Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16)

Avoiding Optimal Mean Robust PCA/2DPCA with Non-Greedy ℓ_1-Norm Maximization

Minnan Luo,[1,4] Feiping Nie,[2*] Xiaojun Chang,[3] Yi Yang,[3]
Alexander Hauptmann,[4] Qinghua Zheng[1]
[1] Shaanxi Province Key Lab of Satellite-Terrestrial Network , Department of Computer Science,
Xi'an Jiaotong University, P. R. China.
[2] School of Computer Science and Center for Optical Imagery Analysis and Learning,
Northwestern Polytechnical University, P. R. China.
[3]Centre for Quantum Computation and Intelligent Systems, University of Technology Sydney.
[4]School of Computer Science, Carnegie Mellon University, PA, USA

Abstract

Robust principal component analysis (PCA) is one of the most important dimension reduction techniques to handle high-dimensional data with outliers. However, the existing robust PCA presupposes that the mean of the data is zero and incorrectly utilizes the Euclidean distance based optimal mean for robust PCA with ℓ_1-norm. Some studies consider this issue and integrate the estimation of the optimal mean into the dimension reduction objective, which leads to expensive computation. In this paper, we equivalently reformulate the maximization of variances for robust PCA, such that the optimal projection directions are learned by maximizing the sum of the projected difference between each pair of instances, rather than the difference between each instance and the mean of the data. Based on this reformulation, we propose a novel robust PCA to automatically avoid the calculation of the optimal mean based on ℓ_1-norm distance. This strategy also makes the assumption of centered data unnecessary. Additionally, we intuitively extend the proposed robust PCA to its 2D version for image recognition. Efficient non-greedy algorithms are exploited to solve the proposed robust PCA and 2D robust PCA with fast convergence and low computational complexity. Some experimental results on benchmark data sets demonstrate the effectiveness and superiority of the proposed approaches on image reconstruction and recognition.

1 Introduction

High-dimensional data are frequently generated in many scientific domains, such as image processing, visual description, remote sensing, time series prediction and gene expression. However, it is usually computationally expensive to handle high-dimensional data due to the curse of dimensionality [Parsons *et al.*, 2004; Chang *et al.*, 2015; Nie and Huang, 2016; Nie *et al.*, 2010]. Therefore, dimension reduction techniques are typically used to extract meaningful features from high-dimensional data without degrading performance. Among these methods, principal component analysis (PCA) learns a set of projections that constitute a low-dimensional linear subspace. It has been widely used in many applications for its simplicity and effectiveness [Jolliffe, 2002].

Typically, standard PCA is based on mean square error, and thus it is disproportionately affected by the presence of outliers which occur often in high-dimensional data. For this issue, multiple robust PCA methods have been proposed to enhance the robustness of PCA by replacing the ℓ_2-norm with ℓ_1-norm distance [Wright *et al.*, 2009; Torre and Black, 2001; Jolliffe, 2002; Ke and Kanade, 2005; Chang *et al.*, 2016]. However, ℓ_1-norm based robust PCA methods usually perform worse due to their lack of rotational invariance and expensive computation. To solve this problem, a rotational invariant ℓ_1-norm based robust PCA, namely R_1-PCA, has been proposed to soften the contributions from outliers by re-weighting each data point iteratively [Ding *et al.*, 2006]. This method was further extended to its 2D version in [Huang and Ding, 2008; Yang *et al.*, 2004]. However, the R_1-PCA/2DPCA models are solved with subspace iteration algorithm, which requires a lot of time to achieve convergence [Kwak, 2008]. Kwak proposed an intuitive method to ensure both the robustness and rotational invariance of PCA by maximizing the ℓ_1-norm of variance with a greedy algorithm [Kwak, 2008]. The corresponding 2D and supervised versions can be found in [Liu *et al.*, 2010; Li *et al.*, 2010; Pang *et al.*, 2010]. However, the greedy algorithm optimizes the projection directions one by one, which makes it easy to get stuck in a local solution. For this issue, Nie *et al.* [Nie *et al.*, 2011] exploited an efficient non-greedy optimization algorithm to optimize all projection directions simultaneously for the ℓ_1-norm maximization problem; Kwak [Kwak, 2014] extended the non-greedy algorithm to an ℓ_p-norm based maximization problem. The corresponding robust 2DPCA with

*Corresponding author. Email: feipingnie@gmail.com

non-greedy algorithm can be found in [Wang *et al.*, 2015].

However, the ℓ_1-norm based robust PCA mentioned above usually uses the mean of the data as the optimal mean [Nie *et al.*, 2014]; moreover, it assumes that the data are already centered, i.e., the average of the data is zero. However, this assumption is unreasonable for the following three reasons: (1) It's hard to ensure the zero mean in real-world applications; (2) The outliers in high-dimensional data often make the data mean biased, which degrades the robustness of PCA [He *et al.*, 2011]. (3) It ignores the data mean calculation and incorrectly uses the average of the data as the optimal mean for ℓ_1-norm based robust PCA. However, the average of the data is the optimal mean for conventional PCA based on a Euclidean distance.

There are relatively few studies which take these important issues into consideration. To the best of our knowledge, He *et al.* [He *et al.*, 2011] proposed a robust PCA based on n maximum correntropy criterion and handled non-centered data with an estimation of optimal mean; Nie *et al.* [Nie *et al.*, 2014] introduced a mean variable and exploited a novel robust PCA objective with optimal mean. Nevertheless, both of these methods integrate the mean calculation into the optimization objective and lead to expensive computation.

In this paper, we develop an equivalent reformulation of the maximization of ℓ_2 variances for conventional PCA, such that the optimal projection directions are learned via maximizing the sum of projected difference between each pair of instances instead of the difference between each instance and the mean of the data. Based on this reformulation, we propose a new robust PCA by maximizing the sum of projected differences between each pair of instances based on the ℓ_1-norm distance. This method automatically avoids calculating the ℓ_1-norm based optimal mean and makes the assumption of centered data unnecessary. An efficient non-greedy method is further exploited to maximize the objective with fast convergence in practical application. Intuitively, we also extend the proposed robust PCA to its 2D version for image recognition. It is noteworthy that the proposed algorithms keep linear computation complexity with respect to the number of data points in practical application.

2 Principal Component Analysis Review

Suppose the given data matrix is $X = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n] \in \mathbb{R}^{d\times n}$, where each instance $\mathbf{x}_i$ is represented by a vector with d-dimensionality; n refers to the number of instances. Conventional PCA learns a transformation to map high dimensional data to low dimensional representations. Specifically, let $W = [\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_m] \in \mathbb{R}^{d\times m}$ be a semi-orthogonal transformation matrix, the idea of traditional PCA is formulated by minimizing the reconstruction error based on ℓ_2-norm distance in the original high-dimensional space, *i.e.*

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

where $\mathbf{m}$ is the mean of the data. By setting the derivative of objective function (1) with respect to $\mathbf{m}$ to zero, we obtain the optimal mean of the data as $\mathbf{m} = \bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$. With some evident transformations, optimization problem (1) can be reformulated as the maximization of covariance in the projected space, i.e.,

$$\max_{W^\top W=I} \sum_{i=1}^{n} \|W^\top(\mathbf{x}_i - \bar{\mathbf{x}})\|_2^2. \quad (2)$$

Generally, the mean $\bar{\mathbf{x}}$ is supposed to be zero; otherwise, it is subtracted from each instance of optimization problem (2). In such a way, the orthogonal transformation matrix W can be solved by maximizing the following optimization problem

$$\max_{W^\top W=I} \sum_{i=1}^{n} \|W^\top\mathbf{x}_i\|_2^2, \quad (3)$$

where the instances $\mathbf{x}_i$ $(i = 1, 2, \cdots, n)$ are centered. PCA has been widely applied in many applications for its efficiency and simplicity However, the high computational complexity and the outlier sensitivity induced by ℓ_2-norm make it hard to apply for a large scale data with high dimensionality [Nie *et al.*, 2011]. As a result, robust PCA is proposed by directly substitute ℓ_1-norm for ℓ_2-norm maximization in optimization problem (3)[Kwak, 2008; Galpin and Hawkins, 1987; Nie *et al.*, 2011], i.e.,

$$\max_{W^\top W=I} \sum_{i=1}^{n} \|W^\top\mathbf{x}_i\|_1. \quad (4)$$

Nevertheless, the existing robust PCA methods and its various neglect the optimal mean calculation and incorrectly utilize $\bar{\mathbf{x}}$ as the optimal mean. However, $\bar{\mathbf{x}}$ is definitely the optimal mean in terms of ℓ_2-norm distance, rather than the ℓ_1-norm used in the objective functions of RPCA. Nie *et al.* [Nie *et al.*, 2014] consider this issue and propose a new robust PCA by integrating the optimization of optimal mean into the dimension reduction objective; however, it leads to expensive computation.

3 The Proposed Methodology

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_1-norm distance. This method automatically avoids calculating the optimal mean with ℓ_1-norm distance and makes the assumption of centered data unnecessary. For a better representation, we first introduce the following theorem. It reformulates the objective of conventional PCA as maximizing the sum of the projected difference between each pair of instances.

Theorem 1. *Let $X = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n] \in \mathbb{R}^{d\times n}$ be the data matrix and $\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$ be the mean of X. The solution W of conventional PCA which minimizes the reconstruction error based on Euclidean distance,* i.e.*,*

$$\min_{W^\top W=I,\mathbf{m}} \sum_{i=1}^{n} \|(\mathbf{x}_i - \mathbf{m}) - WW^\top(\mathbf{x}_i - \mathbf{m})\|_2^2, \quad (5)$$

is also the solution of the following optimization problem

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

Proof. With fixed W satisfying $W^\top W = I$, we set the derivative of objective function (5) with respect to variable $\mathbf{m}$ to zero. Then we have $\mathbf{m} = \bar{\mathbf{x}}$. We substitute $\mathbf{m} = \bar{\mathbf{x}}$ into the objective function (5) and reformulate it as the maximization of covariance in the projected space, *i.e.*,

$$\max_{W^\top W=I} \sum_{i=1}^{n} \|W^\top(\mathbf{x}_i - \bar{\mathbf{x}})\|_2^2. \quad (7)$$

We go further to substitute $\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$ into the objective function above and achieve the following equation

$$\begin{aligned}&\sum_{i=1}^{n} \|W^\top(\mathbf{x}_i - \bar{\mathbf{x}})\|_2^2 \\ =&\sum_{i=1}^{n} \mathbf{x}_i^\top W W^\top \mathbf{x}_i - \frac{1}{n}\sum_{i,j}\mathbf{x}_i^\top W W^\top \mathbf{x}_j; \end{aligned} \quad (8)$$

On the other hand, it is evident to reformulate the objective function of optimization problem (6) through equivalent transformation as

$$\begin{aligned}&\sum_{i,j} \|W^\top(\mathbf{x}_i - \mathbf{x}_j)\|_2^2 \\ =& 2n\sum_{i=1}^{n} \mathbf{x}_i^\top W W^\top \mathbf{x}_i - 2\sum_{i,j}\mathbf{x}_i^\top W W^\top \mathbf{x}_j. \end{aligned} \quad (9)$$

According to Eq. (8) and Eq. (9), the proof is completed. □

Based on Theorem 1, we equivalently reformulate the ℓ_2-norm based PCA as the following optimization problem,

$$\max_{W^\top W=I} \sum_{i,j} \|W^\top(\mathbf{x}_i - \mathbf{x}_j)\|_2^2. \quad (10)$$

As opposed to the conventional PCA formulated by optimization problem (5) or (2), the alternative formulation (10) estimate the transformation matrix with the calculation of optimal mean avoided automatically. However, considering the sensitivity of ℓ_2-norm distance to outliers, in this paper, we propose a novel robust PCA based on ℓ_1-norm by solving the following optimization problem

$$\max_{W^\top W=I} \sum_{i,j} \|W^\top(\mathbf{x}_i - \mathbf{x}_j)\|_1. \quad (11)$$

Note that existing robust PCA which directly replaces the ℓ_2 norm in optimization problem (3) with ℓ_1-norm and incorrectly employs $\bar{\mathbf{x}}$ as the optimal mean of ℓ_1-norm based robust PCA. The proposed objective (11) automatically avoids calculating the ℓ_1-norm based optimal mean and makes the assumption on centered data unnecessary.

3.1 Optimization procedure

In this section, we use an efficient iterative re-weighted algorithm [Nie *et al.*, 2011] to solve the non-smooth optimization problem (11). This algorithm is first proposed to solve general optimization problem,

$$\max_{\mathbf{z}\in\mathcal{C}} \mathcal{L}(\mathbf{z}) = f(\mathbf{z}) + \sum_i |g_i(\mathbf{z})| \quad (12)$$

Algorithm 1 Non-greedy ℓ_1-norm maximization.

Initialize: $\mathbf{z}^{(1)} \in \mathcal{C}$, $k = 1$.
1: **while** not converge **do**
2: $v_i^{(k)} = sgn(g_i(\mathbf{z}^{(k)}))$ for each i;
3: $\mathbf{z}^{(k+1)} = \arg\max_{\mathbf{z}\in\mathcal{C}} f(\mathbf{z}) + \sum_i v_i^{(k)} g_i(\mathbf{z})$;
4: $k = k + 1$;
5: **end while**
Output: $\mathbf{z}^{(k)}$.

where $\mathbf{z} \in \mathcal{C}$ is an arbitrary constraint; f and g_i are arbitrary functions defined on $\mathcal{C}$ for each i. Let $v_i = sgn(g(\mathbf{z}))$ with element-wise sign function $sgn(\cdot)$, then the objective function $\mathcal{L}(\mathbf{z})$ can be reformulated as

$$\mathcal{L}(\mathbf{z}) = f(\mathbf{z}) + \sum_i v_i g_i(\mathbf{z}). \quad (13)$$

As a result, the general optimization problem (12) can be solved with an non-greedy re-weighted algorithm which is described in Algorithm 1.

We follow Algorithm 1 and exploit an non-greedy method to solve the proposed optimization problem (11). The key step lies in addressing the following optimization problem

$$\max_{W^\top W=I} \sum_{i,j} \mathbf{v}_{ij}^\top W^\top(\mathbf{x}_i - \mathbf{x}_j), \quad (14)$$

where $\mathbf{v}_{ij} = sgn((W^{(k)})^\top(\mathbf{x}_i - \mathbf{x}_j))$ is a vector with m-dimensionality. Let $R = \sum_{ij}(\mathbf{x}_i - \mathbf{x}_j)\mathbf{v}_{ij}^\top \in \mathbb{R}^{d\times m}$, then we have $\sum_{i,j}\mathbf{v}_{ij}^\top W^\top(\mathbf{x}_i - \mathbf{x}_j) = Tr(W^\top R)$. As a result, optimization problem (14) can be rewritten as

$$\max_{W^\top W=I} Tr(W^\top R). \quad (15)$$

We solve optimization problem (15) based on the following Theorem 2.

Theorem 2. *Suppose the SVD of R is $R = P\Lambda Q^\top$, where $P \in \mathbb{R}^{d\times d}$, $\Lambda \in \mathbb{R}^{d\times m}$ and $Q \in \mathbb{R}^{m\times m}$. The solution of optimization problem (15) is derived as $W = P[I;\mathbf{0}]Q^\top$.*

Proof. Based on the SVD of R, we have

$$\begin{aligned} Tr(W^\top R) &= Tr(W^\top P\Lambda Q^\top) = Tr(\Lambda Q^\top W^\top P) \\ &= Tr(\Lambda\Psi) = \sum_k \lambda_{kk}\psi_{kk}, \end{aligned}$$

where $\Psi = Q^\top W^\top P$; λ_{kk} and ψ_{kk} represent the (k,k)-th element of matrices Λ and Ψ, respectively. Recall that the constraint $W^\top W = I$, so we have $\Psi\Psi^\top = I$ and $\psi_{kk} \leq 1$, where I is an m by m identity matrix. Combining with the fact $\lambda_{kk} \geq 0$ since λ_{kk} is singular value of R, we arrive at the following inequality

$$Tr(W^\top R) = \sum_k \lambda_{kk}\psi_{kk} \leq \sum_k \lambda_{kk}, \quad (16)$$

where the equality holds when $\psi_{kk} = 1$ $(1 \leq k \leq m)$. As a result, the objective function (15) reaches its maximum when $\Psi = [I, \mathbf{0}]$. Recall that $\Psi = Q^\top W^\top P$, the optimal solution to optimization problem (15) is $W = P\Psi^\top Q^\top = P[I;\mathbf{0}]Q^\top$. The proof is completed. □

Algorithm 2 Robust PCA with non-greedy ℓ_1-norm maximization

Input: data set $\{\mathbf{x}_i \in \mathbb{R}^d : i = 1, 2, \cdots, n\}$, m.
Initialize: $W^{(1)} \in \mathbb{R}^{d\times m}$ s.t. $(W^{(1)})^\top W^{(1)} = I, t = 1$.

while not converge **do**
  $\mathbf{v}_{i,j} = sgn((W^{(t)})^\top(\mathbf{x}_i - \mathbf{x}_j))\ (\forall i < j)$;
  $R = \sum_{i,j}(\mathbf{x}_i - \mathbf{x}_j)\mathbf{v}_{ij}^\top$;
  Calculate the SVD of R as $R = P\Lambda Q^\top$, then $W^{(t+1)} = PQ^\top$;
  $t = t + 1$;
end while

Output: $W^t \in \mathbb{R}^{d\times m}$.

In summary, we describe the non-greedy ℓ_1-norm maximization algorithm in Algorithm 2 for optimization problem (14). Theoretical analysis in [Nie *et al.*, 2011] ensures that the proposed non-greedy algorithm will convergence and usually obtain a local maximum solution within ten iterations in practical application.

3.2 Complexity discussion

In this section, we analyze the computational complexity of the proposed Algorithm 2. Given $F = [\mathbf{f}_1, \mathbf{f}_2, \cdots, \mathbf{f}_n] = W^\top X \in \mathbb{R}^{m\times n}$ with computational complexity $\mathcal{O}(ndm)$, we have $\mathbf{v}_{ij} = sgn(\mathbf{f}_i - \mathbf{f}_j) \in \mathbb{R}^m$ for $i, j = 1, 2, \cdots, n$. It seems like the computational cost of $\mathbf{v}_{ij}$ $(i, j = 1, 2, \cdots, n)$ is $\mathcal{O}(mn^2)$. In fact, it can be avoided with some strategies. Note that the computation of R in Algorithm 2 only depends on m-dimensional vectors $\mathbf{v}_{i\cdot} := \sum_j \mathbf{v}_{ij}^\top$ and $\mathbf{v}_{\cdot j} := \sum_i \mathbf{v}_{ij}$ due to the following equivalent transformation:

$$\begin{aligned} R &= \sum_{i,j}(\mathbf{x}_i - \mathbf{x}_j)\mathbf{v}_{ij}^\top = \sum_{i,j}\mathbf{x}_i\mathbf{v}_{ij}^\top - \sum_{i,j}\mathbf{x}_j\mathbf{v}_{ij}^\top \\ &= \sum_i \mathbf{x}_i \sum_j \mathbf{v}_{ij}^\top - \sum_j \mathbf{x}_j \sum_i \mathbf{v}_{ij}^\top \\ &= \sum_i \mathbf{x}_i\mathbf{v}_{i\cdot}^\top - \sum_j \mathbf{x}_j\mathbf{v}_{\cdot j}^\top = \sum_i \mathbf{x}_i(\mathbf{v}_{i\cdot}^\top - \mathbf{v}_{\cdot i}^\top) \end{aligned}$$

Consequently, the computational cost of R is $\mathcal{O}(nmd)$ when $\mathbf{v}_{i\cdot}$ and $\mathbf{v}_{\cdot i}$ are given. Recall that $\mathbf{v}_{ij} = sgn(\mathbf{f}_i - \mathbf{f}_j)$, thus we have $\mathbf{v}_{i\cdot} = \sum_j \mathbf{v}_{ij} = \sum_j(\mathbf{f}_i - \mathbf{f}_j)$. As a result, the k-th entry of vectors $\mathbf{v}_{i\cdot}, \mathbf{v}_{\cdot i} \in \mathbb{R}^m$ can be obtained efficiently by sorting the k-th entries of $\mathbf{f}_1, \mathbf{f}_2, \cdots, \mathbf{f}_n$ with computational complexity $\mathcal{O}(n\log(n))$. Based on this ranking, the entire computational cost over $\mathbf{v}_{i\cdot}$ and $\mathbf{v}_{\cdot i}$ $(i = 1, 2, \cdots, n)$ is $\mathcal{O}(nm\log(n))$. As a result, the computational complexity of Algorithm 2 is $\mathcal{O}(nm(\log(n) + d)t)$. However, since $\log(n)$ is usually much smaller than d in practical application, the computational complexity of Algorithm 2 reduces to $\mathcal{O}(nmdt)$. Therefore, the proposed robust PCA does not require an additional computational cost in contrast to the state of the art robust PCA in [Kwak, 2008; Nie *et al.*, 2011].

4 Extensions to 2D Version of Robust PCA

To keep the structural information of two dimensional (2D) image matrix, 2DPCA is proposed to construct an covariance matrix using the original 2D image matrices directly. In this section, we extend the proposed robust PCA to its 2D version and develop a novel robust 2DPCA with the calculation of the optimal mean avoided automatically.

We suppose that the data is denoted by $X = [X_1, X_2, \cdots, X_n] \in \mathbb{R}^{c\times d\times n}$, where each component $X_i \in \mathbb{R}^{c\times d}$ $(i = 1, 2, \cdots, n)$ refers to a image matrix; n is the number of data. Let $U = [\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_m] \in \mathbb{R}^{d\times m}$ be the collection of projection directions $\mathbf{u}_k \in \mathbb{R}^d$ for $k = 1, 2, \cdots, m$, then conventional 2DPCA learns transformation matrix U by maximizing the following optimization problem with Frobenius norm,

$$\max_{U^\top U = I, M} \sum_{i=1}^n \|(X_i - M)U\|_F^2, \tag{17}$$

where $M = \frac{1}{n}\sum_i X_i$ is the Frobenius norm based optimal mean which is usually assumed to be zero in conventional 2DPCA. However, as mentioned in the precious section, the mean of the data is not always zero in practical application.

Following Theorem 1, we rewrite the objective of optimization problem (17) and learn the directions via maximizing the sum of the projected difference between each pair of images, *i.e.*,

$$\max_{U^\top U = I} \sum_{i,j} \|(X_i - X_j)U\|_F^2. \tag{18}$$

Considering the poor robustness of Euclidean distance, we go further to replace the ℓ_F-norm used in optimization problem (1) with ℓ_1-norm and propose an alternative formulation of robust 2DPCA as

$$\max_{U^\top U = I} \sum_{i,j} \|(X_i - X_j)U\|_1 \tag{19}$$

$$\Longleftrightarrow \max_{U^\top U = I} \sum_{i,j}\sum_{k=1}^c \|U^\top(\mathbf{x}_{ik} - \mathbf{x}_{jk})\|_1 \tag{20}$$

where $\mathbf{x}_{ik}, \mathbf{x}_{jk} \in \mathbb{R}^d$ $(k = 1, 2, \cdots, c)$ are the transpose of the k-th row of image matrices X_i and X_j, respectively. We can also solve the optimization problem (20) through Algorithm 1. Thus, the key step lies in addressing the following optimization problem

$$\max_{U^\top U = I} \sum_{i,j}\sum_{k=1}^c \mathbf{v}_{ijk}^\top U^\top(\mathbf{x}_{ik} - \mathbf{x}_{jk}), \tag{21}$$

where $\mathbf{v}_{ijk} = sgn\left(U^\top(\mathbf{x}_{ik} - \mathbf{x}_{jk})\right)$ is an m-dimensional vector. Denote $S = \sum_{i,j}\sum_{k=1}^c(\mathbf{x}_{ik} - \mathbf{x}_{jk})\mathbf{v}_{ijk}^\top$, the optimization problem (21) can be rewritten as

$$\max_{U^\top U = I} Tr(U^\top S). \tag{22}$$

Assume the SVD of S is $S = ADB^\top$, we have the optimal solution to problem (22) is $U = A[I; \mathbf{0}]B^\top$ according to Theorem 2. We summarize the robust 2DPCA with non-greedy ℓ_1-norm maximization in Algorithm 3. Similar analysis in [Nie *et al.*, 2011; Wang *et al.*, 2015] ensures the convergence of the proposed Algorithm 3. Since $d \gg \log(n)$ and $n \gg m$ in practical applications, we derive the computational complexity of Algorithm 3 as $\mathcal{O}(nmcdt)$ according to similar strategies used for Algorithm 2.

Table 1: Reconstruction error comparison of three robust PCA methods on 5 benchmark data sets with different dimensions. The best reconstruction result under each dimension is bolded.

JAFFE	Dimension	10	15	20	25	30	35	40	45	50
	RPCA	0.9256	**0.8409**	0.7348	0.7380	0.6849	0.6324	0.5749	0.5603	0.5533
	RPCA-OM	0.9381	0.9094	0.8029	0.7141	0.6911	0.6380	0.6093	0.6673	0.5567
	RPCA-AOM	**0.9114**	0.8632	**0.7236**	**0.6704**	**0.6279**	**0.5851**	**0.5605**	**0.5331**	**0.5130**
UMIST	Dimension	10	15	20	25	30	35	40	45	50
	RPCA	0.9271	**0.8434**	**0.7005**	0.7272	0.6995	0.6118	0.5009	0.4459	0.4088
	RPCA-OM	0.9301	0.8722	0.7793	0.6637	0.5547	0.4901	0.4410	0.4128	0.3972
	RPCA-AOM	**0.9254**	0.8505	0.7521	**0.6478**	**0.4787**	**0.4800**	**0.4297**	**0.4056**	**0.3870**
ORL	Dimension	10	15	20	25	30	35	40	45	50
	RPCA	**0.8870**	0.8070	0.6798	0.5657	0.4850	0.4861	0.4176	0.3771	0.3704
	RPCA-OM	0.9644	0.7948	0.6204	0.5668	0.5019	0.4446	0.3833	0.3556	0.3383
	RPCA-AOM	0.9139	**0.6499**	**0.5809**	**0.5509**	**0.4686**	**0.4352**	**0.3638**	**0.3514**	**0.3353**
COIL20	Dimension	10	15	20	25	30	35	40	45	50
	RPCA	**0.6914**	0.6225	0.5334	0.4616	0.4542	0.4247	0.4054	0.3681	0.3425
	RPCA-OM	0.6923	**0.5620**	0.5096	0.4322	0.4082	0.3975	0.3789	0.3425	0.3096
	RPCA-AOM	0.7119	0.6207	**0.4478**	**0.4240**	**0.4038**	**0.3918**	**0.3770**	**0.3375**	**0.2991**
USPS	Dimension	10	15	20	25	30	35	40	45	50
	RPCA	0.6742	0.6150	0.5692	0.5198	0.4825	0.4254	0.4132	0.3525	0.3305
	RPCA-OM	0.6512	0.5909	0.5616	0.5178	0.4916	**0.4159**	0.3951	0.3802	0.3358
	RPCA-AOM	**0.6477**	**0.5888**	**0.5605**	**0.5033**	**0.4749**	0.4225	**0.3893**	**0.3659**	**0.3129**

Algorithm 3 Robust 2DPCA with non-greedy ℓ_1-norm maximization

Input: data set $\{X_i \in \mathbb{R}^{c\times d} : i = 1, 2, \cdots, n\}$, m.
Initialize: $U^{(1)} \in \mathbb{R}^{d\times m}$ s.t. $(U^{(1)})^\top U^{(1)} = I$, $t = 1$.

while not converge **do**
  $\mathbf{v}_{ijk} = sgn\left((U^{(t)})^\top(\mathbf{x}_{ik} - \mathbf{x}_{jk})\right) \in \mathbb{R}^m$ $(\forall i < j, \forall k)$;
  $S = \sum_{ij}\sum_{k=1}^{c}(\mathbf{x}_{ik} - \mathbf{x}_{jk})\mathbf{v}_{ijk}^\top \in \mathbb{R}^{d\times m}$;
  Calculate the SVD of S as $S = ADB^\top$, then

$$U = A[I; \mathbf{0}]B^\top;$$

  $t = t + 1$;
end while
Output: $U^t \in \mathbb{R}^{d\times m}$.

5 Experimental Analysis

In this section, we conduct thorough experimental evaluations of the proposed Robust PCA and Robust 2DPCA with Avoiding Optimal Mean, abbreviated as RPCA-AOM and 2DRPCA-AOM, respectively.

5.1 Reconstruction error comparison for RPCA

Regarding the experiments on reconstruction with different robust PCA methods, we normalize each initial feature of image into $[0, 1]$ and randomly select 20% images to be occluded with randomly place of $1/4$ size for fair comparison. The evaluation metric is defined as the average reconstruction error between an original unoccluded image and the reconstructed image [Nie *et al.*, 2014], *i.e.*, $\frac{1}{n}\sum_{i=1}^{n}\|\mathbf{x}_i^r - \mathbf{x}_i^o\|_2$ where n is the number of images, $\mathbf{x}_i^r$ denotes the reconstructed image and $\mathbf{x}_i^o$ is the original image without occlusion.

We compare the reconstruction error of the proposed RPCA-AOM with robust PCA with non-greedy ℓ_1-norm maximization (RPCA) [Nie *et al.*, 2011] and optimal mean robust PCA (RPCA-OM) [Nie *et al.*, 2014] in Table 1. Specifically, the reconstruction errors with respect to nine different reduced dimensions from 10 to 50 are reported over 5 benchmark data sets, including the Japanese Female Facial Expression Database (JAFFE) [Dailey *et al.*, 2010], UMIST face data set [Wechsler *et al.*, 2012], the ORL database of faces [Cai *et al.*, 2007], Columbia Object Image Library-20 (COIL-20) data set [Nene *et al.*, 1996] and the USPS handwritten digit database [Liu *et al.*, 2003]. All of the image data sets are downloaded from different web sites.

From Table 1, we can observe that: (1) The proposed RPCA-AOM algorithm performs better than RPCA over all data sets, except for slightly worse performance on a few projected dimensions. Note that RPCA also employs the non-greedy ℓ_1-norm maximization algorithm. However, the proposed RPCA-AOM requires no assumption on the zero-mean of the data, while the RPCA method definitely depends on this assumption. (2) The proposed RPCA-AOM method performs better than RPCA-OM, except for projected dimension 15 in ORL data set and projected dimension 35 in USPS data set. Note that both RPCA-OM and PRCA-AOM take the incorrect optimal mean into consideration. However, RPCA-OM integrates the optimization of the optimal mean into

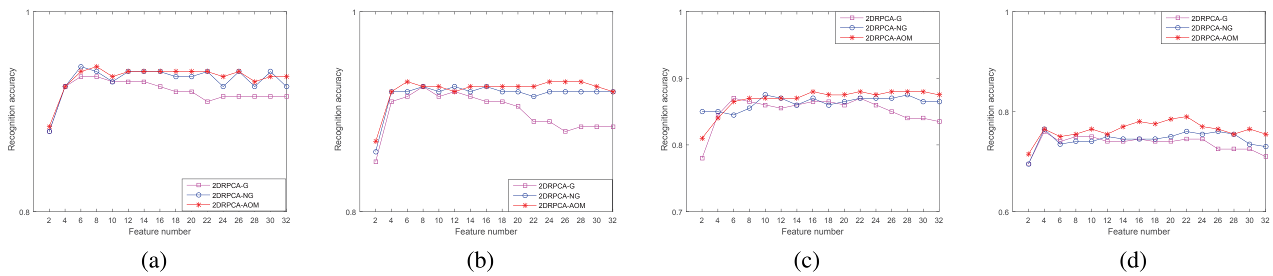

Figure 1: Recognition accuracy comparison over ORL data set. (a) 0% training images with outliers. (b) 20% training images with outliers. (c) 40% training images with outliers. (d) 60% training images with outliers.

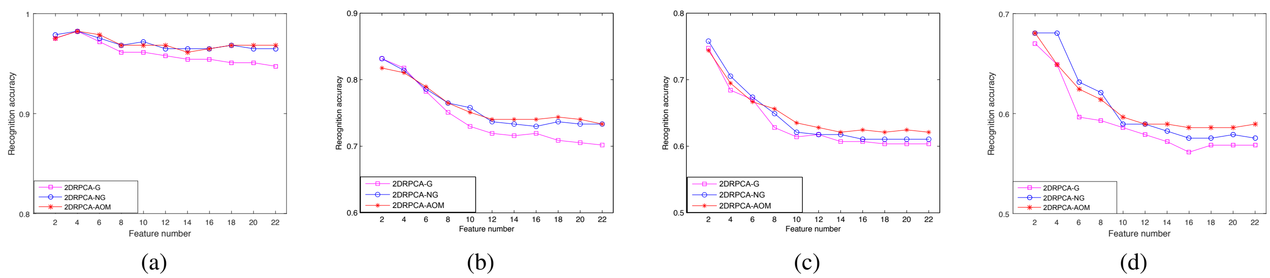

Figure 2: Recognition accuracy comparison over UMIST data set. (a) 0% training images with outliers. (b) 20% training images with outliers. (c) 40% training images with outliers. (d) 60% training images with outliers.

the procedure of dimension reduction, which leads to expensive computational cost. Instead, the proposed PRCA-AOM avoids the calculation of the optimal mean automatically.

5.2 Recognition comparison for robust 2DPCA

Regarding face recognition task, we design a series of experiments to evaluate the performance of different robust 2DPCA methods over the ORL database and UMIST data set. These methods includes robust 2DPCA with greedy algorithm (2DRPCA-G) [Li *et al.*, 2010], robust 2DPCA with non-greedy algorithm (2DRPCA-NG) [Wang *et al.*, 2015] and the proposed 2DRPCA-AOM. Note that both 2DRPCA-G and 2DPCA-NG depend on the assumption of zero-mean of the data and incorrectly employ the mean of the data as the optimal mean of ℓ_1-norm based 2DPCA.

The ORL data set consists of 400 face images of 40 objects, and each object contains ten images. The UMIST data sets consists of 575 face images of 20 objects, and each object contains a varying number of images ranging from 48 to 19. We randomly select half of the images from each object to form the training set and retain the rest as testing set for both of the data sets. To illustrate the robustness of 2DRPCA-AOM, we corrupt a varying percentage of training images with outliers and recognize testing face images in the reduced space with the nearest neighbor (NN) classifier.

With 0, 20, 40 and 60 percentage of images corrupted in training set, Figure 1 and Figure 2 demonstrate the comparisons of recognition accuracy with respect to different methods over ORL data set and UMIST data set, respectively. It indicates that 2DRPCA-G and 2DRPCA-NG achieve worse performance than the proposed 2DRPCA-AOM, especially when the percentage of corrupted training image becomes larger. As a result, the proposed method shows better robustness to outliers. Note that 2DRPCA-NG and 2DRPCA-AOM are based on non-greedy algorithm. They perform much better than 2DRPCA which is based on the greedy algorithm. This illustrates the efficiency and superiority of non-greedy algorithm for ℓ_1-maximization.

6 Conclusion

In this paper, we propose a novel robust PCA to learn the projection directions by maximizing the ℓ_1-norm based projected difference between each pair of instances, instead of the difference between each instance and the mean of data. This method automatically avoids calculating the optimal mean based on the ℓ_1-norm distance and makes any assumption of centered data unnecessary. To solve the proposed non-smooth objective, a non-greedy algorithm is exploited with fast convergence and low computational cost. In a natural way, we extend the proposed method to its 2D version and study the corresponding robust 2DPCA for image recognition. Extensive experimental results illustrate the effectiveness and superiority of the proposed robust PCA and robust 2DPCA on image reconstruction and recognition.

Acknowledgments

This research was funded by the National Science Foundation of China under Grant Nos. 61502377, 61532015, 91418205 and 61532004, the National Science Foundation (NSF) under grant No. IIS-1251187, the U.S. Army Research Office

(W911NF-13-1-0277), China Postdoctoral Science Foundation under Grant No. 2015M582662 and Ministry of Education Innovation Research Team under Grant No. IRT13035.

References

[Cai *et al.*, 2007] Deng Cai, Xiaofei He, Yuxiao Hu, Jiawei Han, and Thomas Huang. Learning a spatially smooth subspace for face recognition. In *CVPR*, 2007.

[Chang *et al.*, 2015] Xiaojun Chang, Feiping Nie, Sen Wang, Yi Yang, Xiaofang Zhou, and Chengqi Zhang. Compound rank-k projections for bilinear analysis. *IEEE Transactions on Neural Networks and Learning System*, 2015.

[Chang *et al.*, 2016] Xiaojun Chang, Feiping Nie, Yi Yang, and Heng Huang. Convex sparse pca for unsupervised feature analysis. *ACM Transactions on Knowledge Discovery from Data*, 2016.

[Dailey *et al.*, 2010] Matthew N Dailey, Carrie Joyce, Michael J Lyons, Miyuki Kamachi, Hanae Ishi, Jiro Gyoba, and Garrison W Cottrell. Evidence and a computational explanation of cultural differences in facial expression recognition. *Emotion*, 10(6):874, 2010.

[Ding *et al.*, 2006] Chris Ding, Ding Zhou, Xiaofeng He, and Hongyuan Zha. R1-PCA: rotational invariant l_1-norm principal component analysis for robust subspace factorization. In *ICML*, 2006.

[Galpin and Hawkins, 1987] Jacqueline S. Galpin and Douglas M. Hawkins. Methods of l_1 estimation of a covariance matrix. *Computational Statistics & Data Analysis*, 5(4):305–319, 1987.

[He *et al.*, 2011] Ran He, Bao-Gang Hu, Wei-Shi Zheng, and Xiangwei Kong. Robust principal component analysis based on maximum correntropy criterion. *IEEE Trans. Image Processing*, 20(6):1485–1494, 2011.

[Huang and Ding, 2008] Heng Huang and Chris Ding. Robust tensor factorization using r_1 norm. In *CVPR*, 2008.

[Jolliffe, 2002] Ian Jolliffe. *Principal component analysis*. 2002.

[Ke and Kanade, 2005] Qifa Ke and Takeo Kanade. Robust l_1 norm factorization in the presence of outliers and missing data by alternative convex programming. In *CVPR*, 2005.

[Kwak, 2008] Nojun Kwak. Principal component analysis based on l_1-norm maximization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(9):1672–1680, 2008.

[Kwak, 2014] Nojun Kwak. Principal component analysis by l_p-norm maximization. *IEEE Transactions on Cybernetics*, 44(5):594–609, 2014.

[Li *et al.*, 2010] Xuelong Li, Yanwei Pang, and Yuan Yuan. L1-norm-based 2dpca. *IEEE Trans. Systems, Man, and Cybernetics, Part B*, 40(4):1170–1175, 2010.

[Liu *et al.*, 2003] Cheng-Lin Liu, Kazuki Nakashima, Hiroshi Sako, and Hiromichi Fujisawa. Handwritten digit recognition: benchmarking of state-of-the-art techniques. *Pattern Recognition*, 36(10):2271–2285, 2003.

[Liu *et al.*, 2010] Yang Liu, Yan Liu, and Keith C. C. Chan. Multilinear maximum distance embedding via l1-norm optimization. In *AAAI*, 2010.

[Nene *et al.*, 1996] Sameer A Nene, Shree K Nayar, Hiroshi Murase, et al. Columbia object image library (coil-20). Technical report, Technical Report CUCS-005-96, 1996.

[Nie and Huang, 2016] Feiping Nie and Heng Huang. Non-greedy l21-norm maximization for principal component analysis. *CoRR*, 2016.

[Nie *et al.*, 2010] Feiping Nie, Heng Huang, Xiao Cai, and Chris H. Q. Ding. Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization. In *NIPS*, 2010.

[Nie *et al.*, 2011] Feiping Nie, Heng Huang, Chris Ding, Dijun Luo, and Hua Wang. Robust principal component analysis with non-greedy l_1-norm maximization. In *IJCAI*, 2011.

[Nie *et al.*, 2014] Feiping Nie, Jianjun Yuan, and Heng Huang. Optimal mean robust principal component analysis. In *ICML*, 2014.

[Pang *et al.*, 2010] Yanwei Pang, Xuelong Li, and Yuan Yuan. Robust tensor analysis with l1-norm. *IEEE Trans. Circuits Syst. Video Techn.*, 20(2):172–178, 2010.

[Parsons *et al.*, 2004] Lance Parsons, Ehtesham Haque, and Huan Liu. Subspace clustering for high dimensional data: a review. *SIGKDD Explorations*, 6(1):90–105, 2004.

[Torre and Black, 2001] Fernando De la Torre and Michael J. Black. Robust principal component analysis for computer vision. In *ICCV*, 2001.

[Wang *et al.*, 2015] Rong Wang, Feiping Nie, Xiaojun Yang, Feifei Gao, and Minli Yao. Robust 2dpca with non-greedy-norm maximization for image analysis. *IEEE Transactions on Cybernetics*, 45(5):1108–1112, 2015.

[Wechsler *et al.*, 2012] Harry Wechsler, Jonathon P Phillips, Vicki Bruce, Francoise Fogelman Soulie, and Thomas S Huang. *Face recognition: From theory to applications*, volume 163. 2012.

[Wright *et al.*, 2009] John Wright, Arvind Ganesh, Shankar R. Rao, YiGang Peng, and Yi Ma. Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization. In *NIPS*, 2009.

[Yang *et al.*, 2004] Jian Yang, David Zhang, Alejandro F. Frangi, and Jing-Yu Yang. Two-dimensional PCA: A new approach to appearance-based face representation and recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(1):131–137, 2004.