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

Robust Formulation for PCA: Avoiding Mean Calculation with $\ell_{2,p}$-Norm Maximization

Shuangli Liao, Jin Li, Yang Liu, Quanxue Gao,* Xinbo Gao
State Key Laboratory of Integrated Services Networks,
Xidian University, Xi'an 710071, Shaanxi, P. R. China

Abstract

Most existing robust principal component analysis (PCA) involve mean estimation for extracting low-dimensional representation. However, they do not get the optimal mean for real data, which include outliers, under the different robust distances metric learning, such as ℓ_1-norm and $\ell_{2,1}$-norm. This affects the robustness of algorithms. Motivated by the fact that the variance of data can be characterized by the variation between each pair of data, we propose a novel robust formulation for PCA. It avoids computing the mean of data in the criterion function. Our method employs $\ell_{2,p}$ norm as the distance metric to measure the variation in the criterion function and aims to seek the projection matrix that maximizes the sum of variation between each pair of the projected data. Both theoretical analysis and experimental results demonstrate that our methods are efficient and superior to most existing robust methods for data reconstruction.

Introduction

In real applications, we usually get the data such as face images and gene expressions with high-dimensionality. If we directly analyze on these high-dimensional data, it will suffer from the curse of dimensionality and cause performance degradation with complexity computation. Thus, how to find an effective representation for high-dimensional data has been an active and fundamental problem in the fields of pattern recognition and machine learning. For this task, dimensionality reduction is an effective and successful approach. It aims to seek a low-dimensional space such that class distribution becomes more apparent which can improve the robustness of subsequent analysis (Jiang et al. 2013; Gao et al. 2017).

To analyze different data types, many dimensionality reduction methods have been developed in the literature, among which principal component analysis (PCA) and linear discriminant analysis (LDA) are two of the most representative methods (Turk and Pentland 1991; Belhumeur, Hespanha, and Kriegman 1997). Moreover, PCA is also usually used as a preprocessing step in almost dimensionality methods including LDA. Thus, in this paper, we focus on how to improve robustness of PCA for dimensionality reduction. PCA aims to seek projection matrix such that the projected data well reconstruct the original data in a least squares sense. It is well known that least square criterion remarkably enlarges the large distance, which results in sensitivity of PCA to outliers and noise (Gao et al. 2013; Shahid et al. 2015).

To handle this problem, many approaches have been developed. These methods can be broadly divided into two categories: nuclear norm based methods and L1-norm based methods. Nuclear-norm based methods aim to seek clean data with low-rank structure. The representative methods include robust PCA (RPCA) (Li, Ma, and Wright 2009) and RPCA with graph (GRPCA) (Shahid et al. 2015). This kind of methods does not get the low-dimensional representation, in other words, they cannot be suitable for dimensionality reduction. Different from nuclear norm based methods, ℓ_1-norm PCA uses ℓ_1-norm instead of squared Euclidean distance as the distance metric in the criterion function of PCA. There are two formulations for ℓ_1-norm based PCA. One is L1-PCA, which minimizes the ℓ_1-norm reconstruction error (Ke and Kanade 2005). It well improves the robustness of PCA to oultiers, but it is difficult to solve L1-PCA and does not have rotational invariance (Ding et al. 2006; Gao et al. 2017; Wang and Gao 2017). Another is PCA-L1 that maximizes ℓ_1-norm covariance and has become an active topic in pattern analysis (Kwak 2008; Wang et al. 2015; Ju et al. 2015; Wang and Wang 2013).

Compared with traditional PCA, the aforementioned ℓ_1-norm PCA methods effectively improve the robustness of algorithms, but they involve the mean estimation of data, which is calculated in the least squared sense. Some works have witnessed that the mean in the least squared sense is not optimal for other different distance metrics such as ℓ_1-norm, $\ell_{2,1}$-norm and nuclear norm (Oh and Kwak 2016; He et al. 2011; Wang et al. 2017). This effects the robustness of algorithms. To tackle this problem, some enhanced robust PCA methods were proposed by simultaneously optimizing mean and projection matrix in the criterion function. For example, Nie *et al.* (Nie, Yuan, and Huang 2014) simultaneously optimized mean and projection matrix by maximizing $\ell_{2,1}$-norm variance, which was extended to robust two-dimensional formulation with F-norm minimization (Wang et al. 2017). However, all of them cannot ob-

*Corresponding author: Q. Gao: qxgao@xidian.edu.cn

tain the global mean (Song, Woodruff, and Zhong 2017; Luo et al. 2016) result in additional computation.

Recently, a novel robust PCA form (RPCA-AOM) was proposed by maximizing the sum of projected differences between each pair of data based on the ℓ_1-norm distance (Luo et al. 2016). It effectively avoids mean computation in solving the projection matrix. However, it has the following limitations. First, RPCA-AOM does not well characterize the geometric structure of data (Gao et al. 2017). Second, it is not only time-consuming but also difficult to solve the local optimal solution of RPCA-AOM. To avoid the aforementioned problems, we develop a novel robust formulation for PCA. Our method employs $\ell_{2,p}$-norm as the distance metric to measure the variation between data points and seeks projection matrix by maximizing the sum of variations between each pair of the projected data. **Our method has the following advantages:**

From the norm point of view, $\ell_{2,p}$-norm and squared ℓ_2-norm (Euclidean distance) have no essential difference. Thus, our method retains traditional PCAs desirable properties. For example, our method has rotational invariance, and the solution is related to the weighted covariance matrix, which well characterizes the geometric structure of data. Furthermore, compared with squared ℓ_2-norm, $\ell_{2,p}$-norm $(0 < p < 2)$ can suppress the effect of outliers in the criterion function, thus our method is robust to outliers. Finally, compared with ℓ_1-norm optimization problem, $\ell_{2,p}$-norm optimization can be easily solved in real applications, and our proposed algorithm has a closed form solution in each iteration.

Principal Component Analysis Review

Assume that we have training images matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n] \in R^{d\times n}$ which includes n samples, where $\mathbf{x}_i \in R^d (i = 1, 2, ..., n)$ denote the ith training image, d is the dimensionality of the samples space. PCA aims to seek a transformation such that the projected data well reconstruct the corresponding original data in a least squares sense. Denote by $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_k] \in R^{d\times k}$ the projection matrix which spans a $k(k < d)$-dimensional subspace, it can be obtained by the model (1).

$$\min_{\mathbf{W}^T\mathbf{W}=\mathbf{I}_k,\mathbf{m}} \sum_{i=1}^{n} \left\|(\mathbf{x}_i - \mathbf{m}) - \mathbf{W}\mathbf{W}^T(\mathbf{x}_i - \mathbf{m})\right\|_2^2 \quad (1)$$

where $\mathbf{m}$ denotes the mean of the training data. $\mathbf{I}_k$ is a k-dimensional identity matrix.

For each data point, we have $\left\|(\mathbf{x}_i - \mathbf{m}) - \mathbf{W}\mathbf{W}^T(\mathbf{x}_i - \mathbf{m})\right\|_2^2 + \left\|\mathbf{W}^T(\mathbf{x}_i - \mathbf{m})\right\|_2^2 = \|(\mathbf{x}_i - \mathbf{m})\|_2^2$. Suppose that the mean $\mathbf{m}$ is known, then $\|(\mathbf{x}_i - \mathbf{m})\|_2^2$ is constant. Thus, by simple algebra, the problem (1) can be reformulated as the maximization of variance in the projected space, i.e. the model (2).

$$\max_{\mathbf{W}^T\mathbf{W}=\mathbf{I}_k,\mathbf{m}} \sum_{i=1}^{n} \left\|\mathbf{W}^T(\mathbf{x}_i - \mathbf{m})\right\|_2^2 \quad (2)$$

As can be seen in the models (1) and (2), the estimation of mean $\mathbf{m}$ is very important for solving projection matrix $\mathbf{W}$. By setting the derivative of the problem (1) with respect to $\mathbf{m}$ to zero, the optimal mean is $\mathbf{m} = \overline{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i$.

In the model (1) or (2), the squared large distance will remarkably dominate the solution. Thus, the objective function (1) or (2) is not robust in the sense that outlying measurements can skew the solution from the desired solution. To handle this problem, many robust methods have been developed for dimensionality reduction, one of the most representative methods is PCA-L1 (Kwak 2008). It aims to solve the projection matrix $\mathbf{W}$ by the model (3).

$$\max_{\mathbf{W}^T\mathbf{W}=\mathbf{I}_k} \sum_{i=1}^{n} \left\|\mathbf{W}^T\overline{\mathbf{x}}_i\right\|_1 \quad (3)$$

where the samples $\overline{\mathbf{x}}_i = \mathbf{x}_i - \overline{\mathbf{x}} (i = 1, 2, \cdots, n)$ denote the centered data corresponding to $\mathbf{x}_i$.

In the model (3), $\overline{\mathbf{x}}$ is estimated under the squared ℓ_2-norm distance metric. However, it is incorrect due to the fact that the optimal mean of training data is different under different distance metrics (Oh and Kwak 2016; Nie, Yuan, and Huang 2014; Wang et al. 2017). This affects the robustness of algorithm. To tackle this problem, Nie *et al*. (Nie, Yuan, and Huang 2014) integrated mean calculation in the criterion function and solved projection matrix by

$$\max_{\mathbf{W}^T\mathbf{W}=\mathbf{I}_k,\mathbf{m}} \sum_{i=1}^{n} \left\|\mathbf{W}^T(\mathbf{x}_i - \mathbf{m})\right\|_2 \quad (4)$$

In the model (4), there are two unknown parameters $\mathbf{W}$ and mean $\mathbf{m}$, which are related to each other, thus it needs to alternatively update them. This usually causes error accumulation. Moreover, it is difficult to get the optimal mean in real applications. To tackle this problem, Luo *et al*. (Luo et al. 2016) proposed RPCA-AOM whose objective function is

$$\max_{\mathbf{W}^T\mathbf{W}=\mathbf{I}_k} \sum_{i,j}^{n} \left\|\mathbf{W}^T(\mathbf{x}_i - \mathbf{x}_j)\right\|_1 \quad (5)$$

Compared with the models (3) and (4), the model (5) effectively avoids the mean computation. However, it has the following limitations. First, RPCA-AOM does not obtain the optimal solution (Markopoulos, Karystinos, and Pados 2014). Second, it needs to solve ℓ_1-norm optimization problem that is difficult. Third, it is not clear whether ℓ_1-norm relates to the covariance matrix, which characterizes the geometric structure.

Robust PCA with non-greedy $l_{2,p}$-norm maximization

Motivation and Objective function

In this section, we consider a general case that the mean of the data is not zero, and propose a novel robust PCA based on $l_{2,p}$-norm distance metric learning which not only avoids mean estimation but also well characterizes the geometric structure. For a better representation, we first introduce the following theorem.

Theorem 1 (Luo et al. 2016): The objective function (1) and the following objective function are equivalent.

$$\max_{\mathbf{W}^T\mathbf{W}=\mathbf{I}} \sum_{i,j} \left\|\mathbf{W}^T(\mathbf{x}_i - \mathbf{x}_j)\right\|_2^2 \quad (6)$$

As can be seen in the model (6), *squared ℓ_2-norm* remarkably enlarges the role of outliers in the criterion function. This makes the model (6) not robust to outliers that deviate significantly from the rest of data. To handle this problem, the contribution of distance metric in the criterion function (6) should reduce the effect of outliers. Moreover, we hope to obtain a robust low-dimensional subspace that well characterizes the geometric structure. Compared with squared ℓ_2-norm, which is used in traditional PCA, both ℓ_2-norm and ℓ_1-norm can weaken the effect of outliers, but, from the norm point of view, *ℓ_2-norm* and squared ℓ_2-norm have no essential difference. This illustrates that ℓ_2-norm not only is robust to outliers but also helps retain PCA's nice properties such as well geometric structure and rotational invariance. Herein, we extend ℓ_2-norm to a generalized form *$\ell_{2,p}$-norm* $(0 < p < 2)$ and then use *$\ell_{2,p}$-norm* as the distance metric.

Combining the aforementioned analysis, we aim to seek the projection matrix $\mathbf{W}$ by solving the model (7).

$$\max_{\mathbf{W}^T\mathbf{W}=\mathbf{I}_k} \sum_{i,j} \left\|\mathbf{W}^T(\mathbf{x}_i - \mathbf{x}_j)\right\|_2^p \tag{7}$$

where $0 < p < 2$.

As can be seen in (7), our proposed objective automatically avoids the estimation of mean. Compared to the model (5), our method needs to solve $\ell_{2,p}$-norm optimization problem, which is easy to solve. Moreover, $\ell_{2,p}$-norm can further suppress the effect of outliers in the criterion function. Thus our method is robust to outliers. Finally, solution of our method relates to weighted covariance matrix, which well characterizes the geometric structure of data.

Algorithm

In this section, we propose an efficient iterative algorithm to solve the problem (7). By simple algebra, we have

$$\begin{aligned}
&\left\|\mathbf{W}^T(\mathbf{x}_i - \mathbf{x}_j)\right\|_2^p \\
=&\sum_{i,j} \left\|\mathbf{W}^T(\mathbf{x}_i - \mathbf{x}_j)\right\|_2^2 s_{ij} \\
=&\sum_{i,j} tr((\mathbf{W}^T(\mathbf{x}_i - \mathbf{x}_j))(\mathbf{W}^T(\mathbf{x}_i - \mathbf{x}_j))^T) s_{ij} \\
=&2tr(\mathbf{W}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{W}) \\
=&2tr(\mathbf{W}^T\mathbf{H})
\end{aligned} \tag{8}$$

where $\mathbf{L}$=$\mathbf{D}$−$\mathbf{S}$, $\mathbf{H} = \mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{W}$. $\mathbf{S}$ is a symmetric matrix whose elements are $s_{ij} = \left\|\mathbf{W}^\mathbf{T}(\mathbf{x}_j - \mathbf{x}_j)\right\|_2^{p-2}$, and $\mathbf{D}$ is diagonal matrix whose diagonal elements are $d_{ii} = \sum_j s_{ij}$.

Substituting Eq. (8) into the objective function (7), our objective function (7) finally becomes

$$\arg\max_{\mathbf{W}^T\mathbf{W}=\mathbf{I}_k} tr(\mathbf{W}^T\mathbf{H}) \tag{9}$$

We can see that matrix $\mathbf{H}$ is dependent on the target variable $\mathbf{W}$, so that Eq. (9) cannot be directly solved. But if $\mathbf{H}$ is known, Eq. (9) can be easily solved. After obtained $\mathbf{W}$, the value of $\mathbf{H}$ can be updated correspondingly, which inspires us to solve the model (9) in an alternative way. Before solving the model (9) with the known $\mathbf{H}$, we first introduce two theorems as follows.

Theorem 2: For the same order matrix $\mathbf{X}$, $\mathbf{Y}$, we have

$$tr(\mathbf{X}^T\mathbf{Y}) \leq \|\mathbf{X}\|_F \|\mathbf{Y}\|_F \tag{10}$$

with equality if and only if $\mathbf{X}$ or $\mathbf{Y}$ is a multiple of the other.

Proof:

$$tr(\mathbf{X}^T\mathbf{Y}) = (vec(\mathbf{X}))^T vec(\mathbf{Y}) \tag{11}$$

According to Cauchy-Schwarz inequality, we have

$$\begin{aligned}
(vec(\mathbf{X}))^T vec(\mathbf{Y}) &\leq \|vec(\mathbf{X})\|_2 \|vec(\mathbf{Y})\|_2 \\
&= \|\mathbf{X}\|_F \|\mathbf{Y}\|_F
\end{aligned} \tag{12}$$

Comparing inequalities (11) and (12) yields

$$tr(\mathbf{X}^T\mathbf{Y}) \leq \|\mathbf{X}\|_F \|\mathbf{Y}\|_F \tag{13}$$

■

Theorem 3: Denote by $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ the compact singular value decomposition of $\mathbf{H} \in \mathbf{R}^{m\times n}$, then $\mathbf{W} = \mathbf{U}\mathbf{V}^T$ is the solution of the following objective function

$$\arg\max_{\mathbf{W}^T\mathbf{W}=\mathbf{I}_k} tr(\mathbf{W}^T\mathbf{H}) \tag{14}$$

where $\mathbf{V}^T\mathbf{V} = \mathbf{U}^T\mathbf{U} = \mathbf{I}_k$, $\mathbf{\Sigma} \in \mathbf{R}^{k\times k}$ is a nonsingular diagonal matrix whose elements on diagonal are singular values λ_j of matrix $\mathbf{H}$. $k = rank(\mathbf{H})$.

Proof:

$$\begin{aligned}
tr(\mathbf{W}^T\mathbf{H}) &= tr(\mathbf{W}^T\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T) \\
&= tr(\mathbf{U}\mathbf{\Sigma}^{1/2}\mathbf{\Sigma}^{1/2}\mathbf{V}^T\mathbf{W}^T)
\end{aligned} \tag{15}$$

According to **Theorem 2**, we have

$$\begin{aligned}
tr(\mathbf{W}^T\mathbf{H}) &\leq \left\|\mathbf{U}\mathbf{\Sigma}^{1/2}\right\|_F \left\|\mathbf{\Sigma}^{1/2}\mathbf{V}^T\mathbf{W}^T\right\|_F \\
&= \left\|\mathbf{\Sigma}^{1/2}\right\|_F \left\|\mathbf{\Sigma}^{1/2}\right\|_F
\end{aligned} \tag{16}$$

Equality holds only when

$$\mathbf{\Sigma}^{1/2}\mathbf{U}^T = \mathbf{\Sigma}^{1/2}\mathbf{V}^T\mathbf{W}^T \tag{17}$$

which satisfies $\mathbf{W} = \mathbf{U}\mathbf{V}^T$. ■

Now, we consider how to solve the objective function (9). Denote the compact singular value decomposition (SVD) of matrix $\mathbf{H}$ by

$$\mathbf{H} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\mathbf{T} \tag{18}$$

where $\mathbf{V}^T\mathbf{V} = \mathbf{U}^T\mathbf{U} = \mathbf{I}_k$, $\mathbf{\Sigma} \in \mathbf{R}^{k\times k}$ is a nonsingular diagonal matrix whose elements on diagonal are singular values λ_j of matrix $\mathbf{H}$.

According to theorem 3, we have that the optimal solution of the objective function (9) is

$$\mathbf{W} = \mathbf{U}\mathbf{V}^T \tag{19}$$

According to the aforementioned analysis, we can see that the solution of our model (7) is dependent on $\mathbf{H} = \mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{W}$, where $\mathbf{X}\mathbf{L}\mathbf{X}^T$ is an adaptive weighted covariance matrix, which well characterizes geometric structure of data. We summarize the pseudo code of solving our model (7) in algorithm 1.

Algorithm 1 Algorithm to solve the objective function (7)

Input: data set $\left\{\mathbf{x}_i \in \mathbf{R}^d : i = 1, 2, \cdots, n\right\}, k$.where $\mathbf{x}_i$ is normalized.
Initialize $\mathbf{W}^{(1)} \in \mathbf{R}^{d\times k}$ which satisfies $\mathbf{W}^T\mathbf{W} = \mathbf{I}_k$, $t = 1$.
repeat
1. For all training samples, calculate $s_{ij}^{(t)}$ and $d_{ii}^{(t)}$ by Eq. (8), i.e., $s_{ij}^{(t)} = \left\|\mathbf{W}^T(\mathbf{x}_i - \mathbf{x}_j)\right\|_2^{p-2}|_{\mathbf{W}=\mathbf{W}^{(t)}}$, $d_{ii}^{(t)} = \sum_j s_{ij}^{(t)}$.
2. Calculate $\mathbf{L}^{(t)} = \mathbf{D}^{(t)} - \mathbf{S}^{(t)}$.
3. Calculate $\mathbf{H}^{(t)} = \mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{W}|_{\mathbf{W}=\mathbf{W}^{(t)}}$.
4. Calculate the singular value decomposition (SVD) of matrix $\mathbf{H}^{(t)}$ by $\mathbf{H}^{(t)} = \mathbf{U}^{(t)}\mathbf{Q}(\mathbf{V}^{(t)})^T$.
5. Solve $\mathbf{W}^{(t+1)} = \arg\max\ tr\left(\mathbf{W}^T\mathbf{H}^{(t)}\right)$ by Eq. (19). i.e. $\mathbf{W}^{(t+1)} = \mathbf{U}^{(t)}(\mathbf{V}^{(t)})^T$.
6. Update t: $t \leftarrow t + 1$.
until $noChange$ is $true$
Output: $\mathbf{W}^{(t+1)} \in \mathbf{R}^{d\times k}$

Convergence analysis

Theorem 4: Algorithm 1 will converge to a local optimal solution of the objective function (7).

Proof: The Lagrangian function of the problem (7) is:

$$L(\mathbf{W}) = \sum_{i,j}\left\|\mathbf{W}^T(\mathbf{x}_i - \mathbf{x}_j)\right\|_2^p - tr(\mathbf{\Lambda}^T(\mathbf{W}^T\mathbf{W} - \mathrm{I})) \tag{20}$$

where the Lagrangian multiplies $\mathbf{\Lambda} = (\mathbf{\Lambda}_{pq})$ for enforcing the orthonormal constrains $\mathbf{W}^T\mathbf{W} = \mathrm{I}$. The KKT condition for optimal solution specifies that the gradient of L must be zero, i.e.,

$$\frac{\partial L}{\partial \mathbf{W}} = p\sum_{i,j}(\mathbf{x}_i\mathbf{x}_i^T\mathbf{W} - \mathbf{x}_i\mathbf{x}_j^T\mathbf{W})s_{ij} - \mathbf{W}\mathbf{\Lambda} = 0 \tag{21}$$

By simple algebra, we have

$$p\sum_{i,j}(\mathbf{x}_i\mathbf{x}_i^T\mathbf{W} - \mathbf{x}_i\mathbf{x}_j^T\mathbf{W})s_{ij} = \mathbf{W}\mathbf{\Lambda} \tag{22}$$

According to step 5 in Algorithm 1, we find the optimal solution of the objective function (9). Thus the converged solution of Algorithm 1 satisfies the KKT condition of the problem (9). The Lagrangian function of the problem (9) is

$$L_2(\mathbf{W}) = 2tr(\mathbf{W}^T\mathbf{H}) - tr(\mathbf{\Lambda}_1^T(\mathbf{W}^T\mathbf{W} - \mathrm{I})) \tag{23}$$

where $\mathbf{H} = \sum_{i,j}(\mathbf{x}_i{\mathbf{x}_i}^T\mathbf{W} - \mathbf{x}_i{\mathbf{x}_j}^T\mathbf{W})s_{ij}$.

Taking the derivative of Eq. (23) w.r.t. $\mathbf{W}$ and setting it to zero, we have

$$\mathbf{H} = \mathbf{W}\mathbf{\Lambda}_1 \tag{24}$$

Eq. (24) is formally similar to Eq. (22). The main difference between Eq. (22) and Eq. (24) is that $\mathbf{W}$ of $\mathbf{H}$ is known in each iteration in Algorithm 1. Suppose we obtain the optimal solution $\mathbf{W}^*$ in the $(t+1)$-th, thus, we have $\mathbf{W}^{(t)} = \mathbf{W}^* = \mathbf{W}^{(t+1)}$. According to the definition of s_{ij}, we can see that Eq. (24) is the same as Eq. (22) in this case. It means that the converged solution of Algorithm 1 satisfies the KKT condition of Eq. (7), i.e.,

$$\left.\frac{\partial L}{\partial \mathbf{W}}\right|_{\mathbf{W}=\mathbf{W}^*} = 0 \tag{25}$$

Thus, the converged solution of Algorithm 1 is a local solution of Eq. (7). ■

Experimental Analysis

In this section, we verify the effectiveness of our method (p=0.5, 1 and 1.5) and compare the reconstruction error of the proposed approach with robust PCA with non-greedy l_1-norm maximization (RPCA) (Nie et al. 2011), optimal mean robust PCA (RPCA-OM) (Nie, Yuan, and Huang 2014), and avoiding optimal mean robust PCA(RPCA-AOM) (Luo et al. 2016), respectively.

Experimental Setup

In the experiments, we validate our approach in four face databases (ORL, COIL20, UMIST and LFWCrop). In each database, we normalize each initial feature of image into [0, 1] and randomly select 20% images and randomly place a 1/4 size of occlusion in the selected images. Furthermore, we use the following reconstruction error to measure the quality of dimensionality reduction method:

$$e = \frac{1}{n}\sum_{i=1}^{n}\left\|\mathbf{x}_i^{clean} - \mathbf{W}\mathbf{W}^T\mathbf{x}_i^{clean}\right\|_2 \tag{26}$$

where n is the number of training data, $\mathbf{W}$ is the learned projection matrix. $\mathbf{x}_i{}^{clean}$ is the i-th clean training sample.

Reconstruction error comparison for robust PCA

The ORL database (Anton et al. 1996) contains ten different images of each of 40 distinct subjects with the resolution 32x32 and illumination changes. For some subjects, the images were taken at different times, facial expressions (open/closed eyes, smiling) and facial details (glasses). All the images were taken against a dark homogeneous background with the subjects in an upright, frontal position (with tolerance for some side movement). Some sample images of this database are shown in Figure 1. In our experiments, we randomly selected half of the images from each object to form the training set and the remaining images as testing set.

The UMIST (Phillips, Bruce, and Soulie 2012) database consists of 380 face images of 20 objects, each category contains 19 pictures. All the images were taken in 19 different postures. For each category, each image was normalized to 112×92 pixels. Some sample images of this database are shown in Figure 2. In this database, we randomly selected 10 images per person as the training set, and the remaining images for testing.

The COIL20 (Nene et al. 1996) database includes 1440 color images of 20 objects (72 images per object). Each

Figure 1: Some samples in the ORL dataset. The second row is noised images.

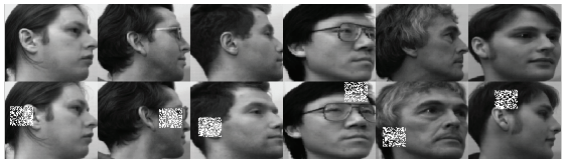

Figure 2: Some samples in the UMIST dataset. The second row is noised images.

Figure 3: Some samples in the COIL20 dataset. The second row is noised images.

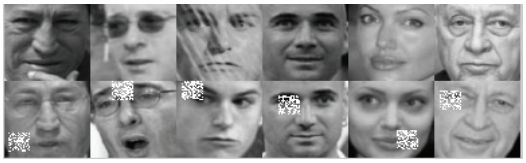

Figure 4: Some samples in the LFWCrop dataset. The second row is noised images.

object, of size 64x64 pixels,was placed in a stable configuration at approximately the center of the turntable. The turntable was then rotated through 360 degrees and 72 images were taken per object, one at every 5 degrees of rotation. Some sample images of this database are shown in Figure 3. In this database, we randomly selected 20 images of each class as the training images, and the remaining images for testing.

The LFWCrop database (Sanderson and Lovell 2009) is a cropped version of the Labeled Faces in the Wild (LFW) dataset, keeping only the center portion of each image (i.e. the face). The extracted area was then scaled to a size of 64x64 pixels. The selection of the bounding box location was based on the positions of 40 randomly selected LFW faces. As the location and size of faces in LFW was determined through the use of an automatic face locator (detector), the cropped faces in LFWCrop exhibit real-life conditions, including misalignment, scale variations, in-plane as well as out-of-plane rotations. Some sample images of this database are shown in Figure 4. In the experiments, we chose person who have more than 20 photos but less than 100 photos as the sub-dataset, which contains 57 classes and 1883 images. In the sub-dataset, we randomly selected ninety percent of images per person for training, and the remaining images for testing.

Table 1 lists reconstruction error versus different dimensions of four methods on four databases. Table 2 lists the average reconstruction error and the corresponding standard deviation of four methods on four databases. Table 3 lists the average time-consuming of each methods on the Coil20 and ORL databases. Figure 5 presents the convergence curve of our method on four databases. Figure 6 presents the reconstruction error of our method under different values of p on the Coil20 dataset.

- RPCA is overall inferior to the other three methods. The main reason is that RPCA use the fixed mean, which is not optimal under the ℓ_1-norm distance metric. This affects the robustness of RPCA.
- RPCA-AOM is overall superior to RPCA-OM. This may be that RPCA-AOM avoids mean computation in the criterion function, while RPCA-OM cannot obtain the optimal mean. However, in some cases, RPCA-AOM is inferior to RPCA-OM. This is probably because that solution of RPCA-OM relates to covariance matrix that characterizes the geometric structure, while RPCA-AOM does not.

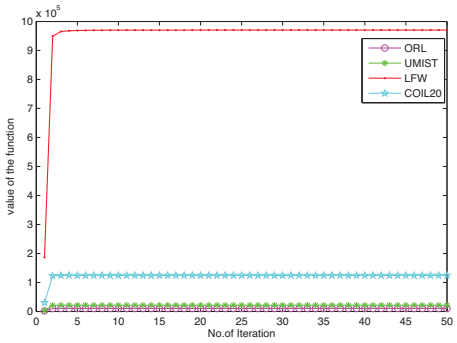

Figure 5: Convergence curve of our method on four databases.

- Our method is overall superior to the other three methods on all databases. This is probably because that our method not only avoids mean estimation but also relates to the adaptive weighted covariance matrix, which well characterizes the geometric structure of data.
- As can be seen in Table 1, Table 2 and Figure 6, our method has better performance when p is small in most cases. This is probably because that it can further suppress the effect of outliers when p is small.
- Figure 5 illustrates that our method quickly converges (about five steps) and has local optimal solution. Table 3 illustrates that our method is faster than RPCA-OM and RPCA-AOM, which involve mean calculation.

Table 1: Reconstruction error versus different dimensions of four methods on four databases. The best reconstruction result under each dimension is bolded.

COIL20	Dimension	10	15	20	25	30	35	40	45	50
	RPCA	0.9709	0.9441	0.9321	0.8983	0.8662	0.8418	0.8118	0.7818	0.7462
	RPCA-OM	0.9080	0.8960	0.8792	0.8704	0.8543	0.8406	0.8184	0.7789	0.7305
	RPCA-AOM	0.9644	0.95.9	0.8850	0.8700	0.8460	0.8335	0.8144	0.7931	0.7578
	P=0.5(our)	**0.8169**	**0.8060**	**0.7909**	**0.7771**	0.7685	0.7534	**0.7206**	0.7082	0.6937
	P=1(our)	0.8407	0.8065	0.7951	0.7779	**0.7663**	**0.7509**	0.7394	0.7190	0.7070
	P=1.5(our)	0.8457	0.8205	0.7984	0.7894	0.7723	0.7514	0.7350	**0.7031**	**0.6894**
ORL	Dimension	10	15	20	25	30	35	40	45	50
	RPCA	0.9874	0.9808	0.9758	0.9511	0.9465	0.9237	0.9047	0.8984	0.8929
	RPCA-OM	0.9731	0.9655	0.9580	0.9439	0.9323	0.9185	0.9143	0.9060	0.8973
	RPCA-AOM	0.9791	0.9372	0.9213	0.8952	0.8758	0.8686	0.8565	0.8524	0.8459
	p=0.5(our)	0.8741	0.8690	0.8632	0.8581	0.8492	0.8434	0.8283	**0.8137**	0.8037
	p=1(our)	**0.8701**	**0.8601**	**0.8546**	**0.8446**	**0.8384**	**0.8311**	0.8267	0.8177	0.8110
	p=1.5(our)	0.8795	0.8736	0.8671	0.8527	0.8446	0.8327	**0.8214**	0.8149	**0.8010**
UMIST	Dimension	10	15	20	25	30	35	40	45	50
	RPCA	0.9748	0.9511	0.9276	0.9155	0.8969	0.8806	0.8565	0.8460	0.8301
	RPCA-OM	0.9494	0.9297	0.9053	0.8995	0.8909	0.8690	0.8541	0.8407	0.8344
	RPCA-AOM	0.9561	0.9444	0.9235	0.9122	0.8758	0.8643	0.8383	0.8268	0.8176
	p=0.5(our)	**0.8873**	**0.8718**	**0.8522**	**0.8465**	**0.8285**	**0.8068**	**0.7989**	**0.7853**	**0.7708**
	p=1(our)	0.9063	0.8979	0.8834	0.8799	0.8731	0.8548	0.8487	0.8230	0.8030
	p=1.5(our)	0.9087	0.8894	0.8774	0.8597	0.8533	0.8275	0.8141	0.8031	0.7855
LFWCrop	Dimension	10	15	20	25	30	35	40	45	50
	RPCA	0.8557	0.8135	0.8052	0.7919	0.7804	0.7543	0.7416	0.7252	0.7062
	RPCA-OM	0.9528	0.9439	0.9174	0.8976	0.8519	0.8152	0.7117	0.6642	0.5480
	RPCA-AOM	0.9463	0.8885	0.8705	0.8342	0.7729	0.7487	0.7202	0.7101	0.6975
	p=0.5(our)	**0.6848**	**0.6645**	**0.6486**	**0.6297**	**0.6100**	0.5939	0.5831	0.5505	**0.5102**
	p=1(our)	0.6872	0.6804	0.6599	0.6502	0.6307	0.6250	0.6002	0.5630	0.5525
	p=1.5(our)	0.6874	0.6762	0.6696	0.6481	0.6210	**0.5836**	**0.5665**	**0.5468**	0.5206

Table 2: The average reconstruction error and the corresponding standard deviation of four methods on four databases.

Methods	RPCA	RPCA-OM	RPCA-AOM	p=0.5(our)	p=1(our)	p=1.5(our)
ORL	0.8882±0.0092	0.8870±0.0099	0.8689±0.0155	0.8141±0.0119	0.8125±0.0119	0.8069±0.0145
COIL20	0.7680±0.0348	0.7389±0.0253	0.7569±0.0128	0.6830±0.0098	0.6888±0.0104	0.6836±0.0235
UMIST	0.8415±0.0130	0.8493±0.0104	0.8314±0.0221	0.8032±0.0206	0.7996±0.0236	0.8069±0.0177
LFWCrop	0.7117±0.0411	0.5458±0.0208	0.6827±0.0209	0.5443±0.0257	0.5290±0.0196	0.5288±0.0279

Table 3: Average Time Consuming Of Each Method On ORL and Coil20 Databases (Measured In Minutes.)

Database	RPCA	RPCA -OM	RPCA -AOM	Our method
Coil20	10.31 ±0.05	19.79 ±0.72	312.12 ±2.08	17.81 ±0.33
ORL	0.44 ±0.02	1.78 ±0.10	14.25 ±0.17	0.92 ±0.02

Conclusion

In this paper, we present a novel robust PCA formulation for dimensionality reduction. Our method employs $l_{2,p}$-norm as the distance metric to measure variance and learns the projection matrix by maximizing the variance between each pair of data. Compared with most existing robust PCA methods, our proposed method not only avoids estimating the optimal mean and but also well characterizes the geometric structure of data. We provide an efficient iteration algorithm, which has a close-form solution in each iteration, to solve the local optimal solution. Experimental results on LFWCrop, ORL, COIL20, and UMIST databases illustrate the superiority of our method.

Acknowledgements

This work is supported by National Natural Science Foundation of China under Grant 61271296 and 61773302, and the 111 Project of China (B08038).

References

Anton, B.; Fein, J.; To, T.; Li, X.; Silberstein, L.; and Evans, C. J. 1996. Immunohistochemical localization of orl-1 in the central nervous system of the rat. *Journal of Comparative Neurology* 368(2):229–251.

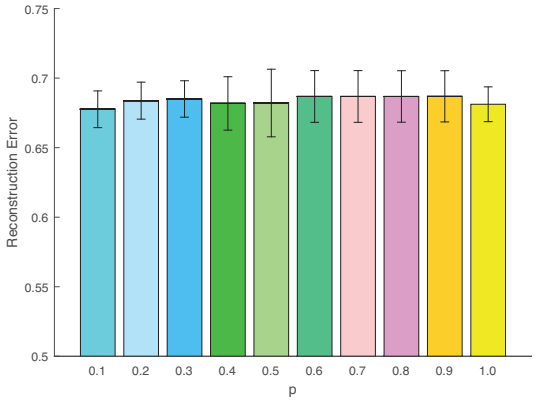

Figure 6: Reconstruction error vs. p on the Coil20 dataset.

Belhumeur, P. N.; Hespanha, J. P.; 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.

Ding, C.; Zhou, D.; He, X.; and Zha, H. 2006. R1-pca:rotational invariant l1-norm principal component analysis for robust subspace factorization. In *International Conference on Machine Learning*, 281–288.

Gao, Q.; Gao, F.; Zhang, H.; Hao, X.; and Wang, X. 2013. Two-dimensional maximum local variation based on image euclidean distance for face recognition. *IEEE Trans. Image Processing* 22(10):3807–3817.

Gao, Q.; Ma, L.; Liu, Y.; Gao, X.; and Nie, F. 2017. Angle 2dpca: A new formulation for 2dpca. *IEEE Transactions on Cybernetics* Digital Object Identifier: 10.1109/TCYB.2017.2712740.

He, R.; Hu, B.; Zheng, W.; and Kong, X. 2011. Robust principal component analysis based on maximum correntropy criterion. *IEEE Transactions on Image Processing* 20(6):1485–1494.

Jiang, B.; Ding, C.; Luo, B.; and Tang, J. 2013. Graph-laplacian pca: Closed-form solution and robustness. In *Computer Vision and Pattern Recognition*, 3492–3498.

Ju, F.; Sun, Y.; Gao, J.; Hu, Y.; and Yin, B. 2015. Image outlier detection and feature extraction via l1-norm-based 2d probabilistic pca. *IEEE Trans. Image Processing* 24(12):4834–4846.

Ke, Q., and Kanade, T. 2005. Robust l1-norm factorization in the presence of outliers and missing data by alternative convex programming. In *In Proc. Computer Vision and Pattern Recognition*, volume 1, 739–746.

Kwak, N. 2008. Principal component analysis based on l1-norm maximization. *IEEE Trans. Pattern Anal. Mach. Intell.* 30(9):1672–1680.

Li, X.; Ma, Y.; and Wright, J. 2009. Robust principal component analysis? *Journal of the ACM* 58(3):1–79.

Luo, M.; Nie, F.; Chang, X.; Yang, Y.; Hauptmann, A.; and Zhang, Q. 2016. Avoiding optimal mean robust pca/2dpca with non-greedy l1-norm maximization. In *International Joint Conference on Artificial Intelligence*, 1802–1808.

Markopoulos, P.; Karystinos, G.; and Pados, D. A. 2014. Optimal algorithms for-subspace signal processing. *IEEE Trans. Signal Processing* 62(19):5046–5058.

Nene, S. A.; Nayar, S. K.; Murase, H.; Nene, S. A.; Nayar, S. K.; and Murase, H. 1996. Columbia object image library (coil-20).

Nie, F.; Huang, H.; Ding, C.; Luo, D.; and Wang, H. 2011. Robust principal component analysis with non-greedy l1-norm maximization. In *International Joint Conference on Artificial Intelligence*, 1433–1438.

Nie, F.; Yuan, J.; and Huang, H. 2014. Optimal mean robust principal component analysis. In *International Conference on Machine Learning*, 1062–1070.

Oh, J., and Kwak, N. 2016. Generalized mean for robust principal component analysis. *Pattern Recognition* 54:116–127.

Phillips, J.; Bruce, V.; and Soulie, F. F. 2012. Face recognition: From theory to applications. *Columbia University*.

Sanderson, C., and Lovell, B. C. 2009. Multi-region probabilistic histograms for robust and scalable identity inference. In *International Conference on Biometrics*, volume 5558, 199–208. Springer.

Shahid, N.; Kalofolias, V.; Bresson, X.; and Bronstein, M. 2015. Robust principal component analysis on graphs. In *IEEE International Conference on Computer Vision*, 2812–2820.

Song, Z.; Woodruff, D. P.; and Zhong, P. 2017. Low rank approximation with entrywise l1-norm error. In *The ACM Sigact Symposium*, 688–701.

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

Wang, Q., and Gao, Q. 2017. Two-dimensional pca with f-norm minimization. In *Association for the Advancement of Artificial Intelligence(AAAI)*, 2718–2724.

Wang, H., and Wang, J. 2013. 2dpca with l1-norm for simultaneously robust and sparse modeling. *Neural Networks* 46(10):190–198.

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.; Gao, Q.; Gao, X.; and Nie, F. 2017. Optimal mean two-dimensional principal component analysis with f-norm minimization. *Pattern Recognition* 68(2):286–294.