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 $\ell_2$-norm distances in the projected subspace. Since squared $\ell_2$-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 $\ell_1$-norm or not-squared $\ell_2$-norm formulations, we propose a robust LPP (rLPP) formulation to minimize the $p$-th order of the $\ell_2$-norm distances, which can better tolerate large outlying data samples because it suppress the introduced biased more than the $\ell_1$-norm or not squared $\ell_2$-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 $\ell_2$-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 "intrinsric" 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 $\ell_1$-norm minimizations or not-squared $\ell_2$-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 $\ell_2$-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., $||x_i - x_j||_p^p$ 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.
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 $\ell_2$-norm distance, which is more general and makes the traditional LPP a special case of our new method when $p = 2$.
- Same as other $\ell_1$-norm or not-squared $\ell_2$-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 rigorously 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\(^1\):

$$
\min_{W^T X D X^T W = I} \sum_{i,j=1}^{n} S_{ij} \left\| W^T x_i - W^T x_j \right\|_2^2. \tag{1}
$$

The objective in Eq. (1) can be written as $tr((W^T X L X^T W)^2)$. Thus the optimal solution $W$ to problem (1) is the $m$ eigenvectors of $(W^T X D X^T W)^{-1} W^T X L X^T W$ 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 an embedded subspace such that the distances of data pairs which are neighbors in the original space are minimized.

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

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:

$$
\min_{W^T X D X^T W = I} \sum_{i,j=1}^{n} S_{ij} \left\| W^T x_i - W^T x_j \right\|_2^p. \tag{2}
$$

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

$$
\mathcal{L}(W) = -\sum_{i,j=1}^{n} S_{ij} \left\| W^T x_i - W^T x_j \right\|_2^p - Tr(\Lambda W^T X D X^T W - I)). \tag{3}
$$

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

$$
\tilde{S}_{ij} = \frac{p}{2} S_{ij} \left\| W^T x_i - W^T x_j \right\|_2^{p-2}. \tag{4}
$$

$\tilde{D}$ 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:

$$
\frac{\partial \mathcal{L}(W)}{\partial W} = X \tilde{L} X^T W - X D X^T W \Lambda = 0, \tag{5}
$$

which indicates that the solution $W$ is the eigenvectors of $(X D X^T)^{-1} X \tilde{L} X^T$. Note that $(X D X^T)^{-1} X \tilde{L} X^T$ 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.
Obviously, when $f(x)$ is 2 times, we have 
\[ f'(x) = 2x - px + p - 2 \leq 0. \]
\[ f''(x) = \frac{p(p - 2)}{2} x^{p - 1}. \]

Lemma 2 For any nonzero vectors $v, v_0$, when $0 < p \leq 2$, the following inequality holds:
\[ \|v\|_2^2 - \frac{p}{2} \|v_0\|_2^2 - \|v\|_2^2 \leq \|v_0\|_2^2 - \frac{p}{2} \|v_0\|_2^2 - \|v_0\|_2^2 \leq 0. \]
\[ \|v\|_2^2 \leq \frac{p}{2} \|v_0\|_2^2 - \|v_0\|_2^2 - \|v_0\|_2^2 \leq 0. \]

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

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 $f(x) = 2x - px + p - 2$, then we have
\[ f'(x) = px - px + p - 2 \leq 0, \]
and
\[ f''(x) = \frac{p(p - 2)}{2} x^{p - 1}. \]

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 $2|x|^p - px^2 + p - 2 \leq 0$. \qed

Lemma 2 For any nonzero vectors $v, v_0$, when $0 < p \leq 2$, the following inequality holds:
\[ \|v\|_2^2 - \frac{p}{2} \|v_0\|_2^2 \leq \|v_0\|_2^2 - \frac{p}{2} \|v_0\|_2^2 - \|v_0\|_2^2 \leq 0. \]
\[ \|v\|_2^2 \leq \frac{p}{2} \|v_0\|_2^2 - \|v_0\|_2^2 - \|v_0\|_2^2 \leq 0. \]

Proof: Consider the case $p = 2$ and $x = 1$.
\[ 2 \left( \|v\|_2^2 - \|v_0\|_2^2 \right) \leq \|v_0\|_2^2 - \|v_0\|_2^2 - \|v_0\|_2^2 \leq 0. \]
\[ \|v\|_2^2 \leq \frac{p}{2} \|v_0\|_2^2 - \|v_0\|_2^2 - \|v_0\|_2^2 \leq 0. \]

where the first inequality is true according to Lemma 1. \qed

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 $\hat{W}$. According to the step 2 in the Algorithm 1, we know that
\[ \hat{W} = \arg\min_{W^T X D X^T W = I} \text{Tr}(W^T X \hat{L} X^T W) \]
\[ = \arg\min_{W^T X D X^T W = I} \sum_{i,j=1}^n \hat{S}_{ij} \|W^T x_i - W^T x_j\|_2^2. \]

Note that $\hat{S}_{ij} = \frac{p}{2} S_{ij} \|W^T x_i - W^T x_j\|_2^p$, so we have
\[ \sum_{i,j=1}^n \frac{p}{2} S_{ij} \|W^T x_i - W^T x_j\|_2^p \leq \sum_{i,j=1}^n \frac{p}{2} S_{ij} \|W^T x_i - W^T x_j\|_2^p \leq \sum_{i,j=1}^n \frac{p}{2} S_{ij} \|W^T x_i - W^T x_j\|_2^p. \]

According to Lemma 2, we have
\[ \sum_{i,j=1}^n S_{ij} \|W^T x_i - W^T x_j\|_2^p \]
\[ \leq \sum_{i,j=1}^n S_{ij} \|W^T x_i - W^T x_j\|_2^p \]
\[ \sum_{i,j=1}^n S_{ij} \|W^T x_i - W^T x_j\|_2^p. \]

Summing Eq. (10) and Eq. (11) in the two sides, we arrive at
\[ \sum_{i,j=1}^n S_{ij} \|W^T x_i - W^T x_j\|_2^p \]
\[ \leq \sum_{i,j=1}^n S_{ij} \|W^T x_i - W^T x_j\|_2^p. \]

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 $\hat{L}$ will satisfy Eq. (5), the KKT condition of problem (2). Therefore, the Algorithm 1 will converge to a local optimum of the problem (2). \qed

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 proposed methods 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 $32 \times 32$ 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 TD2 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.
Theoretically, take a small value to improve the robustness of the proposed method. In the extreme case, when $p \to 0$, the relative difference between different distances becomes smaller. This can be attributed to the fact that, when $p = 0$, the distance measurement will be compromised.

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 intrinsic 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 to the fact that, when $p$ is too small, the distance measurement will be compromised, i.e., the relative difference between different distances becomes 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 method 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,

<table>
<thead>
<tr>
<th>Data set</th>
<th>Number</th>
<th>Dimension</th>
<th>Classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coil20</td>
<td>1440</td>
<td>1024</td>
<td>20</td>
</tr>
<tr>
<td>Vehicle</td>
<td>946</td>
<td>18</td>
<td>4</td>
</tr>
<tr>
<td>PIE face</td>
<td>3329</td>
<td>1024</td>
<td>68</td>
</tr>
<tr>
<td>Reuters21578</td>
<td>8293</td>
<td>1599</td>
<td>65</td>
</tr>
<tr>
<td>TDT2 corpus</td>
<td>9394</td>
<td>3157</td>
<td>30</td>
</tr>
</tbody>
</table>

Table 1: Data sets used in our experiments.
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 settings, 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 $\delta = nf \|X\|_F/\|M\|_F$ 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.

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

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

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

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 $\ell_2$-norm distance, which formulated a non-smooth non-convex minimization problem. The new objective imposes the $p$-th order $\ell_2$-norm distance when computing the pairwise 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


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


