The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Locality Preserving Projection Based on F-norm

Xiangjie Hu,[2] Yanfeng Sun,[1,2,*] Junbin Gao,[3] Yongli Hu,[1,2] Baocai Yin[4]
[1]Beijing Advanced Innovation Center for Future Internet Technology
[2]Faculty of Information Technology, Beijing University of Technology, Beijing, China
[3]Discipline of Business Analytics. The University of Sydney Business School,
University of Sydney, NSW 2006, Australia
[4]Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology, China

Abstract

Locality preserving projection (LPP) is a well-known method for dimensionality reduction in which the neighborhood graph structure of data is preserved. Traditional LPP employ squared F-norm for distance measurement. This may exaggerate more distance errors, and result in a model being sensitive to outliers. In order to deal with this issue, we propose two novel F-norm-based models, termed as F-LPP and F-2DLPP, which are developed for vector-based and matrix-based data, respectively. In F-LPP and F-2DLPP, the distance of data projected to a low dimensional space is measured by F-norm. Thus it is anticipated that both methods can reduce the influence of outliers. To solve the F-norm-based models, we propose an iterative optimization algorithm, and give the convergence analysis of algorithm. The experimental results on three public databases have demonstrated the effectiveness of our proposed methods.

Introduction

High-dimensional data are widely acquired in modern image process and computer vision research (Schadt et al. 2010). The high-dimensionality of data not only increases computational complexity and memory requirements in algorithms, but also adversely affects their performance in real applications. In many practical problems, one can confirm that the high dimensional data in general lie in or close to a low dimensional space or manifold (Wang et al. 2017). Thus how to effectively and properly find low-dimensional embedding of high dimensional data becomes a meaningful and urgent problem in the big data era. The past few decades have witnessed great progress in the research on dimensionality reduction algorithms and models (He et al. 2005; Yan et al. 2005; Xu, Caramanis, and Mannor 2013; Wang, Q. Gao, and Nie 2017).

Principal component analysis (PCA) (Jolliffe 1986) is one of classical dimensionality reduction methods and wide utilized in scientific research. PCA aims to learn a linear transformation by maximizing the variance of projected data or equivalently minimizing the reconstruction error when recovering the data from their low dimensional counterpart.

Linear discriminant analysis (LDA) (Belhumeur, Hespanha, and Kriegman 1997) and locality preserving projection (LPP) (He and Niyogi 2003) are two other popular ways of dimensionality reduction. As a supervised learning approach, LDA aims at learning a discriminant projection matrix by maximizing the ratio of the between-class distance to the inter-class distance. LPP seeks for the low dimensional projection so that the locality structure of data can be well preserved. This has been implemented by building a graph neighborhood information of the data points, then using the Laplacian of the graph to learn a projection matrix which maps the data points to a subspace where the original locality structure is preserved. These methods have been used in object recognition (Yanadume et al. 2004; Iosifidis et al. 2012), face recognition (Turk and Pentland 1991; Zhao and Phillips 1999; Xu et al. 2010) and image recognition (Feng et al. 2006).

However, these dimensionality reduction methods are sensitive to outliers because they all rely on the squared F-norm which magnifies the effect of outliers. It is well-known that the l_1-norm is robust to outliers (Brooks, Dul, and Boone 2013; Meng, Zhao, and Xu 2012). In order to alleviate this effect, Ke and Kanade (Ke and Kanade 2005) proposed the robust PCA called PCA-L1, which uses l_1-norm to minimize the reconstruction error. PCA-L1 achieves robust results, however, it usually gets the local optimal solution due to the discontinuity of the objective function. Motivated by the advantage of using l_1-norm, Zhong and Zhang (Zhong and Zhang 2013) proposed LDA-L1. It maximized the l_1-norm based objective function and a greedy algorithm was suggested to learn a discriminant projection matrix. However, the model is not rotational invariant and the solution seriously depends on the selection of initial point. Yu *et al* (Yu et al. 2011) proposed an enhanced locality preserving projection (ELPP), and this model just focuses on building robust graph structure, and is not really robust to outliers. Recently, F-norm-based methods (Li et al. 2017; Xu et al. 2017), which are derived from squared F-norm, were proposed to be used as an alternative measure in objective functions. It is said that F-norm has less contribution to outliers than the squared F-norm, meanwhile it overcomes the shortcomings of l_1-norm with the rotational invariant.

Although some of the aforementioned methods are robust to outliers, they are concerned with vectorial data. When we have 2D or high-order data, a typical workaround is to vectorize 2D data, which may break their structure and discard valuable spatial information existed in data . In this case, some methods for 2D data are proposed, such as two dimensional principal component analysis (2DPCA) (Yang et al. 2004) and two dimensional locality preserving projection (2DLPP) (Hu, Feng, and Zhou 2007; Zhang, Zhao, and Xiong 2010), which are extended from vectorial data to matrix data. In order to alleviate their sensitivity to outliers, two dimensional l_1-norm based methods were proposed. Wang *et al* (Wang et al. 2015) proposed robust 2DPCA with l_1-norm (denoted by 2DPCA-L1) by using l_1-norm to measure reconstruction error. Pang *et al* (Pang, Li, and Yuan 2010) proposed tensor l_1-norm subspace learning. But these l_1-norm based methods are not rotational invariant.

Motivated by the advantage of robust to outliers of F-norm, Wang and Gao (Wang and Gao 2017) propose F-norm 2DPCA (F-2DPCA), which uses F-norm to replace squared l_2-norm and achieves robustness. However, F-2DPCA only takes into account the global information of data ignoring local structural information among data points.

In this paper, we firstly propose the robust LPP with F-norm termed as F-LPP, and then extend F-LPP for 2D data (F-2DLPP). In F-LPP, the distance is measured in terms of F-norm, and summarizing the distance among different points. In F-2DLPP, the data is represented by matrix for keeping spatial structure. We develop an iterative algorithm to solve F-LPP and F-2DLPP.

This paper has following main contributions:

- Firstly we propose F-norm based LPP model, our method is robust to outliers, meanwhile it keeps the advantage of LPP.
- Secondly, we extend the vector-based F-LPP to matrix-based F-2DLPP, which can make use of structural information within 2D data.
- Finally, an iterative optimization algorithm is given to minimize the distance in projected space. The convergence analysis of the objective function for the proposed algorithm is proved.

The rest of the paper is organized as follows. In Section 2, we review the necessary knowledge for classic Locality Preserving Projections algorithm (LPP). We use F-norm instead of the squared F-norm as metric distance to measure the distance in projected space with LPP in Section 3. In Section 4, the performance of the proposed methods are evaluated on several public facial databases. In the last section, we conclude the paper.

Locality Preserving Projection (LPP)

Let $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\}$ be N training samples, where each $\mathbf{x}_i \in \mathbb{R}^H$ is the i-th sample with dimension H. Locality Preserving Projection (LPP) seeks for a projection vector $\mathbf{a} \in \mathbb{R}^H$, that is, $y_i = \mathbf{a}^T\mathbf{x}_i$ fulfills the following objective function,

$$\min_{\mathbf{a}} \sum_{i,j}^{N} (y_i - y_j)^2 w_{i,j} = \sum_{i,j}^{N} \frac{1}{2} (\mathbf{a}^T\mathbf{x}_i - \mathbf{a}^T\mathbf{x}_j)^2 w_{ij}, \quad (1)$$

where $\mathbf{W} \in \mathbb{R}^{N\times N}$ is a symmetric weight matrix whose elements w_{ij} measure the similarity or affinity of data points. For example, if $\mathbf{x}_i$ and $\mathbf{x}_j$ are connected, its element w_{ij} is defined by so-called Gaussian heat kernel,

$$w_{ij} = \begin{cases} e^{-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{t}} & \text{if } \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ are connected} \\ 0 & \text{otherwise} \end{cases}$$

where t is the kernel width parameter.

As the tradition, we assume the data set $\mathbf{X}$ has been centralized. To avoid a trivial solution for problem (1), it is a common practice to impose a constraint as:

$$\mathbf{a}^T\mathbf{X}\mathbf{D}\mathbf{X}^T\mathbf{a} = 1,$$

where $\mathbf{D}$ is a diagonal matrix whose elements are column sum of matrix $\mathbf{W}$ with $d_{ii} = \sum_j w_{ij}$. Thus the final optimal problem becomes

$$\min_{\mathbf{a}} \sum_{i,j}^{N} \frac{1}{2} (\mathbf{a}^T\mathbf{x}_i - \mathbf{a}^T\mathbf{x}_j)^2 w_{ij} = \mathbf{a}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{a},$$
$$\text{s.t. } \mathbf{a}^T\mathbf{X}\mathbf{D}\mathbf{X}^T\mathbf{a} = 1.$$

Where $\mathbf{L} = \mathbf{D} - \mathbf{W}$ is the Laplace matrix.

It is easy to show that the optimal projection vector $\mathbf{a}$ can be solved by the minimum eigenvalue solution,

$$\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{a} = \lambda\mathbf{X}\mathbf{D}\mathbf{X}^T\mathbf{a}. \quad (2)$$

In fact, we can reformulate the above optimal problem as the one for a number of optimal projection as a projection matrix $\mathbf{A}$. That is, we can seek for h projection vectors so that the data with H dimension can reduce to h. The overall LPP model is defined as

$$\min_{\mathbf{A}} \sum_{i,j}^{N} \frac{1}{2} \|\mathbf{A}^T\mathbf{x}_i - \mathbf{A}^T\mathbf{x}_j\|^2 w_{ij} = \mathbf{A}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{A},$$
$$\text{s.t. } \mathbf{A}^T\mathbf{X}\mathbf{D}\mathbf{X}^T\mathbf{A} = \mathbf{I}.$$

Similarly, all the projection vectors in $\mathbf{A}$ can be solved from the generalized eigenvector problem as defined in equation (2).

The LPP Model Based on F-norm

In this section, we firstly propose the LPP model based on F-norm distance for vector variate, and then extend it to the case for matrix variate.

The LPP Model Based on F-norm for Vector Variate (F-LPP)

Although the conventional LPP based on squared F-norm distance measurement has been successful for many applications, it is vulnerable to outliers because the large error can be exaggerated by the use of squared F-norm. Compared with squared F-norm, F-norm can weaken the effect of outliers. Inspired by this, it is reasonable to replace the squared form with just F-norm in the classical LPP model, and get following model.

$$\min_{\mathbf{A}} \sum_{i,j} \frac{1}{2} ||\mathbf{A}^T\mathbf{x}_i - \mathbf{A}^T\mathbf{x}_j||_F \cdot w_{ij}. \quad (3)$$

Model (3) is so called F-LPP, whose distance is measured by F-norm. Compared with the classical LPP with squared F-norm, the novel F-LPP is not vulnerable to outliers. In fact, the objective function of model (3) is derived from the following form,

$$\begin{aligned}&\sum_{i,j}\frac{1}{2}||\mathbf{A}^T\mathbf{x}_i-\mathbf{A}^T\mathbf{x}_j||_F\cdot w_{ij}\\&=\frac{1}{2}\sum_{i,j}\frac{||\mathbf{A}^T\mathbf{x}_i-\mathbf{A}^T\mathbf{x}_j||_F^2 w_{ij}}{||\mathbf{A}^T\mathbf{x}_i-\mathbf{A}^T\mathbf{x}_j||_F}.\end{aligned}$$

Let

$$w'_{ij}=\frac{w_{ij}}{||\mathbf{A}^T\mathbf{x}_i-\mathbf{A}^T\mathbf{x}_j||_F}. \tag{4}$$

Thus, the above objective function can be written as

$$\min_{\mathbf{A}}\frac{1}{2}\sum_{i,j}||\mathbf{A}^T\mathbf{x}_i-\mathbf{A}^T\mathbf{x}_j||_F^2 w'_{ij}. \tag{5}$$

However the objective in (5) is significantly different from the original LPP objective function in (1) because w'_{ij} is in fact the function of $\mathbf{A}$.

In order to learn the projection matrix $\mathbf{A}$, similar to the classic LPP, we impose the constraint $\mathbf{A}^T\mathbf{X}\mathbf{D}\mathbf{X}^T\mathbf{A}=\mathbf{I}$. Finally, F-LPP model (3) can be transformed to following form

$$\begin{aligned}&\min_{\mathbf{A},\mathbf{W}'}\frac{1}{2}\sum_{i,j}||\mathbf{A}^T\mathbf{x}_i-\mathbf{A}^T\mathbf{x}_j||_F^2 w'_{ij},\\&\text{s.t. } \mathbf{A}^T\mathbf{S}_r\mathbf{A}=\mathbf{I};\end{aligned} \tag{6}$$

where $\mathbf{S}_r=\mathbf{X}\mathbf{D}\mathbf{X}^T$ determined by the original w_{ij}.

We are considering a new optimization with respect to two unknown variables $\mathbf{A}$ and $\mathbf{W}'$ in problem (6). Thus by decoupling the function relation between $\mathbf{W}'$ and $\mathbf{A}$, we end up with a relaxed optimization problem. Two variables will be optimized in an alternative way by optimizing one variable while fixing the other.

Specifically, when $\mathbf{W}'$ is known, we can learn $\mathbf{A}$ by minimizing equation (6) which is equivalent to a classic LPP problem. In fact, treating $\mathbf{W}'$ as a constant matrix and denoting diagonal matrix $\mathbf{D}'$ with its element $d'_{ii}=\sum_j w'_{ij}$ and Laplace matrix $\mathbf{L}'=\mathbf{D}'-\mathbf{W}'$, the optimization problem becomes

$$\begin{aligned}&\min_{\mathbf{A}}\frac{1}{2}\text{tr}(\mathbf{A}^T\mathbf{X}\mathbf{L}'\mathbf{X}^T\mathbf{A}),\\&\text{s.t. } \mathbf{A}^T\mathbf{S}_r\mathbf{A}=\mathbf{I}.\end{aligned} \tag{7}$$

Hence $\mathbf{A}$ is solved by the following eigenvalue problem

$$\mathbf{X}\mathbf{L}'\mathbf{X}^T\mathbf{A}=\mathbf{\Lambda}\mathbf{S}_r\mathbf{A}, \tag{8}$$

where $\mathbf{\Lambda}$ is a diagonal matrix consisting of all the generalized eigenvalues.

When $\mathbf{A}$ is fixed, we will update the new affinity matrix $\mathbf{W}'$ by equation (4).

The iterative optimization algorithm for F-LPP is summarized in Algorithm 1.

Algorithm 1 Optimization Algorithm for F-LPP Model

Require:
Training set $\mathbf{X}=\{\mathbf{x}_i\}, i=1,\cdots,N$, and the maximal iteration number Itr.
Initialize projection matrices $\mathbf{A}$ and Laplace matrix $\mathbf{W}$ and $\mathbf{S}_r=\mathbf{X}\mathbf{D}\mathbf{X}^T$.

```
for t = 1 to Itr do
    For all training points, calculate W' by equation (4) and D', L' = D' − W'.
    Solve (7) for A by generalized eigenvector problem (8) such that A's columns consist of the eigenvectors corresponding to the h minimum eigenvalues.
    if convergence then
        break;
    end if
end for
```

Ensure:
Projection matrix $\mathbf{A}$

The LPP Model Based on F-norm for Matrix Variate (F-2DLPP)

LPP algorithm works only for vectorial variate. For 2D data, a typical approach is to vectorize them. This may destroy the spatial structure of data and ignore the valuable information about the spatial relationships. This section will extend F-LPP model to the case for 2D data, namely F-2DLPP.

Given N training samples $\mathcal{X}=\{\mathbf{X}_i\}$ $(i=1,\cdots,N)$ with $\mathbf{X}_i\in\mathbb{R}^{m\times n}$ being matrix data. $\mathbf{W}\in\mathbb{R}^{N\times N}$ is the affinity matrix of data. F-2DLPP aims to find a projection matrix $\mathbf{U}$ by minimizing the F-norm distance in a projected space. That is

$$\min_{\mathbf{U}}\sum_{i,j}\frac{1}{2}||\mathbf{U}^T\mathbf{X}_i-\mathbf{U}^T\mathbf{X}_j||_F\cdot w_{ij},$$

where $\mathbf{U}\in\mathbb{R}^{m\times l}$ $(l\le m)$ is the projection matrix.

Similar to (5), the objective function of F-2DLPP is equivalent to following formula

$$\begin{aligned}&\sum_{i,j}\frac{1}{2}||\mathbf{U}^T\mathbf{X}_i-\mathbf{U}^T\mathbf{X}_j||_F\cdot w_{ij}\\&=\frac{1}{2}\sum_{i,j}\frac{\text{tr}(\mathbf{U}^T(\mathbf{X}_i-\mathbf{X}_j)(\mathbf{X}_i-\mathbf{X}_j)^T\mathbf{U})w_{ij}}{||\mathbf{U}^T\mathbf{X}_i-\mathbf{U}^T\mathbf{X}_j||_F}\\&=\frac{1}{2}\sum_{i,j}\text{tr}(\mathbf{U}^T(\mathbf{X}_i-\mathbf{X}_j)(\mathbf{X}_i-\mathbf{X}_j)^T\mathbf{U})w'_{ij}.\end{aligned}$$

Adding a constraint condition, the F-2DLPP model can be rewritten as

$$\begin{aligned}&\min_{\mathbf{U}}\frac{1}{2}\text{tr}(\mathbf{U}^T\mathbf{T}_l\mathbf{U}),\\&\text{s.t. } \mathbf{U}^T\mathbf{T}_r\mathbf{U}=\mathbf{I};\end{aligned} \tag{9}$$

where

$$\begin{aligned}\mathbf{T}_l&=\sum_{i,j}(\mathbf{X}_i-\mathbf{X}_j)(\mathbf{X}_i-\mathbf{X}_j)^T w'_{ij},\\\mathbf{T}_r&=\sum_{i=1}^{N}\mathbf{X}_i d_{ii}\mathbf{X}_i^T.\end{aligned} \tag{10}$$

In order to learn matrix $\mathbf{U}$, we still use the two-step iteration algorithm. So the matrix $\mathbf{U}$ is obtained by the eigenvalue decomposition of $\mathbf{T}_l\mathbf{U} = \mathbf{\Lambda}\mathbf{T}_r\mathbf{U}$ with the l minimum eigenvalues, then $\mathbf{W}'$ is updated by fixing $\mathbf{U}$. The optimization algorithm of F-2DLPP is summarized in Algorithm 2.

Algorithm 2 Optimization Algorithm for F-2DLPP Model

Require:
Training set $\mathcal{X} = \{\mathbf{X}_i\}, i = 1, \cdots, N$, and the maximal iteration number Itr.
Initialize projection matrix $\mathbf{U}$ and Laplace matrix $\mathbf{W}$ and $\mathbf{T}_r$ by equation (10).
Updating:

1: **for** $t = 1$ to Itr **do**
2: For all training points, calculate $\mathbf{W}'$ by $w'_{ij} = w_{ij}/||\mathbf{U}^T\mathbf{X}_i - \mathbf{U}^T\mathbf{X}_j||_F$ and $\mathbf{D}', \mathbf{L}' = \mathbf{D}' - \mathbf{W}'$.
3: Solving (9) for $\mathbf{U}$ by generalized eigenvector problem $\mathbf{T}_l\mathbf{U} = \lambda\mathbf{T}_r\mathbf{U}$, such that $\mathbf{U}$'s columns consist of the eigenvectors corresponding to the l minimum eigenvalues.
4: **if** $convergence$ **then**
5: $break$;
6: **end if**
7: **end for**

Ensure:
Projection matrix $\mathbf{U}$.

Convergence Analysis

Now we provide a convergence analysis for Algorithm 1. The main result is presented in the following theorem.

Theorem 1 The objective function value sequence produced by the iterative optimization algorithm in Algorithm 1 is monotonically decreased, that is

$$\begin{aligned} &\sum_{i,j}^{N} ||(\mathbf{A}^{(t+1)})^T\mathbf{x}_i - (\mathbf{A}^{(t+1)})^T\mathbf{x}_j||_F \cdot w_{ij} \\ &\leq \sum_{i,j}^{N} ||(\mathbf{A}^{(t)})^T\mathbf{x}_i - (\mathbf{A}^{(t)})^T\mathbf{x}_j||_F \cdot w_{ij}. \end{aligned} \tag{11}$$

Proof: In each iteration, minimizing (6) by the eigenvalue problem $\mathbf{X}\mathbf{L}'\mathbf{X}^T\mathbf{A} = \mathbf{\Lambda}\mathbf{S}_r\mathbf{A}$, we have

$$\begin{aligned} &\sum_{i,j}^{N} \frac{||(\mathbf{A}^{(t+1)})^T\mathbf{x}_i - (\mathbf{A}^{(t+1)})^T\mathbf{x}_j||_F^2}{||(\mathbf{A}^{(t)})^T\mathbf{x}_i - (\mathbf{A}^{(t)})^T\mathbf{x}_j||_F} w_{ij} \\ &\leq \sum_{i,j}^{N} \frac{||(\mathbf{A}^{t})^T\mathbf{x}_i - (\mathbf{A}^{t})^T\mathbf{x}_j||_F^2}{||(\mathbf{A}^{(t)})^T\mathbf{x}_i - (\mathbf{A}^{(t)})^T\mathbf{x}_j||_F} w_{ij}. \end{aligned} \tag{12}$$

As $w_{ij} \geq 0$, we can absorb w_{ij} inside the norm operators and equation (12) becomes

$$\begin{aligned} &\sum_{i,j}^{N} \frac{||[(\mathbf{A}^{(t+1)})^T\mathbf{x}_i - (\mathbf{A}^{(t+1)})^T\mathbf{x}_j]w_{ij}||_F^2}{||[(\mathbf{A}^{(t)})^T\mathbf{x}_i - (\mathbf{A}^{(t)})^T\mathbf{x}_j]w_{ij}||_F} \\ &\leq \sum_{i,j}^{N} \frac{||[(\mathbf{A}^{t})^T\mathbf{x}_i - (\mathbf{A}^{t})^T\mathbf{x}_j]w_{ij}||_F^2}{||[(\mathbf{A}^{(t)})^T\mathbf{x}_i - (\mathbf{A}^{(t)})^T\mathbf{x}_j]w_{ij}||_F}. \end{aligned} \tag{13}$$

According to $a^2 + b^2 \geq 2ab$, i.e. for any $a > 0$

$$\frac{b^2}{a} \geq 2b - a,$$

we can get inequality,

$$\begin{aligned} &2||[(\mathbf{A}^{(t+1)})^T\mathbf{x}_i - (\mathbf{A}^{(t+1)})^T\mathbf{x}_j]w_{ij}||_F \\ &- ||[(\mathbf{A}^{(t)})^T\mathbf{x}_i - (\mathbf{A}^{(t)})^T\mathbf{x}_j]w_{ij}||_F \\ &\leq \frac{||[(\mathbf{A}^{(t+1)})^T\mathbf{x}_i - (\mathbf{A}^{(t+1)})^T\mathbf{x}_j]w_{ij}||_F^2}{||[(\mathbf{A}^{(t)})^T\mathbf{x}_i - (\mathbf{A}^{(t)})^T\mathbf{x}_j]w_{ij}||_F}. \end{aligned} \tag{14}$$

Summing up equation (14) for each pair i, j, we have

$$\begin{aligned} &\sum_{i,j}^{N} 2||[(\mathbf{A}^{(t+1)})^T\mathbf{x}_i - (\mathbf{A}^{(t+1)})^T\mathbf{x}_j]w_{ij}||_F \\ &- \sum_{i,j}^{N} ||[(\mathbf{A}^{(t)})^T\mathbf{x}_i - (\mathbf{A}^{(t)})^T\mathbf{x}_j]w_{ij}||_F \\ &\leq \sum_{i,j}^{N} \frac{||[(\mathbf{A}^{(t+1)})^T\mathbf{x}_i - (\mathbf{A}^{(t+1)})^T\mathbf{x}_j]w_{ij}||_F^2}{||[(\mathbf{A}^{(t)})^T\mathbf{x}_i - (\mathbf{A}^{(t)})^T\mathbf{x}_j]w_{ij}||_F}. \end{aligned} \tag{15}$$

Combining (13) and (15), we have

$$\begin{aligned} &\sum_{i,j}^{N} 2||[(\mathbf{A}^{(t+1)})^T\mathbf{x}_i - (\mathbf{A}^{(t+1)})^T\mathbf{x}_j]w_{ij}||_F \\ &- \sum_{i,j}^{N} ||[(\mathbf{A}^{(t)})^T\mathbf{x}_i - (\mathbf{A}^{(t)})^T\mathbf{x}_j]w_{ij}||_F \\ &\leq \sum_{i,j}^{N} \frac{||[(\mathbf{A}^{(t)})^T\mathbf{x}_i - (\mathbf{A}^{(t)})^T\mathbf{x}_j]w_{ij}||_F^2}{||[(\mathbf{A}^{(t)})^T\mathbf{x}_i - (\mathbf{A}^{(t)})^T\mathbf{x}_j]w_{ij}||_F}. \end{aligned} \tag{16}$$

Simplifying (16), we obtain the following formula

$$\begin{aligned} &\sum_{i,j}^{N} ||(\mathbf{A}^{(t+1)})^T\mathbf{x}_i - (\mathbf{A}^{(t+1)})^T\mathbf{x}_j||_F \cdot w_{ij} \\ &\leq \sum_{i,j}^{N} ||(\mathbf{A}^{(t)})^T\mathbf{x}_i - (\mathbf{A}^{(t)})^T\mathbf{x}_j||_F \cdot w_{ij}. \end{aligned}$$

Thus, the proof of Theorem 1 is completed.

Theorem 2 Suppose that the sequence $\{\mathbf{A}^{(t)}\}_{t=1}^{\infty}$ generated by Algorithm 1 converges to $\mathbf{A}^*$, then $\mathbf{A}^*$ satisfies the KKT conditions of problem (3) with the constraint $\mathbf{A}^{*T}\mathbf{S}_r\mathbf{A}^* = \mathbf{I}$.

Proof: Under the constraint $\mathbf{A}^T\mathbf{S}_r\mathbf{A} = \mathbf{I}$, the Lagrangian function of problem (3) is

$$L_1(\mathbf{A}) = \frac{1}{2}\sum_{i,j}^{N} \|\mathbf{A}^T\mathbf{x}_i - \mathbf{A}^T\mathbf{x}_j\|_F w_{ij} - \text{tr}(\Lambda(\mathbf{A}^T\mathbf{S}_r\mathbf{A} - \mathbf{I})),$$

where Λ is the Lagrangian multiplier.

Also the Lagrangian function of problem (7) (i.e. (6)) is given by

$$L_2(\mathbf{A}) = \frac{1}{2}\sum_{i,j}^{N} \|\mathbf{A}^T\mathbf{x}_i - \mathbf{A}^T\mathbf{x}_j\|_F^2 w'_{ij} - \text{tr}(\Lambda(\mathbf{A}^T\mathbf{S}_r\mathbf{A} - \mathbf{I})).$$

According to step 3 in Algorithm 1, $\mathbf{A}^{(t+1)}$ is the stationary point of $L_2(\mathbf{A})$ where w'_{ij} is defined by $\mathbf{A}^{(t)}$, i.e.,

$$\sum_{i,j}^{N} w'_{ij}(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T\mathbf{A}^{(t+1)} = \Lambda\mathbf{S}_r\mathbf{A}^{(t+1)}. \quad (17)$$

and $w'_{ij} = w_{ij}/\|\mathbf{A^{(t)}}^T\mathbf{x}_i - \mathbf{A^{(t)}}^T\mathbf{x}_j\|_F$. According to Theorem assumption and w'_{ij} is a continuous function of $\mathbf{A^{(t)}}$, taking $t \to \infty$ on both sides of (17) gives

$$\sum_{i,j}^{N} \frac{w_{ij}}{\|\mathbf{A}^{*T}\mathbf{x}_i - \mathbf{A}^{*T}\mathbf{x}_j\|_F}(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T\mathbf{A}^* = \Lambda\mathbf{S}_r\mathbf{A}^*,$$

which means $\mathbf{A}^*$ is a stationary point of $L_1(\mathbf{A})$. Hence $\mathbf{A}^*$ is a KKT point of problem (3).

This completes the proof of Theorem 2.

Remark 1: Both Theorems 1 and 2 ensure that the converged sequence produced by Algorithm 1 is a local minimum of the F-LPP problem (3). Also we shall note the sequence generated by Algorithm 1 is bounded according to the constraint condition, hence there exists at least an accumulation point for the sequence, i.e., a subsequence $\{A^{(t_k)}\}_{k=1}^{\infty}$ which is convergent.

Remark 2: Theorem 1 shows that, for the sequence $\{(\mathbf{A}^{(t)}, \mathbf{W}'^{(t)})\}$, the objective value is declining towards the optimal value.

In experiment, we empirically show the phenomenon of the objective convergence.

Remark 3: A similar convergence conclusion can be established for F-2DLPP in Algorithm 2. To save space, we simply skip it.

Experimental Results and Analysis

In this section, we conduct several experiments to evaluate the proposed F-LPP and F-2DLPP models on three public face databases: Extended Yale B, CMU-PIE and AR. These experiments are designed to demonstrate the robustness for outliers and the performance of the two proposed models in feature extraction by comparing with other related algorithms. In experiments, we use 1 nearest neighbor (1NN) classifier.

Figure 1: Sample images with/without outliers from Extended Yale B database.

The Extended Yale B is face database [1] , which includes 38 individuals with 64 frontal-face images of each individual. The 64 photos of each one describe intra-class variations, such as illumination, expression and poses. The 64 images of one person are taken from 5 different angles and divided into 5 subsets. In the experiment, 31 individuals are randomly chosen, 40 images of each individual are used for the training data and the rest of images are used for testing. In addition, some training images and test images of each person have white or black dots added as noise. The positions of outliers are randomly distributed and their sizes range from 5×5 to 7×7. Figure 1 shows some samples with or without outliers of one person. All images are in gray-scale and resized to 32×32.

The CMU-PIE face database [2] contains 1632 frontal-face images of 68 individuals, which describe the intra-class variation, such as illumination and expression. In our experiment, we choose 21 images of each individual, among which we use 15 images, a total of 1020 images for training and 6 images, a total of 408 images for testing. We randomly select training and test images of each person to add occlusion as that in Extended Yale B database. Each image is in gray-scale and normalized to 32×32.

The AR face database [3] contains over 4,000 color images corresponding to 126 people's faces. The images were taken in two sessions. Each session consists of 13 images, which is 7 frontal face images and 6 images with occlusion, as scarf or glasses. Each session presents the variation of expression, illumination condition and occlusion. Sample images of one person are shown in Figure 2. In our experiment, we randomly select 13 images of each individual for a total of 1300 images as the training set and the rest for testing. All images are normalized to 32×32.

We firstly do experiments to evaluate the robustness of F-LPP to outliers.

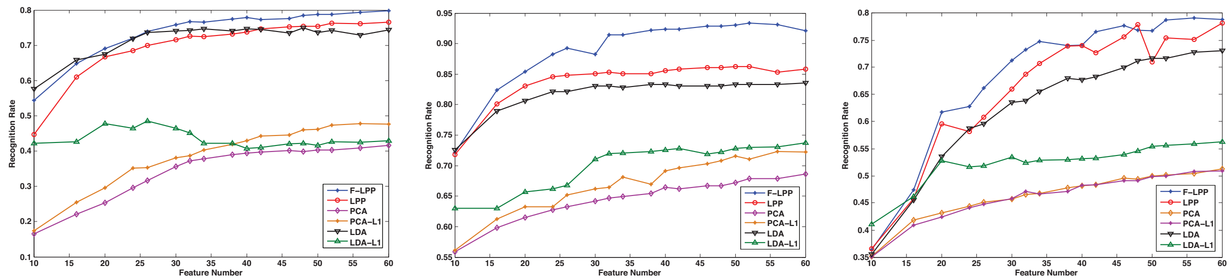
To test robustness, we compare the recognition rate of LPP and F-LPP with different percentage of outliers in training images on Extended Yale B and CMU-PIE databases. As shown in Table 1 that our method is generally higher than LPP. Though the recognition rates of LPP and F-LPP are all decreasing with the increase of outliers percentage, our

[1]http://vision.ucsd.edu/~leekc/ExtYaleDatabase/ExtYaleB.html

[2]http://www.cs.cmu.edu/afs/cs/project/PIE/MultiPie/Multi-Pie/Home.html

[3]http://www2.ece.ohio-state.edu/~aleix/ARdatabase.html

Figure 2: Sample images of one person under two sessions on AR face database.



(a) Extended Yale B (b) CMU-PIE (c) AR

Figure 3: Recognition rate of F-LPP under the different feature numbers on three public databases.

PCT	Extended Yale B		CMU-PIE	
	LPP	F-LPP	LPP	F-LPP
0	88.58	**89.22±0.32**	1	1
5	87.23	**87.58±0.11**	88.97	**94.27±0.13**
10	83.46	**85.03±0.22**	88.24	**93.87±0.17**
20	76.21	**81.67±0.28**	87.01	**93.63±0.31**
30	70.17	**79.44±0.14**	86.52	**93.14±0.26**

Table 1: The first column is the percentage (PCT) of outliers in image data. The last four columns are average recognition rate(%) and variances on Extended Yale B and CMU-PIE databases.

method has a lower decreasing speed than LPP, which illustrates that F-LPP can weak the effect of outliers and is more robust to outliers than squared F-norm.

Besides the robustness, the F-LPP still keeps the advantage of classical LPP, making the "close" data points still close after mapping.

Then we compare the recognition performance at different feature numbers of F-LPP with five other algorithms: LPP, PCA, PCA-L1, LDA and LDA-L1 on three databases. Each algorithm is run 10 times and the mean value of recognition rate and covariance is recorded. The curve shows the trend of recognition rate in Figure 3. From these results, we find that F-LPP is generally superior to the other five algorithms, even if there exists outliers in images on Extended Yale B, CMU-PIE and AR databases, respectively. In addition, Figure 4 illustrates the convergence of our F-LPP in three databases.

Table 2 lists the average recognition rates and variances of six methods at a given feature number on Extended Yale B, CMU-PIE and AR databases, respectively. It can be seen

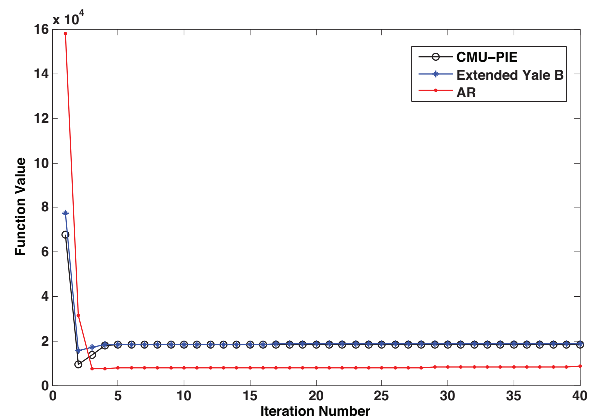
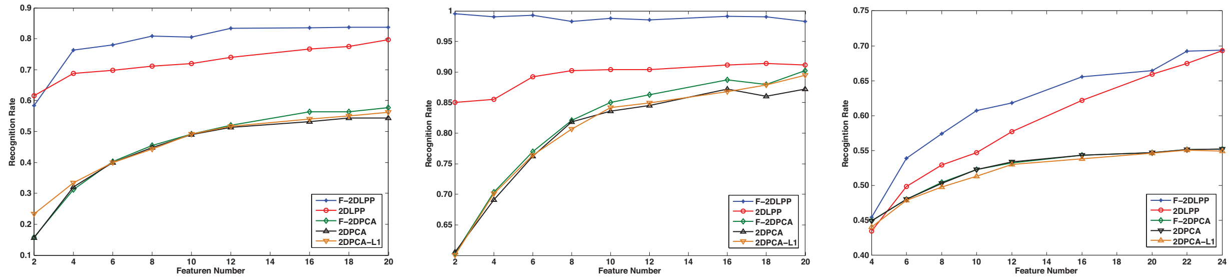


Figure 4: The convergence curves produced by F-LPP on three public databases.

Method	Databases		
	Extended Yale B	CMU-PIE	AR
PCA	41.67±0	68.63±0	51.46±0
PCA-L1	47.90±0.24	73.04±0.79	50.91±0.20
LDA	74.33±0	83.58±0	73.08±0
LDA-L1	43.01±0	73.77±0	56.23±0
LPP	76.48±0	85.78±0	78.15±0
F-LPP	**79.93±0.16**	**92.30±0.21**	**79.01±0.22**

Table 2: The mean recognition rates (%) and variances of six vector variate based methods on three public databases.



(a) Extended Yale B (b) CMU-PIE (c) AR

Figure 5: Recognition rate of F-2DLPP under the different feature numbers on three public databases.

Method	Runtime (s)		
	Extended Yale B	CMU-PIE	AR
PCA	1.30±0.25	0.96±0.03	1.18±0.09
PCA-L1	54.90±0.73	36.14±0.42	55.24±0.39
LDA	9.16±0.33	71.18±0.46	9.78±0.17
LDA-L1	110.97± 1.65	13.28±0.22	117.16± 4.25
LPP	1.38±0.06	1.21±0.04	0.50±0.01
F-LPP	3.41±0.14	4.60±0.08	14.16±0.32

Table 3: The runtime(s) of six vector based methods on the three public databases.

Method	Databases		
	Extended Yale B	CMU-PIE	AR
2DPCA	54.44±0	87.25±0	55.23±0
2DPCA-L1	57.50±0.06	89.80±0.22	55.14±0.06
F-2DPCA	59.27±0	90.39±0	55.23±0
2DLPP	79.70±0	97.55±0	69.31±0
F-2DLPP	**83.76±0.48**	**99.11±0.01**	**69.41±0.27**

Table 4: The average recognition rates (%) and variances of 2D variate based methods on three public databases.

that the F-LPP achieves the highest recognition rates with acceptable variances.

Table 3 shows the runtime of six vector variate based methods on the three public databases. Though our F-LPP is a little slower than classical LPP and PCA, it is comparable on runtime with other methods.

The second group experiment evaluates the performance of feature extraction of F-2DLPP by comparing with 2DLPP, F-2DPCA, 2DPCA and 2DPCA-L1. All algorithms employ 2D data representation. Figure 5 gives the recognition rates of these algorithm on three different databases. These recognition curves have shown that F-2DLPP is generally superior to the other methods.

Table 4 presents the average recognition rates and variances of five 2D dimension reduction methods with a given feature number on the three public databases. Our F-2DLPP also has the best recognition rates compared with other algorithms.

Conclusions

In this paper, we proposed two new LPP dimensionality reduction models based on F-norm for vector and matrix variables, named by F-LPP and F-2DLPP. Different from traditional LPP which uses squared F-norm for distance measurement of data points, the proposed models employ F-norm. Thus the new models keep the merits of classical LPP, and will be more robust to outliers due to the F-norm contributing less to outliers. We proposed an iteration optimization algorithm to solve the F-norm-based models. And the weaker convergence analysis has been given for the algorithms. The experimental results on the several public databases demonstrated the effectiveness of our proposed methods.

Acknowledgements

This research was supported by National Natural Science Foundation of China (Grant No.61772048, 61672071, 61370119, 61390510), Beijing Natural Science Foundation (Grant No.4172003, 4162010, 4152009), Australian Research Council Discovery Projects funding scheme (Project No. DP140102270). Junbin Gao is also supported by the University of Sydney Business School ARC Bridging Fund (2017).

References

Belhumeur, P. N.; Hespanha, J.; and Kriegman, D. J. 1997. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19(7):711–720.

Brooks, J. P.; Dul, J. H.; and Boone, E. L. 2013. A pure l1-norm principal component analysis. *Computational Statistics and Data Analysis* 61(4):83–98.

Feng, G.; Hu, D.; Zhang, D.; and Zhou, Z. 2006. Letters: An alternative formulation of kernel LPP with application to image recognition. *Neurocomputing* 69(13):1733–1738.

He, X., and Niyogi, P. 2003. Locality preserving projections. *Advances in Neural Information Processing Systems* 16(1):186–197.

He, X.; Cai, D.; Yan, S.; and Zhang, H. J. 2005. Neighborhood preserving embedding. In *Proceedings of 10th IEEE International Conference on Computer Vision*, 1208–1213.

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

Iosifidis, A.; Tefas, A.; Nikolaidis, N.; and Pitas, I. 2012. Multi-view human movement recognition based on fuzzy distances and linear discriminant analysis. *Computer Vision and Image Understanding* 116(3):347–360.

Jolliffe, I. T. 1986. *Principal Component Analysis*. Springer-Verlag.

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

Li, T.; Li, M.; Gao, Q.; and Xie, D. 2017. F-norm distance metric based robust 2DPCA and face recognition. *Neural Networks* 94:204–211.

Meng, D.; Zhao, Q.; and Xu, Z. 2012. Improve robustness of sparse PCA by l1-norm maximization. *Pattern Recognition* 45(1):487–497.

Pang, Y.; Li, X.; and Yuan, Y. 2010. Robust tensor analysis with l1-norm. *IEEE Transactions on Circuits and Systems for Video Technology* 20(2):172–178.

Schadt, E. E.; Linderman, M. D.; Sorenson, J.; Lee, L.; and Nolan, G. P. 2010. Computational solutions to large-scale data management and analysis. *Nature Reviews Genetics* 11(9):647–650.

Turk, M., and Pentland, A. 1991. Eigenfaces for recognition. *Journal of Cognitive Neuroscience* 3(1):71–86.

Wang, Q., and Gao, Q. 2017. Two-dimensional PCA with f-norm minimization. In *Proceedings of the 31th AAAI Conference on Artificial Intelligence*, 180–186.

Wang, R.; Nie, F.; Yang, X.; Gao, F.; and Yao, M. 2015. Robust 2DPCA with non-greedy l1-norm maximization for image analysis. *IEEE Transactions on Cybernetics* 45(5):1108–1112.

Wang, Q.; Hu, Y.; Gao, J.; and Sun, Y. 2017. Locality Preserving Projections for Grassmann manifold. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence Main track*, 2893–2900.

Wang, Q.; Q. Gao, X. G.; and Nie, F. 2017. Angle Principal Component Analysis. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence Main track*, 2936–2942.

Xu, Y.; Zhong, A.; Yang, J.; and Zhang, D. 2010. LPP solution schemes for use with face recognition. *Pattern Recognition* 43(12):4165–4176.

Xu, J.; Han, J.; Nie, F.; and Li, X. 2017. Re-weighted discriminatively embedded k-means for multi-view clustering. *IEEE Transactions on Image Processing* 26(6):3016–3027.

Xu, H.; Caramanis, C.; and Mannor, S. 2013. Outlier-Robust PCA: The high-dimensional case. *IEEE Transactions on Information Theory* 59(1):546–572.

Yan, S.; Xu, D.; Zhang, B.; and Zhang, H. J. 2005. Graph embedding: a general framework for dimensionality reduction. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 830–837.

Yanadume, S.; Mekada, Y.; Ide, I.; and Murase, H. 2004. *Recognition of Very Low-Resolution Characters from Motion Images Captured by a Portable Digital Camera*. Springer Berlin Heidelberg.

Yang, J.; Zhang, D.; Frangi, A. F.; and Yang, J. Y. 2004. Two-dimensional PCA: a new approach to appearance-based face representation and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 26(1):131–137.

Yu, G.; Peng, H.; Wei, J.; and Ma, Q. 2011. Enhanced locality preserving projections using robust path based similarity. *Neurocomputing* 74(4):598–605.

Zhang, E.; Zhao, Y.; and Xiong, W. 2010. Active energy image plus 2DLPP for gait recognition. *Signal Processing* 90(7):2295–2302.

Zhao, W., and Phillips, P. J. 1999. Subspace linear discriminant analysis for face recognition. *Technical Report CAR-TR-914, Center for Automation Research,University of Maryland, College Park*.

Zhong, F., and Zhang, J. 2013. Linear discriminant analysis based on l1-norm maximization. *IEEE Transactions on Image Processing* 22(8):3018–3027.