learning a projection $\mathbf{W} \in \Re^{d \times r}$ from the input data $\mathbf{X}$, which maps $\mathbf{x}_i$ in the high d-dimensional space to a vector $\mathbf{y}_i$ in a lower r-dimensional space by computing $\mathbf{y}_i = \mathbf{W}^T \mathbf{x}_i$ where $r < p$, such that the overall variance of the input data in the projected spaced $\Re^r$ is maximized.

Formally, denote the global mean vector of the input data $\mathbf{X}$ as:

$$\mathbf{m}_0 = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \ , \tag{1}$$

we can compute its covariance matrix $\mathbf{S}_G$ as following:

$$\begin{aligned} \mathbf{S}_G &= \sum_{i=1}^{n} (\mathbf{x}_i - \mathbf{m}_0)(\mathbf{x}_i - \mathbf{m}_0)^T \\ &= \mathbf{X} \left(\mathbf{I} - \frac{\mathbf{e}\mathbf{e}^T}{n} \right) \mathbf{X}^T \ , \end{aligned} \tag{2}$$

where the constant factor $\frac{1}{n}$ is removed for brevity. Here, without ambiguity, $\mathbf{I}$ denotes the identity matrix with proper size and $\mathbf{e}$ denotes a constant vector with all entries to be 1 with proper length. Therefore, $\left(\mathbf{I} - \frac{\mathbf{e}\mathbf{e}^T}{n} \right)$ is the centering matrix, which is idempotent. Then PCA seeks the projection $\mathbf{W}$ by maximizing the following objective:

$$\begin{aligned} J_{\text{Global}}(\mathbf{W}) &= \mathbf{tr}\left(\mathbf{W}^T \mathbf{S}_G \mathbf{W} \right) , \\ s.t. \quad & \mathbf{W}^T \mathbf{W} = \mathbf{I} \ , \end{aligned} \tag{3}$$

which can be solved by picking up the eigenvectors of $\mathbf{S}_G$ corresponding to the r largest eigenvalues. Because $\mathbf{S}_G$ maximizes the global variance of the input data in the projected space, the learned projection $\mathbf{W}$ by PCA is globally consistent with respect to the input data.

Learning local consistency via neighborhood variances. Besides taking into account the global variances of an input data set, we further consider its local geometric structures in the projected space. Ideally, if the input data are partitioned into a number of clusters and different clusters are clearly separated in the projected space, nearby data points should belong to the same cluster whilst distant data points should belong to different clusters. Namely, in contrast to maximizing the projected global variances, we want to minimize the local variances in the projected space as much as possible. Mathematically, denote the K-nearest neighbors of $\mathbf{x}_i$ as $\mathcal{N}_i$ and the local mean vector of $\mathbf{x}_i$ as

$$\mathbf{m}_i = \frac{1}{K+1} \sum_{\mathbf{x}_j \in \mathcal{N}_i \cup \{\mathbf{x}_i\}} \mathbf{x}_j \ , \tag{4}$$

we can compute the overall local variances as following:

$$\mathbf{S}_L = \sum_{i=1}^{n} \mathbf{S}_{Li} \ , \tag{5}$$

where

$$\mathbf{S}_{Li} = \sum_{\mathbf{x}_j \in \mathcal{N}_i \cup \{\mathbf{x}_i\}} (\mathbf{x}_j - \mathbf{m}_i)(\mathbf{x}_j - \mathbf{m}_i)^T = \mathbf{X}_i \left(\mathbf{I} - \frac{\mathbf{e}\mathbf{e}^T}{K+1} \right) \mathbf{X}_i^T \ , \tag{6}$$

where $\mathbf{X}_i \in \Re^{d \times (K+1)}$ consists of $\mathbf{x}_i$ and the vectors in $\mathcal{N}_i$ as its columns. Here, the constant factor $\frac{1}{K+1}$ is omitted for brevity. Therefore, we can minimize following objective to achieve the local consistency in the projected space:

$$J_{\text{Local}}(\mathbf{W}) = \textbf{tr}\left(\mathbf{W}^T \mathbf{S}_L \mathbf{W}\right), \quad s.t.\ \mathbf{W}^T\mathbf{W} = \mathbf{I} \ . \tag{7}$$

Note that, in Eqs. (5–7) we did not use the LPP objective (He and Niyogi 2004) to capture the localities of the input data due to its notorious performance sensitivity with respect to the parameters. LPP has two parameters. The first one is the graph construction parameter, which is K when constructing a K-nearest neighbor graph (He and Niyogi 2004) or ϵ when constructing an ϵ-ball graph (He and Niyogi 2004). This parameter acts the same as the parameter K in our method when we compute the local variances in Eqs. (5–6). The second parameter is σ, when we use a Gaussian kernel to compute the affinity between a pair of vertices on the input graph. It is generally recognized that (Nie et al. 2010) the quality of the learned projection for subsequent learning tasks is very sensitive to the parameter σ. This weakness can be remedied in supervised and semi-supervised learning tasks via cross-validation, which, however, is not generally feasible in unsupervised learning tasks because of the unavailability of data labels. As a result, our new objective in Eq. (7) is more advantageous in that it has less parameter and easier to fine tune, which, as the first contribution of this work, is obviously of crucial importance in practical unsupervised learning tasks.

Objective to learn the global and local consistencies. Armed with the objectives that can capture the global and local consistencies of an input data set separately, we can build a combined objective to capture both of them simultaneously. Among several possible ways to combine the two objectives, we choose to formulate our new objective using the ratio of trace method (Jia, Nie, and Zhang 2009), which minimizes the following objective:

$$J_0(\mathbf{W}) = \frac{\textbf{tr}\left(\mathbf{W}^T \mathbf{S}_L \mathbf{W}\right)}{\textbf{tr}\left(\mathbf{W}^T \mathbf{S}_G \mathbf{W}\right)}, \quad s.t.\ \mathbf{W}^T\mathbf{W} = \mathbf{I} \ . \tag{8}$$

Compared to the trace difference method (Wang, Huang, and Ding 2010a), the ratio of trace method has less parameters; compared to the trace of ratio method (Jia, Nie, and Zhang 2009), the ratio of trace method has better interpretability and learning performance.

Upon solution to the optimization problem in Eq. (8), the learned projection $\mathbf{W}$ not only preserves the global variance of the input data set but also rewards its local structures, which thereby is both globally and locally consistent. We call Eq. (8) as the proposed Globally and Locally consistent Unsupervised Projection (GLUP) method.

An Efficient Algorithm to Solve the General Trace Ratio Problem

Despite its clear intuition, Eq. (8) is difficult to solve, which ends up a trace ratio minimization problem (Jia, Nie, and Zhang 2009). Recently, several successful attempts have been made to solve this challenging problem (Guo et al. 2003; Wang et al. 2007; Jia, Nie, and Zhang 2009). Motivated by these prior works, as one of the most important contribution of this paper, we will derive an efficient algorithm to solve the general trace ratio problem.

Useful Theorems

Before we proceed to deriving the solution algorithm to our objective in Eq. (8), we first prove the following useful theorems.

Theorem 1 *The global solution of the following general optimization problem:*

$$\min_{\mathbf{v} \in \mathcal{C}} \frac{f(\mathbf{v})}{g(\mathbf{v})} \ , \quad \textit{where } g(\mathbf{v}) \geq 0 \ (\forall\ \mathbf{v} \in \mathcal{C}) \ , \tag{9}$$

is given by the root of the following function:

$$h(\lambda) = \min_{\mathbf{v} \in \mathcal{C}} f(\mathbf{v}) - \lambda g(\mathbf{v}) \ . \tag{10}$$

Proof. Suppose $\mathbf{v}^*$ is the global solution of the problem in Eq. (9), and λ^* is the corresponding global minimal objective value, the following holds

$$\frac{f(\mathbf{v}^*)}{g(\mathbf{v}^*)} = \lambda^* \ . \tag{11}$$

Thus $\forall\ \mathbf{v} \in \mathcal{C}$, we can derive:

$$\frac{f(\mathbf{v})}{g(\mathbf{v})} \geq \lambda^* \ . \tag{12}$$

Because we know that $g(\mathbf{v}) \geq 0$, we have:

$$f(\mathbf{v}) - \lambda^* g(\mathbf{v}) \geq 0 \ , \tag{13}$$

which means:

$$\min_{\mathbf{v}\ \in \mathcal{C}} f(\mathbf{v}) - \lambda^* g(\mathbf{v}) = 0 \quad \Longleftrightarrow \quad h(\lambda^*) = 0 \ . \tag{14}$$

That is, the global minimal objective value λ^* of the problem in Eq. (9) is the root of the function $h(\lambda)$, which completes the proof of Theorem 1. ■

Theorem 2 *Algorithm 1 decreases the objective value of the problem in Eq.* (9) *in each iteration till converges.*

Proof. In Algorithm 1, from step 1 we know that

$$f(\mathbf{v}_t) - \lambda_t g(\mathbf{v}_t) = 0 \ . \tag{15}$$

According to step 2, we know that

$$f(\mathbf{v}_{t+1}) - \lambda_t g(\mathbf{v}_{t+1}) \leq f(\mathbf{v}_t) - \lambda_t g(\mathbf{v}_t) \ . \tag{16}$$

Combining the above two inequalities, we have:

$$f(\mathbf{v}_{t+1}) - \lambda_t g(\mathbf{v}_{t+1}) \leq 0 \ , \tag{17}$$

which indicates

$$\frac{f(\mathbf{v}_{t+1})}{g(\mathbf{v}_{t+1})} \leq \lambda_t = \frac{f(\mathbf{v}_t)}{g(\mathbf{v}_t)} \ . \tag{18}$$

That is, Algorithm 1 decreases the objective value of Eq. (9) in each iteration, which completes the proof of Theorem 2. ■

Algorithm 1: The algorithm to solve Eq. (9).

```
$t = 1$. Initialize $\mathbf{v}_t \in \mathcal{C}$.
while not converge do
    1. Calculate $\lambda_t = \frac{f(\mathbf{v}_t)}{g(\mathbf{v}_t)}$.
    2. Calculate
        $\mathbf{v}_{t+1} = \arg\min_{\mathbf{v} \in \mathcal{C}} f(\mathbf{v}) - \lambda_t g(\mathbf{v})$ .    (19)
    3. $t = t + 1$.
```

Theorem 3 *Algorithm 1 is a Newton's method to find the root of the function $h(\lambda)$ in Eq.* (10).

Proof. From step 2 in Algorithm 1 we know that

$$h(\lambda_t) = f(\mathbf{v}_{t+1}) - \lambda_t g(\mathbf{v}_{t+1}) \ . \tag{20}$$

Thus

$$h'(\lambda_t) = -g(\mathbf{v}_{t+1}) \ . \tag{21}$$

In Newton's method, the updated solution should be

$$\begin{aligned} \lambda_{t+1} &= \lambda_t - \frac{h(\lambda_t)}{h'(\lambda_t)} \\ &= \lambda_t - \frac{f(\mathbf{v}_{t+1}) - \lambda_t g(\mathbf{v}_{t+1})}{-g(\mathbf{v}_{t+1})} \\ &= \frac{f(\mathbf{v}_{t+1})}{g(\mathbf{v}_{t+1})} \ , \end{aligned} \tag{22}$$

which is exactly the step 1 in Algorithm 1. That is, Algorithm 1 is a Newton's method to find the root of the function $h(\lambda)$. ■

Theorem 3 indicates that Algorithm 1 converges very fast and the convergence rate is quadratic convergence, *i.e.*, the difference between the current objective value and the optimal objective value is smaller than $\frac{1}{c^{c^t}}$($c > 1$ is a certain constant) at the t-th iteration. Therefore, Algorithm 1 scales well to large data sets in real world learning tasks, which adds to its practical value.

Theorem 1–3 present a complete framework to solve the general optimization problem in Eq. (9), where an efficient iterative algorithm is supplied in Algorithm 1 with rigorously proved convergence and satisfactory computational efficiency. It is worth to noting that, besides applying it to solve the general trace ratio minimization problem as in our objective in Eq. (8), we can also employ this framework to efficiently solve many other more complicated optimization problem in machine learning, such as the simultaneous ℓ_1-norm minimization and maximization problem by which a robust distance metric can be learned (Wang, Nie, and Huang 2014). Therefore, we consider Theorem 1–3 as the most important theoretical contribution of this work.

Derivation of the Solution Algorithm to Eq. (8)

Equipped with the optimization framework of Theorem 1–3, we can derive the solution algorithm to the optimization problem in Eq. (8).

Because the problem in Eq. (8) is a special case of the general optimization problem in Eq. (9), we can derive the solution algorithm to Eq. (8) using Algorithm 1, in which the key step is to solve the problem in step 2. Given a computed λ_t by step 1 of Algorithm 1, according to step 2 of Algorithm 1 we turn to solve the following problem for our target optimization problem in Eq. (8):

$$\begin{aligned} \min \ \mathbf{tr}\left(\mathbf{W}^T \mathbf{S}_L \mathbf{W}\right) - \lambda_t \, \mathbf{tr}\left(\mathbf{W}^T \mathbf{S}_G \mathbf{W}\right) \\ s.t. \ \mathbf{W}^T \mathbf{W} = \mathbf{I} \ , \end{aligned} \tag{23}$$

which is known to have optimal solution with eigenvalue decomposition of $\mathbf{S}_L - \lambda_t \mathbf{S}_G$.

Finally, the whole algorithm to solve our objective in Eq. (8) is summarized in Algorithm 2. Obviously, Algorithm 2 is guaranteed to converge due to Theorem 2 and converge fast due to Theorem 3.

Algorithm 2: An efficient iterative algorithm to solve the general trace ratio minimization problem in Eq. (8).

```
Input: Matrices $\mathbf{S}_L$ and $\mathbf{S}_G$.
1. Set $t = 1$ and initialize $\mathbf{W}_t$ by a random guess.
while not converge do
    2. Compute
        $\lambda_t = \frac{\mathbf{tr}\left(\mathbf{W}_t^T \mathbf{S}_L \mathbf{W}_t\right)}{\mathbf{tr}\left(\mathbf{W}_t^T \mathbf{S}_G \mathbf{W}_t\right)}$ .    (24)
    3. Compute the solution of Eq. (23) by the eigenvalue
       decomposition of $\mathbf{S}_L - \lambda_t \mathbf{S}_G$.
    4. $t = t + 1$.
Output: The learned projection matrix $\mathbf{W}$.
```

Experimental Results

In this section, we experimentally study a variety of aspects of the proposed GLUP method in clustering tasks. Our goal is to evaluate the dimensionality reduction capability of the proposed method in unsupervised learning.

Data Sets and Experimental Procedures

Six benchmark data sets are used in our experiments, including:

- two UCI data sets: the **Dermatology** and **Ecoli** data sets;
- one object data set: the **COIL-20** data set;
- one digit and character data sets: the **Binalpha** data set;
- and two face data sets: the **UMIST** and **AR** data sets.

The images in the two face data sets are resized following the standard experimental procedures in computer vision studies. Table 1 summarizes the details of the six experimental data sets used in this study. We use PCA as a preprocessing to remove the null space of all the data sets.

For each data set, we first learn the projections by our method as well as the compared methods. Then we map the data onto the learned subspaces/submanifolds, on which we perform clustering using the K-means clustering algorithm.

Table 1: Description of the experimental data sets.

Data set	Number	Dimension	Class
Dermatology	366	34	6
Ecoli	336	343	8
Coil20	1440	1024	20
BinAlpha	1854	256	10
UMIST	575	644	20
AR	840	768	120

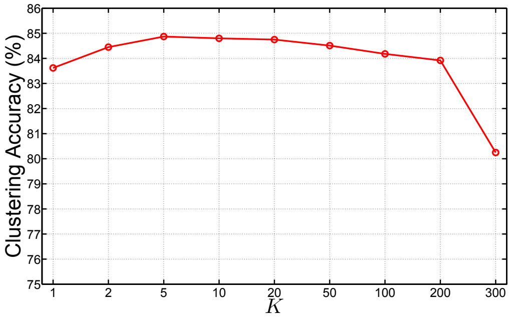

Figure 1: Clustering performance on the Dermatology data set in the projected subspace learned by the proposed GLUP method when the value of K varies.

Study of Parameter K for Local Variance

Before empirically evaluating the clustering performance of the proposed GLUP method, we first study its parameter K. The parameter K controls to which extent we measure the overall local variance defined by $\mathbf{S}_L$ in Eq. (5). When $K = 1$, $\mathbf{S}_{Li}$ computes the covariance over each data point and its nearest neighbor, which is of smallest granularity. Although data variance over small granularity should be more smooth, it also can be dominated by outlier samples. Therefore, a bigger K could potentially lead to a better assessment of the local variances. However, on the other hand, when K grows, the local variances start to approach the global variance. At the extreme case, when $K = n$, the $\mathbf{S}_L$ computed by Eq. (5) is very close to the global variance. As a result, the local property of input data can no longer be captured. Theoretically, an optimal K can be select to be smaller than n_k, where n_k is the number of data points of the smallest data cluster. To verify this hypothesis, we perform clustering on the Dermatology data set by varying the value of K over a large range from 1 to 300. The former is the smallest possible value of K, while the latter is a very big value with respect to the size of the Dermatology data set, because the number of data point of this data set is 366. We repeat the experiments for 100 times to alleviate the impact of the random initialization of both our iterative algorithm and the K-means clustering method. The average clustering performances for different values of K measured by clustering accuracy are reported in Figure 1.

A first glance at the results in Figure 1 shows that the clustering performance of the proposed GLUP method is very stable in a considerably large parameter selection range of K, which makes the parameter fine tuning of our method is easy and adds to its value for practical use. By a more careful examination on the results shown in Figure 1, we can see that the clustering performance on the Dermatology data set reaches its maximum when K is selected in the range of $[5, 100]$, which is the proximal to $366/6 = 61$, *i.e.*, the average number of data points in each cluster. This observation clearly justifies our hypothesis, which is also consistent with our theoretical analysis as discussed above.

Based upon the above observations, in practice we empirically set $K = 30$ for simplicity, because we usually do not know the number of data points in each clusters of an input data set in a priori.

Clustering Performance of the Proposed GLUP Method

Now we experimentally evaluate the dimensionality reduction capability of the proposed method under the setting of unsupervised learning. Besides comparing our new method to its most closely related method, *i.e.*, the UDP method (Yang et al. 2007), we also compare our method against the Kernel Laplacian Embedding (KLE) method (Wang, Huang, and Ding 2010b), which is one of the most recent unsupervised learning method by integrating attribute data and pairwise similarity data, and has demonstrated state-of-the-art dimensionality reduction performance. In addition, we also report the clustering results in the PCA subspace (Jolliffe 2002) and LPP subspace (He and Niyogi 2004) as baselines. For KLE method and LPP method, we construct nearest-neighbor graph for each data set and set the neighborhood size for graph construction as 10 following (He and Niyogi 2004). The reduced dimension of all the compared data are searched in the range of $[1, d/2]$ and we report the best clustering performance, where d is dimensionality of the original experimental data set. Note that, we do not compare our new method to supervised dimensionality reduction methods, because our method is designed for unsupervised settings.

Because the results of the K-means clustering algorithm depend on the initialization, to reduce the statistical variety, we independently repeat the clustering procedures on the projected subspaces learned by all compared methods for 100 times with random initializations, and then we report the results corresponding to the best objective values.

The clustering performance measured by clustering accuracy and normalized mutual information (NMI) are reported in Table 2 and Table 3 respectively. From these experimental results we can see that the proposed method outperforms the compared methods with obvious margins, which demonstrate its effectiveness in the task of data clustering.

Conclusions

In this paper, we proposed an unsupervised projection method for feature extraction to maintain the both global and local consistencies of the input data. Different from traditional unsupervised feature extraction methods such as Principal Component Analysis (PCA) and Locality Preserving Projections (LPP) that can only explore either the global or local geometric structure in the data, our new method learns an optimal projection to maximize the global covariance

Table 2: Comparison of clustering performances measured by clustering accuracy

DATA SET	PCA	LPP	UDP	KLE	GLUP
DERMATOLOGY	78.64%	76.86%	73.58%	83.98%	**85.78%**
ECOLI	65.83%	64.91%	63.88%	45.24%	**67.51%**
COIL20	66.22%	69.45%	63.81%	79.98%	**81.61%**
BINALPHA	44.45%	48.34%	48.48%	47.97%	**49.81%**
UMIST	47.69%	47.87%	49.89%	62.15%	**66.26%**
AR	28.45%	30.19%	25.57%	39.24%	**40.12%**

Table 3: Comparison of clustering performances measured by NMI.

DATA SET	PCA	LPP	UDP	KLE	GLUP
DERMATOLOGY	88.24%	88.01%	87.92%	85.34%	**89.26%**
ECOLI	50.84%	56.97%	53.86%	37.54%	**59.74%**
COIL20	79.15%	79.23%	77.54%	88.31%	**89.91%**
BINALPHA	60.14%	61.24%	61.17%	59.02%	**62.88%**
UMIST	67.58%	67.94%	66.33%	78.64%	**77.61%**
AR	63.64%	65.03%	60.84%	73.19%	**74.38%**

matrix and minimize the newly proposed local covariance matrices simultaneously. To solve the formulated objective which is a trace ratio minimization problem, we presented a simple yet effective optimization method to solve this problem, whose convergence is rigorously guaranteed. Promising experimental results on six benchmark data sets with a variety of experimental settings have demonstrated the effectiveness of the proposed method.

References

Belkin, M., and Niyogi, P. 2002. Laplacian eigenmaps and spectral techniques for embedding and clustering. *NIPS'02*.

Deng, W.; Hu, J.; Guo, J.; Zhang, H.; and Zhang, C. 2008. Comments on globally maximizing, locally minimizing: unsupervised discriminant projection with application to face and palm biometrics. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 30(8):1503–1504.

Fukunaga, K. 1990. *Introduction to statistical pattern recognition*. Academic Press.

Guo, Y.-F.; Li, S.-J.; Yang, J.-Y.; Shu, T.-T.; and Wu, L.-D. 2003. A generalized foley-sammon transform based on generalized fisher discriminant criterion and its application to face recognition. *Pattern Recognition Letter* 24(1-3):147–158.

He, X., and Niyogi, P. 2004. Locality preserving projections. In *NIPS*.

Jia, Y.; Nie, F.; and Zhang, C. 2009. Trace ratio problem revisited. *IEEE Transactions on Neural Networks* 20(4):729–735.

Jolliffe, I. 2002. *Principal component analysis*. Springer.

Nie, F.; Xu, D.; Tsang, I. W.-H.; and Zhang, C. 2010. Flexible manifold embedding: A framework for semi-supervised and unsupervised dimension reduction. *Image Processing, IEEE Transactions on* 19(7):1921–1932.

Roweis, S., and Saul, L. 2000. Nonlinear dimensionality reduction by locally linear embedding. *Science*.

Tenenbaum, J.; Silva, V.; and Langford, J. 2000. A global geometric framework for nonlinear dimensionality reduction. *Science*.

Turk, M., and Pentland, A. 1991. Face recognition using eigenfaces. In *The Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (IEEE CVPR)*, 586–591.

Wang, H.; Yan, S.; Xu, D.; Tang, X.; and Huang, T. S. 2007. Trace ratio vs. ratio trace for dimensionality reduction. In *The Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (IEEE CVPR)*.

Wang, H.; Huang, H.; and Ding, C. 2010a. Discriminant laplacian embedding. In *The Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI 2010)*.

Wang, H.; Huang, H.; and Ding, C. 2010b. Multi-label feature transform for image classifications. In *The Proceedings of The 11th European Conference on Computer Vision (ECCV)*, 793–806.

Wang, H.; Nie, F.; and Huang, H. 2014. Robust distance metric learning via simultaneous ℓ_1-norm minimization and maximization. In *The Proceedings of The 31st International Conference on Machine Learning (ICML 2014)*.

Yang, J.; Zhang, D.; Yang, J.; and Niu, B. 2007. Globally maximizing, locally minimizing: unsupervised discriminant projection with applications to face and palm biometrics. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29(4):650–664.