Learning Robust Locality Preserving Projection via p-Order Minimization

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

Locality preserving projection (LPP) is an effective dimensionality reduction method based on manifold learning, which is defined over the graph weighted squared ℓ_2-norm distances in the projected subspace. Since squared ℓ_2-norm distance is prone to outliers, it is desirable to develop a robust LPP method. In this paper, motivated by existing studies that improve the robustness of statistical learning models via ℓ_1-norm or not-squared ℓ_2-norm formulations, we propose a robust LPP (rLPP) formulation to minimize the p-th order of the ℓ_2-norm distances, which can better tolerate large outlying data samples because it suppress the introduced biased more than the ℓ_1-norm or not squared ℓ_2-norm minimizations. However, solving the formulated objective is very challenging because it not only non-smooth but also non-convex. As an important theoretical contribution of this work, we systematically derive an efficient iterative algorithm to solve the general p-th order ℓ_2-norm minimization problem, which, to the best of our knowledge, is solved for the first time in literature. Extensive empirical evaluations on the proposed rLPP method have been performed, in which our new method outperforms the related state-of-the-art methods in a variety of experimental settings and demonstrate its effectiveness in seeking better subspaces on both noiseless and noisy data.

Introduction

Dimensionality reduction (DR) algorithms 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, DR algorithms assume that the input data objects are homogeneous but not relational, and thereby are devised to deal with a set of attributes in the format of fixed length vectors. For example, Principal Component Analysis (PCA) (Jolliffe 2005) attempts to maximize the covariance among data points, while Linear Discriminant Analysis (LDA) (Fukunaga 1990) aims at maximizing the class separability. In recent years, manifold learning has motivated many DR algorithms using pairwise similarities between data objects, either computed from data attributes or directly obtained from experimental observations. These algorithms, such as ISOMAP (Tenenbaum, De Silva, and Langford 2000), Locally Linear Embedding (LLE) (Roweis and Saul 2000), Laplacian Eigenmap (Belkin and Niyogi 2001), Locality Preserving Projection (LPP) (Niyogi 2004), and discriminant Laplacian embedding (Wang, Huang, and Ding 2010b), *etc.*, generally assume that the observed data are sampled from an underlying sub-manifold which is embedded in a high-dimensional observation space. Due to this reasonable assumption, manifold based learning methods have achieved great success to solve problems in a large number of real-world applications. Among these manifold learning based projection methods, LPP has been widely accepted because it is able to learn a linear projection from the original data, such that it can be easily applied to not only training data but also out-of-sample data points.

In this paper, we address the issue of robustness of LPP in the presence of outlier samples, which is defined as the data points that deviates significantly from the rest majority of the data points. In classical LPP approach, the learning objective is formulated using the squared distance in the projected subspace, which, same as other least square based learning objectives in statistical learning, could be significantly influenced by outlying observations. That is, the traditional LPP formulation becomes inappropriate at contaminated data sets, because large errors squared dominate the sum. Many previous works have been done to improve the robustness of the linear dimensionality reduction methods via using the ℓ_1-norm minimizations or not-squared ℓ_2-norm minimizations (Baccini, Besse, and Falguerolles 1996; Ding et al. 2006; Gao 2008; Ke and Kanade 2005; Kwak 2008; Wright et al. 2009; Nie et al. 2010; 2011; Wang, Nie, and Huang 2014). The key motivation of these prior works lies in that the not-squared ℓ_2-norm distance or error functions, say $\|x_i - x_j\|_2$ for the two vectors x_i and x_j, can generally better tolerate the biases caused by the outlying data samples, especially when the outlier data samples are far away from the normal data distributions. Following the same intuition, we recognize that the distance with lower orders, *e.g.*, $\|x_i - x_j\|_2^p$ where $0 < p < 1$, can achieve the same goal for robustness with potentially better results (Wang et al. 2013), because the bias caused by outlying data samples can be further suppressed when $p(< 1)$ is selected

[*] To whom all correspondence should be addressed. This work was partially supported by NSF-IIS 1117965, NSF-IIS 1302675, NSF-IIS 1344152, NSF-DBI 1356628, NSF-IIS 1423591.

as a very small number. Based on this recognition, we propose a new LPP objective using the p-th order distance, and call the resulted objective as the p-order minimization problem. Because the learned projection by our new method is robust against outlier data samples, we call it robust locality preserving projection (rLPP) method, which is interesting from a number of perspectives as following.

- Our new objective is defined over the p-th ($0 < p \leq 2$) order of the ℓ_2-norm distance, which is more general and makes the traditional LPP a special case of our new method when $p = 2$.
- Same as other ℓ_1-norm or not-squared ℓ_2-norm based learning objectives, our new method is robust against outlier data samples, which is particularly true when $p < 1$. The smaller the value of p is, the better robustness our new method can achieve.
- Despite its clear intuitions and nice theoretical properties, the resulted objective of the proposed rLPP method is difficult to solve, because it is not only non-smooth but also non-convex. To solve the problem, we propose an efficient iterative solution algorithm, whose convergence is rigourously proved.
- Extensive empirical studies have been performed to evaluate a variety aspects of the proposed rLPP method, which clearly demonstrate the effectiveness of the proposed method on not only noiseless data sets but also noisy data sets with outlier samples, especially for the latter case.

Motivation of the Proposed Problem

Suppose we have n data points $\{x_1, \cdots, x_n \in \mathbb{R}^{d\times 1}\}$, we construct a graph using the data with the similarity matrix $S \in \mathbb{R}^{n\times n}$. The Laplacian matrix L is defined as $L = D - S$, where D is a diagonal matrix with the i-th diagonal element as $\sum_j S_{ij}$.

Recently, Locality Preserving Projection (LPP) and its variants have been successfully applied for dimensionality reduction. The projection matrix $W \in \mathbb{R}^{d\times m}(m < d)$ in LPP is obtained by solving the following problem[1]:

$$\min_{W^TXDX^TW=I} \sum_{i,j=1}^{n} S_{ij} \left\|W^Tx_i - W^Tx_j\right\|_2^2. \quad (1)$$

The objective in Eq.(1) can be written as $tr(W^TXLX^TW)$. Thus the optimal solution W to problem (1) is the m eigenvectors of $(W^TXDX^TW)^{-1}W^TXLX^TW$ corresponding to the m smallest eigenvalues.

The basic idea of LPP is to preserve the neighborhood relationship between data points. Specifically, as shown in Eq.(1), LPP tries to find a embedded subspace such that the distances of data pairs which are neighbors in the original space are minimized.

The squared distances used in Eq.(1) do not tolerate large value of distance, thus makes the distances in the embedded subspace tend to be even, *i.e.*, not too large but also not too small. Therefore, the squared distances used in LPP would makes the method can not find the optimal subspace such that most of the distances of local data pairs are minimized but a few of them are large. In this paper, we propose to solve the following problem to find the optimal subspace:

$$\min_{W^TXDX^TW=I} \sum_{i,j=1}^{n} S_{ij} \left\|W^Tx_i - W^Tx_j\right\|_2^p. \quad (2)$$

where $0 < p \leq 2$. Obviously, LPP is a special case of the proposed new method when $p = 2$. More importantly, by setting $p \leq 1$, the method will focus on minimizing most of the distances of local data pairs.

Although the motivation of Eq. (2) is clear, it is a non-smooth objective and difficult to be solved efficiently. Thus, in the next section, we will introduce an iterative algorithm to solve the problem (2). We will show that the original weight matrix W would be adaptively re-weighted to capture clearer cluster structures after each iteration.

Optimization Algorithm to the Proposed Method

The Lagrangian function of the problem (2) is

$$\begin{aligned}\mathcal{L}(W) = &\sum_{i,j=1}^{n} S_{ij} \left\|W^Tx_i - W^Tx_j\right\|_2^p \\ &-Tr(\Lambda(W^TXDX^TW - I)).\end{aligned} \quad (3)$$

Denote a Laplacian matrix $\tilde{L} = \tilde{D} - \tilde{S}$, where $\tilde{S}$ is a re-weighted weight matrix defined by

$$\tilde{S}_{ij} = \frac{p}{2} S_{ij} \left\|W^Tx_i - W^Tx_j\right\|_2^{p-2}, \quad (4)$$

$\tilde{D}$ is a diagonal matrix with the i-th diagonal element as $\sum_j \tilde{S}_{ij}$. Taking the derivative of $\mathcal{L}(W)$ w.r.t W, and setting the derivative to zero, we have:

$$\frac{\partial \mathcal{L}(W)}{\partial W} = X\tilde{L}X^TW - XDX^TW\Lambda = \mathbf{0}, \quad (5)$$

which indicates that the solution W is the eigenvectors of $\left(XDX^T\right)^{-1} X\tilde{L}X^T$. Note that $\left(XDX^T\right)^{-1} X\tilde{L}X^T$ is dependent on W, we propose an iterative algorithm to obtain the solution W such that Eq. (5) is satisfied. The algorithm is guaranteed to converge to a local optimum, which will be proved in the next subsection.

The algorithm is described in Algorithm 1. In each iteration, $\tilde{L}$ is calculated with the current solution W, then the solution W is updated according to the current calculated $\tilde{L}$. The iteration procedure is repeated until converges. From the algorithm we can see, the original weight matrix S is adaptively re-weighted to minimize the objective in Eq. (2) during the iteration.

[1] In practice, to ensure the learned projection W is shift-invariant, D in Eq.(1) should be $L_d = D - D\mathbf{1}\mathbf{1}^TD$, or the training data X must be centered (Nie et al. 2009; Nie, Cai, and Huang 2014)

```
Input: Training data $X \in \mathbb{R}^{d \times n}$. The original weight
       matrix $S \in \mathbb{R}^{n \times n}$. $D$ is a diagonal matrix with
       the $i$-th diagonal element as $\sum_j S_{ij}$.
Initialize $W \in \mathbb{R}^{d \times m}$ such that $W^T X D X^T W = I$ ;
while not converge do
    1. Calculate $\tilde{L}_t = \tilde{D}_t - \tilde{S}_t$, where
    $\tilde{S}_{ij} = \frac{p}{2} S_{ij} \left\| W^T x_i - W^T x_j \right\|_2^{p-2}$, $\tilde{D}$ is a diagonal
    matrix with the $i$-th diagonal element as $\sum_j (\tilde{S})_{ij}$ ;
    2. Update $Q$. The columns of the updated $Q$ are the
    first $m$ eigenvectors of $\left(XDX^T\right)^{-1} X \tilde{L} X^T$
    corresponding to the first $m$ smallest eigenvalues ;
end
Output: $W \in \mathbb{R}^{d \times m}$.
```

Algorithm 1: The algorithm to solve the problem (2).

Convergence Analysis

To prove the convergence of the Algorithm 1, we need the following lemmas:

Lemma 1 *For any scalar x, when $0 < p \leq 2$, we have* $2|x|^p - px^2 + p - 2 \leq 0$.

Proof: Denote $f(x) = 2x^{\frac{p}{2}} - px + p - 2$, then we have

$$f'(x) = p(x^{\frac{p-2}{2}} - 1), \tag{6}$$

and

$$f''(x) = \frac{p(p-2)}{2} x^{\frac{p-4}{2}}. \tag{7}$$

Obviously, when $x > 0$ and $0 < p \leq 2$, then $f''(x) \leq 0$ and $x = 1$ is the only point that $f'(x) = 0$. Note that $f(1) = 0$, thus when $x > 0$ and $0 < p \leq 2$, then $f(x) \leq 0$. Thus $f(x^2) \leq 0$, which indicates $2|x|^p - px^2 + p - 2 \leq 0$. □

Lemma 2 *For any nonzero vectors v, v_0, when $0 < p \leq 2$, the following inequality holds:*

$$\|v\|_2^p - \frac{p}{2} \|v_0\|_2^{p-2} \|v\|_2^2 \leq \|v_0\|_2^p - \frac{p}{2} \|v_0\|_2^{p-2} \|v_0\|_2^2. \tag{8}$$

Proof:

$$\begin{array}{l} 2\left(\frac{\|v\|_2}{\|v_0\|_2}\right)^p - p\left(\frac{\|v\|_2}{\|v_0\|_2}\right)^2 + p - 2 \leq 0 \\ \Rightarrow 2\|v\|_2^p - p\|v_0\|_2^{p-2}\|v\|_2^2 \leq (2-p)\|v_0\|_2^p \\ \Rightarrow \|v\|_2^p - \frac{p}{2}\|v_0\|_2^{p-2}\|v\|_2^2 \leq \|v_0\|_2^p - \frac{p}{2}\|v_0\|_2^{p-2}\|v_0\|_2^2, \end{array}$$

where the first inequality is true according to Lemma 1. □

Now we have the following theorem:

Theorem 1 *The Algorithm 1 will monotonically decrease the objective of the problem (2) in each iteration, and converge to a local optimum of the problem.*

Proof: Suppose the updated W is $\tilde{W}$. According to the step 2 in the Algorithm 1, we know that

$$\begin{array}{l} \tilde{W} = \arg \min\limits_{W^T X D X^T W = I} Tr(W^T X \tilde{L} X^T W) \\ = \arg \min\limits_{W^T X D X^T W = I} \sum\limits_{i,j=1}^{n} \tilde{S}_{ij} \left\| W^T x_i - W^T x_j \right\|_2^2. \end{array} \tag{9}$$

Note that $\tilde{S}_{ij} = \frac{p}{2} S_{ij} \left\| W^T x_i - W^T x_j \right\|_2^{p-2}$, so we have

$$\begin{array}{l} \sum\limits_{i,j=1}^{n} \frac{p}{2} S_{ij} \left\| W^T x_i - W^T x_j \right\|_2^{p-2} \left\| \tilde{W}^T x_i - \tilde{W}^T x_j \right\|_2^2 \\ \leq \sum\limits_{i,j=1}^{n} \frac{p}{2} S_{ij} \left\| W^T x_i - W^T x_j \right\|_2^{p-2} \left\| W^T x_i - W^T x_j \right\|_2^2. \end{array} \tag{10}$$

According to Lemma 2, we have

$$\begin{array}{l} \sum\limits_{i,j=1}^{n} S_{ij} \left\| \tilde{W}^T x_i - \tilde{W}^T x_j \right\|_2^p - \\ \sum\limits_{i,j=1}^{n} \frac{p}{2} S_{ij} \left\| W^T x_i - W^T x_j \right\|_2^{p-2} \left\| \tilde{W}^T x_i - \tilde{W}^T x_j \right\|_2^2 \\ \leq \sum\limits_{i,j=1}^{n} S_{ij} \left\| W^T x_i - W^T x_j \right\|_2^p - \\ \sum\limits_{i,j=1}^{n} \frac{p}{2} S_{ij} \left\| W^T x_i - W^T x_j \right\|_2^{p-2} \left\| \tilde{W}^T x_i - \tilde{W}^T x_j \right\|_2^2. \end{array} \tag{11}$$

Summing Eq. (10) and Eq. (11) in the two sides, we arrive at

$$\begin{array}{l} \sum\limits_{i,j=1}^{n} S_{ij} \left\| \tilde{W}^T x_i - \tilde{W}^T x_j \right\|_2^p \\ \leq \sum\limits_{i,j=1}^{n} S_{ij} \left\| W^T x_i - W^T x_j \right\|_2^p. \end{array} \tag{12}$$

Thus the Algorithm 1 will monotonically decrease the objective of the problem (2) in each iteration t until the algorithm converges. In the convergence, the equality in Eq. (12) holds, thus W and $\tilde{L}$ will satisfy Eq. (5), the KKT condition of problem (2). Therefore, the Algorithm 1 will converge to a local optimum of the problem (2). □

Experimental Results

In this section, we empirically study the proposed robust locality preserving projection (rLPP) method, where our goal is to examine its robustness under the conditions when data outliers are present.

Data Descriptions

We evaluate the proposedmethods on five widely used benchmark data sets in machine learning studies. The data descriptions are summarized in Table 1. The first two data sets are obtained from the UCI machine learning data repository. For the CMU PIE (Face Pose, Illumination, and Expression) face data set, all the face images are resized to 32 × 32 following standard computer vision experimental conventions to reduce the misalignment effects. For the two document data sets, following previous studies, for Reuters21578 data set, we remove the keywords appearing less than 50 times and end up with 1599 features; for TDT2 corpus data set, we remove the keywords appearing less than 100 times, and end up with 3157 features, respectively.

Study of the Parameter of the Proposed Method

The proposed method has only one parameter, *i.e.*, p in Eq. (2), which controls to which extent we suppress the bias introduced by the outlier data samples. Thus, we first evaluate its impacts to the learned projections.

Table 1: Data sets used in our experiments.

Data set	Number	Dimension	Classes
Coil20	1440	1024	20
Vehicle	946	18	4
PIE face	3329	1024	68
Reuters21578	8293	1599	65
TDT2 corpus	9394	3157	30

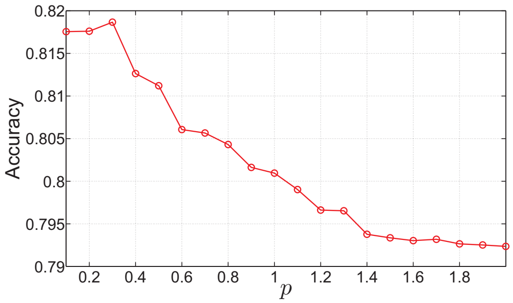


Figure 1: Clustering accuracy in the learned projected spaces by the proposed method *vs.* p on the Coil20 data set.

Experimental setups. We experiment with the Coil20 data set. An important property of the Coil20 data set is that the pictures in the data set were taken repeatedly for one same object from different viewing angles. As a result, the images for the same object indeed reside on an intrinsical manifold. Our goal is to cluster the object images in the projected space learned by the proposed method, by which we examine whether the manifold structures can be discovered such that the clustering performance can be improved. We vary p of the proposed objective in the range of 0.1 to 2 to study its impacts to the clustering performance. We perform the clustering using the K-means clustering method in the projected subspaces, where K is set to the true cluster numbers. The clustering performances with different parameter settings measured by the clustering accuracies are reported in Figure 1. To alleviate the randomness due to the initializations for the K-means clustering method, we repeat the experiment at each parameter setting for 50 times and report the average clustering accuracy in Figure 1.

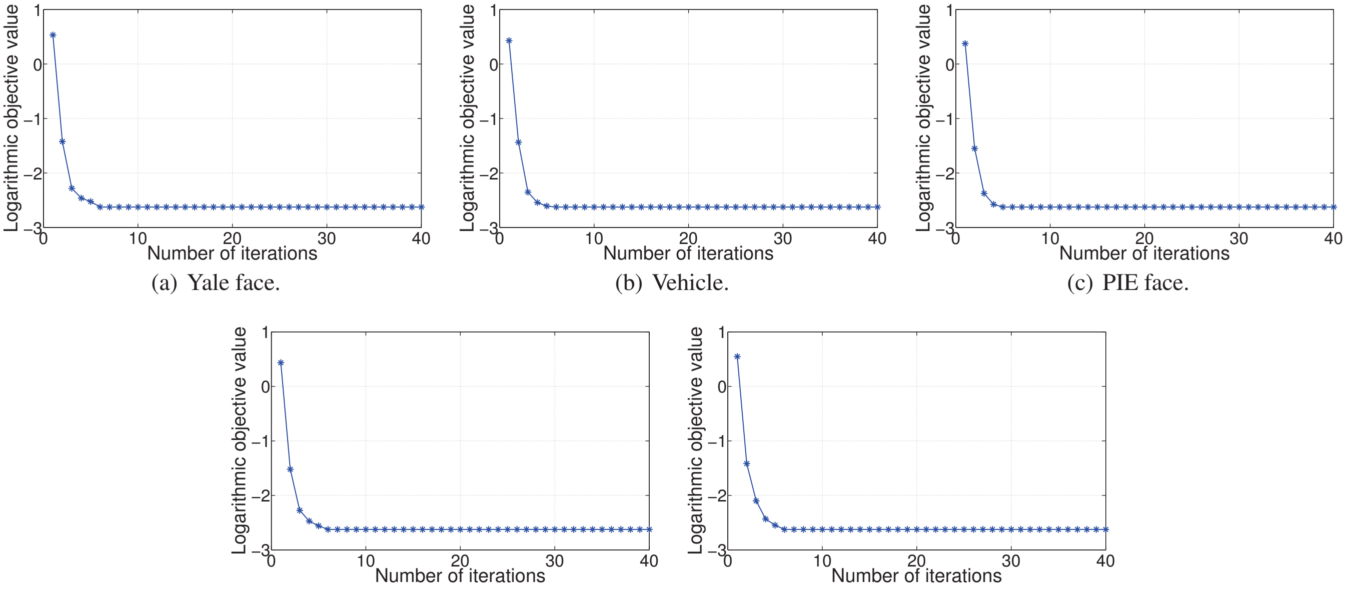
Experimental results. From Figure 1 we can see that, smaller p leads to better clustering accuracy, *i.e.*, the projected subspace learned by our new method with smaller p can better find the intrinsic data structures, which clearly confirms the correctness to use p-order minimization to learn the locality preserved projections. We also notice that when p is very small, *e.g.*, when $p = 0.1$ and $p = 0.2$, the clustering performances is not as good as that when $p = 0.3$. This can be attributed that, when p is too small, the distance measurement will be compromised, *i.e.*, the relative difference between different distances become smaller. In the extreme case, when $p \to 0$, the distance between any data pairs will turn to be the same. Theoretically, p should take a small value to improve the robustness of the proposed objective against outlier data samples; meanwhile p should not be too small to invalidate the distance measurement in the Euclidean space. Upon the results in Figure 1, empirically, we set $p = 0.3$ in all our subsequent experiments, unless otherwise stated.

Convergence Study of the Solution Algorithm

Because the proposed rLPP method employs an iterative solution algorithm, an important issue is its convergence property. We have theoretically proved the convergence of the algorithm, and now we empirically study the convergence property of the proposed iterative algorithm. The objective values of our algorithm on the five data sets in each iteration are plotted in the sub-figures of Figure 2, which show that the objective values of our algorithm keep to decrease along with iterative processes, which is perfectly in accordance with our earlier theoretical analysis. Moreover, the algorithm typically converges to the asymptotic limit within 7 iterations, which demonstrates that our solution algorithm is very computationally efficient. As a result, our new algorithm scales well to large-scale data sets and adds its value for practical use. Upon these experimental results, empirically, we select a stopping threshold of 10^{-5} in all our following experiments, which is sufficient to achieve satisfactory results in terms of convergence.

Experimental Results on Benchmark Noiseless Data

Experimental setups. When we construct the data graph only using data similarity, the proposed rLPP method is an unsupervised method. Thus we compare it against the following unsupervised dimensionality reduction methods: (1) principal component analysis (PCA) (Jolliffe 2005), (2) robust principal component analysis (rPCA) (Wright et al. 2009) (this method is the most recently published robust PCA method with better performance than others (Gao 2008; Ke and Kanade 2005; Ding et al. 2006; Kwak 2008)), (3) locality preserving projections (LPP) (Niyogi 2004) which is the non-robust counterpart of the proposed method, and (4) the Kernel Laplacian Embedding (KLE) method (Wang, Huang, and Ding 2010a). In addition, as a baseline, we also report the clustering results by (5) the K-means clustering method in the original feature space. For PCA, we reduce the dimensionality of the input data such that 90% of data variance is preserved. For rPCA, following (Wright et al. 2009), we set $\lambda = d^{-1/2}$. We empirically select the reduced dimensionality of LPP method to be $c - 1$ where c is the number of clusters of a data set, and use the codes published by the authors (Niyogi 2004). The KLE method 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. We implement the KLE method following its original work (Wang, Huang, and Ding 2010a). For the proposed method, as well as the KLE method and the LPP method, we construct the nearest-neighbor graph for each data set and set the neighborhood size for graph construction as 10 following (Niyogi 2004). Except for rPCA,



(a) Yale face.

(b) Vehicle.

(c) PIE face.

(d) Reuters21578.

(e) TDT2.

Figure 2: Number of iterations vs. the objective value of the proposed rLPP method.

once the projection matrix is obtained by a dimensionality reduction method, K-means clustering method is used to cluster the data points in the projected subspace, where we set K as the true class numbers. Because rPCA method does not produce the projection matrix but the projected data, we use its immediate output for clustering.

For our method, we experiment with two different setting, where we set the parameter p to be 0.3 and 1. Note that, when $p = 2$, the proposed method is exactly the traditional LPP method proposed in (Niyogi 2004); when $p = 1$ or $p = 0.3$, the proposed method is expected to suppress the impacts of the outlier data samples.

Because the results of the K-means clustering algorithm depend on the initialization, to reduce the statistical variety, we independently repeat the clustering procedures in the projected subspaces learned by all compared methods for 50 times with random initializations, and then we report the results corresponding to the best objective values. The clustering performance measured by clustering accuracy are reported in Table 2.

Experimental results. From the experimental results in Table 2 we can see that our method consistently outperforms all other compared methods, which demonstrate the effectiveness of our methods in discovering the inherent manifold structures of the input data and thereby improving the clustering performance. Most importantly, as expected, when p is smaller, the clustering performance of the proposed method is better, which again demonstrate the effectiveness of using p-order minimization in learning LPP subspace.

Robustness Against Outlier Samples

Experimental setups. Because the main advantage of the proposed rLPP method lies in its robustness against outlier data samples, we further evaluate the it on noisy data with outlier samples.

To emulate the outlier samples, given the input data matrix X, we corrupt it by a noise matrix M whose element are i.i.d. standard Gaussian variables. Then we carry out the same procedures as those in the previous subsection for projection learning on $X + \delta M$, where $\delta = nf \frac{\|X\|_F}{\|M\|_F}$ and nf is a given noise factor. We set $nf = 0.1$ in all our studies. We compare our new rLPP method against the same unsupervised dimensionality reduction methods as before and report the clustering results in Table 3.

Experimental results. First, the proposed rLPP method is consistently better than all other compared methods on all five experimental data sets, which demonstrate that our new method is able to effectively learn a subspace to improve the clustering performance on noisy data with outlier data samples. Second, although the improvements by our method over the competing methods on the original data without noise are mediocre as shown in Table 2 in the last subsection, the improvements by our new method on the contaminated data with outlier data samples in this subsection are considerably large. For example, on the Coil20 data set with outlier samples, our new rLPP method improves the clustering accuracy over the baseline PCA method by $42.49\% = (0.721 - 0.506)/0.506$. In contrast, the improvement of clustering accuracy on the same data set by our method over the PCA method under the noiseless condition is only $13.44\% = (0.751 - 0.662)/0.662$. The same observations can be seen on all the other experimental data sets, which show that the proposed method has better capability to learn a more effective subspace for clustering on contaminated data, and confirms its robustness against outlier data

Table 2: Clustering accuracy of the compared methods on the four benchmark data sets without noise.

Data	PCA	rPCA	LPP	KLE	rLPP (p = 1)	rLPP (p = 0.3)
Coil20	0.662	0.659	0.694	0.704	0.749	**0.751**
Vehicle	0.666	0.675	0.709	0.712	0.724	**0.741**
PIE face	0.669	0.676	0.675	0.681	0.694	**0.715**
Reuters21578	0.868	0.883	0.910	0.907	0.921	**0.933**
TDT2 corpus	0.894	0.932	0.947	0.927	0.964	**0.971**

Table 3: Clustering accuracy of the compared methods on the four benchmark data sets with outlier data samples.

Data	PCA	rPCA	LPP	KLE	rLPP (p = 1)	rLPP (p = 0.3)
Coil20	0.506	0.699	0.694	0.691	0.704	**0.721**
Vehicle	0.589	0.633	0.525	0.611	0.673	**0.720**
PIE face	0.591	0.649	0.614	0.611	0.647	**0.689**
Reuters21578	0.801	0.869	0.852	0.841	0.884	**0.912**
TDT2 corpus	0.813	0.918	0.878	0.907	0.923	**0.944**

samples.

Finally, we also notice that when p takes smaller value, the proposed method can achieve better clustering results, which provide one more concrete evidence to support the usefulness of the p-order minimization when learning a LPP subspaces.

In summary, the proposed method is able to achieve better clustering performance on not only noiseless data but also noisy data with outlier samples, which is particularly true for the latter case. These observations are consistent with our motivations to theoretically formulate our new objective in that the p-order minimization can alleviate the negative impacts of outliers data samples.

Conclusions

We proposed a robust LPP method based on the p-th order of ℓ_2-norm distance, which formulated a non-smooth non-convex minimization problem. The new objective imposes the p-th order ℓ_2-norm distance when computing the pair-wise data affinities, which makes the resulted objective very robust against outlier data samples. However, the new objective brings the much more challenging optimization problem than that in traditional LPP. To solve the problem, we introduced an efficient iterative algorithm and provided the rigorous theoretical analysis on the convergence of our algorithm. The new algorithm is easy to be implemented and fast to converge in practice, because we have closed form solution in each iteration. We performed extensive experiments on both noiseless and noisy data, and all results have clearly shown that the proposed method is more effective and robust to outlier samples than traditional methods.

References

Baccini, A.; Besse, P.; and Falguerolles, A. 1996. Al 1-norm pca and a heuristic approach. In *Ordinal and symbolic data analysis*. Springer. 359–368.

Belkin, M., and Niyogi, P. 2001. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *NIPS*, volume 14, 585–591.

Ding, C.; Zhou, D.; He, X.; and Zha, H. 2006. R 1-pca: rotational invariant l 1-norm principal component analysis for robust subspace factorization. In *Proceedings of the 23rd international conference on Machine learning*, 281–288. ACM.

Fukunaga, K. 1990. *Introduction to statistical pattern recognition*. Access Online via Elsevier.

Gao, J. 2008. Robust l1 principal component analysis and its bayesian variational inference. *Neural computation* 20(2):555–572.

Jolliffe, I. 2005. *Principal component analysis*. Wiley Online Library.

Ke, Q., and Kanade, T. 2005. Robust ℓ_1 norm factorization in the presence of outliers and missing data by alternative convex programming. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, 739–746. IEEE.

Kwak, N. 2008. Principal component analysis based on l1-norm maximization. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 30(9):1672–1680.

Nie, F.; Xiang, S.; Song, Y.; and Zhang, C. 2009. Orthogonal locality minimizing globality maximizing projections for feature extraction. *Optical Engineering* 48(1):017202–017202.

Nie, F.; Huang, H.; Cai, X.; and Ding, C. H. 2010. Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization. In *Advances in Neural Information Processing Systems*, 1813–1821.

Nie, F.; Wang, H.; Huang, H.; and Ding, C. 2011. Unsupervised and semi-supervised learning via ℓ_1-norm graph. In *2011 IEEE International Conference on Computer Vision (ICCV 2011)*, 2268–2273. IEEE.

Nie, F.; Cai, X.; and Huang, H. 2014. Flexible shift-invariant locality and globality preserving projections. In *Machine*

Learning and Knowledge Discovery in Databases. Springer. 485–500.

Niyogi, X. 2004. Locality preserving projections. In *Neural information processing systems*, volume 16, 153.

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

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

Wang, H.; Nie, F.; Cai, W.; and Huang, H. 2013. Semi-supervised robust dictionary learning via efficient $\ell_{2,0^+}$-norms minimization. In *2013 IEEE International Conference on Computer Vision (ICCV 2013)*, 1145–1152.

Wang, H.; Huang, H.; and Ding, C. 2010a. Multi-label feature transform for image classifications. In *ECCV 2010*. Springer. 793–806.

Wang, H.; Huang, H.; and Ding, C. H. 2010b. Discriminant laplacian embedding. In *AAAI 2010*.

Wang, H.; Nie, F.; and Huang, H. 2014. Robust distance metric learning via simultaneous l1-norm minimization and maximization. In *Proceedings of the 31st International Conference on Machine Learning (ICML 2014)*, 1836–1844.

Wright, J.; Ganesh, A.; Rao, S.; Peng, Y.; and Ma, Y. 2009. Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization. In *Advances in neural information processing systems*, 2080–2088.